SSA/SIPP/IRS Synthetic Beta File

Analytic Evaluation

Working Papers

Urban Institute/NORC Evaluation Team

March 31, 2009

Acknowledgements

This report is completed in accordance with Social Security Contract No. SS00-06-60113, Order No. SS00-07-31032. The scope of the work the Urban Institute (UI) and the National Opinion Research Center (NORC) undertook is confined solely to the analytic usefulness of the resulting file produced by the Census Bureau. The original contract language is included in Appendix A.

During the course of the evaluation several minor and one major problem were uncovered with the file. The major problem (involving weights) in the Beta File was repaired by the Census Bureau, resulting in much improved results. Our evaluation is based on this revised file.

The principal investigators at the Urban Institute were Karen Smith and Doug Wissoker. At NORC the team consisted of Louise Woodburn, Edward Mulrow, and Fritz Scheuren.

At Census offering support were Martha Stinson, Gary Benedetto, and, from Cornell, Professor John Abowd. We would also like to acknowledge the support of Nick Greenia and Tom Petska at the Internal Revenue Service.

Important review comments were provided by Jim Sears and Bill Davis. We particularly like and largely agree with the Davis observations –

> **"It seems that the internal comparisons (Chapters 3 and 4) were successful for the most part while the external comparisons were not (Chapter 5). To me, this gives support to the data synthesis - including the modeling process.**
>
> **I think the real problem is Census construction of the Completed file. The Synthetic file produces similar results for the descriptive and modeling techniques that were used [in Chapters 3 and 4]. The major problems were found in [the Chapter 5] analysis, but these problems exist because of mistakes made in constructing the Completed file."**

March 31, 2009

Table of Contents

Highlights

Background

Since as early as the 1960's, there has been an ongoing collaboration between the Social Security Administration (SSA), Internal Revenue Service (IRS) and the Census Bureau to combine administrative records and Census records, in order to create files that provide more in-depth and broader analysis capabilities. The early history of this relationship is captured in the SSA reports from the series ***Studies from Interagency Data Linkages*** (Scheuren et al. 1975-80).

For most of the **period** of this productive relationship, due to the sensitive nature of the input data, these files have remained either for internal use only, or with access that was tightly controlled and monitored. There was interest in creating a Public Use file (PUF) that included variables from the Census Bureau's Survey of Income and Program Participation (SIPP), IRS' individual lifetime earnings data, and SSA's individual benefit data. 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. The intended user community for the PUF would be scholars primarily interested in national SSA retirement and disability programs.

Because of the amount and sensitive nature of the data, there has always been a fine balance between analytical depths versus confidentiality concerns. As stated in the Census Beta Test File report to SSA –

> 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.

Thus, instead of containing the actual values from the administrative records or SIPP responses, the PUF contains synthesized values. In general, individuals for five SIPP panels, 1990, 1991, 1992, 1993, and 1996 were included. A handful of variables were <u>not</u> synthesized. These variables are gender, marital status, link to spouse, type of initial benefit, and type of benefit in year 2000.

The analyses included in this report are based on the Gold file, version 4 of the Completed files and version 4.1 (reweighted) of the Synthetic Beta files.

Evaluation

Some major observations are made here based on our data analysis of the SIPP Beta Test File and our comparisons of it with the original or Gold Standard and Completed SIPP files. These will be developed and expanded in the working papers that come later:

(1) There is much to praise in the work done by Census and its experts. Many univariate distributions were "spot on" and QQ-plots produced for the analytic evaluation illustrate this. Unweighted regression analyses had some problems and results for them were more mixed.

(2) In policy simulation modeling there were many instances of differences between the Synthetic and Completed or Gold Standard file that exceeded sampling (synthetic) error that would have led researchers, had they not been discovered, to wrong conclusions.

(3) The actual sampling/synthetic weights on the Synthetic Beta File were themselves part of the problem. These were improved after this was first noticed during the early stages of the evaluation analyses. While better, these new weights may still need more work.

(4) Employment rates on the Synthetic files are too low between 1978 and 1999. The error appears to be due to inadequate adjustments in the completion process associated with the change in earnings data available from the Detailed Earnings Record beginning in 1978. It appears as though once an individual is wrongly assigned to not work in 1978, it permanently lowers the probability that they or others like them will work after 1978. It appears as though what ever error caused the low employment rates in 1978 persists through 1999. The employment problems modestly affect average indexed monthly earnings (AIMEs) but the effect is different by cohort. Low employment rates affect late career earnings for older cohorts, middle career earnings for middle cohorts, and early career earnings for later cohorts.

(5) We have identified significant problems with the health insurance coverage variables that are due to the failure to complete the data in the non-panel years. Census treated "any health insurance" and "employer health insurance" differently, leading to different problems on the Synthetic file. Furthermore, the errors in the health insurance variables allow users to uncover the panel of origin on the Synthetic files, even though this information was supposed to be excluded to avoid disclosure.

(6) We have identified problems with the assignment of longitudinal earnings among immigrants such that a significantly larger share of immigrants on the Synthetic files have earnings before immigrating to the U.S. compared to the Gold file.

(7) We have evaluated marriage durations and found that the Synthetic files have significantly longer marriage durations compared to the Gold file and the differences vary by race and cohort. The longer marriage durations do not appear to be due to the imputation of marriage histories for unmatched cases. This difference significantly increases the share of individuals reporting a marriage of ten or more years on the Synthetic files compared to the Gold file.

(8) We have evaluated the correlation of SIPP-based earnings and administrative earnings by year. None of the files appear to use information from the administrative earnings in assigning SIPP-based earnings. Specifically, we observe no higher correlation in SIPP earnings with administrative earnings among cases that matched SSA data and those that did not match SSA data. Also, we observe no higher correlation in SIPP earnings with administrative earning on the Synthetic files than on the Gold file.

(9) The total net worth values available on the Synthetic file have only limited usefulness because the Synthetic file does not indicate the actual year to which the data refer. Total net worth inflated to 2000 price adjusted dollars is available on the Synthetic files but is collected in different years depending on the source panel. Failure to account for time prevents the user from teasing out the life-cycle saving and spend-down behavior of different cohorts, nor does it allow the user to relate the net worth to the included longitudinal earnings that is the strength of the Synthetic file. Individuals typically save when they are young and working to support consumption when they are older and retired. We can observe this life-cycle saving behavior on the Gold file by examining net worth by cohort and panel, but without more panel detail on the Synthetic file, one can only obtain vague conclusions about the relationship between lifetime earnings and asset accumulation. A similar issue exists for other measures such as home ownership and financial assets, pension coverage, disability status, and, even, number of children.

(10) An overarching point we surmise is that the effort to synthesize data on such a large scale was “a bridge too far,” given how early the whole profession is in creating and using Synthetic data.

(11) A second overarching point is that the Census staff and its consultants did not understand the Social Security data well enough. Subtle, and sometimes not so subtle, problems arose that, while mostly minor, marred the data. We had hoped at one point that Census would address these errors but other priorities prevented this, even though we waited several months.

(12) A third major point is that most of what we found in need of improvement in the Synthetic data seems to be due to concerns also found in the Complete File. In fact, as we detail in Chapter 3 and 4 of this report, the Synthetic File tracked the Complete File very well. The weaknesses that exist, then, are largely in the Complete File and not in the algorithms that were used to do the synthetic file creation.

(13) Bottom line: While we remain optimistic that the synthetic approach being attempted by Census could well succeed eventually, it has not done so as yet.

Recommendations

Permit us to make several recommendations that might be undertaken by others as this work moves forward to its eventual outcome:

(1) Study the weaknesses more than we have. We have performed consistency checks and statistical comparisons for a set of key Synthetic Beta File variables, but much more is needed. As researchers become more familiar with synthetic data, a quantitative toolbox

of comparison tools will begin to define best practices for analytic evaluations of synthetic data, e.g. see Yu (2009).

(2) Add to the set of validity checks comparisons of the Completed file with the Gold file and other administrative data. These comparisons could well have revealed errors in the completion process that were not found when looking only at the Completed and Synthetic data.

(3) Start with a smaller problem to solve. The Census Bureau's approach to synthetic data follows methodology proposed by Rubin (1993) and further developed by Reiter (e.g., 2003). This approach focuses on synthetic modeling of nearly the full data file. Little (1993) discusses a more limited approach that is complex, but potentially easier to implement. A fully synthetic data file should still be the goal, but taking a more limited approach at first enables one to learn more about the data before implementing a full synthetic approach. Additi onal references related to synthetic data for disclosure avoidance are provided at the end of these working notes.

(4) Consider a "data enclave" concept (http://www.norc.org/dataenclave) as a complement to a synthetic Public Use File. A data enclave provides a confidential, protected environment within which authorized researchers can access sensitive microdata remotely from their offices or at the data enclave host offices (Scheuren and Mulcahy 2009). Researchers would perform exploratory analysis with synthetic data, followed by a confirmatory step, using the actual data through an enclave. This, in fact, is what we did, with SSA hosting a project-limited enclave.

(5) In the world of microdata for policy analyses, it is useful to not only compare the actual data to the Synthetic file. Here, in some instances, we have also compared the Synthetic file to Social Security's Microsim data. For the most part, though, we have been checking the Synthetic file for its enumerative properties – how well the SIPP versions (Gold and Completed) and the Synthetic Beta File resemble each other.

(6) In many policy settings, not only is access to actual data unavailable but also older methods have historically been employed, like the use of statistically matched datasets, which, typically, make (too?) strong conditional independence assumptions (e.g., Scheuren and Moriarity 2003, 2005). While we have not done so, synthetic datasets for policy modeling can be predicted, *ceteris paribus*, to do better than has proved to be the case in policy settings where the files, as at Treasury and elsewhere in HHS, were constructed by statistical matching methods. We would even go so far as to conjecture that synthetic data sets will dominate data files created with traditional statistical matching methods when both can be done (D'Orazio, M., Dizio, M. and Scanu, M. (2006; Scheuren 2009) (7) Because the review was time limited and incomplete, there are many more questions we would have asked if a longer period of comparison had been possible. We can, therefore, only make quite qualified statements about the Synthetic Beta File's usefulness.

(8) We may never be able to make unqualified recommendations as to any Synthetic file's usefulness. This would be true in our case, even if workarounds were made for the

current file's known problems. No matter how thorough the comparisons are, the need for an eventual check of specific results against an enclave dataset is viewed as essential, certainly at this early stage of this most promising but still emerging technology.

Chapter 1

Introduction

Fritz Scheuren NORC

The working papers in this volume are separately written pieces, authored by the individual members of the Urban Institute/ NORC team who did the evaluation. There are five chapters, beginning with this introduction. There are also several appendices that go into matters more deeply or put certain key work materials in a more readily accessible form, for possible future reference. Since these are working papers, the reader is assumed to be very familiar with the basic setting and terminology, including even acronyms.

The formal summary deliverable, provided in February and labeled here as "Highlights," starts off the volume. It has been augmented, however, since delivery to add some conjectures about the use of synthetic data sets in the "compared-to-what" world of policy alternatives. Our original charge, as we envisioned it, was to check the Synthetic Beta File for its (enumerative) faithfulness against the Gold and Completed files. It has failed this test in several respects. A better test might have been to compare the Synthetic file approach to existing data file alternatives where policy research is carried out. Here we conjecture it would fare better.

After this current introductory chapter in these working notes, we provide some background (in Chapter 2), largely quoted from the Census authors who developed the Synthetic Beta File, about how they did their work. There are many references to the growing (perhaps even exploding) literature on such methods and we could have quoted even more. We feel pleased to have had the unique opportunity to do an arms length assessment of the approach. That we found problems might not be a source of surprise, especially at this early stage of this new technology. We would caution all those who come after us to develop sound checking methods, not only of software tools but also of actual results. We thought the Census Bureau's internal data checking was thorough and very revealing. It did not alleviate the need for user checks too. The Census Bureau's comparisons, which were limited to the Synthetic versus Completed data, failed to reveal errors generated in the completion process. The final synthetic data product can only ever be as good as the completed data.

In Chapter 3, we revisit and analyze the checks that Census made of its approach. Here we employed a Q-Q graphical approach that contrasted the actual versus synthetic datasets on a wide range of variables. Most of these comparisons indicated that the method was working as well as might be expected. For the major results the synthetic data were very close. For minor small sub-domains the results were less successful. Still, even in these latter cases, the greater "sampling" error that could be calculated should have warned potential users against over-reliance. The typically greater attenuation of sampling error for the synthetic results came out quite well in the Q-Q plots. To address this attenuation, the number of implicates might be increased, a suggestion that was originally made for earlier forms of imputation (e.g., Colledge, et al. 1978, Scheuren 1983, Woodburn and Scheuren 1990). More implicates would also increase the stability of the variance estimates – a point that might be very important for estimating the very large variance/covariance matrices that can arise in complex regression analyses.

Chapter 4 makes comparisons of Synthetic Beta File regression results for employment and earnings with those from the Gold and Completed Files. Here, as is typical for regression model estimation, the results are all done unweighted. The early decision, not to use weighted regressions, proved a happy one, as the weights on the Synthetic Beta File later had to be revised. This chapter recommends that providing general purpose variance software for inference in the more complicated analytic settings could usefully augment the existing Census Bureau approach. This extra effort may well be a role for an early user, not the Census Bureau, however.

Chapter 5 looks at results commonly considered when simulating SSA policy alternatives. The emphasis in the chapter is on the longitudinal strengths and weaknesses of the Synthetic Beta File. It was in this area that many of the early problems were uncovered. The reweighting of the file helped greatly. Some problems still remain on the Synthetic File, including much lower employment rates compared to administrative sources and errors in the health insurance coverage and immigrant earnings. In addition, there are conceptual issues with the provided measures of wealth, assets, and other time-varying variables that significantly limit their practical use. We caution potential users of the Synthetic Beta File that more problems may have been uncovered had there been time to do further checks.

Chapter 2

Background on Synthetic Data and Associated Methodology

Louise Woodburn, NORC

2.1 Overview

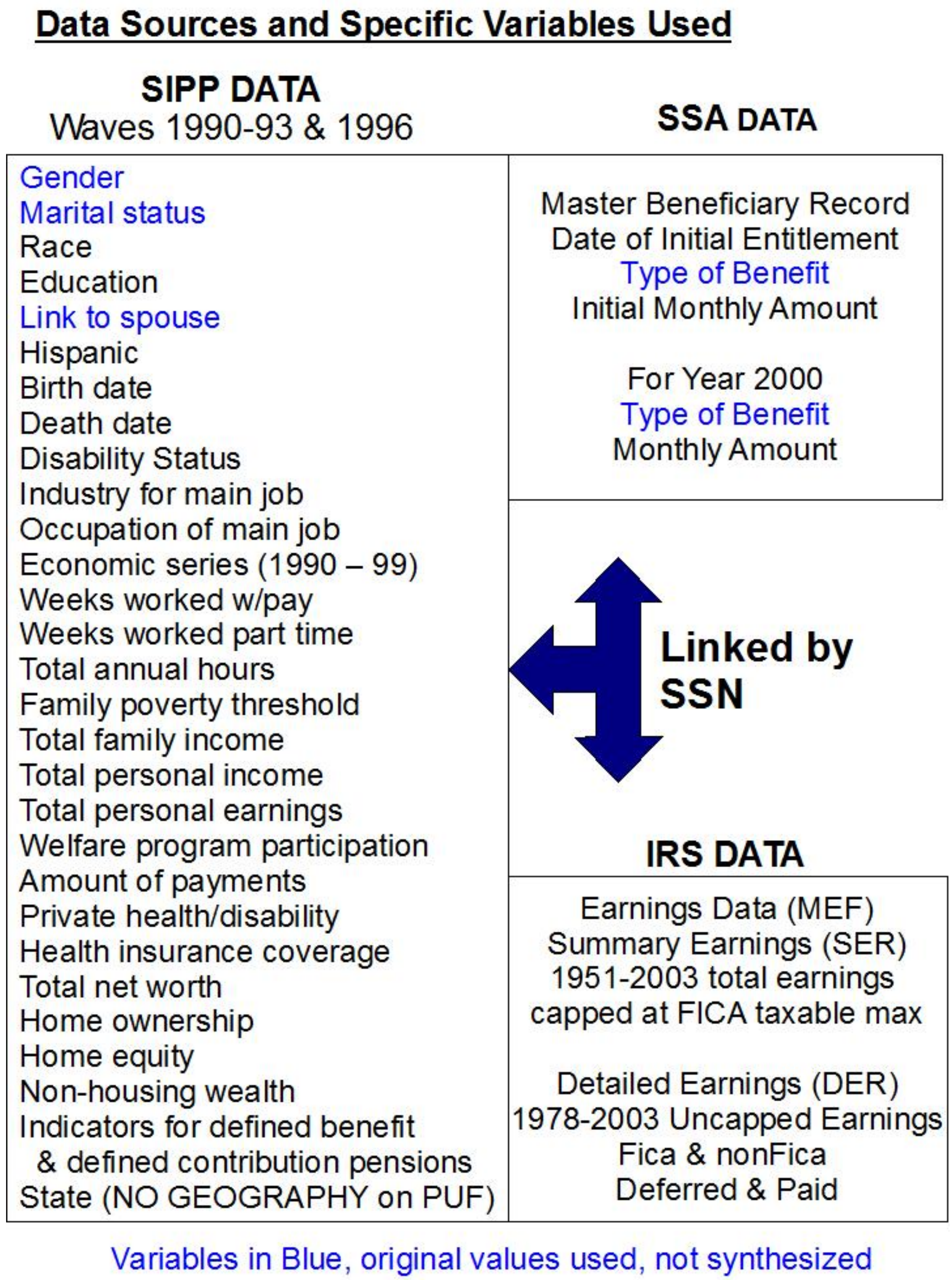
There has long been interest in creating a Public Use file (PUF) that included variables from the Census Bureau's Survey of Income and Program Participation (SIPP), The Internal Revenue Service (IRS) individual lifetime earnings data, and Social Security Administration (SSA) individual benefit data. Census files of survey and IRS-SSA data were routinely produced earlier (Scheuren et al. 1980). Changes in confidentiality concerns, however, led to a long cessation (e.g., see Scheuren and Mulrow2001) for background).

The idea of a Synthetic Public Use File (PUF) seemed an option worth exploring. The goal was to give wide access to a data base that could be policy analyzed both within and outside government 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. The intended user community for the PUF was envisioned as one that would be primarily interested in national retirement and disability programs. As stated in the Census report to SSA,

> "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."

Thus, instead of containing the actual values from the administrative records or SIPP responses, the PUF contains synthesized values. The included data from each data source are shown in Figure 2-1. In general, individuals for five SIPP panels, 1990, 1991, 1992, 1993, and 1996 were included. A handful of variables were not synthesized. These variables are gender, marital status, link to spouse, type of initial benefit, and type of benefit in year 2000. Initially, race and a categorical education variable from SIPP were going to remain unsynthesized, but in order to include the original type of benefit variables, it was necessary to release fewer unsynthesized variables.

Figure 2-1: Data Sources and Specific Variables Used



Note it was not the charge of the UI/NORC team (hereinafter "team") to check the file being created to see if it passed Census-IRS-SSA confidentiality protection requirements. The Team's goal, rather, was to see how well the Synthetic File could reproduce actual results.

We did, as a matter of course, look at the variables that were not synthesized. Since there are only two SIPP variables, gender and marital status, a total of 8 categories – using only these two variables, all categories had ample sample size, the smallest with 2,815. Looking at the combination of gender, marital status, type of benefit in 2000, and type of initial benefit, there were 12 that were unique, when these actual values were all concatenated and hence could operate as identifiers. Looking at married records, we find that 25 percent of the file has a spouse with data also on the file. Of these 67,331 married couples, there are only 40 combinations of gender, type of benefit 2000, type of initial benefit, spouse type of benefit 2000, spouse type of initial benefit that contain just one couple. There are 134 married couples that fall into combination categories that have less than 5 couples. In theory it would be possible, then, to link the actual and synthetic data together for these cases. When a related issue arose in the public release of the earlier CPS-IRS-SSA Exact Match File, the problem was overcome by a distribution-preserving randomized reassignment of codes, something we recommend here for future Synteirc data sets.

The Census staff did an assessment of disclosure risk by the two methods of probabilistic record linking, using the Census' internal record linking software, maintained by the Statistical Research Division. The results from their assessment showed that there were no data segments that had a true match rate over 1 percent. They also assessed disclosure risk by distance-based record linkage (e.g. Domingo-Ferrer, Abowd and Torra 2006). Using this method, they also found a potential match rate of around 1 percent. For further details see the technical description of the file.

In order to understand how the PUF was created, it is important to understand the origin of the data and the step-by-step process of combining, imputing and synthesizing. There are basically three files created. First, the "Gold Standard" file was created by extracting variables from the five SIPP panels and merging them onto the SSA-provided administrative data from the Summary Earnings Records (SER), Detailed Earnings Records (DER), and the Master Beneficiary Record (MBR) by Social Security Number (SSN).[1] The Gold Standard represents the available confidential micro-data that would be used for analysis by an authorized researcher working in a restricted-access facility. Second, all missing data were multiply-imputed four times to create four Pseudo-Complete Files, e.g., Rubin (1987). The final step, was to use create 4 replicates of synthetic data for each completed file, e.g. Rubin (1993) and Raghunathan, Reiter and Rubin (2003), resulting in 16 synthetic data files. As described by the Census staff:

> "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).

[1] The SER contains individual Social Security covered earnings up to the annual Social Security taxable maximum from 1951 to 2003. The DER contains an expanded set of earnings variables from 1978 to 2003. This expanded set includes Social Security covered and uncovered earnings, including earnings above the taxable maximum. It also contains data on tax deferred earnings (mostly employee contributions to tax deferred retirement accounts like 401(k)s and 403(b)s. The MBR contains individual Social Security benefit data, including benefit type and amount for both initial and current benefits.

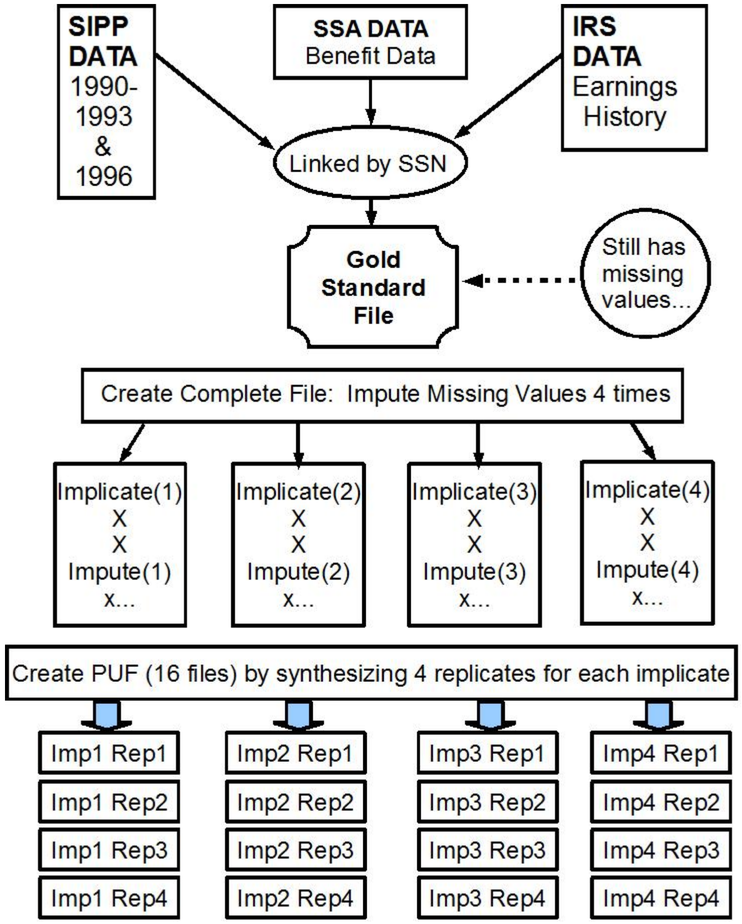
All statistical models are estimated separately for detailed groups of individuals based on the values of categorical variables that include both demographic and economic controls. Logistic and linear regressions also include additional linear controls that are selected from a long list of potential 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 Final SIPP/SSA/IRS-PUF."

Note that although State is included on the Gold Standard file, there is no geography data on the PUF.

Figure 2-2 contains a diagram showing the different steps and files for the creation of the PUF.

Figure 2-2: Schematic of Data File Creation



2.2 Item Missingness

The imputation models for item missing data were based on Bayesian bootstrap and Sequential Regression Multivariate Imputation methods for estimating and sampling from multivariate posterior predictive distributions. The techniques for multiple imputations follow from Rubin (1987), using an approach originally proposed by Rubin in work he did for SSA. The method, as applied here, creates multiple imputations for missing data but then extends the idea to create synthetic data as well.

There were three main types of item missing data, of which the second and third were multiply imputed:

1) Traditional survey non-response, for items not answered by SIPP respondents.
2) All SIPP respondents were made to have the same series of data, e.g. values for years from 1990 – 1999 that extend before and after the observed SIPP panel. Of course, depending on the corresponding survey for a given SIPP respondent, some of these years of data would be missing due to out-of-scope survey years.

3) SIPP respondents in the Gold Standard file for whom the Census Bureau does not have validated SSNs are missing all data items whose linkage depends upon the SSN; that is, all earnings, benefit, and administrative birth and death data.

An additional source of missing data in the Gold Standard file is due to structurally missing data. That is, data 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 missing in the Completed and Synthetic data implicates that constitute the SIPP/SSA/IRS-PUF.

2.3 Creation of the Synthetic Data File

The methods used to create the synthetic data file have seen other applications as well, e.g. Kennickell (1997). In general, the Census synthetic file was built around 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. The 1993 method was enhanced to include the application of Sequential Regression Multivariate Imputation (SRMI) to synthetic data from Raghunathan, Reiter and Rubin (2003). The formal inference methods for multiple-imputation-based partially synthetic data are taken from Little (1993) and Reiter (2003), as well as formal inference methods for multiple-imputation based partially synthetic data that also have missing data from Reiter (2004). As described in the Census report, to implement Bayesian bootstrap (BB) and SRMI:

> ... [Census applies] the principle of estimating the conditional distribution of group of variables (columns of Y) conditional on all other columns D. 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.

Chapter 3

Mean and Distributional Data Comparisons of Synthesized and Complete Data

Louise Woodburn, NORC

3.1 Background

The analyses in this chapter are based on the results provided in the Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File (Abowd, Stinson and Benedetto 2006).[2] The Public Use File combines data 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. The file was created as partially synthetic data with multiple imputation of missing items. The person-level records contain 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.

The Social Security Administration first provided the Synthetic Beta files to the evaluation team in the Fall of 2007. Abowd, Stinson, and Benedetto (2006) summarized variables and distributions from the Completed and Synthesized data in over 60 tables devoted to investigating the analytical validity of the synthetically produced data set. Analysis using the first set of files revealed some significant problems with population counts and employment rates compared to administrative data. The Census Bureau subsequently addressed these problems by reweighting the Synthetic files. Census provided the Team updated Synthetic files that incorporated adjusted weights to the research team in October 2008. The Team never received the weight-adjusted Completed data over the term of the evaluation contract due to confidential data transfer delays at the Census Bureau. Only the analysis in this chapter, which were based on tables provided by the Census Bureau, uses the re-weighted Completed data.

For the reweighted file, Gary Benedetto at Census re-ran Table 29 and Table 61, as requested by NORC staff. Census Table 29 compares the mean monthly Social Security benefit amount among beneficiaries by sex, race, and education for the Completed and Synthetic files and shows confidence intervals. Table 61 compares the distributions of the full set of variables on the Synthetic files to the Completed files.

Initially, we summarize the results from the revised files where we explore the Census provided analytical tables to support the validity analysis. We also compare the results of the revised weights to the initial results. The initial file results are included in Appendix B. Finally, we compare the weight distribution of the original and weight-adjusted Synthetic file (in

[2] Two versions of these results were prepared: first when the project began and later after the weights had been adjusted. The results shown here are from the second version, although both have been analyzed and are available in the Appendices.

Appendix C). Only the revised versions are analyzed here, even though both were examined in the course of our evaluation

The first part of our analysis compares the synthesized data versus the completed data by studying the percentile distributions of all the file variables provided in Table 61. In this table, there are percentile points for both data sets (Completed and Synthetic) for 219 variables. We use quantile-quantile (Q-Q) plots as a tool to investigate differences. We chose Table 29 as a representative table, to evaluate means, variances, and confidence intervals for important variables broken down by various demographic categories.

3.2 Analysis of Percentile Distributions

An example of the data in Census Table 61 is shown in Table 3-1. The full table is included in Appendix B. It shows percentiles, 1, 5, 10, 25, 50, 75, 90, 95, and 99 for each variable for both the Completed and Synthetic data sets. In order to really compare the distributions we use Q-Q plots.

Table 3-1: Example of data from Census Table 61, Percentile Points for Synthesized and Complete Data

Final Revised Data File

Variable	Type	P01	P05	P10	P25	Median	P75	P90	P95	P99
Home Equity	Completed	-9,125	4,000	8,000	21,375	50,000	100,000	160,500	213,250	319,250
	Synthesized	-17,874	2,430	6,275	19,886	49,731	99,550	171,295	231,991	336,918
Nonhousehold Wealth	Completed	-6,000	1,000	2,000	6,000	17,000	60,000	180,000	314,500	761,500
	Synthesized	-48,703	369	1,336	5,204	15,994	56,948	171,958	307,796	831,669
Total Net Worth	Completed	-33,000	-6,000	1,000	9,000	51,000	140,000	292,750	446,500	920,000
	Synthesized	-34,630	-5,855	-132	8,365	50,257	135,722	277,707	414,051	835,547

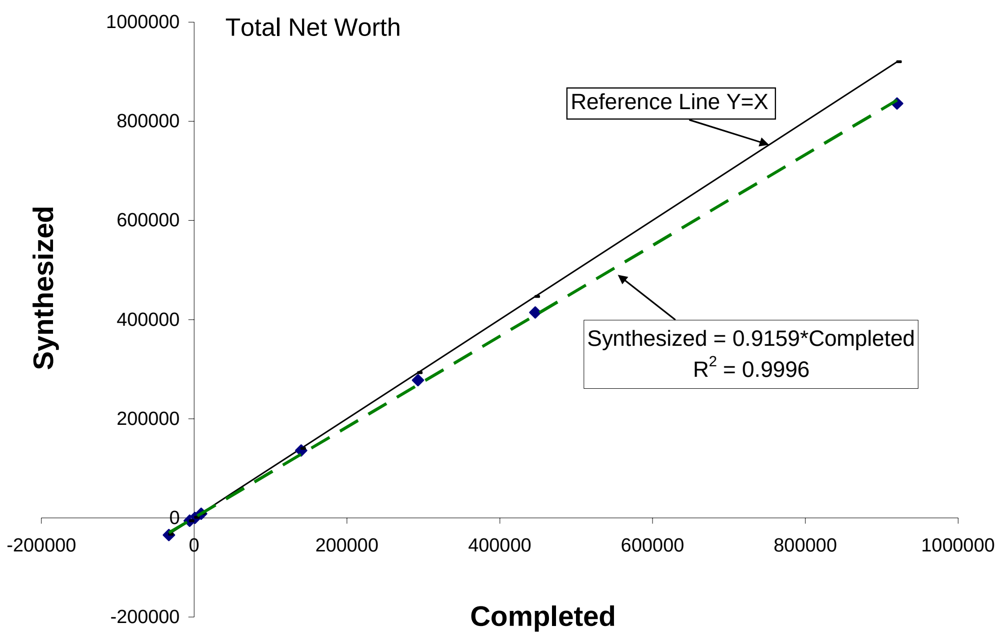
Quantile-Quantile (Q-Q) plots are a powerful visualization tool for judging whether or not two independent samples come from the same statistical distribution (i.e. populations). If two samples are from the same distribution then a scatter plot of paired quantile estimates from each sample will show a linear pattern that falls along a 45-degree reference line through the origin, i.e. a line with slope one and zero intercept. When the pattern of the plotted points deviates from the reference line, one can ascertain how the two distributions differ based on features of the pattern.

If the plotted points have a linear pattern that does not follow the reference line, differences in the means and variances between the two distributions are visually apparent. In comparing the synthesized to the completed data in this way, we can **"see"** exactly how the distributions differ. An intercept value greater than (less than) zero indicates that the sample plotted on the vertical axis comes from a distribution with a larger (smaller) mean. For our comparison of the synthesized versus completed data, we assume that they have the same mean and thus a zero intercept.

Slopes of a fitted linear line in a Q-Q plot reflect the relative size of the variances of the two distributions. A slope greater than one indicates that the distribution from the synthesized data (plotted on the vertical axis) comes from a distribution with a larger variance, than the distribution from the completed data. Alternatively, a slope less than one indicates that the distribution from the synthesized data (plotted on the vertical axis) comes from a distribution with a smaller variance than the distribution from the completed data. If the plotted points have a curved-shape, including an "S" shape, it indicates that one distribution has more or less extreme values. Which distribution has more extreme values in either tails of the distribution is indicated by how the Q-Q plot curves with respect to the reference line

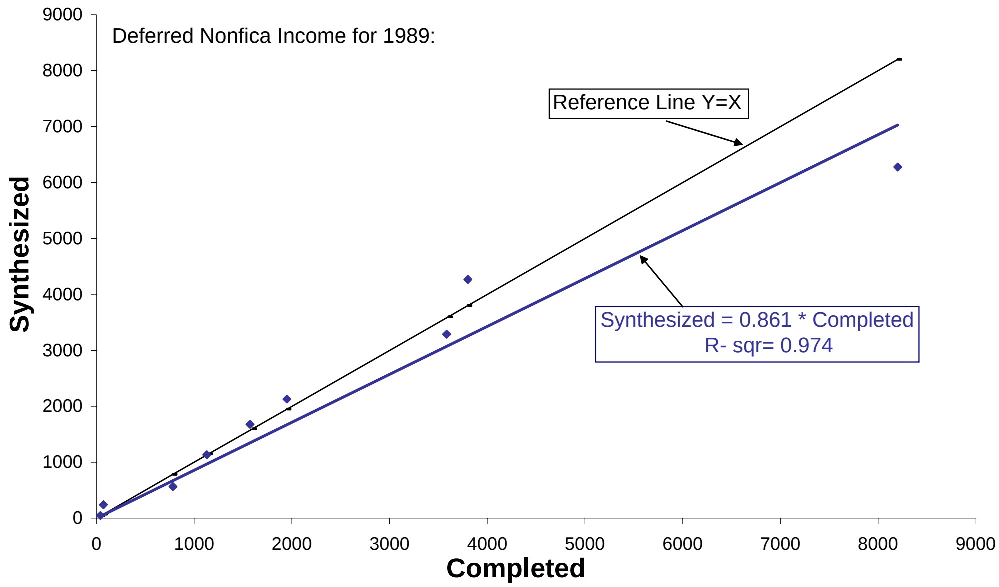
Consider the Q-Q plot for total net worth, as shown in Figure 3-1. As we found with most of the variables, the Completed and Synthetic distributions match up quite nicely. The percentile points for the synthesized data are graphed on the vertical axis versus the completed data that is plotted on the horizontal axis. Figure 3-2 compares deferred non-FICA earnings in 1989 between the Completed and Synthetic data. The solid line is y=x, the guide for distributions that are exactly the same. The dashed line represents a least squares linear regression with a zero intercept for Figure 3-1 and a colored line for Figure 3-2.

Figure 3-1: Q-Q Plot of Total Net Worth
Synthesized Data Percentiles Plotted vs. Completed Data Percentiles.
Excellent distributional agreement – fitted line close to reference Y=X line



A note is in order about the use of R-squared (R^2) here and elsewhere. First, R-squared is calculated algebraically, as in conventional usage. It cannot be interpreted as a function that can be converted, e.g., $R^2/(1-R^2)$, into a function of the F distribution to test for statistical significance of the independent regressors. Because independent order statistics are being used, R-squared does not have a direct inferential interpretation. Still the quantity R-squared does measure the linearness of the relationship and hence the degree to which the two files are distributionally similar.

Figure 3-2: Q-Q Plot of Deferred Non--FICA Earnings in 1989
Synthesized Data Percentiles Plotted vs. Completed Data Percentiles



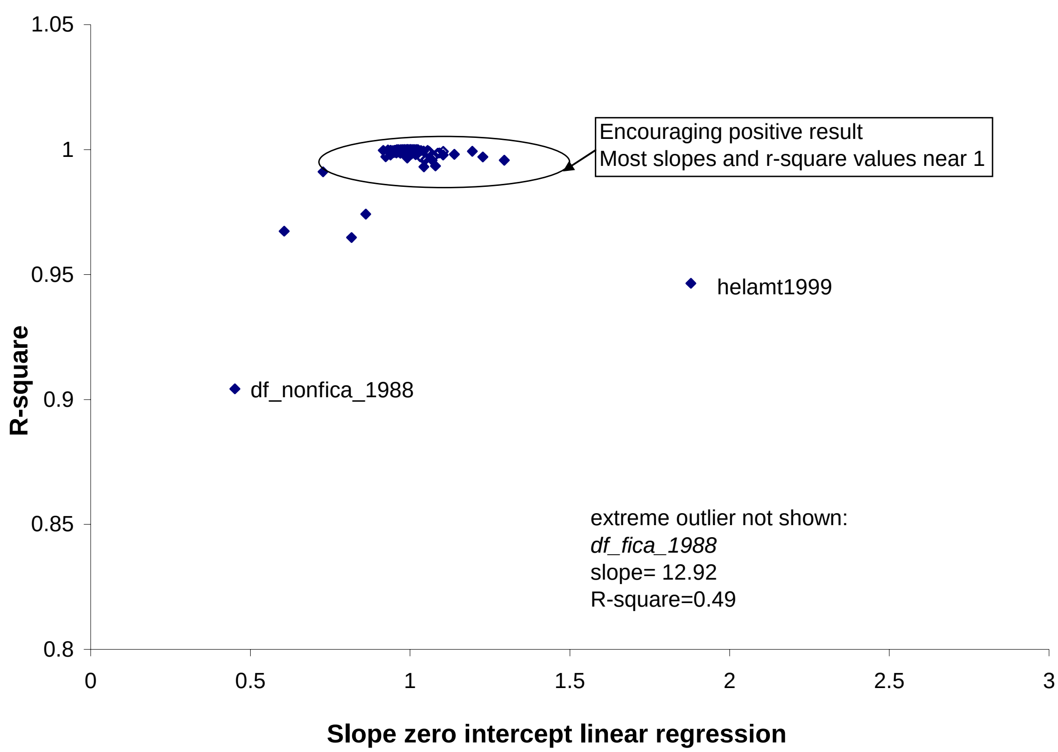
In order to evaluate all variables, we first estimated the slope coefficients for zero-intercept linear regressions for all variables. From the regressions we look at both the computed R-squared value as well as the slope coefficients to determine outcomes where the Completed and synthesized distributions are not similar. A frequency distribution of the slope coefficients is shown in Table 3-2. We found that around 80 percent of the coefficients fall between 0.95 and 1.05.

Table 3-2: Frequency of Slope Coefficients from Linear Regressions of Q-Q Plots

Regression Slope	Frequency	Percent
< 0.8	3	1.4
0.8 - < 0.85	1	0.5
0.85 - < 0.90	1	0.5
0.90 -< 0.95	13	6
0.95 - < 1.00	111	52
1.00 - < 1.05	61	29
1.05 - < 1.10	13	6.1
1.10 - < 1.15	4	1.9
1.15 - < 1.20	1	0.5
>= 1.20	4	1.9

In order to spot outlying values, we plotted R-squared values versus the slope coefficients. These are shown in Figure 3-3.

Figure 3-3: R-squared Values vs. Slope Coefficients



These results were very encouraging. Most of the regression lines for the variables had a slope near 1 and very high R-squared values. There were only 3 variables that indicate some outlying issues. The individual slope and R-squared results for these variables are detailed in Table 3-3.

Table 3-3: Slope Coefficients and R-Squared Values for Outlying Results

Variable Description	Slope	R-Squared
Deferred Non-FICA Wages 1988	0.452	0.904
Total Health Benefit* Dollars 1999	1.880	0.947
Deferred FICA Wages 1988	12.920	0.490

*Total Health Benefit Dollars refers to total combined benefit dollars from workers compensation, own sickness (disability), and veteran disability benefits.

Communications with Census Bureau staff suggest that these inconsistent results could be due to small sample size. Deferred wages were not so common 20 years ago. Additionally, the total health benefit dollar variable contained payments due to disability from private and non-government sources, also is not too common. The Q-Q plots for the variables in Table 3-3 are shown in Figures 4-6 below.

Figure 3-4: Q-Q Plot of Deferred Non-FICA Wages for 1988
Synthesized Data Percentiles Plotted versus Completed Data Percentiles
Outlying data point for completed data suggests more extreme data in Complete Data set.

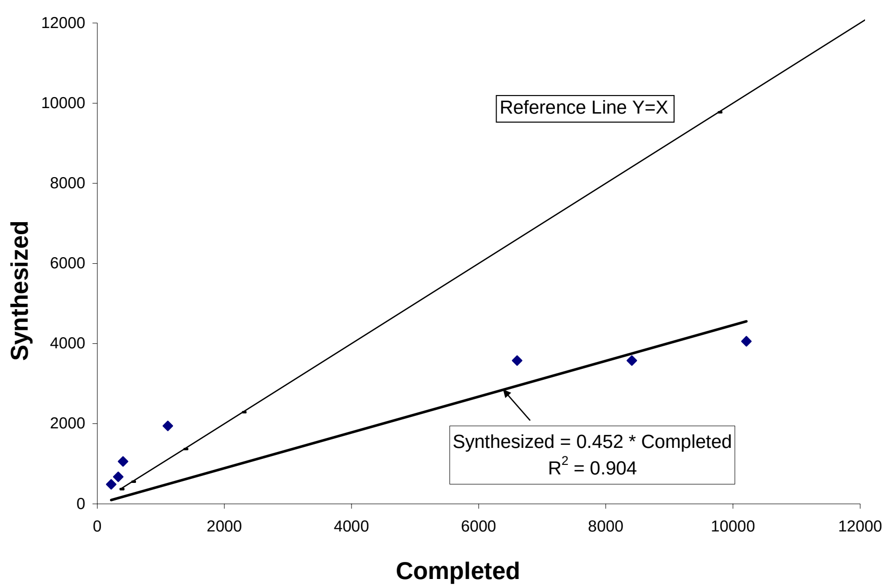


Figure 5: Total Health Benefit Dollars 1999
Synthesized Data Percentiles Plotted versus Completed Data Percentiles

Outlying data point for synthesized data suggests more extreme data in Synthesized Data set.

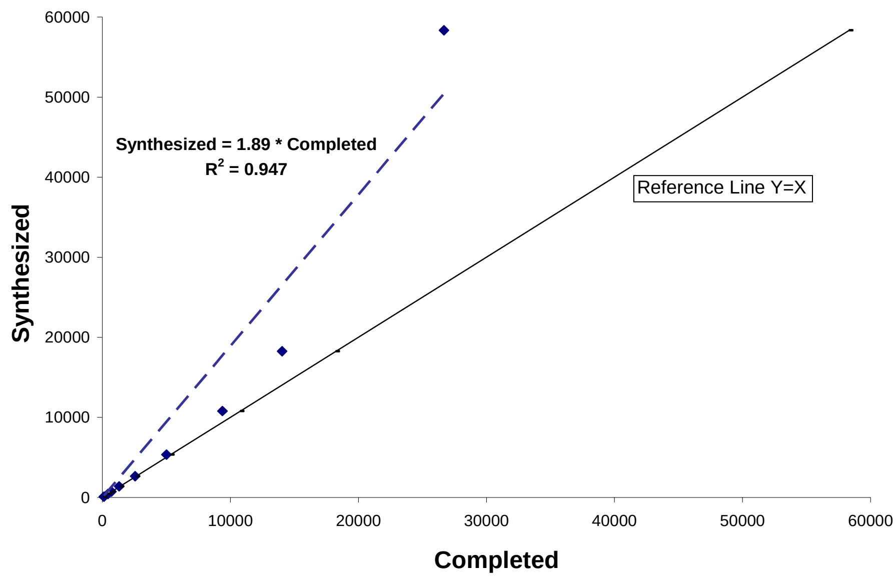
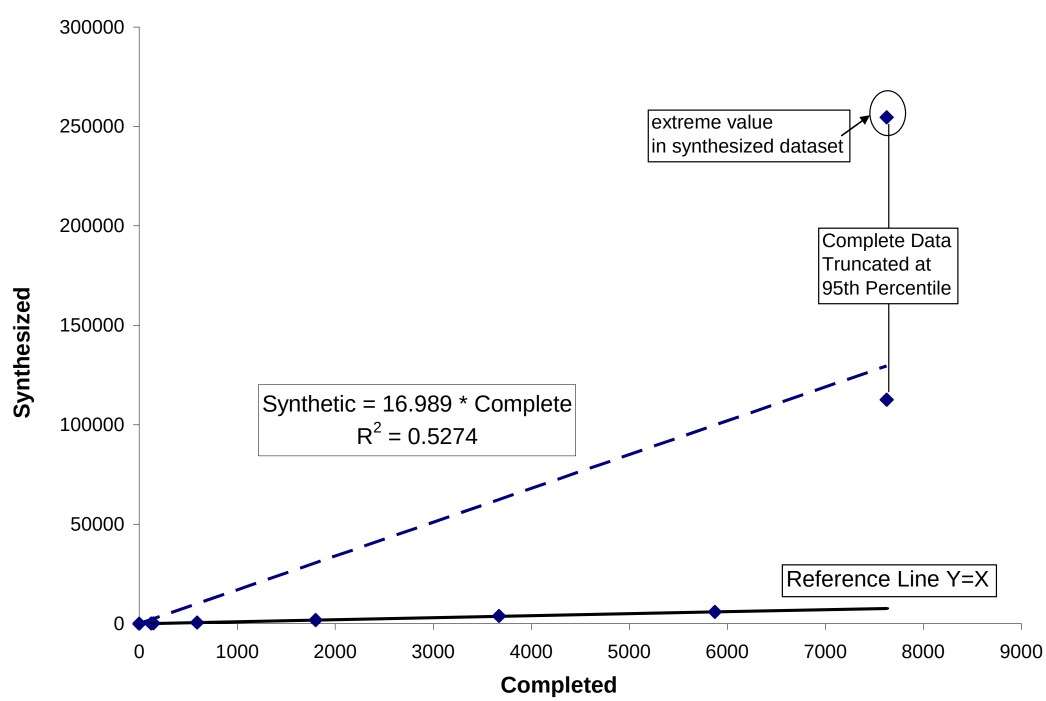


Figure 3-6: Deferred FICA Wages 1988
Synthesized Data Percentiles Plotted versus Completed Data Percentiles

Outlying data point for synthesized data plus truncation at 95th percentile contribute to the skewed relationship



3.3 Analysis of Means, Variances and Confidence Intervals

The second type of analytical validity tables provided in the Final Census Report contained estimates of means, variances and confidence intervals. These estimates are broken down by various demographic variables with variance results generated using all of the implicate and replicate structure in the complete and synthesized data sets. In its response to the RFP to analyze the PUF, UI/NORC suggested that the overall distributions should compare well, but for small subdomains (e.g., cases where the occurrences on the Gold Standard File are few) that the Synthetic Beta File might not be generally reliable. From looking at Census Table 29 (original version in Appendix C, revised in Table 3-4 below), that includes the estimate of Monthly Benefit Amount by demographic group and education, we found that there are some differences in the small subdomains, but overall, the results are consistent.

The estimates from revised Census Table 29 are shown in Table 3-4 below. One metric to evaluate is whether or not the confidence intervals computed for the synthetic and complete data overlap and to what extent.

Table 3-4: Estimates of Mean Monthly Benefit Amount for Year 2000, Confidence Intervals, and Variance from Census Table 29
By Demographic and Educational Group.

Demographic Group	Education Category	Mean		Confidence Interval				Total Variance	
		Synthetic	Complete	Synthetic		Completed		Synthetic	Completed
white females	no HS	585	567	534	637	559	574	490	19
	HS	589	598	574	604	589	608	48	25
	Some Coll	573	596	548	597	587	606	128	29
	College	563	585	512	614	569	600	524	80
	Graduate	609	640	570	649	623	656	372	100
black females	no HS	490	469	460	520	451	487	202	100
	HS	450	444	432	467	429	460	93	78
	Some Coll	440	449	416	463	423	475	176	212
	College	457	446	426	487	404	487	317	635
	Graduate	537	614	371	704	552	676	5,770	1,326
white males	no HS	726	715	673	779	698	731	520	69
	HS	732	742	702	761	730	755	162	45
	Some Coll	726	743	702	751	731	755	131	51
	College	803	812	787	819	796	828	91	98
	Graduate	871	885	836	906	867	903	259	120
black males	no HS	591	578	561	620	558	599	204	140
	HS	533	524	489	577	503	544	410	155
	Some Coll	489	484	442	536	456	512	549	284
	College	545	578	455	634	496	661	2,150	2,202
	Graduate	679	632	584	774	559	705	2,706	1,908

The graphical display of the results in Table 3-4 are very encouraging. For all of the categories in Table 3-4, the synthetic confidence intervals largely overlap with the intervals based on the Completed data. For most but not all of the categories in the graphs for both females (Figure 3-7) and males (Figure 3-8), the confidence intervals are wider for the Synthetic data than for the Completed data and wider for blacks than for whites. Larger confidence intervals are expected among small subgroups on both the Completed and Synthetic files. When confidence intervals are large on the Completed files, they are also tend to be large on the Synthetic files and vise versa

Figure 3-7: Confidence Interval for Mean Monthly Benefit Amount 2000 Females by Race, White / Black, and Education Category

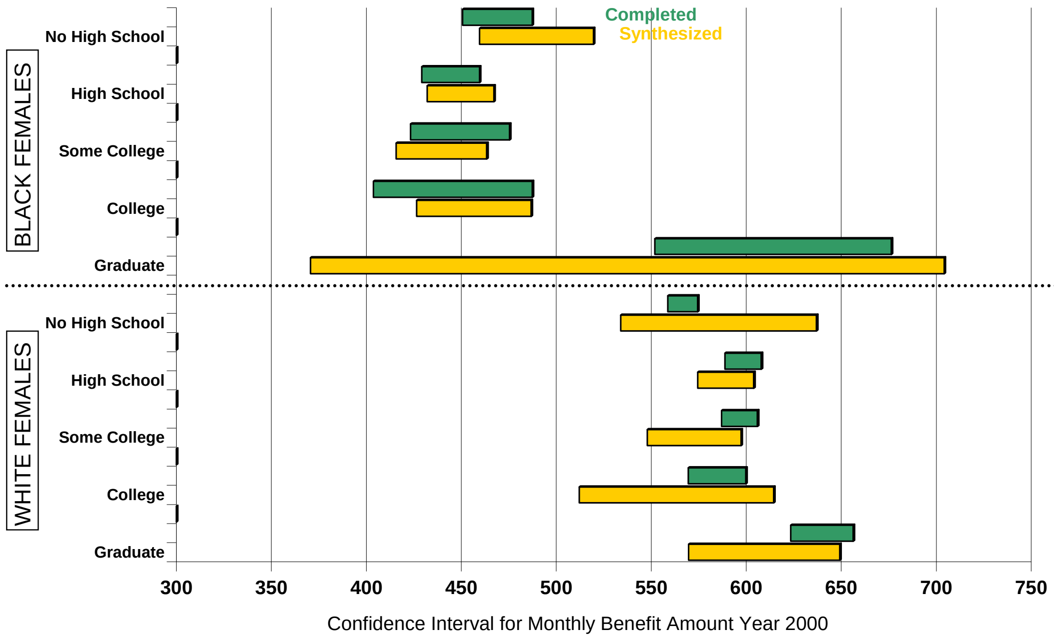
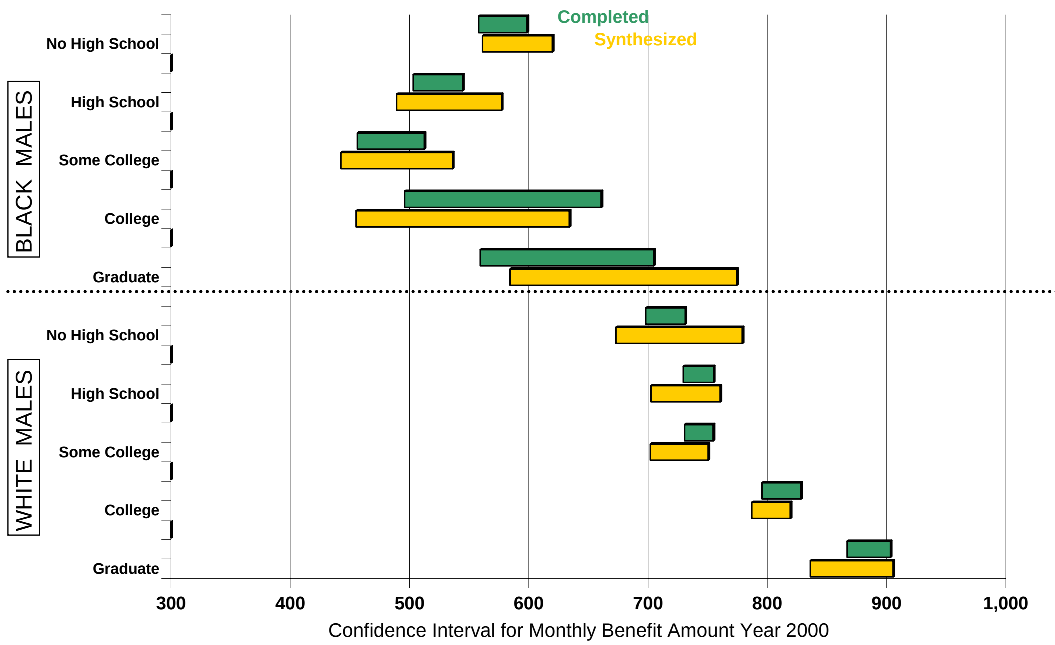


Figure 3-8: Confidence Interval for Mean Monthly Benefit Amount 2000 Males by Race, White / Black, and Education Category



3.4 Weight Comparison of Original and Revised Files

The Census Bureau adjusted the weights on the revised file so that some important employment statistics would match the Gold Standard proportions. In particular, Census added a few variables to the post-stratification/raking procedure, including the employment status in year 2000 and type of benefit in year 2000. The result of the revisions is that the weights on the revised Synthetic file are less extreme. Comparing the distribution of the weights for the 1st implicate/1st replicate files for original and revised, we see that the maximum weight used to be over 30,000, now it is around 14,000. The details are shown in Table 3-5.

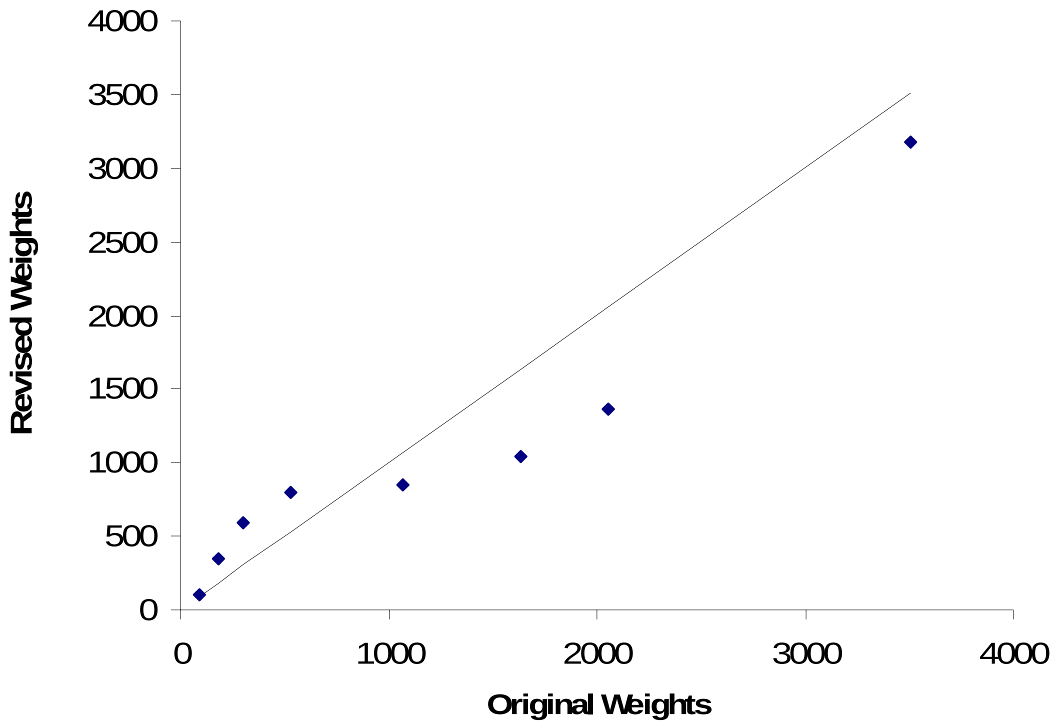
Table 3-5: Comparison of Univariate Distribution of Original and Revised Weights

Percentile	Original	Revised
MAXIMUM	35,094	13,729
99%	3,510	3,171
95%	2,054	1,361
90%	1,636	1,042
75%	1,069	845
50%	531	798
25%	304	596
10%	185	345
5%	92	101
1%	0	0
MINIMUM	0	0

Source: Author's tabulations of the original and revised Synthetic Beta file weights.

Additional insight into the weight changes can be seen by looking at a Q-Q plot of the weights. Such a Q-Q plot is included as Figure 3-9. The lower 50 percent of weights actually increased, while the upper 50 percent of weights decreased. This amounts to adjusting all of the weights closer to the mean weight of 773.

Figure 3-9: Q-Q Plot of Revised vs. Original Weights through the 99th Percentile



Chapter 4

Comparison of Earnings and Employment Models Estimated in Gold Standard, Completed, and Synthetic Beta Files

Doug Wissoker, Urban Institute

4.1 Chapter Summary

In this chapter, we compare regression models of earnings and employment for non-immigrant non-beneficiaries across Gold Standard, Completed, and Synthetic Beta files. Included are adults age 25 and older based on data from 1990 to 1999. The sample is unweighted, unlike other chapters in these working notes. Full details are available in Appendix D. There are five basic findings.

1. The Gold Standard and Completed files generally provide similar coefficients for models of log earnings and the probability of employment for those who worked in the previous year, particularly for men.

2. The Gold Standard and Completed files yield noticeably different estimates of the coefficients for the probability of employment for those who did not work in the previous year. Key differences include both the model constant, the effects of an additional year of work history, and effects of education.

3. The difference in employment probability models for those not working last year fits with finding of different average employment rates between the Gold Standard and Completed files (Chapter 5, sections 5.4 and 5.7 of this report) . For those not working last year (and used in the models of transition to employment), the mean employment rate this year is quite different between these files: In the Gold Standard file, among males not working last year, 27 percent are working this year as compared with 15 percent in the Completed files.

4. The Completed and Synthetic Beta files yield regression coefficients that are broadly similar for earnings and employment. However, for the earnings models some relationships (e.g., education effects) appear to be moderated in the Synthetic Beta files. In the models of employment for those working in the prior year, the age differentials for older workers are smaller in the Synthetic data as compared with the Completed data.

5. The procedures to obtain standard errors for the Synthetic Beta files took a non-negligible amount of effort. If the goal is to have all users implement these formulas, users should be provided with examples that estimate standard errors for regression models (rather than just individual statistics) and include the preferred adjustments for when the unadjusted formulas calculate negative variances or near-zero degrees of freedom.

4.2 Analysis Approach and Methodology

We estimate here simplified versions of the models of earnings and employment for non-immigrant non-beneficiaries developed by Urban Institute researchers for SSA's Polisim microsimulation model. They represent realistic tests of estimated relationships for which the Synthetic Beta file is particularly well suited. The models estimated here exclude many of the interactions used in Polisim; this improves our ability to compare models and understand the differences observed across the files.

The analysis involves two sets of comparisons. First, we compare the estimates from the Gold Standard file with those of the Completed files. This addresses whether the imputation for non-matching of administrative records leads to substantially different results. Even if completion of the Gold Standard file leads to substantially different results from those on the uncompleted Gold Standard file, this does not necessarily imply a problem, since the Gold Standard file may suffer from selection bias in which cases can be matched. Second, we compare the estimates from the Completed file with those based on the Synthetic Beta files, which addresses the effects of synthesizing the data to mask the actual observations.

In general, we find that for the earnings models and models of staying employed, the Completed files and the Gold Standard files provide similar coefficients. However, we find some large differences between the Gold Standard and Completed files in the coefficients of the models of becoming employed, particularly in the constant and coefficients on education and years of employment.

The Synthetic files produce coefficients from the earnings and employment models that are broadly similar to those based on the Completed files. In the earnings models some relationships (e.g., education effects) are moderated in the Synthetic Beta files. In the models of employment among those working in the prior year, the older worker effects are smaller in the synthesized data as compared with the completed data.

Below, we provide an explanation of the procedures used, followed by a discussion of the results.

4.2.1 A review of the basic procedure for analyzing data in the three sets of files

The procedure for estimating regression model parameters in the various samples follows the procedures presented in the SIPP/SSA/IRS final report (Abowd, Stinson, and Benedetto 2006).

The appropriate method depends on the data file being used. For models estimated using the Gold Standard file, model coefficients and their standard errors are used directly. For models estimated using the Completed files, the reported coefficient estimates are the average of the coefficients from the four implicates. The standard error is given as the square root of the sum of 1) the average of the squared standard error over the four implicates; and 2) 1.25 times the variance of the coefficient estimates across implicates. The standard errors are larger than the average from each of the four completed implicates.

The Synthetic Beta file contains 16 files – four synthesized files for each of the four completed implicates. The coefficient estimates are obtained as the average of the coefficient estimates from the 16 implicates. The squared standard error of an estimate equals 1) the average of the squared standard errors plus 2) the 1.25 times variance of the average coefficients across the completed implicates, minus 3) one fourth of the average of the variance of the coefficients within the four subsets of completed implicates. The subtraction of this last term can (and sometimes does) lead to a negative calculated variance.

4.2.2 Models Estimated

Earnings and employment transition models are estimated using data from 1990 to 1999. Three sets of models are reported here: log of earnings relative to the national average earnings for those with positive earnings, the probability of employment at time t given employed at t-1, and the probability of employment at time t given not employed at t-1.[3]

The earnings models can be written in the form:

$$y_{it} = \alpha + \underline{x}_{it}'\underline{\beta} + \delta_i + \varepsilon_{it}$$

$$\varepsilon_{it} = \rho\ \varepsilon_{it-1} + \upsilon_{it}$$

where y_{it} is the dependent variable; $\underline{x}_{it}$ is a vector of independent variables for person *i* at time *t*; δ_i is a random effect, assumed uncorrelated with $\underline{x}_{it}$; ε_{it} is the transitory component of the disturbance; and ρ is the autocorrelation parameter for the transitory component.

The earnings models are estimated in two stages. We use ordinary least squares regression to describe the relationship between the earnings measure and the set of independent variables. Then, we estimate the variance structure of the disturbances based on the residuals from this regression model. In Polisim, the variance structure is the main source of covariance in earnings over time in projected earnings.

The variance structure consists of three components (as compared with up to five components estimated in Polisim): 1) the standard deviation of the random effect (δ_i); 2) the standard deviation of the transitory component (ε_{it}); 3) an autocorrelation parameter(ρ).

We estimate the three variance components using the Stata procedure *xtregar*, which is intended to estimate cross-section-time-series regression models allowing fixed or random effects and autocorrelation. To obtain the standard deviation of the random effect, we regress the residual on a vector of ones and estimate a fixed-effects model with autocorrelation. The standard deviation of the fixed effect provides the estimate of the standard deviation of the random effect. The procedure directly provides an estimate of the autocorrelation procedure.

To obtain a standard error of the autocorrelation parameter, we use a second method. The method is as follows: Calculate a new variable equal to the residual (which varies by time

[3] The log of earnings is defined as log(earnings/national average earnings + 0.25). The addition of 0.25 serves to offset the shift in the distribution that occurs when the data are logged. This shift improves the linear estimator when the data (such as earnings) are highly skewed.

and person) minus the mean residual for each person. For cases with residuals in consecutive periods, regress the demeaned residual on the lag of the demeaned residual. The resulting coefficient on the lagged residual is the autocorrelation parameter, with a standard error given by the regression standard error.

The employment transition models are estimated using a probit model.

4.2.3 Analytic Subsamples

Each set of models is estimated for non-immigrant males and non-immigrant females. Samples include all annual records between 1990 and 1999 for which there is a positive weight and the individual was 1) at least age 25, and 2) not an OASI or DI beneficiary. The positive weight indicates that the person did not die by April 2000 and, in the case of the Gold Standard file, matched the administrative data. For the earnings analysis, we excluded all cases with earnings at least 40 times the national average. Although a minimum age is used to restrict inclusion in the equation, no restriction is placed on the maximum age of non-beneficiaries used in the estimation.

4.2.4 Independent variables

The independent variables in all models are:

1) Age is capped at age 85 and then measured as a spline for 25 – 40, 41 – 55, and 56 and above, with an additional 0-1 indicator for ages 62 – 64, 65 – 69, and 70 and above;

2) Indicators for five 10-year birth cohorts;

2) Whether black;

3) Whether Hispanic;

4) Education, indicators for high school, some college, college degree, and graduate degree;

5) Work history, including indicators of whether worked recently, number of the past ten years worked, and whether worked for all of the previous ten years.

For the comparisons between the Completed and Synthetic files, we also include three indicators of marital status (married, divorced, and widowed) and two indicators of change in marital status during the calendar year (got married and got divorced or widowed).

The specification of the independent variables is simpler than that used in Polisim. In particular, we eliminate many interactions (e.g., we exclude an interaction between education and the age spline) to make it easier to compare model parameters.

4.2.5 Additional notes on estimation of models

As noted earlier, all models used in the comparison across files are estimated without weights.[4] This corresponds to the estimation approach used in Polisim. The weights only appear in the analyses in that we exclude cases with a zero weight.[5]

[4] All models were estimated using Stata version 10.

To calculate the standard errors for models of log earnings and employment transitions from a single implicate, we use robust-cluster standard errors. This method calculates standard errors without imposing an assumption that the disturbances from the multiple observations per individual are uncorrelated. The standard errors from each implicate are then averaged across all implicates and added to the cross-implicate variation in coefficients as described above.

Standard error calculations for the models from the Completed and Synthetic files are performed using the Stata matrix language, Mata. As a test of the program, we calculated the mean and confidence interval for the proportion of Hispanics for those who completed high school without attending college and compared the results with those obtained by the SIPP/SSA/IRS team. The mean and confidence intervals both match (after accounting for an update to the weights).

The standard error calculations were further refined to deal with situations in which the formula led to a negative variance estimate. This occurred only in the Synthetic file earnings regressions for 5 out 29 of the coefficients for men and 1 out of 29 coefficients for women. (This is indicated by a zero in the column headed "Positive Variance" in the tables of Synthetic earnings estimates.) In this situation, we followed the recommendation of the SIPP/SSA/IRS team of setting the "average within-completed-implicate variance" equal to zero (this is the term that is subtracted off in the variance calculation). One difficulty with this solution is that it leads to a discontinuity in the calculation: A small change in the average within-completed-implicate variance can lead to a large change in the estimated variance. Another approach, which would seem preferable (although not taken in this work), would be to require the estimated variance to be at least as large as the average of the squared standard error from the individual runs. This approach would impose that the synthesis and completion process cannot improve upon the precision obtained from treating the data as though actual.[6]

An additional problem occurred when the degrees of freedom was calculated to be below three. In this case, we followed the suggestion of SIPP/SSA/IRS team and set the degrees of freedom equal to three. This occurred in the Synthetic file earnings regressions for 8 out of 29 coefficients for men and 13 out of 29 coefficients for women.

4.2.6 Method of comparing parameter differences across files

For each model and subgroup, we calculate the coefficients and confidence intervals for each of the three sets of data files. The confidence intervals are compared across data sources to see if they overlap. In addition, we calculate whether the point estimates from the Gold Standard

---

[5] We performed a side analysis to investigate whether weighting was likely to substantially affect the comparisons across files. In this analysis, we estimated weighted and unweighted versions of the models from the synthetic beta files. The confidence intervals for the coefficients of the weighted and unweighted earnings models always overlap; the intervals overlap for 94 percent of the coefficients in the employment models. A scan of the coefficients shows that coefficients are generally similar in the employment models. Exceptions include indicators of cohort, older age and Hispanic ethnicity.

[6] In the analyses based on the completed files, we occasionally find the disconcerting finding that the standard errors for the imputed data are smaller than from the Gold Standard data. This occurs when the amount of additional data added by completion of the data leads to an increase in precision that exceeds the variation added to the predictions by imputation.

file are within the confidence intervals based on the Completed file and whether those from the Completed file are in the confidence intervals of the estimates from the Synthetic Beta file. No attempt is made to properly calculate the standard error of the difference in the estimates across the multiple data sets.

4.3 Findings: Comparison of Earnings Model Coefficient Estimates across Files

We report the comparison of the models of log earnings across the Gold Standard and Completed files for non-immigrant males and females in Appendix D Tables 4-1 and 4-2.[7] In each table, we present in the columns from left to right: coefficients by source, 95 percent confidence intervals by source, indicators of whether the Gold Standard estimate is within the coefficient confidence interval for the Completed file and whether the coefficient confidence intervals overlap, standard errors by source. The excel version of the file also includes the degrees of freedom from the Completed file.

The comparisons across the completed and Synthetic Beta files for men and women are reported in Appendix D Tables 4-3 and 4-4. These tables present information parallel to that in Tables 4-1 and 4-2, supplemented by an indicator that the calculation of the standard errors led to a negative variance. Cases with a problem are marked by a zero in the column labeled "positive variance". The tables also include an indicator of a problem in the calculation of the degrees of freedom (indicated by a one in the column labeled "Small DOF"), and the degrees of freedom by source.

4.3.1 Comparison of earnings models in the Gold Standard and Completed files

The Gold Standard and Completed files yield models of log earnings for non-immigrant male and female non-beneficiaries that are generally similar across data source. Of the 46 regression Gold Standard coefficients reported across the two subgroups, 5 are outside of the confidence interval of the completed coefficients and 4 do not have overlapping confidence intervals. The largest differences are observed for work history measures in the model for males and are not typically very large in magnitude. For example, for males, the coefficient on the number of years worked out of the previous 10 years has a coefficient of 0.058 in the Gold Standard file as compared with 0.039 in the Completed files.

The pattern of variance components is also similar across models. The autocorrelation coefficients do not have overlapping confidence intervals; however, the differences are modest. For females, the autocorrelation from the Gold Standard file is 0.401 versus using 0.385 from the Completed file. In addition, the shares of variance associated with the random effect are quite similar across data sources.

4.3.2 Earnings Models in the Completed and Synthetic Files

In general, the completed and Synthetic files show similar patterns of earnings. However, many of the coefficients from the Completed file do not lie within the confidence interval of the Synthetic file coefficients and some of the confidence intervals do not overlap.

[7] All tables referred to in Chapter 4 are found in Appendix D.

For instance, in the regressions for non-immigrant males, 8 of 28 coefficients from the Completed files are outside of the confidence intervals for the Synthetic files and for 5 of these coefficients, the confidence intervals do not overlap. Some of the observed differences are fairly small or non-systematic.

Some notable differences are:

1. Coefficients on indicators for ages 62 to 64 and 65 to 69 are larger in the Completed file than in the Synthetic Beta file, for both men and women;

2. The effects of education are consistently smaller in the Synthetic file, especially for indicator of a graduate degree.

3. For women, the effects of recent employment are weaker in the Synthetic file.

The variables related to marital status, which are not available for the entire 1990 – 1999 period in the Gold Standard file, generally show similar patterns between the completed and synthesized files for men and women.

The pattern of the variance components shifts slightly across data sources. The proportion of variance associated with the random effect is somewhat smaller in the Synthetic files. This holds for both men and women. However, the confidence intervals of the autocorrelation coefficients from the completed and Synthetic files overlap for both men and women.

4.4 Findings: Comparison of Models of the Probability of Staying Employed

The models of the probability of staying employed – that is, the probability of employment for those working in the previous year – are reported in Appendix D Tables 4-5 to 4-8. The models based on the Gold Standard and Completed files are reported in Appendix D Tables 5 and 6, while the models from the completed and Synthetic Beta files are reported in Appendix D Tables 4-7 and 4-8.

4.4.1 Models of Staying Employed Using the Gold and Completed Files

With the Completed files, we can reasonably reproduce the Gold Standard models of the probability of staying employed for both non-immigrant males and females.

For males, the model results look very good: Only two Gold Standard coefficients are not within the confidence interval of the Completed file coefficient. The only variable for which the confidence intervals do not overlap is an indicator that the person worked two quarters earlier, which has a more positive effect in the Gold Standard model. (In fact, we might have expected more differences than this, given that the Completed file has 25 percent more observations due to the completion of the Gold Standard file.)

In the regression for females, many Gold Standard coefficients are out of the completed coefficient confidence interval. However, all but one of the Gold Standard coefficients has an overlapping confidence interval with the coefficients from the Completed file. More importantly, the differences are not substantively very large. For instance, the indicator for high school graduation has a coefficient of 0.184 in the Gold Standard file versus 0.155 in the Completed file, which translates into a difference in the effects of less than a half percentage point.[8]

4.4.2 Models of Staying Employed Using the Completed and Synthetic Beta files

Again, we see fairly good model agreements. However, as was the case with the earnings models, some differences are seen between completed and Synthetic Beta files: Approximately 30 percent of the coefficients from the Completed files are out of the confidence interval of the Synthetic Beta coefficients; 13 percent of coefficients do not have overlapping confidence intervals. These differences are observed for both men and women.

Among both men and women, we see coefficients on indicators of age showing a different pattern across files. In the Completed file, the indicators for ages 62 to 64, 65 to 69 and 70 plus show relatively flat coefficients; in the Synthetic file, these same indicators show a sharply increasing pattern. Each is measured relative to an age trend that is fairly similar across files.

Among males, the Synthetic Beta data show smaller effects of graduate degree, and to some extent college, than those in the completed data. For females, the effects of marriage, divorce and widowhood are much more negative in the Completed file than in the Synthetic file, while the effect of getting married is substantially larger in the Completed file.

4.5 Findings: Comparison of Models of the Probability of Becoming Employed

Finally, we to turn models of the probability of becoming employed – that is, the probability of employment for those not working in the previous year. The models based on the Gold Standard and Completed files are reported in Appendix D Tables 4-9 and 4-10, while the models from the completed and Synthetic Beta files are reported in Appendix D Tables 4-11 and 4-12.

4.5.1 Models of Becoming Employed Using the Gold and Completed files

The Gold Standard and Completed files show some important differences in the models of the transition to employment. This is to be expected as this sample of completed cases contains a disproportionate number of cases that did not match the administrative data. The completion process imputes earnings data for these non-matched cases and a disproportionate number of cases were assigned to be not employed.

This results in very different proportions of non-employed persons who become employed in the following year in the Gold Standard and completed samples. Among men, 27 percent of the Gold Standard sample and 15 percent of the completed sample become employed

[8] In interpreting the observed differences, the reader should keep in mind that these are coefficients on probit models and, therefore, translate into much smaller differences in probabilities. For example, a probit coefficient of .14 on high school completion (from one of the 16 implicates) translates into a .021 increase in the probability of employment.

in the following year. Among women, 20 percent of the Gold Standard sample and 14 percent of the completed sample become employed in the following year.

The model coefficients reported in Appendix D Tables 4-9 and 4-10 show very large differences in the constant for both men and women; this is consistent with the observed differences in proportion of unemployed persons who become employed between the samples. In addition, several other notable differences are observed among the covariates:

1) For both men and women, the effects of education beyond dropping out of high school are larger in the Gold Standard file than in the Completed file;

2) The coefficient on numbers of years in the labor market is much smaller in the Gold Standard file than in the Completed file;

3) For males, the coefficient on race is much more negative in the Gold Standard file than in the Completed file; and

4) For men and to some extent women, the coefficients on the age spline in the Gold Standard file show a somewhat more negative effect than are observed in the Completed file.

4.5.2 Models of Becoming Employed Using the Completed and Synthetic Beta Files

As can be seen in Appendix D Tables 4-11 and 4-12, the models from the Completed files are reproduced fairly well using the Synthetic Beta files. For males, the constant is less negative in the Synthetic file than in the Completed file, in keeping with the slightly lower proportion returning to employment among the synthetic sample. The differences between files on the coefficients on the covariates appear idiosyncratic and do not indicate a problem.

Chapter 5

Cross-Sectional and Longitudinal Comparisons of Completed and Synthetic Beta Files with Microsim and Gold

Karen Smith, Urban Institute

This chapter makes both cross-sectional and longitudinal comparisons between the Gold, Completed, and Synthetic Beta Files. One concern in this comparison is that we cannot know whether any differences between the Gold and Completed files are due to biases in the Gold file due to the incomplete nature of the data. These bias could be corrected in the completion process used to create the Completed files. To deal with this, we use the Microsim file, allowing it to serve as referee: Microsim is complete (within its universe) and thus provides a good outside benchmark for the comparisons.

We examine cross-sectional employment rates by age, sex, and year, as well as the distribution of cross-sectional earnings by year and sex. We also examine longitudinal earnings distributions using both AIME and a hypothetical individual account. We emphasize the longitudinal aspects of the comparability of the Synthetic Beta File with the Microsim, Gold, and Completed files. Much of the analysis is based on an extensive set of graphical comparisons that are included in Appendices E and F.

This chapter also examines an assortment of additional variables of interest for Social Security purposes (marriage duration, immigration age, disability status) and variables where simple summary statistics indicated significant differences between the Gold and Synthetic variables (health insurance coverage, poverty thresholds).

This chapter also discusses a significant limitation of **all** the time-varying variables available on the Synthetic files that are included at a single point in time. Finally, the chapter compares the correlation of SIPP self-reported with administrative annual earnings.

5.1 Background on Microsim

We use the Microsim dataset for evaluating earnings distributions on the Completed and Synthetic files. Microsim is a 1 in a 1000 sample of the Current Work History Sample (CWHS) for 2003 created by the Social Security Administration Actuaries. It is a restricted use file that was made available to the team for this evaluation. Like the Gold, Completed, and Synthetic files, Microsim includes Social Security covered earnings from the SER (1951-2005), total earnings (1978-2005), deferred earnings (1978-2005), and uncovered earnings (1978-2005) from the DER, and Social Security benefits from the MBR. Microsim, which draws its sample from the CWHS, does not suffer from bias associated with differential administrative earnings match rates. It provides a good target for cross-sectional and longitudinal earnings statistics for the Completed and Synthetic files. Its data are limited to basic demographic information, longitudinal earnings, and Social Security benefits.

There are some important differences in the Microsim and Completed samples that complicate the comparisons. First, Microsim represents the Social Security Area Population. The Completed and Synthetic files represent the U.S. resident non-institutionalized population. In many cross-sectional earnings comparisons, we look only at workers. Since most institutionalize individuals will not have earnings, the inclusion of the institutionalized population in Microsim will have a minimal effect. The inclusion of the non-U.S. resident population (military overseas, Puerto Rico, and other territories) is more insidious and will introduce some bias in the comparisons. Second, the Gold and Completed files are based on the U.S. resident population as of April 2000. They include no immigrants that arrive after the 1996 SIPP sample. They do not include individuals that die before the SIPP panels. To adjust for this sample inclusion, we limit the Microsim sample to U.S. residents in 1996 (based on immigration year) that survive to 2000.[9] Despite these differences, Microsim still provides the best available data source for comparisons of the Completed and Synthetic employment and earnings.

5.2 Overview of Results

Many of the comparisons presented in this report indicate the Completed and Synthetic files accurately reflect cross-sectional and longitudinal earnings by sex and year. However, in the course of this evaluation, we have discovered a few troubling results:

- Between 1978 and 2000, employment rates on the Completed and Synthetic files are lower than the employment rates on the Gold and Microsim files. The differences are larger for men than for women and are larger at younger ages than older ages.
- The Synthetic file includes several time-varying variables that are reported at a single point in time. Because interview date is not included on the Synthetic file, this omission prevents the user from teasing out the relationship between longitudinal earnings or age and the time-varying variables that are available at a single point in time: these include total net worth, home ownership, home equity, non-housing wealth, pension coverage, industry, occupation, disability status, and number of children under age 18, and to some extent education.
- The 2003 taxable maximum on all four Completed files and all 16 Synthetic files is $55,892. It should be $87,000.
- The 1958 taxable maximum on all 16 Synthetic files should be $4200. In a handful of cases on all 16 Synthetic files, the maximum is too high ranging from $4367.26 on Synthetic file 2.2 (implicate 2, replicate 2) to $4370.87 on Synthetic file 4.3 (implicate 4, replicate 3)
- The distribution of earnings between 1978 and 1984 are variable across the various data sets, reflecting the transition of only Social Security covered earnings from the SER to total earnings from the DER beginning in 1978. This transition is poorly captured in the Completed and Synthetic data.
- There is significant variation in the maximum annual earnings across files. In many cases, the annual maximum values on the SIPP-based files far exceed the maximum values on the Gold and Microsim files.
- There are some significant errors with Synthetic health insurance coverage variables that unintentionally reveal the SIPP panel and generate incorrect health insurance coverage rates.

[9] Microsim does not have an indicator for residency. It includes an indicator for the year of immigration for immigrants. It does not include indicators for military overseas, Puerto Rico, or territories residence.

- Synthetic file earnings are not consistent with the synthesized immigration year, causing a significantly larger share of immigrants than observed on the Gold file to have earnings prior to entering the United States.
- The Synthetic file overstates the share of individuals with marriages that last 10 or more years (the number of years of marriage needed to qualify for Social Security auxiliary benefits).
- The poverty thresholds on the Synthetic file are monthly values, rather than the annual values described in the documentation.

5.3 Evaluation of Population by Sex and Age

Appendix E Figure 5-1 compares the weighted number of men by age in 2000 for the Gold, pooled Completed, and revised pooled Synthetic files, and the 2008 Office of the Chief Actuary (OCACT) population values for 2000.[10] The observations are weighted using the provided 2000 decennial census weight (Decen_SIPP_wgt_04_01_2000). The population distribution by age for men is quite similar on all of the files: Gold, Completed, Synthetic files, and OCACT. The population size is lower on the SIPP-based files than OCACT at most ages. This largely reflects a difference in the universe between the files. The SIPP-based files represent the non-institutionalized, U.S. resident population while OCACT represents the Social Security Area Population. The Social Security Area population includes institutionalized, military personnel living over seas, Puerto Rico, and U.S. foreign territories that are not included in the SIPP population. The male population size is fairly variable by single year of age on all of the SIPP data compared to OCACT and the differences are larger at younger ages than older ages. The Synthetic file has more variation in the population size within single age groups than the Gold or Completed file (higher highs and lower lows). There is also a notable dip in the male population on the Synthetic files between ages 62 to 64 that is not present on the Gold or Completed files.

Appendix E Figure 5-2 shows the population age distribution for women in 2000. Like men, the number of women on the SIPP-based files is lower than OCACT, reflecting differences in the universe. The differences between the SIPP-base files and OCACT are larger at younger ages than older ages. The difference between the Synthetic and Completed files compared to the Gold file are also larger at younger ages than at older ages. The gap between the SIPP and OCACT population is larger for young men than for young women, reflecting the larger share of men than women that are institutionalized or in the military (included in the Social Security Area population but not in the noninstitutionalized U.S resident population). The opposite is true for older women who are more likely to be institutionalized (in a nursing home) than older men.

The Completed files over-rely on the birth date from the administrative data. In about one percent of cases the administrative birth date is more than four years different from the self-reported birth date (see Table 5-1). This can happen if a respondent misreports a Social Security number or the name and address match the wrong person in the household (e.g. match the son to the father's Social Security number). The minimum age difference is negative 101 (administrative birth date is more than a 100 years earlier than the SIPP birth date) and the

[10] Pooled Synthetic values are the average of the 16 Synthetic files, and pooled Completed values are the average of the four Completed files.

maximum age difference is 70. While this is a small percent of cases, they almost certainly are match errors and will cause some peculiar relationships between the SIPP-based variables and the administrative earnings variables.

Table 5-1. Distribution of the Difference in Administrative Birthdate and the SIPP Birthdate in Years on the Gold File.

Administrative-SIPP Birth Date (Years)	**Unweighted Count**	**Weighted Count (thousands)**	**Unweighted Column Percent**	**Weighted Column Percent**
Missing Administrative Birthdate	43,241	-	16.39	0.00
<= -5	2,252	920	0.85	0.45
-4	616	275	0.23	0.14
-3	733	393	0.28	0.19
-2	1,158	700	0.44	0.34
-1	5,745	4,076	2.18	2.00
0	206,180	194,491	78.16	95.32
1	2,366	1,916	0.90	0.94
2	532	461	0.20	0.23
3	261	246	0.10	0.12
4	158	126	0.06	0.06
5+	551	441	0.21	0.22
Total	263,793	204,045	100.00	100.00

Source: Urban Institute tabulations of the Gold file. Birth date differences are calculated in days and displayed in years.

5.4 Evaluation of Employment Rates

In this section, we compare the number of workers across files by gender and year. We also compare the employment rates by gender, year, and age. Appendix E Figure 5-3a shows the number of men with non-zero annual earnings (workers) from 1951 to 2003 for the Gold, Microsim, Completed, and Synthetic files. These calculations are limited to U.S. resident survivors to 2000, so they are not representative of all workers each year. The number of male workers is very similar across files from 1951 to 1977, but then deviates significantly after 1978. 1978 is the first year that DER-based earnings become available. The DER earnings include earnings in non-Social Security covered jobs (federal government, some State and Local government, workers covered by Railroad Retirement, and some student employment) that are not included on the SER. Both the Microsim and Gold files show an increase in the number of workers beginning in 1978 that is absent in the Completed and Synthetic data. The gap between the number of workers in the Gold versus Synthetic files declines from 8 percent in 1978 to 0 percent in 2000, and widens again after 2000. Microsim finds slightly more male workers than the Gold file due to its larger population coverage (Social Security Area population versus U.S. resident population). We believe that the close match of the Synthetic and Gold file employment rates in 2000 results directly from the reweighting procedure Census adopted in 2008 to improve the employment match to the Gold file. This adjustment forced the 2000 employment in the Synthetic and Gold files to line up. This weight adjustment improved the employment comparisons in 2000, but did little to close the gap in 1978. (The Synthetic data before the weight adjustment had employment rates similar to the unadjusted Completed rates shown in these figures).

The gap in employment beginning in 1978 is not due to the addition of uncovered earnings. Appendix E Figure 5-3b shows the number of men with covered earnings (based on totearn_ser_1951-totearn_ser_2003) from 1951 to 2003 among survivors to 2000. Covered employment rates drop sharply in 1978 in both the Completed and Synthetic files compared to both Gold and Microsim files. The Synthetic file has 21 percent fewer male covered workers in 1980 than the Gold or Microsim. As with total employment, the gap between Synthetic and Gold file narrows from 1978 to 2000, but remains below both Gold and Microsim in every year.

Appendix E Figure 5-3c shows the number of women with non-zero earnings from 1951 to 2003. Like men, the sharp increase in the number of workers observed on the Gold and Microsim files in 1978 are not present for either the Completed or Synthetic files. The gap between the Synthetic and Gold files narrows from 8 percent in 1978 to 0 percent in 2000, and widens again after 2000. As with men, female covered employment drops sharply in 1978 in both the Completed and Synthetic files, while it rises smoothly in the Gold and Microsim files from 1951 to 2000 (see Appendix E Figure 5-3d). Note also that the gap between the number of Microsim and Gold workers is different for men than for women. Microsim has more male workers than Gold, but fewer female workers than Gold. The result for men can be explained by the larger universe represented in Microsim, but the result for women is troubling.

Appendix E Figure 5-3e shows male total employment rates by age in 2000 for the Gold, Microsim, Completed, Synthetic, and 2000 March Current Population Survey (CPS). In all cases, employment is defined as having positive total earnings in 2000.[11] There is a fair amount of variation in employment rates among the files. The Synthetic and Gold files have quite similar age-specific employment rates, though the Synthetic files have significantly higher employment rates between ages 61 and 63 compared to the Gold file. The Completed files have significantly lower employment rates at all ages compared to all of the alternate files. Census adjusted the weights on Synthetic files to better account for employment. This adjustment significantly improved the employment rates on the Synthetic files, but the weights and employment rates remain unchanged in the Completed files. Employment rates at young ages vary a lot across data sources. Recall error for short-term employment contributes to the low values on the CPS, but Microsim also shows significantly lower employment rates at young ages compared to Gold. This can partly be explained by differences in the universe: Institutionalized individuals are less likely to work than non-institutionalized individuals. The non-institutionalized are included in the Microsim sample but not the SIPP sample. Comparing employment rates in prior years confirms this bias in the employment rates of young persons. For example, if a 20-year-old is living in a college dormitory in 1990, but has reentered the community by 2000, we observe the

[11] Total earnings is the sum of deferred and non-deferred FICA and deferred and non-deferred non-fica earnings from the DER. It includes Social Security covered and uncovered earnings from wage and salary and self-employment jobs and includes earnings over the Social Security taxable maximum.

30-year-old in the SIPP in 2000 but not in 1990. The longitudinal earnings fill in the historic data for the institutionalized in a way that we cannot observe in the 2000 cross-section.

Appendix E Figure 5-3f shows male employment rates by age in 1979 among survivors to 2000. Here employment rates on the Synthetic file are significantly lower than the employment rates on the Gold and Microsim files, and the differences are fairly uniform across ages. The Census Bureau's Synthetic file weight adjustment did little to correct the low employment rates in 1979. Comparisons of age-specific employment rates between 1978 and 2000 show that the gap in aggregate employment shown in Appendix E Figure 5-3a are due to lower employment at all ages as is seen in Appendix E Figure 5-3f, but the gap narrows between 1978 and 2000 as the aggregate employment gap narrows. Note also that the difference in 1979 employment rates between Microsim and Gold at young ages is gone as described above.

Male Synthetic employment rates align much more closely with Gold employment rates before 1978 than after 1978. Appendix E Figure 5-3g shows age-specific male employment rates in 1975 among survivors to 2000. Employment rates among the four data sources align quite closely, though Synthetic file rates are low compared to Gold at before age 24 and Microsim rates are low compared to Gold after age 60. The significant employment problems on the Synthetic and Completed files start in 1978 and persist though 2000, but at a declining rate on the Synthetic file as a result of the reweighting.

Appendix E Figures 5-3h, 5-3i, and 5-3j show age-specific female employment rates in 2000, 1979, and 1975 respectively. As with men, female age-specific employment rates on the Synthetic files in 2000 align closely with the Gold file except between ages 62 through 64 where the Synthetic files have slightly higher rates, but they are significantly lower in 1979. Employment rates on the Completed files are significantly lower than on the Gold or Microsim files in all years after 1977. Female employment rates before 1978 align quite closely among the four data sources. As with men, Synthetic female employment rates tend to be low compared to Gold before age 23, but the differences compared to Microsim at older ages for men are not present for women.

5.5 Evaluation of Cross-sectional Earnings

In this section, we compare the distribution of cross-sectional earnings of workers by gender and year for the Gold, Microsim, Completed, and Synthetic files. Generally, the cross-sectional earnings of workers on these files are quite similar for both men and women. However, there are some important differences with maximum values, especially for capped taxable earnings.

5.5.1 Male Total Earnings

Appendix E Figure 5-4a shows the distribution of total earnings (covered and uncovered) divided by the Social Security average earnings of male workers in 2000 for the Gold, pooled Completed, pooled Synthetic, and Microsim files. The distribution of total earnings is very similar through the 98th percentile. The distribution is slightly higher on the Synthetic file compared to the Gold and slightly lower on the Microsim file above the 75th percentile, but the differences are small. Appendix E Figure 5-4b shows the same distribution but for 1995. Here we limit the sample to U.S. residents that survive to 2000. Again, the distributions are similar. As with 2000, total earnings are slightly lower in Microsim and slightly higher on the Synthetic file compared to the Gold file. Here the differences in the Synthetic file compared to the Gold are larger below the 40th percentile and the gap above the 75th percentile is mostly gone.

Displaying the full set of graphical distributions for each age, sex, and data source would quickly become untenable. We calculate the footrule distance as follows to facilitate the comparisons:

$$d = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|$$

where d is the footrule distance, i is the base file, j is the alternate file, k ranges from 1 to 98 and represents the earnings at the kth percentile. The more similar the distributions are, the smaller the footrule distance will be. Identical distributions would generate a zero distance. The footrule distance calculation is sensitive to outliers with differences in maximum values having a large impact on the reported distance. We exclude values above the 98th percentile to reduce the impact of these outliers. We make four separate comparisons: Gold versus Synthetic, Microsim versus Synthetic, Gold versus Microsim, and Gold versus Completed. When the footrule distance reveals significant differences, we display the full distribution to facilitate interpretation of the results.[12]

Appendix E Figure 5-5 shows the footrule distance by year for men's total earnings relative to the average wage of workers from 1951 to 2003. In each cross section, the sample is limited to U.S. resident survivors to 2000. In virtually all years, the distribution of total earnings is most similar between the Gold and Completed files. This implies that the act of completing the file did not substantially alter the distribution of cross-sectional earnings. The footrule distance for the Gold compared to the Synthetic is always higher than the Gold compared to the Completed. This implies that the Synthetic files systematically have more differences in the earnings distribution to the Gold than do the Completed files. In other words, completing the file did little to change the distribution compared to the Gold, but synthesizing the file has systematically altered the distribution compared to the Gold and Completed files, but generally the differences are small.

The footrule distance increases sharply in 1978. This reflects the availability of deferred and uncovered earnings from the DER beginning in 1978. The higher earnings also significantly increases the differences in the distributions across the files. There is also a shift up in the footrule distance in the Gold compared to Completed file beginning in 1990 that reflects the

---

12 We also calculated the Spearman distance, Euclidian distance, and Bray Curtis distance similarity measures (Teknomo 2006). The alternate measures have different advantages and disadvantages, but the set of similarity measures agreed on the general goodness of fit identified using the footrule distance.

availability of less and uncensored Medicare earnings on the DER.[13] The differences are larger between the Synthetic files compared to both the Gold and Microsim than for Gold compared to Completed. Before 1978, the Synthetic male total earnings distribution is more similar to Microsim than to Gold. After 1978, the differences in the Synthetic distribution are more variable and usually larger compared to Microsim than to Gold.

The peak in the footrule distance in male total earnings in 1978 reflects differences in earnings primarily around the Social Security taxable maximum (see Appendix E Figure 5-6), which is markedly higher on the Synthetic file compared to the other data sources. Note also that the earnings are lower in Microsim along the earnings distribution compared to the other data sources. The gap between Gold and Microsim is larger in 1978 than in other years. The basic pattern of the male total earnings distribution shown in Appendix E Figure 5-6 is present in all years from 1978 to 1982, though the hump in the Synthetic file compared to the Gold shrinks and rises as the taxable maximum relative to the average wage increased from 1.67 in 1978 to 2.23 in 1982. The synthesizing process does not fully capturing the distribution of employment and earnings around the taxable maximum in the several years after uncapped and uncovered earnings first become available. As the footrule distance shows, other than the disturbance around the DER seam, the distribution of male total earnings on the Synthetic files reasonably closely align with both the Gold and Microsim files.

5.5.2 Female Total Earnings

Compared to male workers, the distribution of total income of female workers on the Synthetic file is more similar to the Gold and Microsim distributions after 1978 but more dissimilar before 1978. Appendix E Figure 5-7 shows the distribution of total earnings for female workers in 2000 on the Gold, Microsim, pooled Completed, and pooled Synthetic files. The population is limited to U.S. resident survivors to 2000. Female total earnings in 2000 are quite similar across the data files. As with men, female total earnings are slightly higher on the Synthetic files and slightly lower on Microsim compared to Gold, but the differences are quite small.

Appendix E Figure 5-8 shows the footrule distance for cross-sectional total earnings of female workers relative to the Social Security average wage from 1951 to 2003. Completed and Microsim female earnings are quite similar to the Gold female earnings. The differences are larger for the Synthetic and Gold earnings than for the Completed and Gold, but are smaller for women's earnings than for men's earnings. The distance increases in 1978 for the Gold compare to the Completed, reflecting the addition of deferred and uncovered earnings on the DER beginning in 1978. Interestingly, compared to Microsim and Gold, the differences in Synthetic earnings of workers are larger for women before 1978 than after 1978. The opposite is true for men. The maximum difference in female total cross-sectional earnings is in 1980, not 1978 as it was for men. Appendix E Figure 5-9 shows the full distribution of earnings of female workers relative to the Social Security average wage in 1980. The Gold and Completed file earnings are very similar, but the Synthetic earnings are higher than Gold and Microsim earnings across the distribution. The relationship changes in earlier years. Appendix E Figure 5-10 shows the

[13] Before 1991, Medicare and OASDI earnings were capped. Beginning in 1991, the cap for Medicare earnings was increases compared to OASDI earnings and in 1994, the cap for Medicare earnings was eliminated. The DER includes the higher Medicare earnings.

distribution of earnings of female workers relative to the average wage in 1951 among survivors to 2000. The Synthetic earnings are lower than Gold, Completed, and Microsim earnings above the 20th percentile (about 22 percent lower at the median compared to the Gold). The low Synthetic female earnings are present in virtually every year from 1951 to 1977.

5.5.3 Male Taxable Earnings

Social Security benefits are based only on taxable earnings: these are earnings in Social Security covered employment up to the taxable maximum. These earnings exclude covered earnings above the maximum and earnings in uncovered jobs. The administrative earnings on all of the comparison datasets are limited to taxable earnings before 1978. Beginning in 1978, administrative data from IRS tax forms become available. These data include earnings in both covered and uncovered jobs, and earnings above the Social Security taxable maximum. For Social Security purposes, only taxable (covered) earnings matter. Appendix E Figure 5-11 shows the footrule distance by year for taxable earnings among U.S. resident male workers that survive to 2000. As with total earnings, differences in taxable earnings among the alternate data sources increase significantly in 1978 and remain high between 1978 and 1983. Despite the fact that taxable earnings are capped, thus constraining differences among high earnings, the footrule distance is even higher for taxable earnings than for total earnings between 1978 and 1982. The footrule distance also rises sharply in 2003.

The difference in 2003 reflects a significant error in the taxable maximum on the Completed and Synthetic files. The taxable maximum in 2003 was $87,000 (2.55 times the average wage). No one on the Completed or Synthetic files has taxable earning in 2003 above $55,892 (1.64 times the average wage). Appendix E Figure 5-12 shows the full distribution of taxable earnings for men in 2003. The distributions on the Gold, Completed, and Synthetic files are quite similar up to the erroneously low cap on the Completed and Synthetic files. Appendix E Figure 5-12 also shows that taxable earnings are higher on the Gold, Completed, and Synthetic files compared to Microsim. The gap between Microsim and Synthetic taxable earnings begins to widen in about 2000, and it increases between 2000 and 2003.

Appendix E Figure 5-13 shows the full distribution of taxable earnings of male workers relative to the average wage in 1978. As with total earnings, the Synthetic taxable earnings distribution is significantly higher than on all of the alternate data sources. The higher Synthetic earnings distribution persists from 1978 to 1983 and is reflected in the higher than average footrule distance values for these years. But unlike 2003, which simply applied the wrong taxable maximum, 1978 to 1983 Synthetic taxable earnings are higher along the distribution than both the Gold and Microsim earnings. Covered employment rates in 1978 are also 10 percent lower on the Synthetic and Completed files than on the Gold and Microsim files in 1978. Most years the distributions of taxable earnings are markedly closer than those shown in Appendix E Figure 5-13. Appendix E Figure 5-14 shows the distribution of taxable income relative to the average wage in 1995. In 1995, the distribution of taxable earnings is fairly similar across files, though Synthetic taxable earnings are higher than the alternate files below the 40th percentile and closely match Gold earnings above the 40th percentile. Taxable earnings in Microsim are lower than the alternate files along the distribution.

5.5.4 Female Taxable Earnings

Appendix E Figure 5-15 shows the footrule distance for taxable earnings divided by the average wage for female workers for each of the comparison data sources. The values are the same as for total earnings before 1978 as the total income is limited to taxable income before 1978. As with male taxable earnings, the footrule distance show significant differences in the distribution of female taxable earnings between 1978 and 1983, and the sharp increase in 2003 is also present. The Synthetic and Completed files used the wrong taxable maximum in 2003. Because fewer women have capped earnings than men, the impact of this error in the overall distribution is smaller for women than for men. Appendix E Figure 5-16 shows the distribution of taxable earnings in 2003 of female workers relative to the average wage for each of the comparison data files. The distributions are all similar through the 90th percentile, but the Synthetic and Completed files are capped at 1.64 times the average wage rather than the correct 2.55 times the average wage ($87,000).

Appendix E Figure 5-17 shows the same distribution, but for 1979. Here, the high footrule distance in 1979 reflects the significantly higher distribution of taxable earnings of female workers on the Synthetic file compared to the other data sources. Female employment rates in covered jobs are also 12 percent lower in 1979 on the Synthetic files than on the Gold and Microsim files.

Overall, except for the assignment of some specific taxable maximum values, the Completed and Gold files have very similar cross-sectional distributions of total and taxable earnings of both male and female workers. This implies that the individuals without an administrative match receive similar cross-sectional earnings as those with a match.

However, employment rates are different between the Gold and Completed files and the differences in earnings distributions are larger when we account for employment differences. The largest differences in the cross-sectional earnings occur between 1978 and 1983 and appear to be related to a failure of the completion process to fully account for the differences in employment and earnings between the SER- and DER-based earnings.

The differences in the earnings distribution on the Synthetic files compared to Gold are almost always larger than the differences between the Completed and Gold files. Microsim, which provides a measure of earnings distribution without regard to the survey match, tends to have lower earnings than Gold, but in many years, the cross-sectional Synthetic total earnings fall between Gold and Microsim along the distribution.

5.5.5 Comparison of the Distribution of Total Taxable Earnings by Birth Year

Appendix E Figure 5-18a shows the 20th, 40th, 50th, 60th, and 80th percentile total taxable earnings on the Gold and Synthetic files in thousands of dollars (g indicates Gold and s indicates Synthetic in the key). The number of years of earnings included in this calculation differs by cohort. Individuals born in 1901 are age 50 in 1951. They have only 16 years of potential earnings by age 65 to include in their total earnings. Each later cohort is one year younger than the preceding cohort and includes one more year of earnings in the total. Individuals born in

1938 have earnings from age 16 in 1954 to age 65 in 2003. Individuals born in after 1938 have one fewer year of earnings in the total than the preceding cohort because they are one year younger. Individuals born in 1980 have longitudinal earnings for only 8 years (age 16 in 1996 to age 23 in 2003).

Total taxable earnings increases with cohort for cohorts born in or before 1901 to those born in the mid to late 1940s, and then falls for later cohorts. Total taxable earnings are slightly lower in the Synthetic files than the Gold file for each cohort at each segment of the total earnings distribution. This implies that the Synthetic longitudinal earnings are systematically lower than the Gold longitudinal earnings, but the differences are small. Differences in the total earnings on the Synthetic compared to Microsim are smaller than Synthetic compared to Gold (see Appendix E Figure 5-18b). This implies that the synthesized earnings may even improve the total longitudinal earnings distribution compared to the Gold by adding earnings to individuals without an administrative data match.

5.6 Evaluation of AIME by Birth Year and Sex

In this section, we examine longitudinal earnings by sex and cohort. We use a measure of longitudinal earnings modeled after the average indexed monthly earnings (AIME) calculation used to determine Social Security benefits, but we modify it slightly to facilitate comparisons of the disabled with non-disabled and cohorts born before and after 1917. The AIME calculation includes the average of the top 35-years of wage-indexed taxable earnings through age 62, constrained by the 1951 to 2003 data limitation.[14] Unlike the official AIME, our version does not adjust the number of computation years for the disabled. It also calculates AIME for individuals born before 1917, even though benefits for these individuals are based on the old average monthly earnings (AME) formula.

AIMEs generally increase with birth cohort, reflecting the wage indexing of earnings in the formula, but decline for cohorts born after 1941 (62-year-olds in 2003) because they have not completed a full career of earnings (they have not yet attained their peak career earnings). Appendix E Figure 5-19a shows the 20th, 40th, 50th, 60th, and 80th percentile of AIME by birth year for the Gold and pooled Synthetic files among survivors to 2000. As with total taxable earnings, AIMEs are lower on the Synthetic files compared to the Gold file, but the differences are small. The differences are larger for the middle cohorts that include more years of earnings than early and late cohorts. Synthetic AIMEs also have a different pattern compared to the Gold for individuals born from 1971 to 1980. Gold AIMEs monotonically fall for these cohorts, but Synthetic AIMEs jump up for the 1971 cohort, followed by a sharp drop for the 1973 cohort and then fall, but at a shallower slope compared to the Gold file. Appendix E Figure 5-19b shows the same distribution but for Microsim compared to the Synthetic files. As with total taxable earnings, the differences between Microsim and Synthetic are smaller than the differences between Gold and Synthetic. The peculiar AIME pattern for individuals born after 1970 is even more pronounced compared to Microsim.

---

[14] Earnings are wage indexed to the year the individual turns age 60 or to 2001 for individuals younger than age 60 in 2003. AIME includes nominal earnings after age 60 or 2001. The computation years are the number of years from age 16 to the age attained in 2003 up to a maximum of 35 years.

5.6.1 Male AIME

Appendix E Figures 5-20a through 20k show the full distribution of AIME for Microsim, Gold, pooled Completed, and pooled Synthetic for men that survive to 2000 by five-year birth cohort groups from 1926 to 1980.

For men born from 1926 to 1930 (Appendix E Figure 20a), Synthetic male AIMEs are significantly higher than the comparison data sources below the 30th percentile, but then fall below the comparison data sources above the 42nd percentile. Synthetic AIME is about 65 percent higher than Gold at the 10th percentile and about 7 percent lower than both Gold and Microsim at the 70th percentile. The maximum AIME is constrained by the Social Security taxable maximum.

The differences are slightly smaller for men born from 1931 to 1935 (Appendix E Figure 5-20b), but Synthetic AIMEs are still higher than the Gold and Microsim below the 28th percentile and lower than Gold and Microsim above the 28th percentile. Synthetic male AIME is about 46 percent higher than Gold at the 10th percentile and about 7 percent lower than Gold and Microsim at the 70th percentile.

The gap between Synthetic and Gold male AIMEs at the bottom of the distribution narrows for subsequent cohort, and Synthetic male AIMEs fall below Gold AIMEs for cohorts born after 1950. Synthetic male AIMEs remain below Gold above the 28th percentile in virtually all cohorts, but the gap above the 28th percentile narrows and is generally within 5 percent of the Gold values for cohorts born after 1950. The gap between Microsim and Gold increases for the middle cohorts (those born from 1941 to 1960) compared to cohorts born before or after. This reflects the larger number of years included in the calculation for these cohorts. Generally, Synthetic male AIMEs fall between the Gold and Microsim values along the distribution for these middle cohorts.

For men born from 1976 to 1980, Synthetic AIMEs are significantly higher than Gold AIMEs throughout the distribution (11 percent higher at the 20th percentile and 12 percent higher at the 80th percentile).

5.6.2 Female AIME

Appendix E Figures 5-21a through 5-21k show the full distribution of AIME for Microsim, Gold, pooled Completed, and pooled Synthetic for women that survive to 2000 by five-year birth cohort groups from 1926 to 1980.

For women born from 1926 to 1930 (Appendix E Figure 5-21a), Synthetic female AIMEs are fairly similar to the comparison data sources throughout the AIME distribution, though Synthetic AIMEs for women born from 1926 to 1930 are higher than Gold AIMEs below the 60th percentile and lower than Gold AIMEs above the 60th percentile. Gold and Microsim AIME distributions quite similar. Synthetic values drop below both Gold and Microsim above the 75th percentile, with Synthetic female AIMEs about 10 percent lower than Gold AIMEs between the 90th and 98th percentiles.

AIME distributions for women born from 1931 to 1935 and from 1936 to 1940, like those born from 1926 to 1930, are higher than Gold at the bottom of the distribution and lower than Gold above at the above about the 20th percentile. The gap between Synthetic and Gold AIME distributions for women born after 1940 narrows compared to earlier cohorts with Synthetic AIMEs generally lower than Gold AIMEs but usually by less than 10 percent.[15] As with men, Synthetic female AIMEs for individuals born from 1976 to 1980 are significantly higher than both Gold and Microsim AIMEs.

Appendix E Figures 5-22a and 5-22b show Q-Q plots for male and female AIMEs respectively. The Q-Q plots compare the AIME distribution of the Synthetic file compared to both Microsim and Gold for 5-year cohort groups. Each Q-Q plot includes the OLS estimated slope and R-squared. The Q-Q plots confirm the significantly higher male AIMEs on the Synthetic file in the bottom third of the distribution compared to both Microsim and Gold for early cohorts (born before 1941). Again, we see that Synthetic AIME distributions for men born from 1941 to 1975 align closely with both Microsim and Gold, and Synthetic male AIMEs for men born from 1976 to 1981 are significantly higher than both Microsim and Gold.

Table 5-2 shows a summary of Q-Q plot distributions for Synthetic AIMEs compared to Microsim and Gold AIMEs by cohort and sex. All of the R-squared values are very close to one. Compared to Microsim, the Synthetic Q-Q slopes are less than one (Synthetic AIMEs lower than Microsim AIMEs) for females born before 1955 and greater than one for females born after 1955 (Synthetic AIMEs higher than Gold AIMEs). The slope increases significantly for women born from 1976 to 1981. The Q-Q statistics are similar for Synthetic versus Gold as for Synthetic versus Microsim. The slope deviates more from one for early cohorts compared to the Gold distributions than the Microsim distributions. The patterns are similar for men as for women including the dramatic increase in slope for men born between 1976 and 1981.

[15] The percent differences are larger at the bottom of the distribution (below the 20th percentile), but the denominators are small.

Table 5-2:Slope and R-Squared Values for Q-Q Plots for Synthetic AIME Distributions Compared to Microsim and Gold Files by Birth Year and Sex

	Microsim		Gold	
Birth Year	**Slope**	**R-Squared**	**Slope**	**R-Squared**
Female				
1926 – 1930	0.954	0.9975	0.929	0.9983
1931 – 1935	0.932	0.9978	0.932	0.9994
1936 – 1940	0.966	0.9989	0.942	0.9992
1941 – 1945	0.962	0.9992	0.949	0.9994
1946 – 1950	0.968	0.9997	0.948	0.9996
1951 – 1955	0.966	0.9999	0.956	0.9996
1956 – 1960	1.008	0.9997	0.982	0.9996
1961 – 1965	1.002	0.9985	1.004	0.9995
1966 – 1970	1.022	0.9993	1.000	0.9979
1971 – 1975	1.037	0.9992	1.013	0.9990
1976 – 1981	1.208	0.9946	1.200	0.9866
Male				
1926 – 1930	0.968	0.9652	0.956	0.9809
1931 – 1935	0.964	0.9912	0.952	0.9904
1936 – 1940	0.987	0.9958	0.973	0.9971
1941 – 1945	0.980	0.9989	0.957	0.9985
1946 – 1950	1.001	0.9958	0.949	0.9987
1951 – 1955	0.995	0.9980	0.966	0.9995
1956 – 1960	1.011	0.9959	0.978	0.9994
1961 – 1965	1.050	0.9982	0.988	0.9990
1966 – 1970	1.044	0.9992	0.998	0.9968
1971 – 1975	1.054	0.9982	1.001	0.9974
1976 – 1981	1.182	0.9971	1.133	0.9988

Source: Urban Institute tabulations of pooled Synthetic, Microsim, and Gold file AIME Q-Q distributions. The fitted lines use an intercept of 0.

5.6.3 Longitudinal Measure of Earnings Not Captured by the AIME Calculation

To further examine longitudinal earnings, we generated a hypothetical individual account to measure the effect of both the level and timing of earnings that are not captured in the AIME calculation.

Suppose B is the account balance, r is the annual rate of return on the balance, c is the annual contribution, t is the current period.

$$B_t = \sum_{t=1951}^{2003}((B_{t-1} * r_{t-1}) + c_t)$$

While the account balance will depend on the rate of return and contribution rate, for comparison purposes, the specific values do not matter. We have set r at a 3.3 percent nominal return and c to be 4 percent of the social security taxable earning. The calculation is limited to include only earnings from age 16 to 65 (inclusive) between 1951 and 2003. Unlike AIME, the individual account is not limited to the top 35 years of earnings. It also does not index earnings to wage growth. Individual accounts based on earnings early in a career have more years to compound than earnings late in a career. Because the individual account is not limited to the top 35 years of earnings, a long career worker will have a higher individual account than a similar worker with a shorter work history. Individuals with earnings early in their lifetime will have a higher individual account balance than a similar individual with earnings late in their lifetime.

Individual accounts generally increase with birth cohort, reflecting the higher number of years of earnings in the formula, but decline for cohorts born after 1941 (62-year-olds in 2003) because they have not completed a full career of earnings (they have not yet attained their peak career earnings). Appendix E Figure 5-23a shows the 20th, 40th, 50th, 60th, and 80th percentile of individual account by birth year for the Gold and pooled Synthetic files among survivors to 2000. As with total taxable earnings and AIMEs, Individual accounts are lower on the Synthetic files compared to the Gold file, but the differences are small. The differences are larger for the middle cohorts that include more years of earnings than early and late cohorts. Synthetic individual accounts do not have the funny pattern compared to the Gold for individuals born from 1971 to 1980 that was present for AIMEs (Appendix E Figure 5-19a and 5-19b). Instead, Gold and Synthetic file individual accounts both monotonically fall for these later cohorts. Appendix E Figure 5-23b shows the same distribution but for Microsim compared to the Synthetic files. As with total taxable earnings and AIME, the differences between Microsim and Synthetic individual accounts are smaller than the differences between Gold and Synthetic accounts.

Examination of the full distribution of individual accounts by sex and cohort mimic the results found for AIME distributions. While the metric is different, the patterns are the same. Synthetic individual account distributions tend to be high at bottom of the distribution (below the 34th percent) and low at the middle of the distribution compared to Gold and Microsim for cohorts born before 1940. The differences are larger for individuals born before 1941 and after 1975 than for individuals born between 1941 and 1975. Synthetic individual account balances are generally within 10 percent of the Gold balances throughout the distribution. Synthetic individual accounts are about 7 percent lower than Gold among individuals born before 1976, but then shift to being about 15 percent higher than Gold for individuals born from 1976 to 1980.

Appendix E Figure 5-24a and 5-24b displays Q-Q plots of individual account balances for both men and women by cohort group respectively. These Q-Q plots compare Synthetic individual account distributions with both Microsim and Gold distributions. Each plot shows the estimated linear regression slope and R-squared value. Table 5-3 summarizes the slope and R-squared for each cohort group and sex. Most R-squared values are close to one. The slope tends to be less than one for women born before 1951 and greater than one for women born after 1960. The slope significantly deviates from one for women born from 1967 to 1981 compared to both the Microsim and Gold files. The patterns are similar for men. As with AIMEs, the Q-Q plots of individual accounts significantly deviate from the fitted line at the bottom of the distribution for men born before 1945. Synthetic individual account distributions for women quite similar to Microsim and Gold distributions for all but the first and last cohort groups that include the least amount of data.

Table 5-3:Slope and R-Squared Values for Q-Q Plots for Synthetic Individual Account Balance Distributions Compared to Microsim and Gold Files by Birth Year and Sex

Birth Year	Microsim		Gold	
Female	Slope	R-Squared	Slope	R-Squared
1926 - 1930	0.976	0.9927	0.928	0.9987
1931 - 1935	0.944	0.9916	0.931	0.9978
1936 - 1940	0.971	0.9930	0.925	0.9995
1941 - 1945	0.950	0.9969	0.934	0.9986
1946 - 1950	0.973	0.9937	0.930	0.9999
1951 - 1955	0.978	0.9971	0.957	0.9996
1956 - 1960	1.031	0.9975	0.978	0.9998
1961 - 1965	1.037	0.9862	1.010	0.9992
1966 - 1970	1.070	0.9754	1.015	0.9953
1971 - 1975	1.128	0.9668	1.033	0.9982
1976 - 1981	1.355	0.9043	1.196	0.9952
Male				
1926 - 1930	0.975	0.9524	0.951	0.9769
1931 - 1935	0.964	0.9831	0.945	0.9899
1936 - 1940	0.977	0.9933	0.962	0.9953
1941 - 1945	0.973	0.9970	0.950	0.9974
1946 - 1950	1.001	0.9953	0.947	0.9991
1951 - 1955	1.000	0.9984	0.968	0.9997
1956 - 1960	1.030	0.9946	0.981	0.9995
1961 - 1965	1.085	0.9877	0.997	0.9988
1966 - 1970	1.080	0.9906	1.017	0.9917
1971 - 1975	1.151	0.9541	0.972	0.9846
1976 - 1981	1.310	0.9658	1.111	0.9974

Source: Urban Institute tabulations of pooled Synthetic, Microsim, and Gold file individual account Q-Q distributions. The OLS fitted lines use an intercept of 0.

5.7 Evaluation of Number of Work Years

In this section, we examine the distribution of work years from age 16 to 65 (inclusive) between 1951 and 2003 by cohort, sex, and source file among U.S. resident survivors to 2000. Work years are based on the number of years with non-zero Social Security covered earnings. Early cohorts have their calculated number of work years limited by the 1951 data constraint, and later cohorts have work years limited by their age.

5.7.1 Male Work Years

Appendix E Figures 5-25a though 5-25k show the distribution of total male covered work years by cohort for the Gold, Microsim, pooled Synthetic, and pooled Completed data files. Appendix E Figure 5-25a shows the distribution of work years for men born from 1926 to 1930 (ages 25 to 21 respectively in 1951). The distribution of work years across data sources is similar. The Synthetic files have fewer men than both Gold and Microsim with fewer than 10 years of covered earnings needed to qualify for Social Security benefits, fewer men with 40 or more years of earnings, but more men with 25 to 34 years of earnings. This employment difference should lower calculated Synthetic AIMEs for men at the bottom of the AIME distribution compared to Gold, but calculated Synthetic AIMEs are in fact significantly higher (Appendix E Figure 5-20a). This implies that the differences in AIMEs of men born from 1926 to 1930 are due to both errors in the work years and errors in the earning amounts.

Differences in the distribution of male work years between the Synthetic and Gold files are smaller for men born after 1935 than before 1935. In fact the distribution of Synthetic male work years generally aligns more closely with Microsim than with Gold, but Microsim and Gold align very closely, so when there are differences between Gold and Synthetic work years there are also differences between Gold and Microsim. The general trend is for the Synthetic files to have too few long career workers and too many mid-length career workers across all of the cohorts compared to both Microsim and Gold, but the differences are small.

5.7.2 Female Work Years

Appendix E Figures 5-26a through 5-26k show the distribution of total female covered work years by cohort for the Gold, Microsim, pooled Synthetic, and pooled Completed data files. Like Appendix E Figures 5-25a to 5-25j, they are based on the number of years with positive covered earnings from age 16 to 65 inclusive among U.S. resident survivors to 2000. As with male work years, the distribution of female work years on the Synthetic files aligns fairly closely with both Gold and Microsim. While male work years are highly skewed towards long careers, the opposite is true for female work years among early cohorts. The share of women with longer careers increases for women born after 1930 compared to women born before 1930, reflecting the increased female labor force participation among women over time. While the distribution of male work years was very similar between Microsim and Gold, the differences are larger for female work years. Generally, the distribution of Synthetic female work years more closely aligns with Microsim than with Gold. The difference in the Synthetic cumulative distribution of female work years compared to both Microsim and Gold is generally less than 4 percentage points for the various cohorts. The exceptions are for women born between 1961 and 1970 where the differences are larger; the Synthetic files tend to have more short-career female workers than both Gold and Microsim. This reflects the persistently lower 1978 to 1999 employment rates on the Synthetic file compared to the Gold and Microsim files.

5.7.3 Reemployment Hazard

The errors in Synthetic file employment rates described above appear to be due to low reemployment hazard rates beginning in 1978. Figure 5-27a and 5-27b show the share of men and women respectively that work each year given that they did not work in the prior year (reemployment hazard) from 1952 to 2003 among men that survive to 2000 for the Gold, Microsim, Completed, and Synthetic files. Prior to 1978, employment is based only on Social Security covered earnings available from the SER. Beginning in 1978, employment is based on total earnings available from the DER. The reemployment hazard includes spikes in 1955 and 1957 that reflect changes in Social Security coverage due to the addition of self employed, active duty military, and some state and local government workers to the covered work force.[16] The spike in 1978 reflects the change in the data time series that added uncovered earnings. Uncovered workers in 1977 would have been classified as not working in 1977, but working in 1978. All four data files find this spike, but the size of the spike is about 3 percentage points lower for men in the Completed and Synthetic files than in the Microsim and Gold files. After 1978, the reemployment hazard is about four percentage points lower for men (2 percent for women) on the Synthetic and Completed files compared to the Gold and Microsim files. This lower reemployment hazard contributes to the lower post 1978 employment rates found on the Synthetic files and contribute to longer unemployment spells on the Synthetic files compared to the Gold file.

These longer unemployment spells have only a small impact on AIME, individual account, and work year distributions. Synthetic work histories are generally shortened by one or two years. These censored years may have no or very little impact on AIMEs due to drop-out years. They do lower calculated work years and individual account balances and we see this in the distribution comparisons, but the impact over a lifetime is relatively small. While the overall longitudinal impact is small, this error should be corrected in future versions of the synthetic data.

5.8 Other Key Comparisons: Health Insurance, Immigration Status, Marriage Duration, Poverty, Disability.

5.8.1 Health Insurance Coverage

The values for health insurance coverage on the Synthetic and Completed files have significant errors that understate the annual share of individuals in the U.S. with health insurance coverage. The health insurance variables have not been completed on the Completed files for years beyond the range of the actual SIPP panel years. The errors are different for the two types of insurance coverage on the files: any coverage and employer coverage.

For **any** coverage (Hicovannual1990-Hicovannual1999), the Completed file contains missing values for the non-panel years (see Appendix E Table 5-4a). Annual health insurance coverage rates based on the non-missing values are similar between the Gold, Completed, and Synthetic files (see Appendix E Table 5-4b), but the Completed files include non-missing data for less than half the population in most years. The missing values in the Completed files are then repeated with noise added on the Synthetic files. This means that the attentive user can generally identify SIPP source panel on the Synthetic files.

[16] See Table 2.A.1 of the Annual Statistical Supplement to the Social Security Bulletin, 2001. Social Security Administration, Washington DC.

For **employer** coverage (Hiempannual1990-Hiempannual1999), the Completed file contains zeros rather than missing values for all non-panel years (see Appendix E Table 5-5a). Average coverage rates for **employer** coverage are significantly understated on both the Completed and Synthetic files as they contain a disproportionate number of false zeros (see Appendix Tables 5-5b). For example, 45 percent of surviving respondents in 1990 have employer-sponsored health insurance on the Gold file, but only 8 percent do on the Completed. The errors in the Completed files are repeated with noise on the Synthetic files, yielding Synthetic annual employer coverage rates of between 10 and 23 percent instead of the 38 to 45 percent observed on the Gold file.

5.8.2 Immigrants

A much higher share of immigrants have earnings prior to their immigration year on the Completed and Synthetic files compared to the Gold file. About 7 percent of immigrants have pre-arrival earnings on the Gold file, while about 35 percent do on the Synthetic files (see Appendix E Table 5-6).[17] For individuals immigrating to the United States for the first time, we expect to see no earnings in the U.S. before the immigration year. The 7 percent of immigrants with pre-immigration earnings on the Gold file could reflect multiple immigration spells in the U.S., with the SIPP reported immigration year reflecting the most recent immigration spell. It could also reflect errors on the Gold file due to Census imputations for missing values that did not account for historic earnings. Regardless of the cause of pre-immigration earnings in the Gold, the rate is significantly higher on both the Completed and Synthetic files. This error will overstate Social Security eligibility and benefits among immigrants, as the data include more covered quarters and higher average earnings for these workers. It also harms the face validity of the file.

5.8.3 Marriage Duration

The Completed and Synthetic files overstate the share of individuals with a marriage that lasted 10 or more years compared to the Gold file. Marriage duration is important for Social Security since individuals married for at least ten years are qualified to claim spouse and survivor benefits based on their former spouse's earnings record. Overall the difference in the share of individuals with a qualified marriage is about 3 percentage points higher on the Completed and Synthetic files compared to the Gold file, but certain subgroups have significantly larger differences. Blacks have 11 percentage points more qualified marriages on the Completed and Synthetic files than the Gold file (see Appendix E Figure 5-28a). Individuals between ages 30 and 49 have about 8 percentage points more qualified marriages in the Completed and Synthetic files compared to the Gold file (see Appendix E Figure 5-28b Appendix E Table 5-7).

The trend for longer marriages on the Completed file compared to the Gold file is not solely due the completion process. The marriage durations on the Completed file are different and usually longer than on the Gold file, even among records with a valid marriage history

[17] Immigration year is given as a range of years. These calculations are based on having historic earnings before the earliest year in the range.

topical module on the SIPP. Rather than simply replicating the values observed from the marriage history topical module, the Completed file replaced (synthesized) the marriage history. In many cases, the differences between marriage durations on the Gold and Completed files are very large. The age at first marriage generally agrees between the Gold and Completed files, but age at marriage terminations and subsequent marriages are different. Some marriages on the Completed files reflect dates after 2003.

Large shifts in the synthesized marital histories compared to actual histories can have large implications for both annual family income and Social Security benefits. Because the marital status of an individual on the Synthetic file does not necessarily coincide with that on the Gold file, family earnings records will not accurately reflect the earnings of individuals within the household. The Synthetic file would not accurately reflect the effects of divorce, marriage, or widowhood on individual earnings. Overstating the share of qualified marriages will overestimate the number of individuals eligible for Social Security spouse and survivor benefits.

5.8.4 Poverty Thresholds

The poverty thresholds on the Gold, Completed, and Synthetic files do not make sense with respect to the poverty thresholds determined by the Census. On both the Gold and Completed files, the poverty thresholds are extremely high and look as if the annual value was multiplied by 12. The Synthetic file has the opposite problem—the values on the file appear much too low. We assume the values represent monthly values, rather than annual values. We adjusted the annual poverty thresholds by dividing the reported poverty thresholds by 12 on the Completed and Gold files and by multiplying the thresholds by 12 on the Synthetic files (see Appendix E Table 5-8).

5.8.5 Disability

The SIPP collects a self-reported measure of health limitations and among those reporting a limitation, the SIPP asks if the limitation prevents work. The Synthetic and Completed files accurately reflect work limitation rates compared to the Gold file by age, but less so by lifetime earnings. Appendix E Figure 5-29 shows the share of individuals that report a health condition that limits the amount or type of work in aggregate and by age. Overall 11 percent of the samples report a work limitation on the Gold, Synthetic, and Completed files. The difference between the Gold and alternate files is no more than one percentage point in any broad age group.

By lifetime earnings quintile, the differences in the work limitation rates between the Gold and Synthetic files are larger. The Synthetic and Completed files have a significantly smaller share of individuals with a work limitation in the bottom lifetime earnings quintile and a higher share of individuals with a work limitation in the higher lifetime earnings quintiles (Appendix E Figure 5-30).[18] For example, 23 percent of individuals in the bottom lifetime earnings quintile have a work limitation on the Gold file, while only 17 percent do on the Synthetic files (18 percent on the Completed files). While work limitation rates decline as lifetime earnings rise on all of the files, work limitation rates are about two percentage points higher on the Synthetic file compared to the Gold file in the highest two earnings quintiles.

---

[18] Lifetime earnings quintiles are based on the sum of Social Security taxable earnings from age 16 to 65. The quintile is calculated based on the cohort-specific distribution.

Among individuals that report having a work limitation, the share whose limitation prevents work is lower on the Synthetic files compared to the Gold file, both by age and by lifetime earnings quintile. Half of individuals who report a work limitation on the Gold file report that their limitation prevents them from working. This share is only 43 percent on the Synthetic files. The difference between the Gold and Synthetic rates persists throughout the age distribution (see Appendix E Figure 5-31). The difference is also present by lifetime earnings quintile. The share whose health limitation prevents work is higher for low lifetime earners than high lifetime earners. The share whose work limitation prevents work is lower on the Synthetic file compared to the Gold file in all lifetime earnings groups, but the gap as a percent of the limited population is larger for individuals in the bottom and top earnings groups than in the middle of the distribution (see Appendix E Figure 5-32)

5.9 Wealth and Time-Varying Variables.

The Synthetic files include a set of time-varying variables that are reported at a single point in time. Unfortunately, the Synthetic files do not include the information on when in the decade the variables are measured, making them significantly less useful than they otherwise could be. (Although the variables are ascribed to the year 2000, the Completed file data actually come from the year between 1990 and 1999 when the survey data for the individual was collected; they are assumed to apply to the year 2000, with correction for inflation.) This section describes the problem using total net worth, but this problem is true for all of the time-varying variables. These include net worth, home equity, non-housing wealth, pension coverage, industry, occupation, disability status, and number of children under age 18.

The prime example of this issue is total net worth. Total net worth inflated to 2000 price adjusted dollars is available on the Synthetic files but is collected in different years depending on the source panel. Failure to account for time prevents the user from teasing out the life-cycle saving and spend-down behavior of different cohorts, nor does it allow the user to relate the net worth to the included longitudinal earnings that are the strength of the Synthetic file. Individuals typically save when they are young and working to support consumption when they are older and retired. We can observe this life-cycle saving behavior on the Gold file by examining net worth by cohort and panel. Younger individuals have lower net worth in 1990 than in 1996, reflecting their asset accumulation as they age.[19] Similarly, older individuals have higher net worth in 1990 than in 1996, reflecting their asset spend-down in retirement (see Appendix E Figure 5-33a). Real median net worth (in 2000 price-adjusted dollars) falls over time for individuals born before 1926 and rises over time for cohorts born between 1935 and 1965.

The peculiar declining net worth for individuals born from 1971 to 1975 reflects the change in living arrangements of these individuals between 1990 and 1996. The net worth variable included on the SIPP data is household net worth. Individuals born in 1975 are 15 years old in 1990 and 21 years old in 1996. The 1990 SIPP panel data reflects the household assets including these young individuals' parents. Their household assets decline over time as these younger individuals age and leave home. The net worth value in the later panels are more likely

[19] Net worth on the Gold file is in nominal dollars. We price adjusted net worth on the Gold file using the same adjustment factors used to adjust the Completed files by panel.

to include only the lower assets of the independent younger individual. Indeed, by the 1996 panel, the median net worth of the youngest cohorts is lower than their predecessors as expected.

Appendix E Figure 5-33b and 5-33c show the same information as Appendix E Figure 5-33a, but for the Completed and Synthetic files respectively. The Completed and Synthetic files show the same saving behavior as the Gold file over time, though the pattern is much less prevalent on the Synthetic file than either the Gold or Completed files. Unfortunately, the panel variable is not included on the Synthetic file, so there is no way to tease out the time component of saving and spend-down based on the data included on the Synthetic file. This is true of **all** the time varying variables that are reported at a single point in time on the Synthetic files. Not only do these variables not vary over time on the Synthetic files, the user does not know what point in time the variables represent. It is not possible to accurately relate the status of these variables with the longitudinal earnings data that are the strength of the Synthetic data files.

Appendix E Table 5-9 shows the ratio of total net worth on the pooled Completed and pooled Synthetic files compared to the Gold file by panel and in aggregate. The ratios are far from one in most cohorts and in most years. The ratios for the Completed and Synthetic are also different from each other. If the difference between the Completed and Gold were due to the completion of the file, then we would expect that the Synthetic and Completed files would have similar ratios relative to the Gold file. The Completed and Synthetic ratios, however, are quite different. For example, the ratio of Completed to Gold for individuals born from 1941 to 1945 pooled by panel is 1.05. It is 0.88 for the Synthetic files. For individuals born from 1976 to 1981 the ratio is 1.20 for the Completed files and 1.41 on the Synthetic files. Across panels, the ratio of Synthetic to Gold for individuals born from 1956 to 1960 ranges from 1.88 for 1990 panel observations to 0.51 for 1993 panel observations. Without accounting for changing living arrangements of younger cohorts, the Synthetic file will make individuals born after 1970 look richer than individuals born from 1960 to 1970.

5.10 Correlation of Self-Reported and Administrative Earnings.

The correlation of administrative earnings and self-reported earnings are similar between the Gold, Completed, and Synthetic files. The completion and synthesizing process did not change this relationship; however, the correlation is much less than one on all files. This says that self-reported earnings tend to be more compressed than those on the administrative records.

At each core interview, respondents are asked to report monthly income from a variety of sources including wage and salary, self-employment jobs, and occasional earnings in the past four months. The SIPP beta files include vectors for individual earned income (totearn1990-totearn1999) based on summing the 12 monthly values of earned income from the longitudinal core SIPP data.

One would hope that the self-reported earnings would be highly correlated with the administrative earnings data. However, there are a number of reasons why self-reported earnings may differ from administrative earnings:

- Respondents report net earnings rather than gross earnings;

- Respondents have variable monthly earnings and incorrectly account for the variation;
- Respondents have unreported earnings (underground economy);
- Respondents report ballpark amounts while the administrative data report actual amounts;
- Proxy respondents may have incomplete knowledge of earnings;
- Respondents may be out of sample for some waves of the survey;
- Census imputes non-responses without regard to the administrative data.

We calculated the Spearman correlation coefficient for annual self-reported total earnings and administrative total earnings.[20] To reduce the impact of outliers, we capped earnings at 3 times the economy-wide average earnings. The correlation coefficient for Gold earnings ranges from 0.70 to 0.62 from 1990 to 1992 (see Appendix E Table 5-10). The correlation on the Gold file is lower from observations based on the 1996 panel than earlier panels and the correlation declines as the panel ages (from 1996 to 1999).[21] The correlation is also significantly lower in 1995, which is based on fewer data observations than other years due to the SIPP panel design.[22]

At the outset, it is not clear what the correlation should be on the Completed and Synthetic files compared to the Gold file. If self-reported earnings were used to fill-in for missing administrative earnings, the correlation could be significantly higher on the Completed file than on the Gold file. The correlation of self-reported and administrative earnings on the Completed file is similar to that on the Gold file. It is slightly lower or the same before 1995 and slightly higher after 1995. There is little variation in the correlation among the four Completed implicates. This suggests that self reported earnings were not directly used to impute missing administrative data.

The correlation between self-reported and administrative earnings on the Synthetic files is lower than the correlation on the Gold file in most years and most implicates. This presumably results from the process of synthesizing the data. The Synthetic correlations are about one to three percent lower than the Gold correlations, though the correlation is as much as 7 percent lower in some Synthetic implicates in 1999 (0.58. for implicate 3.4 compared to 0.62 on the Gold file). As with the Gold file, the correlation is lower for 1996 panel years than earlier panel years. There is more variation in the correlation among the 16 Synthetic implicates than among the four Completed implicates.

---

[20] Administrative earnings are the sum of defer_der_fica_{year}, nondefer_der_fica_{year}, defer_der_nonfica_{year}, nondefer_der_nonfica_{year}.
[21] The administrative match rate was lower for the 1996 SIPP panel compared with earlier SIPP panels.
[22] Census did not field a 1994 or 1995 SIPP panel. Instead, Census used the saved funds to support a larger and longer 1996 SIPP panel.

References

Abowd, J., Stinson, M. and Benedetto, G. (2006). "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project", November 5, 2006.

Bar-Yossef Z., (2005). "Algorithms for Large Data Sets: Rank Aggregation" http://www.ee.technion.ac.il/courses/049011/spring05/lectures/lecture7.pdf.

Chang, C. and Chen, J., (2008). "System and Method of Information Retrieval Engine Evaluation Using Human Judgment Input" ***World Intellectual Property Organization.*** http://www.wipo.int/pctdb/en/wo.jsp?IA=WO2008014262&wo=2008014262&DISPLAY=DESC

D'Orazio, M., Dizio, M. and Scanu, M. (2006). ***Statistical Matching: Theory and Practice.*** Chichester,WestSussex,U.K.:Wiley,2006.

Dwork, C., Kumar, R., Naor, M., and Sivakumar, D. "Rank Aggregation Methods for the Web," http://www.wisdom.weizmann.ac.il/~naor/PAPERS/rank_www10.html.

Higham, N.J., (1996). ***Accuracy and Stability of Numerical Algorithms.*** SIAM.

Invited Paper Meeting, Session 27 (2009). "Uncertainty in Statistical Matching," ***57th Session of the International Statistical Institute***, August 2009, Durban, South Africa. 1. Marco Di Zio with Marcello D'Orazio and Mauro Scanu, 2. Christopher Moriarity, 3. Suzanne Raessler with Hans Kiesl, 4. William Winkler, Disc: Fritz Scheuren.

Kennickell, A.B., (1997). "Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances," ***Record Linkage Techniques 1997***, Washington, DC: National Academy Press. Pp. 248 -267 (available at http://www.fcsm.gov under Methodology reports).

Little, R. J. A. (1993). "Statistical Analysis of Masked Data," ***Journal of Official Statistics***, 9, 407-426.

Moriarity, C., and Scheuren, F. (2001), "Statistical Matching: A Paradigm of Assessing the Uncertainty in the Procedure," ***Journal of Official Statistics***, 17, 407–422.

Moriarity, C., and Scheuren, F. (2003), "A Note on Rubin's Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputation," ***Journal of Business & Economic Statistics***, 21, 65–73.

Mulrow, J.M. and Scheuren, F.J., (1998). "The Confidentiality Beasties: A Fable About the Elephant, the Duck, and the Pig," ***1998-1999 SOI Turning Administrative Systems into Information Systems.***

Oh H.L., Scheuren F. (1983) "Weighting adjustments for unit nonresponse," in: ***Incomplete Data in Sample Surveys***, Madow W.G., Olkin I., Rubin D., (Eds.), Academic Press, New York, 2, 143-184..

Parreira, J.X., Donato, D., Castillo, C., and Weikum, G., (2007). "Computing Trusted Authority Scores in Peer-to-Peer Web Search Networks," *ACM: AIRweb*, http://airweb.cse.lehigh.edu/2007/papers/paper_108.pdf.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). "Multiple imputation for statistical disclosure limitation," ***Journal of Official Statistics***, 19, 1-16.

Reiter, J. P. (2003). "Inference for partially synthetic, public use microdata sets," ***Survey Methodology,*** 181-189.

Reiter, J. P. (2004). "Simultaneous use of multiple imputation for missing data and disclosure limitation," ***Survey Methodology***, 30, 235-242.

Reiter, J. P. (2005). "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study," ***Journal of the Royal Statistical Society***, Series A 168, 185-205.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, Inc.

Rubin, D. B. (1993). "Discussion: Statistical disclosure limitation," ***Journal of Official Statistics,*** 9, 462-468.

Scheuren, et al. (1975-1980) ***Studies from Interagency Data Linkages***, Reports 1 – 11.

Scheuren, F.J. (2007). "Book Review: *Statistical Matching: Theory and Practice,*" Journal of the American Statistical Association, Vol. 102, No. 479: 1076–1077.

Scheuren, F.J. and Mulcahy, T. (2009). "Binary Star Systems: Employing Both Data Enclaves and Synthetic Data at Inference." Paper presented in Brussels at NTTS Conference on February 20, 2009.

Scheuren, F.J. (2009) Presentation on Statistical Matching and other Synthetic and Semi-Synthetic Data Sets: Strengths and Limitations, International Statistical Institute Meetings in Durbin, South Africa

Teknomo, K. (2006) "Similarity Measurement," http://people.revoledu.com/kardi/tutorial/Similarity/Normalization.html

Vlachou1, A., Vazirgiannis1, M., and Berberich, K. (2008) "Representing and Quantifying Rank - Change for the Web Graph," ***Algorithms and Models for the Web-Graph***, Lecture Notes in Computer Science, Springer, Berlin / Heidelberg.

Yu, M. (2008). "Disclosure Risk Assessments and Control," Doctoral Dissertation, University of Michigan Program in Survey Methodology and the Joint Program in Survey Methodology.

Yu, M. (2009). "Fully-Synthetic Data for Disclosure Control," *Washington Statistical Society Seminar*, February 18, 2009, Washington, DC.

Appendix A

Task Order for an Independent Evaluation of the SIPP/SSA/IRS Synthetic Beta File

Evaluation of Strengths and Weaknesses of the SIPP/SSA/IRS Synthetic Beta File

A. Purpose of Task Order Contract

Responses to the Survey of Income and Program Participation (SIPP) are linked with individual benefit data from the Social Security Administration (SSA) and lifetime earnings records from the Internal Revenue Service (IRS). The resulting matched data are potentially useful for addressing questions about savings and retirement behavior, but access to the matched data must be severely limited to protect confidentiality. With support from the Social Security Administration, the Census Bureau has undertaken development of synthetic data that are based on records from SIPP, SSA, and IRS but that cannot be linked back to particular individuals. The preliminary version of the synthetic data file that is now available is known as the SIPP/SSA/IRS Synthetic Beta File. The purpose of this Task Order Contract is to provide an independent evaluation of this Synthetic Beta File. The Task Order report will provide an evaluation of strengths and weaknesses of the synthetic data and will document any anomalies or other special features encountered in the data. Findings from this Task Order Contract will inform the decision of whether to release the Synthetic Beta File as a public use file.

The Census Bureau chose a technique called partially synthetic data with multiple imputation of missing items for creation of the Synthetic Beta File. Construction began with a standardized extract of about 125 variables from each wave of the 1990, 1991, 1992, 1993, and 1996 SIPP panels. SSA benefit histories, annual earnings histories, and birth and death dates were added from administrative records. These SIPP and administrative data together form the Gold Standard File. Missing data were imputed four separate times to form the Completed Gold Standard File. Finally, four predicted values of individual-level records were produced based on each imputation implicate, yielding a total of 16 synthetic data implicates in the Synthetic Beta File. Detailed information about the Gold Standard, Completed Gold Standard, and Synthetic Beta Files is provided in the *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project*, which is included here as Appendix A.

The Synthetic Beta File was designed so that its records have an acceptably low probability of being linked back to regular SIPP data. The Census Bureau's Disclosure Review Board has already approved the Synthetic Beta File. Therefore, this Task Order shall not focus on possible disclosure concerns. While synthetic data appear well suited for protecting sensitive information, there is no guarantee that conclusions based on the Synthetic Beta File will be the same as those based on confidential micro-data. Under this Task Order, the Contractor shall create statistics and conduct regression analyses using confidential data and also using the Synthetic Beta File. Based on comparisons of these statistics and regression coefficients, the Contractor shall recommend in writing types of analyses for which the Synthetic Beta File appears well suited and identify important areas where it might produce misleading or inconsistent results.

Concurrently, the Census Bureau will allow controlled researcher access to the Synthetic Beta File via a remote access site. To avoid delaying the potential public release of the Synthetic Beta File, work under this Task Order will be limited to analyses that can be carried out fairly quickly. Therefore, the Contractor shall concentrate on evaluating the Synthetic Beta File and not attempt to evaluate the complex methodology used to create it. To the extent that data analysis reveals areas where improvements might be made in a future synthetic file, the Contractor is invited to recommend such improvements, but the main focus shall be on providing recommendations for potential users of the current Synthetic Beta File.

Some evaluation of the Synthetic Beta File already exists in *Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project*. This Census Bureau evaluation is limited to comparisons between the Synthetic Beta File and the Completed Gold Standard File. Analyses conducted under this Task Order shall provide an independent evaluation that complements the previous evaluation. In addition to providing comparisons between the Synthetic Beta File and the Completed Gold Standard File, the Contractor shall use data from the Gold Standard File for some comparisons. While the Contractor may choose to repeat some regressions from the previous evaluation to provide comparisons involving both the Gold Standard and Completed Gold Standard Files, the Contractor shall also specify models that differ from those already tested by the Census Bureau.

B. Project Management

The work under this Task Order shall be performed under the general provisions established by the umbrella. The SSA Task Manager for this project will provide technical direction and oversight and will share any relevant information he may have on areas that interest other users of the Synthetic Beta File. The Contractor's manager for this task will be the Principal Investigator (PI), who will have the following responsibilities: 1) day-to-day management of the Task Order, 2) point of contact for the SSA Task Manager, and 3) submission of monthly progress updates.

Brief progress updates will be sent electronically to the SSA Task Manager, Jim Sears, at jim.sears@ssa.gov by the end of the third week of each month. These messages shall note the work accomplished during the previous month under the Task Order Contract, list the names of the individuals who accomplished this work along with the hours each individual worked, and identify any significant problems encountered in its performance. Progress updates shall also include a record of any substantive conclusions from telephone contacts and meetings between the SSA Task Manager and Contractor's representatives. Any developments that would motivate changes to the set of analyses proposed by the Contractor shall be noted in monthly progress updates. For months where no problems have emerged, the progress update need not be more than a few sentences in length. Unresolved problems identified in monthly progress updates will be resolved in meetings called by SSA's Contracting Officer that shall include the SSA Task Manager, and the PI and that may include the Project Officer for the OP Task Order mechanism and other representatives of SSA and the Contractor.

C. Statement of Work

C.1 Overview

The work conducted under this Task Order will provide guidance to potential users of the SIPP/SSA/IRS Synthetic Beta File who do not necessarily have access to confidential data. The Contractor shall compare regression coefficients and other statistics obtained using the Synthetic Beta File with similar statistics from confidential data sources.

Two different sources of confidential data shall be used for comparisons with the Synthetic Beta File. The first is the Completed Gold Standard File. Because this file was the starting point for data synthesis, comparisons with it would be useful for anyone assessing the data synthesis process. However, consistency with the Completed file is not a sufficient test of analytic validity since the completion process itself could introduce bias. The second source of confidential data is the Gold Standard File prior to completion.

The Gold Standard File includes annual versions of dated SIPP variables representing the years 1990 to 1999, but values are missing for any year outside the actual SIPP panel. The Completed Gold Standard File includes imputed values for missing years and also includes a specially formulated weight that allows all of the observations from several different SIPP panels to represent the U.S. population as of April 1, 2000. The Gold Standard File provides a raw SIPP weight and identifies the SIPP panel from which each observation was taken. The Contractor shall use the SIPP weight as a starting point for analyses involving the Gold Standard File. In order to represent the U.S. population using the Gold Standard File, the Contractor may choose to limit attention to separate analyses of records from the 1990 SIPP panel and records from the 1996 SIPP panel. The Contractor could choose to propose an alternate way of representing the U.S. population with the Gold Standard File and might adjust SIPP weights to account for cases that are missing administrative data. Analyses of the Completed Gold Standard File shall be based on records from all available SIPP panels and shall use the specially formulated weight included on the file.

About 15 percent of SIPP respondents are not linked with administrative records. Missing administrative data have been imputed for the Completed Gold Standard File, and analysis of the Completed Gold Standard File shall include cases with imputed data. However, analyses of the Gold Standard File before completion shall be limited to respondents who are matched with administrative records.

The Synthetic Beta File contains a very limited set of variables. Although the Gold Standard File includes a SIPP identifier that would allow linkage with a much richer set of survey and administrative variables, the scope of this Task Order shall be limited to comparisons involving variables included on the Synthetic Beta File. Descriptions of these variables are provided in the *Technical Description of the Creation of the*

SIPP/SSA/IRS Gold Standard Files and SIPP-SSA-IRS PUF, which is included here as Appendix B.

The Government does not impose particular criteria for determining whether statistics from the Synthetic Beta File are qualitatively similar to those from confidential data. The Contractor shall use statistical criteria and other generally accepted evaluation standards in its assessment.

C.2 Description of Tasks Under Task Order

1. Planning Meetings

The Contractor will meet with the SSA Task Manager and other Office of Research, Evaluation, and Statistics (ORES) staff at the ORES Washington, DC offices to discuss Contractor staffing decisions, the data files that SSA will provide, and the specific analyses proposed by the Contractor.

The Contractor will also meet with Census Bureau and ORES staff at the Census offices in Suitland, Maryland to learn more about the Synthetic Beta File and other potential users of this file. The Census Bureau will host a one-day seminar for users of the Synthetic Beta file and will be available to answer any questions the Contractor may have.

The Contractor shall write a brief summary to provide a record of any substantive decisions made at the meeting with ORES. If discussion with Census Bureau and ORES staff or other potential users of the Synthetic Beta File leads to any changes in the set of analyses originally proposed by the Contractor, the summary memo shall contain a complete list of the analyses that the Contractor will undertake.

2. Evaluation of SIPP/SSA IRS Synthetic Beta File

2.1 Draft Report

This report shall provide comparable sets of regression coefficients and other statistics obtained from 1) the SIPP/SSA/IRS Synthetic Beta File, 2) the Completed Gold Standard File, and 3) the Gold Standard File.

Analyses shall include the following:

- Consistency checks among the Synthetic Beta File variables. For example, are dates of earnings consistent with age and with dates of SSA benefit receipt?
- Comparisons of basic univariate statistics for variables in the three analysis files. These statistics shall include medians and other relevant percentiles, means, and standard deviations.
- Multivariate analyses involving sets of key variables for particular populations of interest such as men and women, people with disabilities, widows, people with low-incomes, and racial minorities.

- Simple regressions (e.g, ordinary least squares or probit) with at least ten different dependent variables including measures of earnings, wealth, pension coverage, and SSA benefit amount. Dependent variables should be chosen primarily for relevance to SSA programs and retirement income. Some dependent variables should be chosen to address areas such as disability status or deferred compensation that are not covered by the existing Census Bureau evaluation.

The set of specific analyses shall be proposed by the Contractor and shall be responsive to areas of interest to the research community that are identified by the SSA Task Manager. The examples below are intended to illustrate the types of analyses that the Contractor might propose.

- Investigate the relationship between the Social Security benefit for April 2000 and the Primary Insurance Amount (PIA). To what extent do retired individuals and couples with high PIAs receive higher benefits than those with low PIAs?
- A divorced person may receive Social Security benefits from an ex-spouse's earnings record if that marriage lasted at least 10 years. Calculate whether each individual has a marriage that lasted at least 10 years and ended in divorce. Group observations by gender, current marital status, age, race, wealth, and AIME. Focus comparisons on groups of likely policy interest. For example, how likely are poor, elderly, divorced women to be eligible for Social Security benefits on a former spouse's earnings record?
- Assess the potential for matching earnings histories from the Synthetic Beta File to another data set that lacks longitudinal earnings. Note that past researchers have sometimes needed to statistically create lifetimes earnings histories. See, for example, "The Role of Earnings and Financial Risk in Distributional Analyses of Social Security Reform" by Thomas Hungerford in *Journal of Policy Analysis and Management*, Spring 2006, pages 417-438. Does the Synthetic Beta File appear superior to previously available data for this type of analysis?
- Existing analyses of the Synthetic Beta File include regressions of log earnings for a particular year. Instead of repeating this, analyze earnings received at a particular age. Run separate regressions for men and women by birth cohort using race, education, and marital status as explanatory variables.
- For elderly people, model claiming of Social Security retirement benefits at a particular age (e.g., 62, 63, 64, 65, or over 65) as a function of earnings before age 62, race, gender, marital status, and education.

2.2 Final Report

The Contractor shall meet with the SSA Task Manager and other Government staff at the ORES Washington, DC offices to discuss findings in the Draft Report and solicit feedback. Upon receipt of comments from SSA, the Contractor shall make necessary revisions to the Draft Report and programs and submit the final revised version to the SSA Task Manager.

D. Schedule and Deliverables

In addition to the monthly progress reports referred to in section B, the Contractor shall provide the deliverables cited below. All deliverables other than the monthly progress reports shall be provided in hard copy as well as electronic format.

Task/ Subtask	Description	Due Date*
1	**Planning Meeting**	
	Meet with ORES and Census Bureau staff	2 weeks
	Summary memo	4 weeks
2.1	**Draft Report**	
	5 bound copies and 1 camera ready	30 weeks
2.2	**Final Report**	
	20 bound copies and 1 camera ready	52 weeks

***** Due dates are specified from effective start date of performance

E. Payment Schedule and Milestones

The Task Order shall be a Labor Hours Contract. Payments shall be made no more than monthly in accordance with FAR Part 52.232-7, Payments under Time-and-Materials and Labor-Hour Contracts. Payments will be based on labor hours performed for each task, consistent with the contractor's proposed labor hour costs for each task. Payments shall not exceed the maximum agreed upon cost for the Task Order.

F. Period of Performance

The period of performance for this Task Order is August 30, 2007 through August 29, 2008.

G. Instructions

The work under this Task Order shall represent an independent evaluation of the SIPP/SSA/IRS Synthetic Beta File. Therefore, individuals who worked on the creation of the Synthetic Beta File (also known as the SIPP/SSA/IRS Public Use File) are prohibited from working on this Task Order.

Use of Gold Standard and Completed Gold Standard Files shall take place at the SSA's Washington, DC site. Computer access will be provided to the Contractor for this purpose. Access to the Synthetic Beta File is not limited to the SSA site.

Access to SIPP data matched with IRS earnings records is available only for agents of the U.S. Census Bureau. Any Contractor staff member running analyses on datasets that include earnings records shall become a Sworn Census Agent to gain access to these data. This does not apply to the synthetic earnings data on Synthetic Beta File.

In accordance with your umbrella contract, Section H-4, the Contractor shall follow the clause instructions for preparing and submitting all documents contained in the clause.

*The Contractor shall submit all required *documents within 5 working days of task order award. The below forms are available electronically via the following website:* http://co.ba.ssa.gov/ope/forms/security.htm.

- 2 completed forms FD-258, "Fingerprint Figures*," (The contractor will absorb the costs for obtaining fingerprints.)
- 1 completed FPS 176T (temporary) dated 10/2004 Statement of Personal History for Contract and Childcare Personnel (Expires 10/1/2005).
- 1 completed Optional Form 306, "Declaration for Federal Employment,"
- 1 completed "Fair Credit Reporting Act (FCRA) authorization form**," and

For a Non-U.S. citizen, 1 legible photocopy of the work authorization permit and social security card.

* Preprinted with MD900310Z, SOC SEC ADMIN, PROT SEC BR, BALTIMORE, MD on the form.

** The FCRA, as amended on September 30, 1997, requires that the Government notify each applicant, employee, and contractor (in a document consisting solely of the notice) that a consumer report may be used for employment purposes. The applicant, employee, or contractor must authorize this use in writing before the Government obtains the consumer report. The FCRA also requires that, before taking adverse action relative to an employment decision based on a consumer report, the agency provide the consumer with a copy of the report, and a copy of the Federal Trade Commission's Consumer Rights Notice. To comply with these requirements, SSA requires that the contractor submit each applicant's or employee's signed FCRA authorization form along with the other investigative documents.

H. Project Director and Key Personnel

a) The performance of the services required by this task order shall be conducted under the direction of:

Dr. Fritz Scheuren

The Government reserves the right to disapprove any successor to this individual in accordance with the substitution of key personnel provisions contained in Section G-3 of your umbrella contract.

b) The key personnel under this task order shall be:

1. Ms. Karen Elizabeth Smith

2. Dr. Douglas Wissoker
3. Dr. Sarah M. Hughes
4. Dr. Caroline Ratcliffe

I. Attachments

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

Attachment B: *Technical Description of the Creation of the SIPP/SSA/IRS Gold Standard Files and the SIPP-SSA-IRS PUF*

Attachment C: *SSA Program Analyst Manual*

Appendix B

Tables 29 and 61 from

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

The following tables were used in the analysis described in Chapter 3, Mean and Distributional Data Comparisons of Synthesized and Complete Data.

Table 29: MBA 2000 by demographic group and education

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

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 1

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
					Date variables9						
birthdate9	completed9	1/22/19559	1/12/19139	4/28/1922	9/6/19289	4/21/19439	6/13/19579	4/1/1969	2/1/1977	9/10/1979	4/20/19819
birthdate9	synthesized9	2/17/19559	4/24/19139	8/22/19229	3/23/1929	10/1/19439	7/2/19579	1/27/1969	8/25/19769	6/10/1979	3/7/19819
date_initial_entitle9	completed9	3/9/19889	12/9/19639	1/31/19709	10/9/19739	12/24/19809	10/24/1989	5/24/19 69	6/1/2000	9/9/2001	9/1/20029
date_initial_entitle9	synthesized9	5/17/19889	3/3/19649	4/5/19709	11/21/19739	3/7/19819	12/21/1989	7/30/19 69	6/20/20009	8/31/2001	9/29/20029
deathdate9	completed9	7/5/20019	4/12/20009	5/17/20009	7/16/20009	12/2/20009	7/3/20019	2/17/20029	6/24/20029	8/5/2002	9/14/20029
deathdate9	synthesized9	10/19/20009	2/4/19 39	4/22/19 69	8/6/19 89	7/13/20009	3/18/20019	11/26/20019	6/2/20029	8/28/20029	12/7/20029
					MBA Variables9						
mba_20009	completed9	6439	28	919	1569	3539	611	919	11369	12609	1549
mba_20009	synthesized9	6429	37	949	1559	3509	609	219	11379	1259	15379
mba_initial_real9	completed9	609	349	1059	1659	3379	549	8739	11169	12369	14339
mba_initial_real9	synthesized9	6129	459	1109	1709	339	5519	8809	11209	12379	14319
					Marital History Variables9						
age_mar19	completed9	23.49	15.89	17.29	18.19	19.89	22.39	25.69	309	33.49	42.49
age_mar19	synthesized9	23.19	15.79	179	17.9	19.69	22.19	25.29	29.29	32.49	40.9
duration_end19	completed	9289	09	19	19	39	139	19619	19739	19779	19819
duration_end19	synthesized	9639	09	19	29	49	2429	19489	19759	20059	20709
duration_end29	completed9	10749	09	19	29	49	19339	19609	19689	19729	19789
duration_end29	synthesized9	1059	09	29	49	219	18609	19519	19889	20159	20629
duration_end39	completed9	18149	19	69	19289	19429	19539	19619	19679	1969	19749
duration_end39	synthesized9	18429	19	7239	1929	19449	19549	19619	19669	1969	19739
duration_mar19	completed9	14.59	0.39	1.29	2.29	4.7	9.69	20.19	369	44.79	55.59
duration_mar19	synthesized9	13.49	0.39	1.39	2.29	4.49	8.89	18.29	33.19	42.39	53.69
duration_mar29	completed9	1169	09	29	49		19559	19709	19759	19789	19819
duration_mar29	synthesized9	12009	19	29	59	179	19419	1969	19769	19839	20659
duration_mar39	completed9	12989	09	29	39	109	19539	19649	19709	19739	1979
duration_mar39	synthesized9	12189	29	119	279	1419	18039	19659	19879	20269	21139
duration_mar49	completed9	19559	19259	19369	19419	1949	19569	19629	19679	19709	19729
duration_mar49	synthesized9	19569	19249	19349	19429	19509	19579	19639	19689	19709	19739
					Wealth Variables9						
homeequity9	completed9	723149	-90009	40009	80009	220009	500009	1000009	1636259	2157509	3200009
homeequity9	synthesized9	744919	-262729	13279	52849	18539	489429	1010629	1788729	2493269	3809429
nonhouswealth9	completed9	749259	-70009	10009	20009	60009	170009	610009	1815009	3172509	7650009
nonhouswealth9	synthesized9	729219	-752359	-5139	10569	46959	151819	560429	1776019	3240719	8776949
totnetworth9	completed9	1196329	-330009	-60009	1000	90009	515009	1410009	2945009	449500	9250009
totnetworth9	synthesized9	1131459	-428449	-76319	-9 59	75259	497619	1374189	2870709	4362229	8796409
					DER Earnings Arrays9						
nondefer_der_fica_19789	completed9	132009	49	3119	7389	25409	71889	140269	210729	27729	669669
nondefer_der_fica_19789	synthesized9	141119	1119	516	969	26919	71069	13879	21139	283279	831519
nondefer_der_fica_1979	completed9	132689	569	3459	8139	27459	77309	149339	229009	288969	586539
nondefer_der_fica_1979	synthesized9	143529	69	394	9439	28989	77269	150209	229039	289589	631229
nondefer_der_fica_19809	completed9	119309	589	349	8259	28729	82729	160149	250689	303369	532909
nondefer_der_fica_19809	synthesized9	120419	1719	729	13389	3369	83069	159729	246439	301859	565059
nondefer_der_fica_19819	completed9	126019	619	397	969	329	3159	177119	274059	333749	56929
nondefer_der_fica_19819	synthesized9	125659	109	539	11579	3481	92189	175769	270849	328339	562989
nondefer_der_fica_19829	completed9	13439	709	439	10209	35869	100719	187529	289419	354019	606739
nondefer_der_fica_19829	synthesized9	134269	1709	7319	14159	37679	100079	184009	283269	347829	609439

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 2

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
nondefer_der_fica_19839	completed9	141849	689	4459	10469	36879	104349	196869	305309	375329	652439
nondefer_der_fica_19839	synthesized9	140919	1219	6229	13489	37749	102019	190329	297629	369879	643349
nondefer_der_fica_19849	completed9	150539	779	4719	11049	39759	111059	208629	327309	39 059	69 219
nondefer_der_fica_19849	synthesized9	153939	1809	8189	15719	41429	112629	206229	322669	396659	754069
nondefer_der_fica_19859	completed9	158159	709	479	11419	40449	115149	219209	343769	418009	744779
nondefer_der_fica_19859	synthesized9	159 89	1379	6589	14319	41909	115049	216609	339079	42649	780729
nondefer_der_fica_19869	completed9	166679	789	4929	11549	42219	12019	229309	361129	444789	814209
nondefer_der_fica_19869	synthesized9	172919	1409	6859	14679	43029	118649	228249	35629	442179	876179
nondefer_der_fica_19879	completed9	173929	79	5009	11819	43859	125389	237209	373229	452329	849 49
nondefer_der_fica_19879	synthesized9	176819	1579	7859	16089	46039	122559	232849	37089	460979	891749
nondefer_der_fica_19889	completed9	18259	859	5369	12539	45709	130669	24759	392119	477479	896559
nondefer_der_fica_19889	synthesized9	182549	889	5479	14089	4589	127059	24179	386149	47761	900159
nondefer_der_fica_1989	completed9	18877	949	5729	13409	48629	136419	257449	406019	49613	938859
nondefer_der_fica_1989	synthesized9	187719	1469	7279	15639	47479	133869	253929	401089	49767	933579
nondefer_der_fica_19 09	completed9	19588	909	5969	1429	51229	141809	265059	411439	51300	965439
nondefer_der_fica_19 09	synthesized9	195559	1569	8149	17849	5139	139689	262319	407109	52098	981979
nondefer_der_fica_19 19	completed9	20495	949	59	14439	52959	145489	272119	426159	551369	1095679
nondefer_der_fica_19 19	synthesized9	207549	1309	6819	16149	52449	143549	268649	421419	556389	1140559
nondefer_der_fica_19 29	completed9	21543	919	5819	14109	53569	150339	284439	447409	581309	1204509
nondefer_der_fica_19 29	synthesized9	21869	1439	7359	17249	52779	148079	282189	44329	587749	1214419
nondefer_der_fica_19 39	completed9	22267	919	5789	1459	55959	15449	29289	46379	610939	1277539
nondefer_der_fica_19 39	synthesized9	226319	214	959	2079	58929	153319	289379	460819	617969	1277789
nondefer_der_fica_19 49	completed9	226349	889	5729	14559	55979	156339	295619	467319	614179	1258679
nondefer_der_fica_19 49	synthesized9	22989	215	9809	20679	59319	154289	293259	46729	62059	1305959
nondefer_der_fica_19 59	completed9	23562	909	6009	14889	57719	160829	303229	482589	637019	1319039
nondefer_der_fica_19 59	synthesized9	239189	2769	12209	24209	65809	162159	301609	480159	641709	1350039
nondefer_der_fica_19 69	completed9	25237	979	6709	15879	60009	166559	313019	49 539	664179	1383019
nondefer_der_fica_19 69	synthesized9	251069	201	9859	2209	66589	171609	315879	504119	67309	1461849
nondefer_der_fica_19 79	completed9	258719	1069	7849	18569	65139	176459	329669	528369	706879	1528649
nondefer_der_fica_19 79	synthesized9	263339	180	9229	21969	68289	17719	332129	536919	727859	1539709
nondefer_der_fica_19 89	completed9	277329	133	9249	21709	73659	189289	347849	555109	746459	1597829
nondefer_der_fica_19 89	synthesized9	285549	2729	12609	27049	78409	192859	35359	559749	752839	1636109
nondefer_der_fica_19	completed9	296479	1409	10589	25539	82739	201989	364069	580829	787239	170779
nondefer_der_fica_19	synthesized9	315949	3409	16489	3409	0179	208069	369789	587459	798909	1813589
nondefer_der_fica_20009	completed9	323209	1519	1249	2970	94849	217679	383959	613409	839319	1828949
nondefer_der_fica_20009	synthesized9	338289	3909	19309	39579	102979	225759	393769	63049	87879	1893909
nondefer_der_fica_20019	completed9	330959	1569	12639	32049	103689	23159	401859	638829	869 69	1845849
nondefer_der_fica_20019	synthesized9	347109	3909	19619	4209	109229	2379	408569	646379	882579	1974139
nondefer_der_fica_20029	completed9	336979	1239	12049	32459	106459	240049	414089	657719	88989	1882329
nondefer_der_fica_20029	synthesized9	357029	3439	17539	40479	110189	243959	422379	67194	914139	1956409
nondefer_der_fica_20039	completed9	347579	1329	13239	33649	112089	249 79	428359	67701	912129	193449
nondefer_der_fica_20039	synthesized9	371309	4519	21949	46339	115319	249889	43179	68410	945249	2245619
nondefer_der_nonfica_19789	completed9	85359	19	759	1859	8179	58459	136049	202119	257429	381789
nondefer_der_nonfica_19789	synthesized9	86509	469	2079	3949	1149	57009	135339	198939	249 29	378189
nondefer_der_nonfica_1979	completed9	88459	179	819	1989	869	64119	145459	211449	253209	361519
nondefer_der_nonfica_1979	synthesized	91009	1019	4229	7589	18309	63039	144729	208519	25309	362369
nondefer_der_nonfica_19809	completed	9019	219	1009	2489	1119	73849	16079	230749	28059	401439

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 3

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
nondefer_der_nonfica_19809	synthesized9	101609	1229	4709	8179	1919	6769	163219	235219	285429	41489
nondefer_der_nonfica_19819	completed9	10949	229	869	2119	8719	68449	179129	265569	328489	501079
nondefer_der_nonfica_19819	synthesized9	107739	1069	4629	8189	19109	63269	171689	257179	316939	480109
nondefer_der_nonfica_19829	completed9	112729	169	719	1659	6449	45639	19 839	288269	352139	553069
nondefer_der_nonfica_19829	synthesized9	11236	939	3869	679	15949	47439	190159	285309	351139	532249
nondefer_der_nonfica_19839	completed9	1229	19	739	1889	7659	69459	21869	301979	362849	537159
nondefer_der_nonfica_19839	synthesized9	12079	1079	459	8259	2029	66539	201319	288869	355349	516069
nondefer_der_nonfica_19849	completed9	138539	189	759	169	759	1379	242139	321169	386609	575679
nondefer_der_nonfica_19849	synthesized9	130619	1049	4159	7429	17759	5959	227839	31979	389049	568109
nondefer_der_nonfica_19859	completed9	135029	19	729	1639	7179	68529	246289	334719	397359	57649
nondefer_der_nonfica_19859	synthesized9	154869	89	3719	679	18909	106689	264539	359249	433169	606749
nondefer_der_nonfica_19869	completed9	14329	209	759	1889	7979	72259	255139	345449	419749	665039
nondefer_der_nonfica_19869	synthesized9	14020	909	3749	6869	16779	58489	247039	342019	418759	656839
nondefer_der_nonfica_19879	completed9	14429	239	839	1959	8139	72379	259529	354889	428139	624779
nondefer_der_nonfica_19879	synthesized9	140679	1129	4609	8309	20479	63909	241559	350129	425929	609 19
nondefer_der_nonfica_19889	completed9	15509	219	879	2049	8429	85179	276189	376079	450959	656089
nondefer_der_nonfica_19889	synthesized9	15077	929	3369	5619	12319	66909	267139	373649	455059	681379
nondefer_der_nonfica_1989	completed9	166609	209	839	209	8389	82829	286409	393369	479149	688579
nondefer_der_nonfica_1989	synthesized9	153469	829	3529	6359	14649	60779	273239	392449	482249	706559
nondefer_der_nonfica_19 09	completed9	1619	25	929	218	9029	78929	292569	409219	495279	706949
nondefer_der_nonfica_19 09	synthesized9	154219	1049	4009	6929	15149	54609	278159	406219	492309	703489
nondefer_der_nonfica_19 19	completed9	181969	269		2459	10809	112009	319439	44379	531159	747839
nondefer_der_nonfica_19 19	synthesized9	1749	1049	4609	8379	22689	81529	306539	437009	526769	719589
nondefer_der_nonfica_19 29	completed9	190969	249	1049	2579	11239	118629	332219	463879	555339	807379
nondefer_der_nonfica_19 29	synthesized9	176449	1219	4919	8849	21879	76979	306079	449 79	542749	765859
nondefer_der_nonfica_19 39	completed9	201729	309	1239	3149	12939	115539	34179	487409	59780	924249
nondefer_der_nonfica_19 39	synthesized9	190309	1259	506	9009	21409	8289	322779	47649	59339	27639
nondefer_der_nonfica_19 49	completed9	203489	259	1119	3009	12679	117559	342539	488829	589459	876049
nondefer_der_nonfica_19 49	synthesized9	189709	1339	5739	10879	2817	91039	320239	474279	569739	801009
nondefer_der_nonfica_19 59	completed9	209549	289	1139	2859	13069	117449	347889	504289	623539	1006549
nondefer_der_nonfica_19 59	synthesized9	196039	1309	5659	10139	23049	84609	328549	482379	59353	984819
nondefer_der_nonfica_19 69	completed9	211649	339	1129	2669	12059	112689	353929	510589	62305	951219
nondefer_der_nonfica_19 69	synthesized9	195909	1429	5929	10769	26209	82269	325619	496269	610149	869 39
nondefer_der_nonfica_19 79	completed9	223089	359	1449	3419	15089	122059	365789	530459	660639	1074579
nondefer_der_nonfica_19 79	synthesized9	223949	1819	7539	13989	34449	102219	340139	537229	671439	1076309
nondefer_der_nonfica_19 89	completed9	222589	479	1739	379	14859	117009	368879	53559	662719	1078119
nondefer_der_nonfica_19 89	synthesized9	21579	1969	7979	1419	33179	101019	345849	533569	663389	1035519
nondefer_der_nonfica_19	completed9	235079	409	1779	4359	17619	127879	376589	552879	677889	1072959
nondefer_der_nonfica_19	synthesized9	238659	2509	10319	18559	44189	122179	366069	559089	703169	1100349
nondefer_der_nonfica_20009	completed9	24279	439	1779	4419	1779	129229	390629	57009	69 29	1148639
nondefer_der_nonfica_20009	synthesized9	238869	2739	10309	18139	40889	119 69	37009	572509	716009	1126479
nondefer_der_nonfica_20019	completed9	250259	359	1689	4549	19069	137459	401609	586449	722379	111589
nondefer_der_nonfica_20019	synthesized9	251019	2909	11649	20889	4939	13479	383519	592509	732659	111949
nondefer_der_nonfica_20029	completed9	282479	469	2279	6009	28589	194769	438759	63539	780239	1232959
nondefer_der_nonfica_20029	synthesized9	27569	1649	7649	15009	45789	180929	425609	623509	774219	1149009
nondefer_der_nonfica_20039	completed9	291249	459	2289	649	33719	210969	456659	652839	788069	1249839
nondefer_der_nonfica_20039	synthesized9	332949	4819	20729	3793	9309	281979	486829	671109	82069	117449

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 4

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
defer_der_fica_19879	completed9	11059	4109	4109	4109	8709	11019	12269	16379	20489	20489
defer_der_fica_19879	synthesized	9239	6159	6159	6539	775	9219	10709	11329	11329	12439
defer_der_fica_19889	completed9	77719	09	1229	1419	5909	17619	36729	58749	76279	76279
defer_der_fica_19889	synthesized9	135459	89	609	1559	5859	19429	42179	62569	774539	2896579
defer_der_fica_1989	completed9	32639	109	89	1579	5309	12009	24549	47519	61789	264889
defer_der_fica_1989	synthesized9	25779	359	1359	239	5559	12989	27209	49 39	7479	205509
defer_der_fica_19 09	completed9	21139	359	1489	2619	6009	13279	28009	49 79	66659	84009
defer_der_fica_19 09	synthesized9	19659	309	1319	2479	5609	11889	25489	48449	65149	88039
defer_der_fica_19 19	completed9	23219	449	1549	2669	6109	13709	28219	51279	70249	87509
defer_der_fica_19 19	synthesized9	2119	449	1809	3079	6089	12839	26579	50659	6914	96889
defer_der_fica_19 29	completed9	22809	469	1739	2909	6449	14439	29	5429	73139	87289
defer_der_fica_19 29	synthesized9	22759	69	229	3619	6749	13889	29149	54589	7535	9529
defer_der_fica_19 39	completed9	24489	459	1729	3009	6789	15039	31909	57639	7774	94919
defer_der_fica_19 39	synthesized9	23659	619	219	3619	6969	14589	30919	57539	7916	96939
defer_der_fica_19 49	completed9	27049	449	1819	3129	7019	15849	33349	60039	8051	92409
defer_der_fica_19 49	synthesized9	25359	519	2189	3709	7239	15749	32249	58539	7938	95849
defer_der_fica_19 59	completed9	27759	469	1719	3129	7089	1619	34549	6279	8341	92409
defer_der_fica_19 59	synthesized9	25329	559	2459	4249	8489	15619	33349	61139	8312	95429
defer_der_fica_19 69	completed9	26419	509	1949	3369	7659	17149	36349	66729	8719	5009
defer_der_fica_19 69	synthesized9	2639	59	2439	4249	839	17259	35749	65749	8652	97119
defer_der_fica_19 79	completed9	27969	49	1919	3389	7869	18129	38929	7185	9369	5009
defer_der_fica_19 79	synthesized9	27909	609	2519	4369	8679	18189	38569	7075	9157	96279
defer_der_fica_19 89	completed9	2979	519	2059	3689	849	19279	419	7722	9369	100009
defer_der_fica_19 89	synthesized9	29869	689	2849	483	9419	19419	41129	7548	95579	105919
defer_der_fica_19	completed9	31409	489	2039	3789	89	20249	44539	80969	100009	100009
defer_der_fica_19	synthesized9	31789	59	2439	439	259	20179	43789	7975	97969	126469
defer_der_fica_20009	completed9	3429	459	2059	378	9059	20859	46469	85749	105009	105009
defer_der_fica_20009	synthesized9	34439	709	2979	5079	10249	21619	46889	85429	103269	117419
defer_der_fica_20019	completed9	33749	459	2019	387	939	21549	4824	90339	105009	105009
defer_der_fica_20019	synthesized9	3429	29	3389	5479	10739	22259	48759	89259	103519	111789
defer_der_fica_20029	completed9	3489	459	2009	390	9539	21979	4920	96449	110009	120009
defer_der_fica_20029	synthesized9	35139	839	3109	5159	10379	21949	4856	9489	109739	125229
defer_der_fica_20039	completed9	37209	559	2259	4089	10009	22719	5119	104529	120009	140009
defer_der_fica_20039	synthesized9	40269	1029	4469	6609	12529	25329	54829	107159	120289	142069
defer_der_nonfica_19879	completed9	61949	3649	3649	5489	1369	2286	97719	169759	169759	169759
defer_der_nonfica_19879	synthesized9	34029	709	709	489	20439	34359	4719	58949	58949	68709
defer_der_nonfica_19889	completed9	17619	2209	2209	2209	3319	4079	17419	66069	102129	102129
defer_der_nonfica_19889	synthesized9	17989	509	509	509	8759	12359	25829	42039	42039	43719
defer_der_nonfica_1989	completed9	18729	419	729	7839	11509	16009	19509	36009	38029	82009
defer_der_nonfica_1989	synthesized9	1929	829	4769	6869	11589	1759	22979	35849	44939	52179
defer_der_nonfica_19 09	completed9	38849	769	2939	480	9649	18859	34019	6779	79 59	125679
defer_der_nonfica_19 09	synthesized9	29729	409	2289	430	9169	17529	30669	55939	76389	116959
defer_der_nonfica_19 19	completed9	25869	259	1919	3549	8369	1779	3179	6309	75009	122549
defer_der_nonfica_19 19	synthesized9	25239	729	2669	436	919	17259	30919	59039	78639	119769
defer_der_nonfica_19 29	completed9	24729	79	679	2249	7379	17019	3209	60009	75009	116739
defer_der_nonfica_19 29	synthesized9	23429	449	1729	3069	7249	1539	29809	56409	77459	116649
defer_der_nonfica_19 39	completed9	25319	59	609	2069	6869	16489	33229	60529	75009	125009

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 5

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
defer_der_nonfica_19 39	synthesized9	2319	409	1689	3029	6949	14989	28869	55529	74959	122669
defer_der_nonfica_19 49	completed9	24489	69	819	19	6509	16229	31589	60009	75009	117579
defer_der_nonfica_19 49	synthesized9	22109	39	1539	2649	6209	14159	28239	54319	73879	110859
defer_der_nonfica_19 59	completed9	24859	129	819	1949	6509	16849	32039	61519	75009	122319
defer_der_nonfica_19 59	synthesized9	22219	419	1519	2629	5979	14189	28189	54659	74319	114319
defer_der_nonfica_19 69	completed9	2479	79	609	2089	6509	18009	33409	60319	75009	111149
defer_der_nonfica_19 69	synthesized9	23309	49	1659	2869	6429	15219	30479	57919	75809	112719
defer_der_nonfica_19 79	completed9	26719	79	769	2259	7749	18479	35069	66679	78369	120009
defer_der_nonfica_19 79	synthesized9	25379	459	1819	3189	7319	17059	32689	62589	80129	116509
defer_der_nonfica_19 89	completed9	279	11	979	2589	8449	1939	37629	70089	80009	122409
defer_der_nonfica_19 89	synthesized9	24459	439	1719	3089	709	16409	31479	61089	78419	112989
defer_der_nonfica_19	completed9	28759	119	1019	2369	8179	19719	38749	71259	80359	122409
defer_der_nonfica_19	synthesized9	25079	449	1739	3079	719	16729	32539	62039	80259	113379
defer_der_nonfica_20009	completed9	29279	129	109	2679	8519	20939	39229	73449	80049	120009
defer_der_nonfica_20009	synthesized9	25739	519	1939	3389	7809	17759	33459	62839	80939	110369
defer_der_nonfica_20019	completed9	31429	219	1209	269	8259	21669	43059	78959	85859	125429
defer_der_nonfica_20019	synthesized9	27319	509	2159	3769	8219	18169	36069	68929	85209	111189
defer_der_nonfica_20029	completed9	35529	179	1449	327	9689	24009	49289	85009	110009	139 59
defer_der_nonfica_20029	synthesized9	31589	659	2359	3989	8749	19539	43919	80459	101759	126419
defer_der_nonfica_20039	completed9	37779	219	139	3009	8979	24009	5335	93619	120009	146329
defer_der_nonfica_20039	synthesized9	35239	739	2649	452	9869	21559	47729	85979	112149	144709
				SER Earnings Arrays9							
earn1937_to_19519	completed9	55559		589	1529	6849	28639	80729	148129	195209	304249
earn1937_to_19519	synthesized9	60819	39	119	229	1509	16629	85089	185539	266589	397609
totearn_ser_19519	completed9	1659		489	1129	4439	14639	28219	36009	36009	36009
totearn_ser_19519	synthesized9	15289	159	69	1349	3709	11939	26139	36009	36009	36009
totearn_ser_19529	completed9	17549	109	529	1239	4609	16079	30859	36009	36009	36009
totearn_ser_19529	synthesized9	16239	19	819	1539	3979	13429	28329	36009	36009	36009
totearn_ser_19539	completed9	18569	109	559	1389	5219	17719	33609	36009	36009	36009
totearn_ser_19539	synthesized9	17309	21	909	1669	439	14909	31379	36009	36009	36009
totearn_ser_19549	completed9	18959	109	549	1339	5379	18519	34789	36009	36009	36009
totearn_ser_19549	synthesized9	17649	22	969	1789	4769	15529	31729	36009	36009	36009
totearn_ser_19559	completed9	20609	149	669	1529	5509	18789	36649	42009	42009	42009
totearn_ser_19559	synthesized9	19339	259	1059	1959	509	16459	33589	42009	42009	42009
totearn_ser_19569	completed9	21769	149	749	1719	6459	20809	39169	42009	42009	42009
totearn_ser_19569	synthesized9	20789	279	1189	2189	5959	18839	36679	42009	42009	42009
totearn_ser_19579	completed9	22529	169	879	2149	7749	21909	40659	42009	42009	42009
totearn_ser_19579	synthesized9	2159	329	1349	2519	7039	19 89	37909	42009	42009	42009
totearn_ser_19589	completed9	22829	149	839	2039	7759	2239	41969	42009	42009	42009
totearn_ser_19589	synthesized9	2159	29	1259	2359	6959	20779	37369	4129	41889	42689
totearn_ser_1959	completed9	25059	17	929	219	8139	24039	44639	48009	48009	48009
totearn_ser_1959	synthesized9	24009	309	129	2439	7149	22179	41819	48009	48009	48009
totearn_ser_19609	completed9	2559	189	1039	2389	8579	24739	4629	48009	48009	48009
totearn_ser_19609	synthesized9	24389	359	1429	2619	7289	22539	43469	48009	48009	48009
totearn_ser_19619	completed9	26029	209	1079	2479	8949	25449	47829	48009	48009	48009
totearn_ser_19619	synthesized9	24389	319	1319	2489	7149	22339	44039	48009	48009	48009
totearn_ser_19629	completed9	26759	189	1079	274	9559	26679	48009	48009	48009	48009

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 6

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
totearn_ser_19629	synthesized9	25329	389	1549	2849	8059	24109	45579	48009	48009	48009
totearn_ser_19639	completed9	27189	19	1129	270	969	27669	48009	48009	48009	48009
totearn_ser_19639	synthesized9	25849	359	1489	2789	8269	25259	46239	48009	48009	48009
totearn_ser_19649	completed9	27949	209	1179	2869	10279	29379	48009	48009	48009	48009
totearn_ser_19649	synthesized9	26239	329	1389	2649	8189	26249	47389	48009	48009	48009
totearn_ser_19659	completed9	28339	239	1289	2919	10489	30269	48009	48009	48009	48009
totearn_ser_19659	synthesized9	26919	389	159	2969	8689	2739	48009	48009	48009	48009
totearn_ser_19669	completed9	33839	239	1329	3129	10889	31469	60109	66009	66009	66009
totearn_ser_19669	synthesized9	31909	379	1629	316	9389	28579	55229	66009	66009	66009
totearn_ser_19679	completed9	3479	249	1379	3359	11719	33079	62689	66009	66009	66009
totearn_ser_19679	synthesized9	33039	469	1889	349	79	29 69	58769	66009	66009	66009
totearn_ser_19689	completed9	38979	289	1569	3619	12669	35829	67369	78009	78009	78009
totearn_ser_19689	synthesized9	36829	49	2029	3769	10819	32479	62749	78009	78009	78009
totearn_ser_1969	completed9	40579	29	1769	3989	13509	38189	72489	78009	78009	78009
totearn_ser_1969	synthesized9	38809	49	209	4029	11739	35259	68179	78009	78009	78009
totearn_ser_19709	completed9	42139	289	1779	4019	14369	41069	76609	78009	78009	78009
totearn_ser_19709	synthesized9	40149	539	2229	4229	12439	37659	70989	78009	78009	78009
totearn_ser_19719	completed9	4339	289	1789	4119	15029	43689	78009	78009	78009	78009
totearn_ser_19719	synthesized9	4129	539	2219	4249	12679	39829	739	78009	78009	78009
totearn_ser_19729	completed9	47949	309	1929	4449	15959	46189	8683	9000	9000	90009
totearn_ser_19729	synthesized9	45539	559	2329	4489	13489	42449	8055	9000	9000	90009
totearn_ser_19739	completed9	53989	389	2109	4789	169	4968	93279	108009	108009	108009
totearn_ser_19739	synthesized9	51759	609	259	4969	14879	46609	88749	108009	108009	108009
totearn_ser_19749	completed9	60879	369	2269	5339	19109	5346	9689	132009	132009	132009
totearn_ser_19749	synthesized9	58239	639	289	5659	16769	5026	95769	132009	132009	132009
totearn_ser_19759	completed9	64589	439	2439	5519	19659	56839	106269	141009	141009	141009
totearn_ser_19759	synthesized9	61949	759	3159	6019	17489	53339	101909	141009	141009	141009
totearn_ser_19769	completed9	70019	459	2729	6149	21649	61229	114839	153009	153009	153009
totearn_ser_19769	synthesized9	67209	829	3569	6779	1929	57249	110679	153009	153009	153009
totearn_ser_19779	completed9	75159	489	2879	6579	23169	65409	123529	165009	165009	165009
totearn_ser_19779	synthesized9	70349	859	3669	69	19709	5809	115339	165009	165009	165009
SIPP Arrays9											
famwelamt19 09	completed9	21779	259	1139	2109	5319	11949	28479	57229	77639	114549
famwelamt19 09	synthesized9	23129	429	1729	2989	6249	12879	2979	59319	80209	121839
famwelamt19 19	completed9	26649	369	1549	2879	6689	15409	36259	68129	89049	131589
famwelamt19 19	synthesized9	22389	279	1249	2289	519	11429	27259	60479	83129	130909
famwelamt19 29	completed9	23269	219	1039	2019	5159	12749	31569	60939	80749	123019
famwelamt19 29	synthesized9	22089	289	1259	229	5149	1139	2769	58879	80759	127609
famwelamt19 39	completed9	2469	269	1169	2269	5709	13759	34129	64919	84719	123939
famwelamt19 39	synthesized9	2489	379	1549	2759	6019	13179	3219	6679	0159	132069
famwelamt19 49	completed9	22649	209	1069	2079	5259	1249	30079	59609	77939	118119
famwelamt19 49	synthesized9	20969	279	1209	2179	4779	10349	24909	57159	79329	129069
famwelamt19 59	completed9	19209	21	989	1939	4879	10829	22769	50909	71839	108159
famwelamt19 59	synthesized9	21029	289	119	2149	4789	10479	24069	58819	81669	131209
famwelamt19 69	completed9	28889	339	1279	2449	6219	15259	40159	7476	98469	150979
famwelamt19 69	synthesized9	30269	469	1859	319	6839	1489	39129	79739	10789	174219
famwelamt19 79	completed9	25949	289	1219	2249	5619	1339	33329	7025	94249	134589

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 7

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
famwelamt19 79	synthesized9	29	409	1659	2869	6169	13779	36819	84439	110659	193879
famwelamt19 89	completed9	22439	239	1039	1939	489	1169	27979	60759	82569	128879
famwelamt19 89	synthesized9	25469	369	139	239	5109	11389	29279	72739	104189	164059
famwelamt19	completed9	19409	279	1169	2039	4749	10339	22949	49489	72949	120559
famwelamt19	synthesized9	23449	39	1519	2569	5249	11169	27249	6466	90249	146439
fpov19 09	completed9	1296789	739579	797469	84031	96089	1190339	1572439	1877779	2083409	2722329
fpov19 09	synthesized9	1318629	746779	805819	860479	1017309	1213919	1582449	1889489	2085979	2676949
fpov19 19	completed9	1359129	766759	821739	863529	1038249	1250069	1656339	1973049	2189639	279869
fpov19 19	synthesized9	140209	771839	832879	890939	1065649	128769	168589	2013339	2233989	2831609
fpov19 29	completed9	1391759	792849	840939	883089	1063439	1276029	1702889	202009	2247669	2902619
fpov19 29	synthesized9	1438189	798059	85760	917319	109769	1323069	1731479	2061969	2279659	291539
fpov19 39	completed9	1436049	817829	87016	907329	1091869	1315489	1757189	209 279	2322739	3004789
fpov19 39	synthesized9	1468049	822619	87906	930579	1114119	13479	1777979	2124759	2343279	2964319
fpov19 49	completed9	1466609	814779	87869	25059	1108209	1347509	1800449	2135369	2358509	29 5789
fpov19 49	synthesized9	1487629	823759	88859	43879	1127349	1370949	181409	2150459	2358489	2956359
fpov19 59	completed9	1494389	792589	88736	952079	1117519	1357179	1833849	2185909	2416369	3145519
fpov19 59	synthesized9	1523469	80973	90013	967129	1138289	1391919	1857559	2210239	2447029	3159439
fpov19 69	completed9	1552139	87867	92638	978839	1142869	1398289	1910289	2300749	255309	3330049
fpov19 69	synthesized9	1570469	88733	94111	94479	1169659	1426029	1922279	2300079	2534859	3230779
fpov19 79	completed9	154949	88870	93505	96839	1163229	1395159	1949039	2301449	2555719	3216509
fpov19 79	synthesized9	160714	90227	953359	1008219	1189779	1461719	1977079	2368129	2617609	3297789
fpov19 89	completed9	1589 69	89096	948249	1013829	1182649	1453769	198289	2347559	2593649	3238419
fpov19 89	synthesized9	162224	90015	96439	1021509	1205329	1475059	19 289	2393169	2641339	334169
fpov19	completed9	162718	92542	97229	1037319	1207719	1482339	2025769	2405409	2653069	3343639
fpov19	synthesized9	168355	92148	98489	1044939	1236259	152739	2059 59	2480629	2762569	3582439
ftotinc19 09	completed9	369659	10449	5735	92369	176929	313289	495289	712619	885429	1283979
ftotinc19 09	synthesized9	352719	6589	55489	89789	171569	301379	472069	674339	83269	1212169
ftotinc19 19	completed9	387139	17649	6461	9889	189679	333259	522889	74401	903309	1235239
ftotinc19 19	synthesized9	35629	-689	47439	80009	159 79	295209	483089	711159	882219	1254459
ftotinc19 29	completed9	389139	13019	6146	9669	186369	330949	528819	75954	921749	1259759
ftotinc19 29	synthesized9	368159	229	51229	85329	168639	30669	49 669	73493	90209	1263039
ftotinc19 39	completed9	40269	20809	69219	105679	195559	344379	543759	77512	945709	1296989
ftotinc19 39	synthesized9	37969	859	52839	88109	173459	317819	515139	75370	929619	1298979
ftotinc19 49	completed9	417839	15609	67009	105089	197679	353349	565659	820649	1002879	138129
ftotinc19 49	synthesized9	387889	-1949	51429	87519	173379	32109	529 19	78275	968709	1363579
ftotinc19 59	completed9	430209	12639	66159	107939	204329	370559	586709	834059	1002309	1423819
ftotinc19 59	synthesized9	40829	-9 79	40679	79149	170849	334329	565769	836889	103719	1478419
ftotinc19 69	completed9	467039	10949	65119	104219	202939	369169	605359	898819	1127249	2085889
ftotinc19 69	synthesized9	438979	-879	5679	5809	189349	349089	574159	854909	1071859	1881859
ftotinc19 79	completed9	463589	1219	6322	9549	194649	36169	60471	902289	1143009	2133619
ftotinc19 79	synthesized9	451809	1129	5671	94949	188589	350789	58859	884909	1125909	2082029
ftotinc19 89	completed9	489 29	12359	66239	104829	205479	382839	64043	956039	1208979	2310819
ftotinc19 89	synthesized9	482769	849	5885	98979	197929	37109	62751	951969	1219279	2391909
ftotinc19	completed9	525619	13269	67349	108589	21539	40509	672859	1018179	133239	2745239
ftotinc19	synthesized9	513429	-389	5574	9789	201939	387189	657319	1012919	1337609	2823129
helamt19 09	completed9	22139	389	1549	2859	5929	11489	239	50489	79839	172389
helamt19 09	synthesized9	20519	329	1449	2569	5389	10309	20949	45269	74809	173439

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 8

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
helamt19 19	completed9	28679	519	2219	3969	8149	15279	32119	6527	98409	206339
helamt19 19	synthesized9	28869	579	2289	3949	789	14879	30489	66119	103279	221019
helamt19 29	completed9	3189	559	2119	3889	8349	16889	35079	75079	112879	210159
helamt19 29	synthesized9	29749	519	219	3929	7959	15309	3109	65919	10669	225589
helamt19 39	completed9	29179	479	1949	3609	7949	15989	33279	6650	97769	208239
helamt19 39	synthesized9	30219	559	2269	3959	8049	15549	32019	67689	107739	228259
helamt19 49	completed9	3269	569	2279	4089	8979	17749	37919	75149	111749	225549
helamt19 49	synthesized9	33159	569	2359	4169	869	16979	36319	76069	120019	250779
helamt19 59	completed9	39659	1169	4279	7069	13469	25189	46949	86559	131879	2129
helamt19 59	synthesized9	35949	849	29	5149	10669	20859	40359	77689	129439	232009
helamt19 69	completed9	32879	549	2439	434	9039	17979	35779	67409	10479	277489
helamt19 69	synthesized9	38449	839	3089	517	9859	18429	37359	78489	136959	374469
helamt19 79	completed9	41029	849	2969	5129	10489	21119	45639	89259	136609	242569
helamt19 79	synthesized9	62159	1309	4369	7029	12959	25109	5459	114719	189879	332029
helamt19 89	completed9	47849	1139	3609	6379	1189	23479	4757	97459	143109	268379
helamt19 89	synthesized9	65009	1139	4209	6739	12669	24039	47689	108329	186669	893019
helamt19	completed9	49459	119	4469	7279	1329	25879	5098	95239	142109	265109
helamt19	synthesized9	73809	1209	4679	7639	14539	27439	56049	116659	200549	857269
totearn19 09	completed9	176439	1079	6979	15879	53559	138869	248819	379679	484119	7739
totearn19 09	synthesized9	164659	859	5629	11409	38819	12559	236739	365769	468809	75989
totearn19 19	completed9	185069	1489	8329	17689	5549	143369	259489	40109	513679	836169
totearn19 19	synthesized9	17662	959	579	12429	43009	132369	251639	394419	508939	841929
totearn19 29	completed9	193539	158	9189	19859	60409	149179	269829	418319	541709	870169
totearn19 29	synthesized9	178819	859	6149	13789	4579	133439	253249	398159	517719	853779
totearn19 39	completed9	19519	151	9159	19669	60869	15059	27189	421439	549789	87849
totearn19 39	synthesized9	190169		6319	13979	48129	142189	269509	424159	55377	901149
totearn19 49	completed9	198719	1979	10849	22319	63489	152149	275349	429169	55686	90189
totearn19 49	synthesized9	190769	1129	7029	15539	50619	142609	26869	424689	55240	906229
totearn19 59	completed9	198579	2859	15259	29439	70149	151329	270009	421979	54303	911429
totearn19 59	synthesized9	193859	3469	15689	27869	63649	141039	264389	422509	55008	923259
totearn19 69	completed9	243239	1769	11979	25749	74839	174189	315219	494569	66129	1383329
totearn19 69	synthesized9	243529	2319	11989	24579	71369	171659	315179	495589	665089	1457249
totearn19 79	completed9	245719	2339	13139	28119	78709	178559	319119	497239	653229	1301389
totearn19 79	synthesized9	255929	219	11729	24659	74989	180279	328749	519529	69279	1514869
totearn19 89	completed9	253519	2679	15449	31969	8459	185609	327589	514839	673959	1331949
totearn19 89	synthesized9	270549	2339	12859	27329	81369	189719	343179	551869	737319	169 179
totearn19	completed9	278349	3549	20069	3947	96619	200439	350509	552609	737039	1719 89
totearn19	synthesized9	291309	159	12059	28119	85749	198919	36079	584529	805809	2139529
tothoursannual19 09	completed9	16719	449	1819	346	9729	19459	21979	25959	29159	35379
tothoursannual19 09	synthesized9	15219	19	1039	2209	729	18019	21419	24359	26919	32579
tothoursannual19 19	completed9	16039	479	1789	3279	8529	18719	21459	25279	28439	36089
tothoursannual19 19	synthesized9	15119	19	1159	2359	7249	17629	2129	24359	27179	33179
tothoursannual19 29	completed9	16839	559	2169	419	10729	19549	2169	25179	28129	34509
tothoursannual19 29	synthesized9	15619	409	169	3089	8229	18029	21409	24709	27629	34039
tothoursannual19 39	completed9	17069	539	2059	3869	10109	19709	22089	26739	30039	38059
tothoursannual19 39	synthesized9	15849	439	1679	3039	8159	18369	21669	25429	28429	34789
tothoursannual19 49	completed9	16519	649	2269	405	9769	19239	21539	25589	28559	34789

Table 61: Percentiles of Synthetic and Completed Variables9 Panel

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
tothoursannual19 49	synthesized9	15579	689	219	3689	8389	17629	21289	24609	2749	33329
tothoursannual19 59	completed9	16209	1119	3209	506	9729	18369	21359	24529	27269	32439
tothoursannual19 59	synthesized9	15509	1479	3669	545	9629	16779	2089	23189	25429	30449
tothoursannual19 69	completed9	17439	619	2359	439	11179	19659	22429	27029	30159	36489
tothoursannual19 69	synthesized9	17119	679	239	4319	10789	1929	22269	26459	29439	35469
tothoursannual19 79	completed9	17309	579	2349	4459	11169	19859	22009	26319	29469	36449
tothoursannual19 79	synthesized9	17009	489	1949	3709	10279	19519	22189	26459	29579	36659
tothoursannual19 89	completed9	17269	619	2389	4649	11179	19869	21829	26089	29239	35879
tothoursannual19 89	synthesized9	17119	569	2169	4139	1069	19549	22139	26239	29329	36329
tothoursannual19	completed9	17479	69	2779	5269	11829	19 59	21829	25979	28889	35539
tothoursannual19	synthesized9	16909	459	2019	3939		19249	22029	26559	30029	37789
totinc19 09	completed9	166159	-829	1529	10249	46289	126769	237709	367939	471609	740689
totinc19 09	synthesized9	162829	-7649		8279	38729	117369	234809	373459	481589	768329
totinc19 19	completed9	172849	-6759	2009	10219	47629	129419	245879	385809	495739	806369
totinc19 19	synthesized9	168159	-569	2689	8969	39419	117569	240419	388639	504919	83379
totinc19 29	completed9	177629	-7209	1929	11009	50509	130879	250429	397769	516639	843909
totinc19 29	synthesized9	174689	-5119	3249	10169	44089	121329	246639	402849	52919	876989
totinc19 39	completed9	183009	-6679	3519	14909	54939	136379	256089	404209	524139	857079
totinc19 39	synthesized9	180659	-4949	4579	13309	49739	128309	253409	409169	535329	885609
totinc19 49	completed9	186679	-6839	7349	20409	59669	139429	258669	411169	53059	863489
totinc19 49	synthesized9	187149	-4879	769	18259	53589	132129	25979	424859	55541	910049
totinc19 59	completed9	189 59	-7589	10819	26279	64009	140609	261319	41229	528849	875779
totinc19 59	synthesized9	189 69	-6369	10769	23649	569	130369	259629	428519	56489	4769
totinc19 69	completed9	218419	-7649	8939	25259	68929	155239	287689	457329	593709	1120629
totinc19 69	synthesized9	220119	-322	969	24119	6579	151589	28879	466449	61269	1168309
totinc19 79	completed9	231039	-2909	1089	28739	74449	165529	302859	47749	622969	117509
totinc19 79	synthesized9	227519	-2589	10919	24969	6809	156759	296939	478559	630149	1229
totinc19 89	completed9	238049	-39	12979	32339	78959	172929	312889	49679	647019	1159109
totinc19 89	synthesized9	2379	-2829	1359	29019	72689	163609	307139	502019	666329	1305209
totinc19	completed9	260429	-3739	16789	3669	85209	180879	32969	527389	707209	1634349
totinc19	synthesized9	259889	-2079	16639	33639	78019	172059	327889	537869	726329	1710449
wkspt19 09	completed9	15.89	0.39	1.19	2.29	5.29	12.39	22.59	36.79	44.59	529
wkspt19 09	synthesized9	15.79	0.29	1.29	2.29	5.19	129	22.49	36.79	44.29	51.19
wkspt19 19	completed9	15.89	0.39	1.19	2.19	59	12.19	22.59	37.19	45.79	529
wkspt19 19	synthesized9	15.79	0.39	1.29	2.29	5.19	11.9	22.29	379	45.29	529
wkspt19 29	completed9	16.9	0.39	1.19	2.29	5.69	13.49	24.59	38.29	45.89	539
wkspt19 29	synthesized9	16.39	0.29	1.29	2.29	5.49	12.89	23.69	37.29	44.39	529
wkspt19 39	completed9	16.49	0.39	1.19	2.29	5.39	12.89	23.69	37.89	46.79	529
wkspt19 39	synthesized9	16.29	0.39	1.29	2.39	5.39	12.49	23.19	389	46.59	529
wkspt19 49	completed9	15.79	0.39	1.29	2.19	5.19	12.49	21.49	36.59	45.79	529
wkspt19 49	synthesized9	15.9	0.39	1.49	2.59	5.59	12.79	21.69	36.29	44.79	529
wkspt19 59	completed9	16.59	0.69	2.29	3.89	8.29	13.9	23.59	32.49	409	49.49
wkspt19 59	synthesized9	18.19	0.79	2.39	3.9	8.59	15.79	25.19	36.69	42.29	49.59
wkspt19 69	completed9	22.39	0.79	2.89	4.9	11.29	19.19	33.29	44.69	48.89	529
wkspt19 69	synthesized9	229	0.59	2.39	4.29	10.49	18.9	33.49	44.89	48.69	51.9
wkspt19 79	completed9	19.89	0.49	1.79	3.39	8.19	16.59	29.79	43.29	49.49	539
wkspt19 79	synthesized9	20.69	0.49	1.9	3.59	8.39	17.29	31.49	44.69	49.89	529

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 10

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
wkspt19 89	completed9	209	0.69	2.29	3.9	8.39	16.59	29.79	439	49.49	529
wkspt19 89	synthesized9	21.29	0.69	2.59	4.3	9.29	17.9	329	459	49.9	529
wkspt19	completed9	209	0.59	2.29	3.89	8.69	16.89	29.79	42.39	48.59	529
wkspt19	synthesized9	21.9	0.69	2.79	4.79	109	18.89	32.9	459	49.39	529
wkswp19 09	completed9	419	3.29	109	17.39	35.39	47.59	51.29	51.9	529	529
wkswp19 09	synthesized9	40.79	29	7.59	13.89	34.79	48.59	51.29	51.9	529	529
wkswp19 19	completed9	41.59	3.69	10.49	17.29	35.79	48.89	51.69	529	529	529
wkswp19 19	synthesized9	41.29	2.6	9.19	15.79	34.59	49	51.59	529	529	529
wkswp19 29	completed9	41.9	3.89	11.59	19.49	36.9	47.69	51.59	539	539	539
wkswp19 29	synthesized9	41.59	2.5	9.59	17.19	36.79	48.29	51.39	529	529	529
wkswp19 39	completed9	42.49	3.89	11.19	18.59	38.19	49.39	51.79	529	529	529
wkswp19 39	synthesized9	42.19	2.6	9.69	16.9	37.19	49.79	51.69	529	529	529
wkswp19 49	completed9	429	4.59	12.89	19.39	36.39	48.59	51.69	529	529	529
wkswp19 49	synthesized9	42.29	3.9	12.19	19.19	36.9	49.19	51.49	51.9	529	529
wkswp19 59	completed9	41.29	6.89	15.19	20.29	35.49	46.59	50.59	51.69	51.89	529
wkswp19 59	synthesized9	41.39	4.59	12.49	19.49	35.49	47.59	50.79	51.69	51.89	529
wkswp19 69	completed9	43.29	49	13.59	22.49	39.79	49.39	51.59	529	529	529
wkswp19 69	synthesized9	449	3.79	13.69	249	429	49.9	51.59	529	529	529
wkswp19 79	completed9	43.39	3.69	11.59	209	39	50.9	51.9	529	539	539
wkswp19 79	synthesized9	43.89	2.89	11.19	20.79	41.69	50.89	51.9	529	529	529
wkswp19 89	completed9	449	4.39	14.9	24.9	40.9	50.19	51.9	529	529	529
wkswp19 89	synthesized9	44.79	4.79	15.59	26.49	43.19	50.59	51.89	529	529	529
wkswp19	completed9	43.9	5.89	169	26.19	40.49	49.59	51.89	529	529	529
wkswp19	synthesized9	44.89	4.79	16.19	27.29	439	50.59	51.89	529	529	529

Table 61: Percentiles of Synthetic and Completed Variables9 Panel 11

Variable Name	Type9	Mean9	P019	P059	P109	P259	Median9	P759	P909	P959	P9
					Cardinal Categorical Variables9						
time_arrive_usa9	completed9	5.569	19	19	29	49	69	89	89	89	89
time_arrive_usa9	synthesized9	5.449	19	1	1.19	49	6	8	8	8	89
totfam_kids9	completed9	0.9	09	0	0	09	0	2	3	3	59
totfam_kids9	synthesized9	0.9	09	0	0	09	0	2	3	3	59

Appendix C

Revised Table 61

The following data were produced by Gary Benedetto, U. S. Census Bureau, to replace the data in Table 61 from “Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project.” Revisions were necessary because the weights for the Synthetic Beta File were recalculated. The other table that was used in our analysis, Table 29, was also revised. Table 4 in Chapter 3 of this report is the revised Table 29.

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
age_mar1_C	age_mar1	complete	23.4	15.8	17.3	18.1	19.8	22.3	25.6	29.9	33.3	42.3
age_mar1_S	age_mar1	masked	23.1	15.7	17.1	18	19.6	22.1	25.3	29.3	32.5	41
birthdate_C	birthdate	complete	-1812	-17145	-13760	-11438	-6096	-938	3365	6237	7190	7777
birthdate_S	birthdate	masked	-1834	-17230	-13806	-11495	-6105	-954	3389	6189	7189	7757
dt_int_ent_C	dt_int_ent	complete	10293	1438	3676	5022	7663	10889	13301	14754	15234	15584
dt_int_ent_S	dt_int_ent	masked	10292	1494	3677	5011	7653	10872	13296	14736	15213	15603
deathdate_C	deathdate	complete	15160	14712	14747	14806	14946	15159	15387	15514	15556	15597
deathdate_S	deathdate	masked	14983	13923	14148	14317	14692	15059	15315	15497	15576	15659
df_fica_1987_C	df_fica_1987	complete	1097	410	410	410	870	1101	1226	1637	2048	2048
df_fica_1987_S	df_fica_1987	masked	982	555	555	555	610	910	1348	1451	1451	1451
df_fica_1988_C	df_fica_1988	complete	7851	0	122	144	590	1800	3672	5874	7627	7627
df_fica_1988_S	df_fica_1988	masked	14291	12	90	194	617	1808	3890	5897	112623	254601
df_fica_1989_C	df_fica_1989	complete	3291	9	89	157	527	1200	2435	4742	6178	26488
df_fica_1989_S	df_fica_1989	masked	2390	24	112	210	504	1169	2542	4570	5685	18702
df_fica_1990_C	df_fica_1990	complete	2103	35	147	260	600	1320	2776	4959	6625	8350
df_fica_1990_S	df_fica_1990	masked	2016	32	136	251	569	1229	2618	4935	6645	8766
df_fica_1991_C	df_fica_1991	complete	2301	44	155	266	608	1366	2796	5094	6995	8596
df_fica_1991_S	df_fica_1991	masked	2136	44	175	297	606	1295	2678	5053	6941	9483
df_fica_1992_C	df_fica_1992	complete	2271	46	174	290	640	1437	2980	5407	7287	8728
df_fica_1992_S	df_fica_1992	masked	2240	61	205	332	646	1376	2877	5349	7360	9243
df_fica_1993_C	df_fica_1993	complete	2441	46	173	300	678	1500	3173	5737	7741	9488
df_fica_1993_S	df_fica_1993	masked	2365	56	204	340	678	1447	3073	5673	7755	9567
df_fica_1994_C	df_fica_1994	complete	2700	45	183	313	700	1579	3314	6000	8027	9240
df_fica_1994_S	df_fica_1994	masked	2538	49	206	350	706	1558	3215	5837	7885	9472
df_fica_1995_C	df_fica_1995	complete	2768	47	173	312	705	1614	3435	6254	8313	9240
df_fica_1995_S	df_fica_1995	masked	2589	56	233	398	810	1568	3310	6076	8221	9367
df_fica_1996_C	df_fica_1996	complete	2636	50	195	338	764	1710	3624	6654	8702	9500
df_fica_1996_S	df_fica_1996	masked	2613	58	233	403	816	1702	3542	6499	8566	9598
df_fica_1997_C	df_fica_1997	complete	2792	49	192	338	785	1808	3881	7170	9362	9500
df_fica_1997_S	df_fica_1997	masked	2779	61	246	424	858	1810	3820	7039	9159	9589
df_fica_1998_C	df_fica_1998	complete	2976	52	207	369	849	1923	4196	7711	9934	10000
df_fica_1998_S	df_fica_1998	masked	2964	64	266	454	915	1923	4083	7547	9643	10423
df_fica_1999_C	df_fica_1999	complete	3136	49	206	380	900	2021	4446	8073	10000	10000
df_fica_1999_S	df_fica_1999	masked	3125	57	236	425	914	1998	4315	7913	9830	11649
df_fica_2000_C	df_fica_2000	complete	3417	45	203	375	900	2076	4622	8542	10500	10500
df_fica_2000_S	df_fica_2000	masked	3381	64	268	468	975	2095	4548	8372	10345	11351
df_fica_2001_C	df_fica_2001	complete	3363	46	201	386	936	2144	4808	9013	10500	10500
df_fica_2001_S	df_fica_2001	masked	3362	79	300	500	1028	2169	4719	8748	10367	10938
df_fica_2002_C	df_fica_2002	complete	3482	45	200	390	953	2194	4900	9622	11000	12000

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
df_fica_2002_S	df_fica_2002	masked	3462	74	281	481	1007	2161	4764	9333	10963	12349
df_fica_2003_C	df_fica_2003	complete	3716	56	226	411	1000	2270	5111	10430	12000	14000
df_fica_2003_S	df_fica_2003	masked	3926	91	391	596	1184	2441	5298	10534	12021	14182
df_nfica_1987_C	df_nfica_1987	complete	6226	364	364	548	1369	2286	9771	16975	16975	16975
df_nfica_1987_S	df_nfica_1987	masked	4656	754	754	754	1941	3664	7765	9677	9677	9677
df_nfica_1988_C	df_nfica_1988	complete	1686	220	220	220	331	407	1112	6606	8409	10212
df_nfica_1988_S	df_nfica_1988	masked	1496	487	487	487	672	1060	1945	3577	3577	4058
df_nfica_1989_C	df_nfica_1989	complete	1833	41	72	783	1130	1571	1950	3585	3802	8200
df_nfica_1989_S	df_nfica_1989	masked	1833	49	240	565	1134	1679	2130	3287	4267	6276
df_nfica_1990_C	df_nfica_1990	complete	3810	73	295	482	960	1875	3383	6764	7990	12534
df_nfica_1990_S	df_nfica_1990	masked	3012	40	234	420	907	1737	3026	5706	7639	11882
df_nfica_1991_C	df_nfica_1991	complete	2567	25	186	344	827	1753	3145	6150	7500	12353
df_nfica_1991_S	df_nfica_1991	masked	2506	64	237	414	878	1695	3067	5870	7781	12297
df_nfica_1992_C	df_nfica_1992	complete	2449	7	67	221	719	1685	3172	6000	7500	11400
df_nfica_1992_S	df_nfica_1992	masked	2342	32	147	279	698	1557	2970	5641	7640	11647
df_nfica_1993_C	df_nfica_1993	complete	2512	5	61	209	677	1626	3308	6009	7500	12500
df_nfica_1993_S	df_nfica_1993	masked	2356	31	143	281	684	1531	2943	5629	7555	12298
df_nfica_1994_C	df_nfica_1994	complete	2428	6	85	199	650	1609	3130	6000	7500	11637
df_nfica_1994_S	df_nfica_1994	masked	2243	30	134	250	633	1469	2877	5449	7301	11204
df_nfica_1995_C	df_nfica_1995	complete	2469	12	85	195	644	1679	3190	6088	7500	12116
df_nfica_1995_S	df_nfica_1995	masked	2252	33	134	248	614	1481	2874	5464	7277	11335
df_nfica_1996_C	df_nfica_1996	complete	2467	7	63	218	650	1800	3321	6021	7500	11100
df_nfica_1996_S	df_nfica_1996	masked	2311	37	142	267	630	1553	3042	5642	7420	11112
df_nfica_1997_C	df_nfica_1997	complete	2659	7	77	224	769	1844	3498	6637	7664	12000
df_nfica_1997_S	df_nfica_1997	masked	2541	36	154	293	741	1731	3300	6186	7729	11753
df_nfica_1998_C	df_nfica_1998	complete	2791	11	98	259	833	1932	3756	6974	8000	12240
df_nfica_1998_S	df_nfica_1998	masked	2543	35	154	295	744	1731	3298	6346	7916	11583
df_nfica_1999_C	df_nfica_1999	complete	2871	11	100	237	809	1968	3873	7125	8026	12240
df_nfica_1999_S	df_nfica_1999	masked	2644	37	159	301	768	1794	3462	6561	8141	11682
df_nfica_2000_C	df_nfica_2000	complete	2917	12	108	266	837	2084	3918	7275	8000	11950
df_nfica_2000_S	df_nfica_2000	masked	2637	38	166	322	778	1827	3484	6473	8096	11148
df_nfica_2001_C	df_nfica_2001	complete	3140	22	121	275	828	2161	4296	7862	8552	12542
df_nfica_2001_S	df_nfica_2001	masked	2847	42	184	341	814	1901	3767	7215	8573	11402
df_nfica_2002_C	df_nfica_2002	complete	3547	18	148	334	978	2400	4921	8496	10999	13995
df_nfica_2002_S	df_nfica_2002	masked	3232	47	200	365	849	2034	4434	8097	10490	12874
df_nfica_2003_C	df_nfica_2003	complete	3765	21	140	301	899	2400	5295	9297	12000	14632
df_nfica_2003_S	df_nfica_2003	masked	3570	58	228	413	947	2181	4833	8733	11317	14526
duration_end1_C	duration_end1	complete	927	0	1	1	3	12	1961	1973	1977	1981
duration_end1_S	duration_end1	masked	928	0	1	1	4	126	1950	1974	1996	2046

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
duration_end2_C	duration_end2	complete	1078	0	1	2	4	1934	1960	1968	1972	1978
duration_end2_S	duration_end2	masked	1045	0	2	3	16	1871	1952	1981	2002	2040
duration_end3_C	duration_end3	complete	1813	1	6	1928	1942	1953	1961	1967	1969	1974
duration_end3_S	duration_end3	masked	1823	1	127	1927	1943	1954	1961	1966	1969	1973
duration_mar1_C	duration_mar1	complete	14.4	0.3	1.2	2.2	4.7	9.6	19.9	35.7	44.6	55.4
duration_mar1_S	duration_mar1	masked	13.8	0.4	1.3	2.2	4.5	9	18.9	34.2	43.2	54.3
duration_mar2_C	duration_mar2	complete	1172	0	2	4	9	1956	1970	1975	1978	1981
duration_mar2_S	duration_mar2	masked	1205	1	2	5	16	1945	1969	1976	1983	2045
duration_mar3_C	duration_mar3	complete	1301	0	2	3	11	1953	1964	1970	1973	1979
duration_mar3_S	duration_mar3	masked	1249	2	9	21	113	1850	1966	1984	2014	2087
duration_mar4_C	duration_mar4	complete	1955	1925	1936	1942	1950	1956	1962	1967	1970	1972
duration_mar4_S	duration_mar4	masked	1956	1923	1933	1942	1951	1957	1963	1968	1970	1973
earn37_51_C	earn37_51	complete	5552	9	58	152	682	2862	8064	14810	19513	30408
earn37_51_S	earn37_51	masked	5963	4	22	55	284	1974	8414	17634	25001	37638
fwelamt1990_C	fwelamt1990	complete	2170	25	112	210	528	1188	2839	5712	7746	11430
fwelamt1990_S	fwelamt1990	masked	2266	37	152	269	586	1239	2901	5899	8028	12243
fwelamt1991_C	fwelamt1991	complete	2650	36	152	284	662	1528	3605	6774	8888	13111
fwelamt1991_S	fwelamt1991	masked	2486	33	145	268	604	1346	3184	6536	8820	13578
fwelamt1992_C	fwelamt1992	complete	2319	21	101	200	513	1269	3148	6075	8048	12326
fwelamt1992_S	fwelamt1992	masked	2297	25	115	213	505	1189	2973	6161	8275	12881
fwelamt1993_C	fwelamt1993	complete	2462	26	115	223	565	1369	3412	6483	8438	12374
fwelamt1993_S	fwelamt1993	masked	2463	31	133	244	558	1275	3221	6688	9017	13189
fwelamt1994_C	fwelamt1994	complete	2254	20	105	205	521	1242	2995	5942	7769	11766
fwelamt1994_S	fwelamt1994	masked	2186	25	114	210	482	1087	2653	5988	8195	13114
fwelamt1995_C	fwelamt1995	complete	1905	20	97	191	482	1074	2256	5055	7160	10700
fwelamt1995_S	fwelamt1995	masked	1991	24	107	196	450	996	2230	5606	7848	12584
fwelamt1996_C	fwelamt1996	complete	2869	33	126	242	617	1517	3984	7434	9809	15112
fwelamt1996_S	fwelamt1996	masked	2931	42	171	304	674	1526	3935	7508	10130	16339
fwelamt1997_C	fwelamt1997	complete	2575	29	120	222	558	1332	3299	6977	9348	13343
fwelamt1997_S	fwelamt1997	masked	2535	37	151	266	582	1269	3058	6708	9461	15466
fwelamt1998_C	fwelamt1998	complete	2219	22	103	191	485	1158	2762	6008	8147	12521
fwelamt1998_S	fwelamt1998	masked	2131	31	125	221	487	1072	2451	5703	8221	13791
fwelamt1999_C	fwelamt1999	complete	1919	27	115	201	469	1023	2260	4879	7253	11991
fwelamt1999_S	fwelamt1999	masked	1930	34	137	233	485	1000	2188	4920	7376	12329
fpov1990_C	fpov1990	complete	10831	6166	6652	7013	8324	9954	13135	15666	17364	22600
fpov1990_S	fpov1990	masked	11095	6221	6724	7180	8528	10215	13336	15918	17596	22707
fpov1991_C	fpov1991	complete	11351	6393	6854	7205	8676	10452	13806	16458	18248	23269
fpov1991_S	fpov1991	masked	11701	6434	6941	7393	8877	10747	14063	16811	18651	23687
fpov1992_C	fpov1992	complete	11628	6612	7014	7368	8891	10676	14210	16858	18735	24140

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
fpov1992_S	fpov1992	masked	12058	6651	7135	7603	9156	11076	14494	17312	19182	24624
fpov1993_C	fpov1993	complete	11997	6820	7258	7567	9127	11006	14643	17517	19363	24974
fpov1993_S	fpov1993	masked	12315	6857	7326	7731	9312	11308	14881	17811	19678	24991
fpov1994_C	fpov1994	complete	12255	6796	7328	7715	9266	11277	15020	17822	19665	24931
fpov1994_S	fpov1994	masked	12483	6869	7399	7854	9420	11500	15194	18043	19852	24933
fpov1995_C	fpov1995	complete	12489	6614	7406	7949	9338	11356	15322	18245	20154	26184
fpov1995_S	fpov1995	masked	12780	6723	7498	8069	9514	11671	15581	18562	20561	26694
fpov1996_C	fpov1996	complete	12974	7333	7732	8157	9563	11712	15929	19189	21282	27644
fpov1996_S	fpov1996	masked	13043	7382	7811	8258	9720	11852	15961	19117	21088	26885
fpov1997_C	fpov1997	complete	12947	7424	7802	8319	9698	11683	16271	19202	21307	26748
fpov1997_S	fpov1997	masked	13231	7463	7893	8357	9867	12033	16369	19530	21511	26910
fpov1998_C	fpov1998	complete	13275	7431	7911	8456	9874	12160	16524	19574	21602	26948
fpov1998_S	fpov1998	masked	13410	7460	7997	8483	10022	12237	16583	19755	21712	27192
fpov1999_C	fpov1999	complete	13582	7717	8109	8652	10066	12396	16886	20050	22096	27790
fpov1999_S	fpov1999	masked	13779	7690	8183	8686	10232	12545	16959	20280	22408	28606
ftotinc1990_C	ftotinc1990	complete	37323	1157	5886	9431	17975	31674	49946	71765	89165	128928
ftotinc1990_S	ftotinc1990	masked	36264	1148	6017	9508	17880	31056	48388	68959	85081	123200
ftotinc1991_C	ftotinc1991	complete	39099	1887	6619	10212	19272	33724	52720	74881	90929	124220
ftotinc1991_S	ftotinc1991	masked	37344	827	5654	9085	17534	31424	50412	73221	90150	126176
ftotinc1992_C	ftotinc1992	complete	39329	1402	6303	9890	18979	33525	53410	76495	92803	126621
ftotinc1992_S	ftotinc1992	masked	38605	1021	6050	9582	18378	32550	52236	75739	92380	127798
ftotinc1993_C	ftotinc1993	complete	40716	2204	7091	10801	19911	34911	54949	78120	95156	130430
ftotinc1993_S	ftotinc1993	masked	39319	1329	6335	9895	18686	33336	53092	76589	93912	130144
ftotinc1994_C	ftotinc1994	complete	42267	1690	6887	10749	20130	35851	57184	82766	100927	138823
ftotinc1994_S	ftotinc1994	masked	40711	905	6199	9929	18896	34138	55278	80735	99065	137872
ftotinc1995_C	ftotinc1995	complete	43530	1377	6814	11023	20784	37592	59287	84157	101050	143272
ftotinc1995_S	ftotinc1995	masked	42320	-212	5056	9101	18637	35459	58384	84863	103746	147094
ftotinc1996_C	ftotinc1996	complete	47327	1277	6741	10724	20740	37515	61271	90713	113750	210824
ftotinc1996_S	ftotinc1996	masked	44318	319	6009	9953	19427	35316	57617	85686	107421	194394
ftotinc1997_C	ftotinc1997	complete	47038	1362	6520	10252	19946	36851	61252	91179	115438	216712
ftotinc1997_S	ftotinc1997	masked	45767	676	6224	10122	19783	36117	59486	88531	112136	203282
ftotinc1998_C	ftotinc1998	complete	49700	1385	6838	10791	21029	38978	64931	96515	121985	235057
ftotinc1998_S	ftotinc1998	masked	48476	704	6538	10698	20829	38027	62955	94064	119733	227099
ftotinc1999_C	ftotinc1999	complete	53345	1510	6975	11177	22046	41268	68193	103010	134751	277340
ftotinc1999_S	ftotinc1999	masked	51671	388	6360	10758	21057	39098	65235	99913	132814	281831
helamt1990_C	helamt1990	complete	2200	38	152	283	590	1142	2385	4992	7947	17223
helamt1990_S	helamt1990	masked	2036	30	136	248	531	1029	2110	4548	7498	17101
helamt1991_C	helamt1991	complete	2865	50	221	396	813	1530	3207	6523	9839	20648
helamt1991_S	helamt1991	masked	2764	50	205	364	757	1433	2912	6333	9763	21262

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
helamt1992_C	helamt1992	complete	3167	57	213	389	830	1678	3486	7467	11187	20914
helamt1992_S	helamt1992	masked	3021	49	210	374	785	1537	3191	6858	10910	22320
helamt1993_C	helamt1993	complete	2912	47	193	360	793	1601	3312	6610	9737	20876
helamt1993_S	helamt1993	masked	3035	50	211	371	784	1554	3237	6875	10972	23183
helamt1994_C	helamt1994	complete	3240	53	224	405	892	1768	3773	7465	10996	22209
helamt1994_S	helamt1994	masked	3279	51	219	390	845	1665	3600	7498	11833	24704
helamt1995_C	helamt1995	complete	3956	122	434	708	1345	2517	4707	8602	13137	21231
helamt1995_S	helamt1995	masked	3977	109	395	655	1253	2378	4546	8432	13612	23690
helamt1996_C	helamt1996	complete	3253	53	243	428	899	1778	3549	6643	10353	27637
helamt1996_S	helamt1996	masked	3640	75	299	505	975	1838	3617	7275	11918	36881
helamt1997_C	helamt1997	complete	4095	84	298	510	1050	2114	4555	8855	13554	24132
helamt1997_S	helamt1997	masked	5088	114	387	640	1212	2342	4952	9978	16125	29157
helamt1998_C	helamt1998	complete	4779	113	365	644	1186	2344	4735	9686	14094	26772
helamt1998_S	helamt1998	masked	5791	120	417	683	1288	2443	4742	10180	16056	29428
helamt1999_C	helamt1999	complete	4897	120	442	719	1318	2568	5015	9390	14040	26683
helamt1999_S	helamt1999	masked	6547	107	429	730	1392	2663	5340	10802	18250	58343
homeequity_C	homeequity	complete	71786	-9125	4000	8000	21375	50000	100000	160500	213250	319250
homeequity_S	homeequity	masked	72857	-17874	2430	6275	19886	49731	99550	171295	231991	336918
mba_2000_C	mba_2000	complete	643	28	91	155	352	611	919	1137	1261	1553
mba_2000_S	mba_2000	masked	637	33	90	151	345	603	915	1130	1253	1534
mba_initial_real_C	mba_initial_real	complete	611	35	106	165	337	551	875	1119	1241	1435
mba_initial_real_S	mba_initial_real	masked	608	42	106	167	334	547	874	1114	1232	1433
ndf_fica_1978_C	ndf_fica_1978	complete	13235	51	322	754	2589	7257	14091	21106	27756	67066
ndf_fica_1978_S	ndf_fica_1978	masked	13354	101	490	978	2770	7268	14038	21068	27809	70905
ndf_fica_1979_C	ndf_fica_1979	complete	13389	58	356	838	2805	7817	15005	22900	28939	58699
ndf_fica_1979_S	ndf_fica_1979	masked	13108	75	427	1000	3075	7975	15097	22866	28496	54570
ndf_fica_1980_C	ndf_fica_1980	complete	12013	60	361	848	2938	8357	16093	25106	30369	53486
ndf_fica_1980_S	ndf_fica_1980	masked	11830	153	670	1263	3378	8400	16034	24551	29499	53203
ndf_fica_1981_C	ndf_fica_1981	complete	12680	63	410	1003	3373	9422	17802	27451	33434	57022
ndf_fica_1981_S	ndf_fica_1981	masked	12580	106	546	1186	3600	9412	17696	27037	32625	55340
ndf_fica_1982_C	ndf_fica_1982	complete	13520	72	450	1051	3665	10182	18840	29003	35452	60764
ndf_fica_1982_S	ndf_fica_1982	masked	13323	160	711	1402	3887	10176	18526	28326	34554	58971
ndf_fica_1983_C	ndf_fica_1983	complete	14276	70	462	1073	3784	10559	19798	30597	37642	65436
ndf_fica_1983_S	ndf_fica_1983	masked	14016	108	592	1307	3852	10389	19241	29662	36715	63778
ndf_fica_1984_C	ndf_fica_1984	complete	15159	79	487	1139	4072	11241	21003	32855	40007	70227
ndf_fica_1984_S	ndf_fica_1984	masked	15055	168	787	1535	4279	11363	20737	32147	39263	67856
ndf_fica_1985_C	ndf_fica_1985	complete	15944	75	497	1181	4166	11681	22072	34517	41928	74850
ndf_fica_1985_S	ndf_fica_1985	masked	15853	127	641	1416	4300	11633	21809	33788	41961	75180
ndf_fica_1986_C	ndf_fica_1986	complete	16818	82	512	1201	4342	12204	23099	36301	44605	81790

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
ndf_fica_1986_S	ndf_fica_1986	masked	16857	132	674	1452	4424	12122	22882	35558	43953	81374
ndf_fica_1987_C	ndf_fica_1987	complete	17538	81	522	1230	4507	12744	23884	37479	45403	85118
ndf_fica_1987_S	ndf_fica_1987	masked	17394	146	734	1530	4626	12456	23507	36902	45485	83355
ndf_fica_1988_C	ndf_fica_1988	complete	18432	88	562	1300	4710	13272	24957	39404	47922	90000
ndf_fica_1988_S	ndf_fica_1988	masked	18128	86	553	1407	4687	12953	24463	38535	47361	87483
ndf_fica_1989_C	ndf_fica_1989	complete	19054	97	598	1397	5015	13852	25950	40818	49844	94143
ndf_fica_1989_S	ndf_fica_1989	masked	18870	141	727	1579	4970	13697	25548	40066	49427	92359
ndf_fica_1990_C	ndf_fica_1990	complete	19775	95	626	1493	5276	14423	26729	41363	51402	96779
ndf_fica_1990_S	ndf_fica_1990	masked	19609	147	794	1756	5315	14268	26405	40660	51666	95638
ndf_fica_1991_C	ndf_fica_1991	complete	20703	100	635	1516	5504	14787	27465	42822	55359	110013
ndf_fica_1991_S	ndf_fica_1991	masked	20548	127	703	1661	5430	14671	27015	42002	54931	110801
ndf_fica_1992_C	ndf_fica_1992	complete	21777	97	615	1493	5576	15324	28690	44940	58382	120989
ndf_fica_1992_S	ndf_fica_1992	masked	21747	142	769	1781	5530	15083	28320	44056	57679	118190
ndf_fica_1993_C	ndf_fica_1993	complete	22512	95	608	1542	5806	15730	29575	46638	61371	128330
ndf_fica_1993_S	ndf_fica_1993	masked	22667	204	960	2047	6056	15614	29129	46062	61342	126883
ndf_fica_1994_C	ndf_fica_1994	complete	22910	94	613	1555	5859	15963	29866	46963	61750	126632
ndf_fica_1994_S	ndf_fica_1994	masked	22865	202	946	2025	6097	15765	29517	46580	61561	125411
ndf_fica_1995_C	ndf_fica_1995	complete	23875	96	638	1597	6069	16462	30653	48528	64014	133049
ndf_fica_1995_S	ndf_fica_1995	masked	23788	260	1161	2331	6638	16482	30268	47811	63208	130391
ndf_fica_1996_C	ndf_fica_1996	complete	25611	103	724	1706	6334	17052	31640	50252	66732	139269
ndf_fica_1996_S	ndf_fica_1996	masked	25127	184	972	2206	6839	17251	31506	49801	65868	142079
ndf_fica_1997_C	ndf_fica_1997	complete	26232	116	857	1986	6901	18055	33288	53135	71128	153480
ndf_fica_1997_S	ndf_fica_1997	masked	26373	173	976	2303	7198	18055	33313	53188	71331	151148
ndf_fica_1998_C	ndf_fica_1998	complete	28118	149	1024	2371	7786	19421	35133	55872	75009	160681
ndf_fica_1998_S	ndf_fica_1998	masked	28446	266	1309	2813	8170	19551	35343	55380	73949	159486
ndf_fica_1999_C	ndf_fica_1999	complete	30005	165	1192	2813	8748	20639	36694	58346	79074	171608
ndf_fica_1999_S	ndf_fica_1999	masked	30543	326	1673	3514	9343	20950	36850	57854	78007	169977
ndf_fica_2000_C	ndf_fica_2000	complete	32220	153	1253	2972	9486	21743	38294	61111	83548	182011
ndf_fica_2000_S	ndf_fica_2000	masked	32439	339	1834	3862	10223	22303	38617	60955	83309	176991
ndf_fica_2001_C	ndf_fica_2001	complete	33463	169	1393	3481	10824	23516	40479	64153	87329	185052
ndf_fica_2001_S	ndf_fica_2001	masked	33413	362	1982	4256	11240	23891	40425	63237	85404	180531
ndf_fica_2002_C	ndf_fica_2002	complete	34149	136	1332	3506	11146	24476	41869	66223	89513	189259
ndf_fica_2002_S	ndf_fica_2002	masked	35020	328	1861	4226	11491	24624	41868	65842	88587	186556
ndf_fica_2003_C	ndf_fica_2003	complete	35233	135	1449	3658	11742	25503	43279	68146	91912	194341
ndf_fica_2003_S	ndf_fica_2003	masked	36087	395	2159	4633	12001	25402	43075	67410	91381	201120
ndf_nfica_1978_C	ndf_nfica_1978	complete	8479	19	75	184	814	5793	13486	20077	25509	38022
ndf_nfica_1978_S	ndf_nfica_1978	masked	8450	38	173	340	1070	5646	13267	19492	24321	36576
ndf_nfica_1979_C	ndf_nfica_1979	complete	8793	17	81	199	864	6343	14448	21036	25182	36129
ndf_nfica_1979_S	ndf_nfica_1979	masked	8778	81	328	600	1530	6104	14139	20385	24534	35294

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
ndf_nfica_1980_C	ndf_nfica_1980	complete	9840	21	100	248	1110	7311	15988	22956	27938	40085
ndf_nfica_1980_S	ndf_nfica_1980	masked	9791	103	387	686	1710	6609	15791	22873	27617	39630
ndf_nfica_1981_C	ndf_nfica_1981	complete	10873	22	86	210	862	6759	17789	26429	32703	50026
ndf_nfica_1981_S	ndf_nfica_1981	masked	10359	79	350	638	1579	5948	16618	25115	30761	46800
ndf_nfica_1982_C	ndf_nfica_1982	complete	11131	16	71	164	637	4434	19663	28631	34978	54827
ndf_nfica_1982_S	ndf_nfica_1982	masked	10796	73	305	549	1342	4435	18401	27672	33868	51809
ndf_nfica_1983_C	ndf_nfica_1983	complete	12176	19	74	187	755	6700	21625	30018	36079	53621
ndf_nfica_1983_S	ndf_nfica_1983	masked	11848	91	382	689	1724	6395	19943	28574	34952	51017
ndf_nfica_1984_C	ndf_nfica_1984	complete	13722	18	75	168	745	8826	23985	31997	38473	57127
ndf_nfica_1984_S	ndf_nfica_1984	masked	12762	83	335	599	1488	6017	22239	31272	37914	55739
ndf_nfica_1985_C	ndf_nfica_1985	complete	13364	20	72	161	700	6560	24455	33278	39534	57487
ndf_nfica_1985_S	ndf_nfica_1985	masked	13942	68	289	533	1450	8156	24845	33281	39590	55670
ndf_nfica_1986_C	ndf_nfica_1986	complete	14221	20	75	187	784	7071	25334	34341	41685	66118
ndf_nfica_1986_S	ndf_nfica_1986	masked	13666	73	296	548	1404	5821	24241	33594	40725	63627
ndf_nfica_1987_C	ndf_nfica_1987	complete	14314	23	82	194	801	7007	25732	35257	42612	61774
ndf_nfica_1987_S	ndf_nfica_1987	masked	13787	90	364	661	1706	6316	23963	34423	41681	59228
ndf_nfica_1988_C	ndf_nfica_1988	complete	15372	21	87	204	820	8259	27476	37444	44762	65211
ndf_nfica_1988_S	ndf_nfica_1988	masked	14647	69	258	448	1052	6388	26293	36679	44076	65792
ndf_nfica_1989_C	ndf_nfica_1989	complete	16550	20	84	208	824	8026	28480	39159	47711	68315
ndf_nfica_1989_S	ndf_nfica_1989	masked	15440	67	287	523	1290	6452	27304	38794	47308	69808
ndf_nfica_1990_C	ndf_nfica_1990	complete	16116	24	92	218	892	7780	29105	40768	49403	70388
ndf_nfica_1990_S	ndf_nfica_1990	masked	15249	78	312	554	1309	5781	27486	39836	47869	68538
ndf_nfica_1991_C	ndf_nfica_1991	complete	18104	26	100	244	1064	11062	31836	44290	52970	74421
ndf_nfica_1991_S	ndf_nfica_1991	masked	17179	88	386	710	1973	8416	30035	42793	51670	71193
ndf_nfica_1992_C	ndf_nfica_1992	complete	18915	24	102	255	1090	11454	33017	46223	55413	79754
ndf_nfica_1992_S	ndf_nfica_1992	masked	17787	105	411	755	1958	8372	30699	44741	54110	77245
ndf_nfica_1993_C	ndf_nfica_1993	complete	19983	30	121	314	1264	11429	33897	48486	59262	91816
ndf_nfica_1993_S	ndf_nfica_1993	masked	19126	108	442	795	1963	9091	32401	47214	58354	91361
ndf_nfica_1994_C	ndf_nfica_1994	complete	20219	24	113	303	1251	11577	34076	48661	58763	87251
ndf_nfica_1994_S	ndf_nfica_1994	masked	19291	112	478	917	2463	9866	32342	47301	57085	81943
ndf_nfica_1995_C	ndf_nfica_1995	complete	20904	28	112	287	1310	11685	34738	50387	62180	100293
ndf_nfica_1995_S	ndf_nfica_1995	masked	19876	104	464	841	2037	9420	33224	48384	59472	97931
ndf_nfica_1996_C	ndf_nfica_1996	complete	21167	35	118	274	1228	11384	35438	51028	61935	94749
ndf_nfica_1996_S	ndf_nfica_1996	masked	19962	114	474	883	2257	8988	33171	49631	60821	88907
ndf_nfica_1997_C	ndf_nfica_1997	complete	22375	36	147	350	1553	12318	36620	53068	65992	107607
ndf_nfica_1997_S	ndf_nfica_1997	masked	21073	130	548	1024	2643	9926	34148	51509	64514	99155
ndf_nfica_1998_C	ndf_nfica_1998	complete	22333	49	177	393	1517	11869	36949	53597	66105	107825
ndf_nfica_1998_S	ndf_nfica_1998	masked	21599	152	612	1134	2811	10407	35062	52781	65222	102084
ndf_nfica_1999_C	ndf_nfica_1999	complete	23635	40	182	450	1783	12975	37737	55300	67761	107150

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
ndf_nfica_1999_S	ndf_nfica_1999	masked	22572	157	709	1294	3288	11477	35956	53745	66893	102617
ndf_nfica_2000_C	ndf_nfica_2000	complete	24287	43	177	444	1778	13000	39024	56956	69946	114863
ndf_nfica_2000_S	ndf_nfica_2000	masked	23114	178	728	1344	3286	11608	36591	55688	69394	107134
ndf_nfica_2001_C	ndf_nfica_2001	complete	25217	35	171	465	1965	14086	40340	58824	72293	112430
ndf_nfica_2001_S	ndf_nfica_2001	masked	25051	215	884	1651	4216	13739	38754	58894	72680	110865
ndf_nfica_2002_C	ndf_nfica_2002	complete	28504	46	230	620	3010	19967	44117	63792	78223	123295
ndf_nfica_2002_S	ndf_nfica_2002	masked	27809	137	632	1280	4280	18892	42773	62462	77166	116130
ndf_nfica_2003_C	ndf_nfica_2003	complete	29400	45	238	669	3536	21627	45949	65495	78897	124754
ndf_nfica_2003_S	ndf_nfica_2003	masked	32181	358	1582	2956	8346	26536	47501	66076	80599	119049
nonhouswealth_C	nonhouswealth	complete	74203	-6000	1000	2000	6000	17000	60000	180000	314500	761500
nonhouswealth_S	nonhouswealth	masked	69986	-48703	369	1336	5204	15994	56948	171958	307796	831669
time_arrive_usa_C	time_arrive_usa	complete	5.51	1	1	1.75	4	6	8	8	8	8
time_arrive_usa_S	time_arrive_usa	masked	5.39	1	1	1	4	6	8	8	8	8
totearn1990_C	totearn1990	complete	17882	115	728	1650	5549	14145	25163	38332	48792	78102
totearn1990_S	totearn1990	masked	16957	91	579	1230	4305	13125	24228	37275	47702	77040
totearn1991_C	totearn1991	complete	18755	156	864	1826	5731	14623	26251	40445	51776	84105
totearn1991_S	totearn1991	masked	18347	122	689	1466	4918	14062	25993	40312	51724	85013
totearn1992_C	totearn1992	complete	19624	166	953	2054	6244	15212	27314	42252	54698	87686
totearn1992_S	totearn1992	masked	18677	111	734	1633	5266	14216	26254	40887	53092	86891
totearn1993_C	totearn1993	complete	19790	160	954	2038	6299	15354	27529	42544	55488	88527
totearn1993_S	totearn1993	masked	19474	123	743	1613	5375	14907	27487	42773	55528	89253
totearn1994_C	totearn1994	complete	20148	208	1129	2305	6537	15502	27915	43344	56223	90764
totearn1994_S	totearn1994	masked	19596	136	826	1782	5583	14904	27452	42975	55974	91346
totearn1995_C	totearn1995	complete	20137	308	1596	3046	7195	15401	27375	42634	54894	91742
totearn1995_S	totearn1995	masked	19635	361	1640	2952	6736	14614	26751	42158	54560	91477
totearn1996_C	totearn1996	complete	24722	188	1259	2693	7751	17784	32000	50053	66934	140612
totearn1996_S	totearn1996	masked	23756	211	1187	2455	7075	16919	30899	48443	64638	135364
totearn1997_C	totearn1997	complete	24938	253	1387	2929	8128	18197	32314	50204	66035	131836
totearn1997_S	totearn1997	masked	25031	229	1229	2569	7557	17881	32265	50710	67406	142095
totearn1998_C	totearn1998	complete	25743	296	1644	3342	8737	18900	33187	52038	68040	135278
totearn1998_S	totearn1998	masked	26378	260	1433	2980	8378	18940	33618	53383	70783	150977
totearn1999_C	totearn1999	complete	28248	391	2126	4114	9933	20399	35492	55868	74473	175018
totearn1999_S	totearn1999	masked	28597	219	1530	3346	9209	20129	35612	56783	76938	198424
tot_ser_1951_C	tot_ser_1951	complete	1658	8	47	112	442	1460	2820	3600	3600	3600
tot_ser_1951_S	tot_ser_1951	masked	1565	14	65	132	395	1272	2667	3600	3600	3600
tot_ser_1952_C	tot_ser_1952	complete	1753	10	52	123	460	1604	3083	3600	3600	3600
tot_ser_1952_S	tot_ser_1952	masked	1672	19	81	155	431	1435	2907	3600	3600	3600
tot_ser_1953_C	tot_ser_1953	complete	1854	10	55	138	520	1769	3356	3600	3600	3600
tot_ser_1953_S	tot_ser_1953	masked	1776	19	84	165	474	1592	3206	3600	3600	3600

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
tot_ser_1954_C	tot_ser_1954	complete	1893	10	54	133	536	1849	3475	3600	3600	3600
tot_ser_1954_S	tot_ser_1954	masked	1811	20	87	171	504	1658	3266	3600	3600	3600
tot_ser_1955_C	tot_ser_1955	complete	2058	14	67	152	550	1873	3660	4200	4200	4200
tot_ser_1955_S	tot_ser_1955	masked	1988	22	97	188	535	1750	3474	4200	4200	4200
tot_ser_1956_C	tot_ser_1956	complete	2176	14	74	172	646	2080	3914	4200	4200	4200
tot_ser_1956_S	tot_ser_1956	masked	2125	25	110	213	627	1980	3754	4200	4200	4200
tot_ser_1957_C	tot_ser_1957	complete	2252	16	88	217	776	2190	4063	4200	4200	4200
tot_ser_1957_S	tot_ser_1957	masked	2203	29	124	246	737	2088	3885	4200	4200	4200
tot_ser_1958_C	tot_ser_1958	complete	2282	14	83	204	776	2240	4194	4200	4200	4200
tot_ser_1958_S	tot_ser_1958	masked	2210	27	118	235	741	2168	3855	4143	4190	4253
tot_ser_1959_C	tot_ser_1959	complete	2506	17	93	220	813	2404	4462	4800	4800	4800
tot_ser_1959_S	tot_ser_1959	masked	2446	28	124	246	762	2300	4271	4800	4800	4800
tot_ser_1960_C	tot_ser_1960	complete	2560	18	103	237	857	2475	4631	4800	4800	4800
tot_ser_1960_S	tot_ser_1960	masked	2484	32	135	261	782	2334	4447	4800	4800	4800
tot_ser_1961_C	tot_ser_1961	complete	2604	20	107	247	896	2549	4785	4800	4800	4800
tot_ser_1961_S	tot_ser_1961	masked	2497	30	132	258	782	2342	4520	4800	4800	4800
tot_ser_1962_C	tot_ser_1962	complete	2675	18	107	274	955	2668	4800	4800	4800	4800
tot_ser_1962_S	tot_ser_1962	masked	2587	33	146	287	864	2516	4630	4800	4800	4800
tot_ser_1963_C	tot_ser_1963	complete	2719	19	113	271	996	2766	4800	4800	4800	4800
tot_ser_1963_S	tot_ser_1963	masked	2633	33	146	286	889	2616	4676	4800	4800	4800
tot_ser_1964_C	tot_ser_1964	complete	2796	20	118	287	1028	2939	4800	4800	4800	4800
tot_ser_1964_S	tot_ser_1964	masked	2682	32	142	282	899	2729	4765	4800	4800	4800
tot_ser_1965_C	tot_ser_1965	complete	2833	23	130	292	1048	3026	4800	4800	4800	4800
tot_ser_1965_S	tot_ser_1965	masked	2746	37	160	308	942	2846	4800	4800	4800	4800
tot_ser_1966_C	tot_ser_1966	complete	3383	23	132	313	1088	3148	6009	6600	6600	6600
tot_ser_1966_S	tot_ser_1966	masked	3264	36	164	330	1005	2966	5678	6600	6600	6600
tot_ser_1967_C	tot_ser_1967	complete	3480	24	139	338	1172	3309	6269	6600	6600	6600
tot_ser_1967_S	tot_ser_1967	masked	3368	42	181	354	1067	3105	6004	6600	6600	6600
tot_ser_1968_C	tot_ser_1968	complete	3900	28	158	364	1268	3588	6740	7800	7800	7800
tot_ser_1968_S	tot_ser_1968	masked	3763	46	198	384	1157	3373	6431	7800	7800	7800
tot_ser_1969_C	tot_ser_1969	complete	4061	29	177	399	1354	3824	7253	7800	7800	7800
tot_ser_1969_S	tot_ser_1969	masked	3952	47	208	411	1248	3646	6964	7800	7800	7800
tot_ser_1970_C	tot_ser_1970	complete	4219	28	178	407	1446	4116	7667	7800	7800	7800
tot_ser_1970_S	tot_ser_1970	masked	4084	49	216	425	1316	3888	7262	7800	7800	7800
tot_ser_1971_C	tot_ser_1971	complete	4347	28	179	414	1510	4383	7800	7800	7800	7800
tot_ser_1971_S	tot_ser_1971	masked	4218	51	223	441	1375	4137	7533	7800	7800	7800
tot_ser_1972_C	tot_ser_1972	complete	4803	31	194	449	1601	4630	8692	9000	9000	9000
tot_ser_1972_S	tot_ser_1972	masked	4662	53	237	469	1470	4417	8260	9000	9000	9000
tot_ser_1973_C	tot_ser_1973	complete	5411	38	213	488	1714	4991	9343	10800	10800	10800

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
tot_ser_1973_S	tot_ser_1973	masked	5264	56	252	500	1574	4797	9026	10800	10800	10800
tot_ser_1974_C	tot_ser_1974	complete	6103	36	230	543	1928	5374	9986	13200	13200	13200
tot_ser_1974_S	tot_ser_1974	masked	5912	63	283	565	1738	5130	9733	13200	13200	13200
tot_ser_1975_C	tot_ser_1975	complete	6479	44	249	563	1994	5716	10649	14100	14100	14100
tot_ser_1975_S	tot_ser_1975	masked	6280	69	301	596	1817	5445	10335	14100	14100	14100
tot_ser_1976_C	tot_ser_1976	complete	7028	46	277	626	2197	6163	11512	15300	15300	15300
tot_ser_1976_S	tot_ser_1976	masked	6820	75	346	679	2019	5864	11209	15300	15300	15300
tot_ser_1977_C	tot_ser_1977	complete	7544	50	295	669	2345	6588	12394	16500	16500	16500
tot_ser_1977_S	tot_ser_1977	masked	7223	82	362	714	2127	6099	11814	16500	16500	16500
totfam_kids_C	totfam_kids	complete	0.9	0	0	0	0	0	2	3	3	5
totfam_kids_S	totfam_kids	masked	0.9	0	0	0	0	0	2	3	3	5
tothrs_ann1990_C	tothrs_ann1990	complete	1680	45	184	353	990	1954	2201	2599	2918	3540
tothrs_ann1990_S	tothrs_ann1990	masked	1564	23	116	243	784	1856	2156	2492	2766	3357
tothrs_ann1991_C	tothrs_ann1991	complete	1615	49	182	334	871	1887	2150	2536	2851	3617
tothrs_ann1991_S	tothrs_ann1991	masked	1555	27	138	271	796	1830	2135	2461	2748	3346
tothrs_ann1992_C	tothrs_ann1992	complete	1693	57	220	428	1095	1964	2173	2524	2818	3455
tothrs_ann1992_S	tothrs_ann1992	masked	1590	46	188	344	900	1850	2133	2452	2740	3348
tothrs_ann1993_C	tothrs_ann1993	complete	1719	55	212	396	1035	1980	2214	2681	3010	3812
tothrs_ann1993_S	tothrs_ann1993	masked	1652	52	191	345	914	1913	2194	2620	2934	3684
tothrs_ann1994_C	tothrs_ann1994	complete	1663	65	231	414	996	1937	2158	2565	2865	3486
tothrs_ann1994_S	tothrs_ann1994	masked	1611	74	237	402	917	1850	2145	2510	2801	3395
tothrs_ann1995_C	tothrs_ann1995	complete	1630	114	326	514	989	1850	2139	2461	2735	3252
tothrs_ann1995_S	tothrs_ann1995	masked	1527	122	310	480	881	1699	2084	2318	2550	3057
tothrs_ann1996_C	tothrs_ann1996	complete	1756	62	242	451	1144	1975	2250	2711	3024	3655
tothrs_ann1996_S	tothrs_ann1996	masked	1650	63	220	393	986	1879	2184	2570	2858	3456
tothrs_ann1997_C	tothrs_ann1997	complete	1742	58	242	459	1142	1995	2205	2636	2950	3649
tothrs_ann1997_S	tothrs_ann1997	masked	1632	45	184	350	950	1894	2161	2541	2836	3492
tothrs_ann1998_C	tothrs_ann1998	complete	1738	64	247	477	1140	1993	2186	2615	2929	3590
tothrs_ann1998_S	tothrs_ann1998	masked	1680	53	209	402	1034	1934	2177	2572	2888	3576
tothrs_ann1999_C	tothrs_ann1999	complete	1759	72	285	541	1206	2002	2187	2604	2896	3560
tothrs_ann1999_S	tothrs_ann1999	masked	1668	46	206	412	1011	1914	2169	2573	2887	3601
totinc1990_C	totinc1990	complete	16869	-804	169	1061	4791	12962	24101	37164	47583	74645
totinc1990_S	totinc1990	masked	16406	-844	53	788	4048	12138	23655	37172	47817	75996
totinc1991_C	totinc1991	complete	17565	-633	219	1060	4923	13268	24961	39015	50064	81464
totinc1991_S	totinc1991	masked	17056	-598	266	958	4217	12264	24345	38911	50354	82683
totinc1992_C	totinc1992	complete	18075	-679	215	1158	5218	13431	25446	40285	52237	85280
totinc1992_S	totinc1992	masked	17651	-547	289	1048	4681	12612	24936	40157	52414	86450
totinc1993_C	totinc1993	complete	18619	-614	381	1559	5666	13980	26046	40928	53017	86578
totinc1993_S	totinc1993	masked	18196	-510	442	1395	5197	13261	25570	40741	52858	86726

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
totinc1994_C	totinc1994	complete	18998	-614	795	2131	6148	14277	26328	41623	53739	87122
totinc1994_S	totinc1994	masked	18840	-464	820	1947	5665	13645	26114	42141	54820	90452
totinc1995_C	totinc1995	complete	19337	-697	1166	2738	6590	14401	26580	41763	53537	88652
totinc1995_S	totinc1995	masked	19237	-596	1187	2553	6051	13600	26349	42684	55854	92374
totinc1996_C	totinc1996	complete	22289	-673	990	2668	7133	15936	29326	46414	60281	113725
totinc1996_S	totinc1996	masked	22121	-390	1008	2510	6822	15474	29015	46453	60776	116452
totinc1997_C	totinc1997	complete	23600	-186	1219	3058	7736	17016	30839	48448	63223	120123
totinc1997_S	totinc1997	masked	23169	-233	1194	2756	7229	16290	30148	48098	63217	123293
totinc1998_C	totinc1998	complete	24333	-289	1469	3444	8209	17783	31895	50464	65577	118144
totinc1998_S	totinc1998	masked	23943	-241	1435	3139	7739	17127	31282	50056	65568	120844
totinc1999_C	totinc1999	complete	26640	-257	1867	3916	8888	18613	33643	53599	71842	166974
totinc1999_S	totinc1999	masked	26761	-188	1816	3676	8404	18082	33484	54281	73487	179658
totnetworth_C	totnetworth	complete	118620	-33000	-6000	1000	9000	51000	140000	292750	446500	920000
totnetworth_S	totnetworth	masked	109788	-34630	-5855	-132	8365	50257	135722	277707	414051	835547
wkspt1990_C	wkspt1990	complete	15.8	0.3	1.2	2.2	5.2	12.4	22.6	36.7	44.5	52
wkspt1990_S	wkspt1990	masked	16.3	0.3	1.2	2.3	5.4	12.7	23.4	37.8	45.5	51.8
wkspt1991_C	wkspt1991	complete	15.9	0.3	1.1	2.1	5	12.2	22.6	37.2	45.8	52
wkspt1991_S	wkspt1991	masked	16.4	0.3	1.2	2.3	5.2	12.6	23.4	38.4	46.7	52
wkspt1992_C	wkspt1992	complete	17	0.3	1.1	2.2	5.7	13.5	24.6	38.4	46	53
wkspt1992_S	wkspt1992	masked	17.1	0.3	1.3	2.4	5.8	13.7	25.2	38.7	46.1	52.3
wkspt1993_C	wkspt1993	complete	16.5	0.3	1.1	2.2	5.3	12.8	23.7	38	46.7	52
wkspt1993_S	wkspt1993	masked	16.6	0.2	1.2	2.2	5.3	12.8	24	39	47.6	52
wkspt1994_C	wkspt1994	complete	15.8	0.3	1.2	2.1	5.2	12.4	21.5	36.7	45.9	52
wkspt1994_S	wkspt1994	masked	16.3	0.3	1.3	2.4	5.5	12.8	22.5	38	46.6	52
wkspt1995_C	wkspt1995	complete	16.6	0.6	2.2	3.8	8.2	13.9	23.5	32.5	40.1	49.4
wkspt1995_S	wkspt1995	masked	17.9	0.6	2.1	3.7	8.2	15.5	24.8	36.8	42.9	50.2
wkspt1996_C	wkspt1996	complete	22.4	0.7	2.8	4.9	11.3	19.2	33.3	44.8	48.9	52
wkspt1996_S	wkspt1996	masked	22.1	0.7	2.7	4.7	10.9	19.1	33.2	44.3	48.4	51.5
wkspt1997_C	wkspt1997	complete	19.9	0.4	1.7	3.3	8.1	16.6	29.9	43.4	49.6	53
wkspt1997_S	wkspt1997	masked	20.1	0.4	1.7	3.3	8.1	16.8	30.3	43.4	49.1	52.3
wkspt1998_C	wkspt1998	complete	20	0.6	2.2	4	8.4	16.5	29.9	43.2	49.5	52
wkspt1998_S	wkspt1998	masked	20.1	0.6	2.3	4	8.5	16.8	29.8	42.6	48.7	52
wkspt1999_C	wkspt1999	complete	20.1	0.5	2.2	3.9	8.6	16.9	30	42.5	48.7	52
wkspt1999_S	wkspt1999	masked	20.9	0.6	2.5	4.4	9.4	17.8	31.2	43.1	48.4	52
wkswp1990_C	wkswp1990	complete	41.2	3.3	10.3	17.7	35.6	47.7	51.2	51.9	52	52
wkswp1990_S	wkswp1990	masked	41.3	2.4	8.1	14.8	36.4	48.8	51.3	52	52	52
wkswp1991_C	wkswp1991	complete	41.7	3.7	10.6	17.5	36	48.9	51.6	52	52	52
wkswp1991_S	wkswp1991	masked	41.9	3.1	10.1	17.1	36.3	49.4	51.6	52	52	52
wkswp1992_C	wkswp1992	complete	42	3.9	11.7	19.8	37.2	47.7	51.5	53	53	53

Revised Sythetic Beta File Descriptive Measures

variable	original_name	Type	mean	p01	p05	p10	p25	median	p75	p90	p95	p99
wkswp1992_S	wkswp1992	masked	42.2	3.2	10.9	19	37.8	48.5	51.5	52.3	52.3	52.3
wkswp1993_C	wkswp1993	complete	42.6	3.9	11.3	18.8	38.5	49.4	51.7	52	52	52
wkswp1993_S	wkswp1993	masked	42.7	3	10.5	18.2	38.7	50.1	51.7	52	52	52
wkswp1994_C	wkswp1994	complete	42.1	4.6	13	19.6	36.6	48.6	51.6	52	52	52
wkswp1994_S	wkswp1994	masked	42.9	4.7	13.6	20.8	38.3	49.6	51.5	52	52	52
wkswp1995_C	wkswp1995	complete	41.3	6.8	15.2	20.4	35.6	46.6	50.5	51.6	51.9	52
wkswp1995_S	wkswp1995	masked	42	5.8	14.4	20.8	36.8	47.9	50.8	51.7	51.9	52
wkswp1996_C	wkswp1996	complete	43.4	4.2	13.8	23	40	49.4	51.5	52	52	52
wkswp1996_S	wkswp1996	masked	43.9	4.5	14.8	24.7	41.7	49.6	51.5	51.9	52	52
wkswp1997_C	wkswp1997	complete	43.5	3.6	11.8	20.4	39.4	50.9	52	52	53	53
wkswp1997_S	wkswp1997	masked	43.1	2.8	10.1	19.2	39.3	50.5	51.8	52	52.3	52.3
wkswp1998_C	wkswp1998	complete	44.2	4.4	15.2	25.4	41.2	50.2	51.9	52	52	52
wkswp1998_S	wkswp1998	masked	44.3	4.4	15.3	25.6	41.9	50.1	51.7	52	52	52
wkswp1999_C	wkswp1999	complete	44	5.9	16.2	26.6	40.6	49.6	51.8	52	52	52
wkswp1999_S	wkswp1999	masked	44.6	5	16.2	27.3	42.4	50.1	51.7	52	52	52

Appendix D

Tables for Chapter 4

Comparison of Earnings and Employment Models Estimated in Gold Standard, Completed, and Synthetic Beta Files

Doug Wissoker, Urban Institute

Table 4-1: Regression model of Log(DER Earnings relative to national average+.25), 1990-1999
Comparison of Gold Standard to Completed File
Group: Non-immigrant male non-beneficiaries, 25+

	Coefficient		Confidence Interval				Gold Estimate in	CIs	Standard Error		Degrees of Freedom
	Gold	Completed	Gold		Completed		Completed CI	Overlap	Gold	Completed	Completed
Age spline 25 - 40	**0.0149**	**0.0156**	0.0139	0.0159	0.0144	0.0168	**in range**	**overlap**	0.0006	0.0007	19.1
Age spline 41 - 55	**-0.0005**	**-0.0001**	-0.0017	0.0008	-0.0013	0.0011	**in range**	**overlap**	0.0007	0.0007	464.5
Age spline 56+	**-0.0424**	**-0.0404**	-0.0460	-0.0387	-0.0447	-0.0361	**in range**	**overlap**	0.0022	0.0025	23.9
Age62 - 64, indicator	**0.2568**	**0.2422**	0.2346	0.2789	0.2169	0.2675	**in range**	**overlap**	0.0135	0.0149	29.7
Age65-69, indicator	**0.4644**	**0.4333**	0.3814	0.5473	0.3456	0.5210	**in range**	**overlap**	0.0504	0.0525	61.7
Age 70 or older, indicator	**-0.0449**	**-0.0365**	-0.3178	0.2280	-0.3219	0.2488	**in range**	**overlap**	0.1658	0.1667	23.5
Cohort1931_40	**-0.0635**	**-0.0575**	-0.0949	-0.0321	-0.0892	-0.0259	**in range**	**overlap**	0.0191	0.0191	130.2
Cohort1941_50	**-0.0879**	**-0.0868**	-0.1228	-0.0529	-0.1253	-0.0482	**in range**	**overlap**	0.0212	0.0229	41.5
Cohort1951_60	**-0.1324**	**-0.1317**	-0.1697	-0.0951	-0.1733	-0.0900	**in range**	**overlap**	0.0227	0.0247	37.7
Cohort1961_70	**-0.1682**	**-0.1624**	-0.2070	-0.1294	-0.2046	-0.1201	**in range**	**overlap**	0.0235	0.0251	45.3
Cohort1971_80	**-0.1701**	**-0.1515**	-0.2116	-0.1287	-0.1937	-0.1092	**in range**	**overlap**	0.0252	0.0254	79.1
High School	**0.1714**	**0.1735**	0.1603	0.1825	0.1623	0.1846	**in range**	**overlap**	0.0068	0.0067	51.1
Some college	**0.2907**	**0.2929**	0.2792	0.3023	0.2812	0.3046	**in range**	**overlap**	0.0070	0.0070	52.9
College Degree	**0.5733**	**0.5748**	0.5592	0.5874	0.5601	0.5894	**in range**	**overlap**	0.0086	0.0087	48.9
Graduate Degree	**0.7013**	**0.7046**	0.6861	0.7165	0.6890	0.7202	**in range**	**overlap**	0.0092	0.0094	67.8
Black	**-0.1604**	**-0.1684**	-0.1716	-0.1491	-0.1782	-0.1585	**in range**	**overlap**	0.0068	0.0060	11629.4
Hispanic	**-0.0679**	**-0.0586**	-0.0825	-0.0533	-0.0721	-0.0450	**in range**	**overlap**	0.0089	0.0082	259.0
Worked t-1 and t-2	**0.5915**	**0.5825**	0.5746	0.6083	0.5556	0.6094	**in range**	**overlap**	0.0102	0.0140	6.4
No work t-1, work t-2	**-0.0232**	**-0.0264**	-0.0428	-0.0036	-0.0465	-0.0064	**in range**	**overlap**	0.0119	0.0116	19.7
Work t-1, no work t-2	**0.2585**	**0.2802**	0.2434	0.2737	0.2586	0.3019	**out**	**overlap**	0.0092	0.0115	7.2
Number of years worked out of last 10	**0.0583**	**0.0392**	0.0549	0.0618	0.0322	0.0462	**out**	**no overlap**	0.0021	0.0034	4.7
Worked all of last 10 years	**0.2334**	**0.2606**	0.2228	0.2440	0.2503	0.2709	**out**	**no overlap**	0.0064	0.0061	41.4
Constant	**-1.6615**	**-1.5379**	-1.7212	-1.6017	-1.5992	-1.4766	**out**	**no overlap**	0.0363	0.0365	44.5
Autocorrelation	**0.3661**	**0.3513**	0.3611	0.3712	0.3465	0.3562	**out**	**no overlap**	0.0031	0.0029	253.9
N	**452451**	**572572**									
Std dev of random effect	**0.505**	**0.513**									
Std dev of transitory effect	**0.297**	**0.305**									
Share variance from random effect	**0.743**	**0.739**									
Mean dependent variable	**0.367**	**0.311**									

Table 4-2: Regression model of Log(DER Earnings relative to national average+.25), 1990-1999
Comparison of Gold Standard to Completed File
Group: Non-immigrant female non-beneficiaries, 25+

	Coefficient		Confidence Interval				Gold Estimate in	CIs	Standard Error		Degrees of Freedom
	Gold	Completed	Gold		Completed		Completed CI	Overlap	Gold	Completed	Completed
Age spline 25 - 40	**0.0059**	**0.0057**	0.0049	0.0068	0.0048	0.0065	**in range**	**overlap**	0.0006	0.0005	2030935.1
Age spline 41 - 55	**-0.0044**	**-0.0041**	-0.0055	-0.0034	-0.0051	-0.0031	**in range**	**overlap**	0.0006	0.0006	1010.8
Age spline 56+	**-0.0251**	**-0.0233**	-0.0282	-0.0220	-0.0263	-0.0203	**in range**	**overlap**	0.0019	0.0018	295.0
Age62 - 64, indicator	**0.1610**	**0.1586**	0.1413	0.1806	0.1371	0.1800	**in range**	**overlap**	0.0119	0.0128	46.7
Age65-69, indicator	**0.1548**	**0.1346**	0.0967	0.2129	0.0791	0.1902	**in range**	**overlap**	0.0353	0.0337	389.9
Age 70 or older, indicator	**-0.0861**	**-0.0610**	-0.1748	0.0026	-0.1492	0.0272	**in range**	**overlap**	0.0539	0.0535	265.8
Cohort1931_40	**-0.0542**	**-0.0400**	-0.0834	-0.0250	-0.0672	-0.0128	**in range**	**overlap**	0.0177	0.0165	3286.2
Cohort1941_50	**-0.0557**	**-0.0415**	-0.0877	-0.0237	-0.0711	-0.0119	**in range**	**overlap**	0.0194	0.0180	7597.4
Cohort1951_60	**-0.1081**	**-0.0917**	-0.1421	-0.0742	-0.1231	-0.0602	**in range**	**overlap**	0.0206	0.0191	5144.9
Cohort1961_70	**-0.1590**	**-0.1372**	-0.1944	-0.1237	-0.1701	-0.1043	**in range**	**overlap**	0.0215	0.0200	2938.5
Cohort1971_80	**-0.1224**	**-0.1019**	-0.1601	-0.0847	-0.1363	-0.0676	**in range**	**overlap**	0.0229	0.0209	9381.0
High School	**0.1346**	**0.1356**	0.1250	0.1443	0.1266	0.1447	**in range**	**overlap**	0.0059	0.0055	268.5
Some college	**0.2592**	**0.2573**	0.2490	0.2693	0.2474	0.2672	**in range**	**overlap**	0.0062	0.0060	114.5
College Degree	**0.4560**	**0.4648**	0.4431	0.4688	0.4515	0.4782	**in range**	**overlap**	0.0078	0.0080	50.0
Graduate Degree	**0.6147**	**0.6123**	0.6009	0.6285	0.5981	0.6265	**in range**	**overlap**	0.0084	0.0085	70.0
Black	**-0.0131**	**-0.0176**	-0.0216	-0.0047	-0.0252	-0.0099	**in range**	**overlap**	0.0052	0.0046	1889.0
Hispanic	**0.0212**	**0.0151**	0.0087	0.0337	0.0030	0.0272	**in range**	**overlap**	0.0076	0.0073	90.2
Worked t-1 and t-2	**0.4876**	**0.4824**	0.4792	0.4959	0.4724	0.4923	**in range**	**overlap**	0.0051	0.0057	18.1
No work t-1, work t-2	**-0.0498**	**-0.0543**	-0.0603	-0.0393	-0.0642	-0.0444	**in range**	**overlap**	0.0064	0.0060	114.1
Work t-1, no work t-2	**0.2352**	**0.2392**	0.2283	0.2420	0.2321	0.2464	**in range**	**overlap**	0.0042	0.0043	36.3
Number of years worked out of last 10	**0.0453**	**0.0397**	0.0437	0.0469	0.0375	0.0419	**out**	**no overlap**	0.0010	0.0012	11.8
Worked all of last 10 years	**0.2513**	**0.2538**	0.2439	0.2587	0.2471	0.2604	**in range**	**overlap**	0.0045	0.0040	3908.1
Constant	**-1.4441**	**-1.4035**	-1.4948	-1.3933	-1.4502	-1.3568	**in range**	**overlap**	0.0308	0.0284	1542.0
Autocorrelation	**0.4009**	**0.3853**	0.3963	0.4054	0.3810	0.3897	**out**	**no overlap**	0.0028	0.0027	268.1
N	**437416**	**544971**									
Std dev of random effect	**0.439**	**0.444**									
Std dev of transitory effect	**0.273**	**0.279**									
Share variance from random effect	**0.721**	**0.716**									
Mean dependent variable	**-0.082**	**-0.106**									

Table 4-3: Regression model of Log (DER Earnings relative to national average+.25), 1990-1999
Completed versus Synthetic Files
Group: Non-immigrant male non-beneficiaries, 25+

	Coefficient		Confidence Interval				Completed		Standard Error				Degrees of Freedom	
							Estimate in	CIs			Positive	Small		
	Completed	Synthetic	Completed		Synthetic		Synthetic CI	Overlap	Completed	Synthetic	Variance	DOF	Completed	Synthetic
Age spline 25 - 40	**0.0117**	**0.0102**	0.0105	0.0128	0.0092	0.0112	**out**	**overlap**	0.0007	0.0004	1	1	23	3
Age spline 41 - 55	**0.0000**	**0.0001**	-0.0012	0.0012	-0.0031	0.0033	**in range**	**overlap**	0.0007	0.0014	1	0	401	3
Age spline 56+	**-0.0404**	**-0.0383**	-0.0448	-0.0361	-0.0426	-0.0340	**in range**	**overlap**	0.0025	0.0023	1	0	22	8
Age62 - 64, indicator	**0.2434**	**0.0558**	0.2180	0.2689	0.0344	0.0772	**out**	**no overlap**	0.0150	0.0122	1	0	28	15
Age65-69, indicator	**0.4362**	**0.3416**	0.3482	0.5242	0.2661	0.4172	**out**	**overlap**	0.0526	0.0442	1	0	55	24
Age 70 or older, indicator	**-0.0208**	**0.0564**	-0.3007	0.2591	-0.1446	0.2574	**in range**	**overlap**	0.1634	0.1148	1	0	23	15
Cohort1931_40	**-0.0547**	**-0.0317**	-0.0865	-0.0230	-0.0612	-0.0022	**in range**	**overlap**	0.0191	0.0169	1	0	111	15
Cohort1941_50	**-0.0796**	**-0.0410**	-0.1184	-0.0408	-0.0864	0.0043	**in range**	**overlap**	0.0230	0.0233	1	0	38	6
Cohort1951_60	**-0.1186**	**-0.0896**	-0.1609	-0.0764	-0.1392	-0.0400	**in range**	**overlap**	0.0250	0.0211	1	1	33	3
Cohort1961_70	**-0.1308**	**-0.1055**	-0.1734	-0.0883	-0.1452	-0.0657	**in range**	**overlap**	0.0253	0.0169	1	1	41	3
Cohort1971_80	**-0.0744**	**-0.0604**	-0.1175	-0.0314	-0.1194	-0.0013	**in range**	**overlap**	0.0258	0.0302	0	0	63	6
High School	**0.1709**	**0.1419**	0.1593	0.1824	0.1043	0.1795	**in range**	**overlap**	0.0068	0.0160	1	1	36	3
Some college	**0.2940**	**0.2544**	0.2825	0.3054	0.1944	0.3144	**in range**	**overlap**	0.0069	0.0255	1	1	58	3
College Degree	**0.5812**	**0.4857**	0.5668	0.5956	0.3847	0.5867	**in range**	**overlap**	0.0086	0.0429	1	1	53	3
Graduate Degree	**0.7095**	**0.5593**	0.6938	0.7252	0.4483	0.6703	**out**	**no overlap**	0.0094	0.0472	1	0	57	3
Black	**-0.1525**	**-0.1394**	-0.1624	-0.1426	-0.1582	-0.1207	**in range**	**overlap**	0.0060	0.0086	1	0	1948	4
Hispanic	**-0.0591**	**-0.0748**	-0.0725	-0.0456	-0.0918	-0.0578	**in range**	**overlap**	0.0082	0.0096	1	0	242	14
Married	**0.1871**	**0.1619**	0.1785	0.1956	0.1489	0.1750	**out**	**no overlap**	0.0052	0.0070	1	0	226	8
Divorced	**0.1435**	**0.1302**	0.1307	0.1564	0.1094	0.1510	**in range**	**overlap**	0.0078	0.0108	1	0	2390	6
Widowed	**0.1469**	**0.1344**	0.0788	0.2149	0.0613	0.2075	**in range**	**overlap**	0.0408	0.0408	1	0	63	11
Got married	**0.0711**	**0.0744**	0.0578	0.0845	0.0455	0.1033	**in range**	**overlap**	0.0080	0.0142	1	0	101	5
Got divorced/widowed	**-0.0247**	**0.0100**	-0.0377	-0.0116	-0.0116	0.0316	**out**	**no overlap**	0.0079	0.0119	1	0	1573	10
Worked t-1 and t-2	**0.5828**	**0.5993**	0.5560	0.6096	0.5477	0.6508	**in range**	**overlap**	0.0139	0.0219	0	1	6	3
No work t-1, work t-2	**-0.0230**	**-0.0053**	-0.0428	-0.0031	-0.0309	0.0202	**in range**	**overlap**	0.0115	0.0131	0	0	21	6
Work t-1, no work t-2	**0.2827**	**0.2415**	0.2615	0.3039	0.2229	0.2601	**out**	**no overlap**	0.0113	0.0086	0	0	7	4
Number of years worked out of last 10	**0.0367**	**0.0367**	0.0300	0.0433	0.0320	0.0415	**in range**	**overlap**	0.0033	0.0024	1	0	5	5
Worked all of last 10 years	**0.2515**	**0.2620**	0.2413	0.2616	0.2526	0.2714	**out**	**overlap**	0.0060	0.0052	1	0	46	11
Constant	**-1.5297**	**-1.4682**	-1.5919	-1.4676	-1.5509	-1.3854	**in range**	**overlap**	0.0369	0.0405	0	0	38	5
Autocorrelation	**0.3495**	**0.3573**	0.3447	0.3544	0.2790	0.4356	**in range**	**overlap**	0.0029	0.0333	1	1	220	3
Average Statistics														
N	**572572**	**573504**												
Std dev of random effect	**0.509**	**0.522**												
Std dev of transitory effect	**0.304**	**0.331**												
Share variance from random effect	**0.737**	**0.713**												
Mean dependent variable	**0.311**	**0.287**												

Table 4-4: Regression model of Log (DER Earnings relative to national average+.25), 1990-1999
Completed versus Synthetic Files
Group: Non-immigrant female non-beneficiaries, 25+

	Coefficient		Confidence Interval				Completed Estimate in Synthetic CI	CIs Overlap	Standard Error		Positive Variance	Small DOF	Degrees of Freedom	
	Completed	Synthetic	Completed		Synthetic				Completed	Synthetic			Completed	Synthetic
Age spline 25 - 40	**0.0069**	**0.0066**	0.0060	0.0077	0.0048	0.0084	**in range**	**overlap**	0.0005	0.0008	1	1	9458498	3
Age spline 41 - 55	**-0.0040**	**-0.0036**	-0.0050	-0.0030	-0.0054	-0.0018	**in range**	**overlap**	0.0006	0.0008	1	1	950	3
Age spline 56+	**-0.0233**	**-0.0224**	-0.0263	-0.0203	-0.0253	-0.0195	**in range**	**overlap**	0.0018	0.0017	1	0	287	21
Age62 - 64, indicator	**0.1586**	**0.0277**	0.1371	0.1801	0.0136	0.0417	**out**	**no overlap**	0.0128	0.0084	1	0	45	45
Age65-69, indicator	**0.1347**	**0.0840**	0.0791	0.1904	0.0377	0.1304	**out**	**overlap**	0.0338	0.0277	1	0	362	49
Age 70 or older, indicator	**-0.0637**	**-0.0167**	-0.1520	0.0245	-0.0961	0.0628	**in range**	**overlap**	0.0535	0.0477	1	0	258	77
Cohort1931_40	**-0.0400**	**-0.0326**	-0.0672	-0.0127	-0.0598	-0.0054	**in range**	**overlap**	0.0165	0.0160	1	0	3057	31
Cohort1941_50	**-0.0417**	**-0.0185**	-0.0714	-0.0120	-0.0461	0.0091	**in range**	**overlap**	0.0181	0.0146	1	0	5705	7
Cohort1951_60	**-0.0921**	**-0.0605**	-0.1236	-0.0606	-0.0955	-0.0254	**in range**	**overlap**	0.0192	0.0149	1	1	4840	3
Cohort1961_70	**-0.1405**	**-0.1016**	-0.1735	-0.1075	-0.1474	-0.0558	**in range**	**overlap**	0.0200	0.0195	1	1	2527	3
Cohort1971_80	**-0.1190**	**-0.0917**	-0.1534	-0.0847	-0.1443	-0.0391	**in range**	**overlap**	0.0209	0.0224	1	1	25835	3
High School	**0.1363**	**0.0959**	0.1272	0.1454	0.0669	0.1250	**out**	**no overlap**	0.0055	0.0127	1	0	239	3
Some college	**0.2571**	**0.2012**	0.2471	0.2670	0.1514	0.2510	**out**	**overlap**	0.0060	0.0212	1	1	110	3
College Degree	**0.4611**	**0.3892**	0.4477	0.4745	0.3048	0.4736	**in range**	**overlap**	0.0080	0.0358	1	1	50	3
Graduate Degree	**0.6069**	**0.4683**	0.5927	0.6210	0.3710	0.5656	**out**	**no overlap**	0.0085	0.0420	1	0	75	3
Black	**-0.0261**	**-0.0326**	-0.0338	-0.0183	-0.0473	-0.0180	**in range**	**overlap**	0.0047	0.0062	1	1	2007	3
Hispanic	**0.0139**	**-0.0121**	0.0018	0.0261	-0.0232	-0.0010	**out**	**no overlap**	0.0073	0.0067	1	0	89	152
Married	**-0.0524**	**-0.0527**	-0.0600	-0.0448	-0.0638	-0.0416	**in range**	**overlap**	0.0046	0.0062	1	0	5403	13
Divorced	**-0.0567**	**-0.0524**	-0.0693	-0.0441	-0.0658	-0.0390	**in range**	**overlap**	0.0076	0.0074	1	0	67	9
Widowed	**-0.0815**	**-0.0643**	-0.1194	-0.0437	-0.1012	-0.0274	**in range**	**overlap**	0.0228	0.0219	1	0	107	39
Got married	**0.0348**	**0.0043**	0.0225	0.0471	-0.0144	0.0230	**out**	**overlap**	0.0075	0.0108	1	0	806	19
Got divorced/widowed	**0.0057**	**-0.0058**	-0.0065	0.0180	-0.0274	0.0158	**in range**	**overlap**	0.0074	0.0118	1	0	311	9
Worked t-1 and t-2	**0.4810**	**0.3961**	0.4709	0.4911	0.3638	0.4284	**out**	**no overlap**	0.0058	0.0137	0	1	17	3
No work t-1, work t-2	**-0.0546**	**-0.0507**	-0.0646	-0.0446	-0.0675	-0.0340	**in range**	**overlap**	0.0060	0.0071	1	1	98	3
Work t-1, no work t-2	**0.2391**	**0.1556**	0.2318	0.2464	0.1514	0.1598	**out**	**no overlap**	0.0043	0.0018	1	1	32	3
Number of years worked out of last 10	**0.0398**	**0.0418**	0.0376	0.0420	0.0400	0.0435	**out**	**overlap**	0.0012	0.0009	1	0	11	5
Worked all of last 10 years	**0.2520**	**0.2378**	0.2454	0.2586	0.2221	0.2534	**in range**	**overlap**	0.0040	0.0069	1	0	4421	3
Constant	**-1.3998**	**-1.2859**	-1.4468	-1.3527	-1.3805	-1.1914	**out**	**overlap**	0.0286	0.0402	1	1	1116	3
Autocorrelation	**0.3850**	**0.3711**	0.3806	0.3893	0.2888	0.4534	**in range**	**overlap**	0.0027	0.0350	1	1	256	3
<u>Average Statistics</u>														
N	**544971**	**546372**												
Std dev of random effect	**0.443**	**0.447**												
Std dev of transitory effect	**0.279**	**0.311**												
Share variance from random effect	**0.716**	**0.674**												
Mean dependent variable	**-0.106**	**-0.113**												

Table 4-5: Probit model of employment at t for those employed at t-1, 1990-1999
Comparison of Gold Standard to Completed File
Group: Non-immigrant male non-beneficiaries, 25+

	Coefficient		Confidence Interval				Gold Estimate in	CIs	Standard Error		Degrees of Freedom
	Gold	Completed	Gold		Completed		Completed CI	Overlap	Gold	Completed	Completed
Age spline 25 - 40	**-0.0089**	**-0.0087**	-0.0119	-0.0058	-0.0110	-0.0063	**in range**	**overlap**	0.0018	0.0014	68718.8
Age spline 41 - 55	**-0.0060**	**-0.0028**	-0.0091	-0.0028	-0.0064	0.0008	**in range**	**overlap**	0.0019	0.0021	21.6
Age spline 56+	**-0.0554**	**-0.0521**	-0.0634	-0.0474	-0.0611	-0.0430	**in range**	**overlap**	0.0048	0.0053	28.3
Age62 - 64, indicator	**0.3910**	**0.3732**	0.3224	0.4597	0.2830	0.4633	**in range**	**overlap**	0.0417	0.0512	14.1
Age65-69, indicator	**0.4082**	**0.3656**	0.2709	0.5455	0.2342	0.4969	**in range**	**overlap**	0.0834	0.0795	201.3
Age 70 or older, indicator	**0.5520**	**0.4748**	0.2908	0.8132	0.2475	0.7021	**in range**	**overlap**	0.1586	0.1381	1059.3
Cohort1931_40	**0.0190**	**0.0102**	-0.0416	0.0796	-0.0466	0.0671	**in range**	**overlap**	0.0368	0.0346	862.5
Cohort1941_50	**0.1062**	**0.0890**	0.0354	0.1769	0.0239	0.1541	**in range**	**overlap**	0.0430	0.0396	1904.6
Cohort1951_60	**0.0677**	**0.0480**	-0.0105	0.1460	-0.0239	0.1200	**in range**	**overlap**	0.0475	0.0437	919.6
Cohort1961_70	**0.0671**	**0.0310**	-0.0166	0.1509	-0.0455	0.1075	**in range**	**overlap**	0.0509	0.0464	593.9
Cohort1971_80	**0.2329**	**0.1574**	0.1319	0.3339	0.0729	0.2420	**in range**	**overlap**	0.0613	0.0514	1440.5
High School	**0.1181**	**0.1080**	0.0966	0.1396	0.0885	0.1274	**in range**	**overlap**	0.0130	0.0117	88.1
Some college	**0.1767**	**0.1818**	0.1541	0.1992	0.1626	0.2009	**in range**	**overlap**	0.0137	0.0116	541.3
College Degree	**0.2828**	**0.2799**	0.2558	0.3098	0.2506	0.3092	**in range**	**overlap**	0.0164	0.0170	21.2
Graduate Degree	**0.3050**	**0.3024**	0.2771	0.3328	0.2786	0.3263	**in range**	**overlap**	0.0169	0.0145	1060.6
Black	**-0.0917**	**-0.0981**	-0.1144	-0.0691	-0.1238	-0.0723	**in range**	**overlap**	0.0138	0.0147	14.3
Hispanic	**-0.0245**	**-0.0280**	-0.0555	0.0066	-0.0571	0.0011	**in range**	**overlap**	0.0189	0.0173	40.3
Worked, t-2	**0.4933**	**0.4134**	0.4628	0.5239	0.3871	0.4397	**out**	**no overlap**	0.0186	0.0158	90.1
Worked, t-3	**0.2279**	**0.2146**	0.1971	0.2588	0.1885	0.2407	**in range**	**overlap**	0.0188	0.0157	87.3
Number of years worked out of last 10	**0.0766**	**0.0741**	0.0699	0.0832	0.0669	0.0814	**in range**	**overlap**	0.0040	0.0040	10.4
Worked all of last 10 years	**0.3942**	**0.4217**	0.3727	0.4156	0.3993	0.4442	**out**	**overlap**	0.0130	0.0130	19.9
Constant	**0.4262**	**0.4550**	0.2787	0.5738	0.3327	0.5772	**in range**	**overlap**	0.0896	0.0742	727.9
N	**454672**	**575414**									
Mean dependent variable	**0.973**	**0.965**									

Table 4-6: Probit model of employment at t for those employed at t-1, 1990-1999
Comparison of Gold Standard to Completed File
Group: Non-immigrant female non-beneficiaries, 25+

	Coefficient		Confidence Interval				Gold Estimate in	CIs	Standard Error		Degrees of Freedom
	Gold	Completed	Gold		Completed		Completed CI	Overlap	Gold	Completed	Completed
Age spline 25 - 40	**0.0205**	**0.0157**	0.0182	0.0229	0.0136	0.0179	**out**	**no overlap**	0.0014	0.0013	337.8
Age spline 41 - 55	**-0.0098**	**-0.0082**	-0.0125	-0.0071	-0.0106	-0.0057	**in range**	**overlap**	0.0016	0.0015	580.9
Age spline 56+	**-0.0456**	**-0.0430**	-0.0527	-0.0384	-0.0503	-0.0357	**in range**	**overlap**	0.0043	0.0044	76.8
Age62 - 64, indicator	**0.2661**	**0.2359**	0.2023	0.3300	0.1691	0.3027	**in range**	**overlap**	0.0388	0.0399	54.6
Age65-69, indicator	**0.1823**	**0.1623**	0.0695	0.2951	0.0476	0.2770	**in range**	**overlap**	0.0685	0.0689	79.0
Age 70 or older, indicator	**0.2762**	**0.2856**	0.1099	0.4425	0.1193	0.4519	**in range**	**overlap**	0.1010	0.1005	154.2
Cohort1931_40	**-0.0122**	**-0.0117**	-0.0718	0.0474	-0.0655	0.0420	**in range**	**overlap**	0.0362	0.0327	7487.8
Cohort1941_50	**0.0582**	**0.0537**	-0.0091	0.1255	-0.0067	0.1141	**in range**	**overlap**	0.0409	0.0367	16029.7
Cohort1951_60	**0.0219**	**0.0116**	-0.0512	0.0949	-0.0542	0.0775	**in range**	**overlap**	0.0444	0.0400	3447.1
Cohort1961_70	**0.0551**	**0.0265**	-0.0216	0.1319	-0.0455	0.0985	**in range**	**overlap**	0.0466	0.0436	275.2
Cohort1971_80	**0.2930**	**0.2077**	0.2058	0.3803	0.1253	0.2902	**out**	**overlap**	0.0530	0.0497	109.6
High School	**0.1844**	**0.1553**	0.1655	0.2034	0.1376	0.1731	**out**	**overlap**	0.0115	0.0107	128.2
Some college	**0.2099**	**0.1903**	0.1903	0.2295	0.1708	0.2097	**out**	**overlap**	0.0119	0.0116	59.6
College Degree	**0.2287**	**0.2127**	0.2053	0.2520	0.1904	0.2350	**in range**	**overlap**	0.0142	0.0134	91.6
Graduate Degree	**0.3379**	**0.3105**	0.3119	0.3639	0.2816	0.3394	**in range**	**overlap**	0.0158	0.0169	24.1
Black	**0.0695**	**0.0391**	0.0515	0.0875	0.0236	0.0546	**out**	**overlap**	0.0109	0.0094	467.4
Hispanic	**0.0251**	**0.0108**	0.0001	0.0501	-0.0103	0.0320	**in range**	**overlap**	0.0152	0.0129	5220.9
Worked, t-2	**0.3692**	**0.3317**	0.3471	0.3912	0.3127	0.3507	**out**	**overlap**	0.0134	0.0115	708.6
Worked, t-3	**0.2680**	**0.2516**	0.2465	0.2894	0.2309	0.2723	**in range**	**overlap**	0.0130	0.0124	51.9
Number of years worked out of last 10	**0.0476**	**0.0536**	0.0437	0.0516	0.0498	0.0575	**out**	**overlap**	0.0024	0.0023	46.2
Worked all of last 10 years	**0.2898**	**0.3043**	0.2735	0.3062	0.2881	0.3205	**in range**	**overlap**	0.0099	0.0097	57.8
Constant	**-0.3948**	**-0.2408**	-0.5133	-0.2763	-0.3516	-0.1299	**out**	**overlap**	0.0720	0.0670	160.0
N	**437970**	**545857**									
Mean dependent variable	**0.951**	**0.944**									

Table 4-7: Probit model of employment at t for those employed at t-1, 1990-1999
Completed versus Synthetic Files
Group: Non-immigrant male non-beneficiaries, 25+

	Coefficient		Confidence Interval				Completed Estimate in Synthetic CI	CIs Overlap	Standard Error		Positive Variance	Small DOF	Degrees of Freedom	
	Completed	Synthetic	Completed		Synthetic				Completed	Synthetic			Completed	Synthetic
Age spline 25 - 40	-0.0119	-0.0099	-0.0144	-0.0095	-0.0126	-0.0072	in range	overlap	0.0015	0.0016	1	0	168431	41
Age spline 41 - 55	-0.0030	-0.0054	-0.0065	0.0005	-0.0083	-0.0025	in range	overlap	0.0021	0.0017	1	0	24	57
Age spline 56+	-0.0520	-0.0590	-0.0611	-0.0428	-0.0683	-0.0497	in range	overlap	0.0054	0.0052	1	0	26	13
Age62 - 64, indicator	0.3728	-0.0731	0.2826	0.4630	-0.1208	-0.0253	out	no overlap	0.0512	0.0285	1	0	14	46
Age65-69, indicator	0.3668	0.2240	0.2351	0.4985	0.1094	0.3386	out	overlap	0.0797	0.0683	1	0	181	45
Age 70 or older, indicator	0.4856	0.6570	0.2581	0.7130	0.3997	0.9144	in range	overlap	0.1382	0.1478	1	0	1631	17
Cohort1931_40	0.0127	0.0232	-0.0442	0.0696	-0.0207	0.0670	in range	overlap	0.0345	0.0264	1	0	854	93
Cohort1941_50	0.0947	0.0887	0.0297	0.1597	0.0238	0.1537	in range	overlap	0.0395	0.0380	1	0	2371	24
Cohort1951_60	0.0605	0.0613	-0.0111	0.1321	-0.0080	0.1305	in range	overlap	0.0435	0.0409	1	0	1332	33
Cohort1961_70	0.0607	0.0804	-0.0153	0.1367	0.0053	0.1555	in range	overlap	0.0462	0.0446	1	0	925	40
Cohort1971_80	0.2206	0.2563	0.1358	0.3054	0.1774	0.3352	in range	overlap	0.0515	0.0476	1	0	1712	125
High School	0.1060	0.0956	0.0867	0.1254	0.0687	0.1225	in range	overlap	0.0116	0.0144	1	0	98	8
Some college	0.1827	0.1725	0.1635	0.2019	0.1230	0.2221	in range	overlap	0.0116	0.0233	1	0	533	4
College Degree	0.2850	0.2336	0.2558	0.3142	0.1660	0.3013	in range	overlap	0.0170	0.0313	1	0	22	4
Graduate Degree	0.3065	0.2184	0.2826	0.3304	0.1641	0.2726	out	no overlap	0.0145	0.0261	1	0	1259	4
Black	-0.0828	-0.0903	-0.1090	-0.0566	-0.1208	-0.0598	in range	overlap	0.0149	0.0162	1	0	14	7
Hispanic	-0.0284	-0.0109	-0.0580	0.0011	-0.0359	0.0142	in range	overlap	0.0175	0.0149	1	0	37	46
Married	0.1560	0.1581	0.1407	0.1713	0.1428	0.1734	in range	overlap	0.0093	0.0092	1	0	6874	105
Divorced	0.1147	0.1311	0.0845	0.1448	0.1102	0.1521	in range	overlap	0.0177	0.0126	1	0	28	68
Widowed	0.1283	0.1350	-0.0054	0.2621	0.0232	0.2468	in range	overlap	0.0799	0.0669	1	0	53	57
Got married	0.1257	0.0698	0.0532	0.1982	-0.0187	0.1584	in range	overlap	0.0430	0.0490	1	0	36	10
Got divorced/widowed	-0.0250	0.0321	-0.0842	0.0342	-0.0220	0.0862	out	overlap	0.0360	0.0319	1	0	23275	29
Worked, t-2	0.4129	0.4172	0.3869	0.4389	0.3913	0.4431	in range	overlap	0.0157	0.0153	1	0	105	33
Worked, t-3	0.2141	0.2739	0.1878	0.2405	0.2511	0.2967	out	no overlap	0.0158	0.0138	1	0	77	108
Number of years worked out of last 10	0.0724	0.0656	0.0655	0.0794	0.0604	0.0708	out	overlap	0.0039	0.0030	1	0	12	26
Worked all of last 10 years	0.4151	0.3825	0.3927	0.4376	0.3616	0.4034	out	overlap	0.0130	0.0121	1	0	20	19
Constant	0.4612	0.4415	0.3369	0.5854	0.3289	0.5541	in range	overlap	0.0754	0.0682	1	0	451	228
N	575414	577673												
Mean dependent variable	0.965	0.964												

Table 4-8: Probit model of employment at t for those employed at t-1, 1990-1999
Completed versus Synthetic Files
Group: Non-immigrant female non-beneficiaries, 25+

	Coefficient		Confidence Interval				Completed Estimate in Synthetic CI	CIs Overlap	Standard Error		Positive Variance	Small DOF	Degrees of Freedom	
	Completed	Synthetic	Completed		Synthetic				Completed	Synthetic			Completed	Synthetic
Age spline 25 - 40	**0.0185**	**0.0136**	0.0164	0.0206	0.0109	0.0164	**out**	**overlap**	0.0013	0.0015	1	0	382	13
Age spline 41 - 55	**-0.0078**	**-0.0061**	-0.0103	-0.0054	-0.0090	-0.0031	**in range**	**overlap**	0.0015	0.0017	1	0	685	17
Age spline 56+	**-0.0430**	**-0.0452**	-0.0503	-0.0357	-0.0553	-0.0351	**in range**	**overlap**	0.0044	0.0055	1	0	71	9
Age62 - 64, indicator	**0.2364**	**-0.0637**	0.1692	0.3035	-0.1381	0.0106	**out**	**no overlap**	0.0401	0.0403	1	0	52	8
Age65-69, indicator	**0.1627**	**0.0928**	0.0477	0.2777	-0.0027	0.1883	**in range**	**overlap**	0.0691	0.0570	1	0	76	49
Age 70 or older, indicator	**0.2819**	**0.3843**	0.1156	0.4482	0.1707	0.5978	**in range**	**overlap**	0.1005	0.1191	1	0	148	11
Cohort1931_40	**-0.0097**	**0.0181**	-0.0633	0.0439	-0.0291	0.0653	**in range**	**overlap**	0.0326	0.0284	1	0	13644	98
Cohort1941_50	**0.0579**	**0.1038**	-0.0025	0.1183	0.0462	0.1614	**in range**	**overlap**	0.0367	0.0346	1	0	18343	73
Cohort1951_60	**0.0170**	**0.0767**	-0.0488	0.0829	0.0153	0.1380	**in range**	**overlap**	0.0400	0.0369	1	0	3901	94
Cohort1961_70	**0.0262**	**0.0826**	-0.0462	0.0985	0.0123	0.1530	**in range**	**overlap**	0.0438	0.0417	1	0	242	38
Cohort1971_80	**0.1805**	**0.2884**	0.0973	0.2638	0.2121	0.3647	**out**	**overlap**	0.0501	0.0456	1	0	98	57
High School	**0.1557**	**0.1315**	0.1379	0.1735	0.0843	0.1787	**in range**	**overlap**	0.0108	0.0222	1	0	121	4
Some college	**0.1889**	**0.1991**	0.1693	0.2085	0.1583	0.2399	**in range**	**overlap**	0.0117	0.0197	1	0	55	4
College Degree	**0.2021**	**0.2415**	0.1795	0.2248	0.1860	0.2971	**in range**	**overlap**	0.0136	0.0261	1	0	76	4
Graduate Degree	**0.2967**	**0.2579**	0.2672	0.3263	0.1772	0.3387	**in range**	**overlap**	0.0172	0.0368	1	0	22	4
Black	**0.0171**	**0.0180**	0.0014	0.0329	0.0046	0.0315	**in range**	**overlap**	0.0096	0.0081	1	0	416	131
Hispanic	**0.0060**	**-0.0123**	-0.0152	0.0272	-0.0346	0.0100	**in range**	**overlap**	0.0129	0.0131	1	0	5663	26
Married	**-0.1038**	**-0.0676**	-0.1191	-0.0884	-0.0822	-0.0530	**out**	**no overlap**	0.0093	0.0088	1	0	519	77
Divorced	**-0.1636**	**-0.0431**	-0.1878	-0.1394	-0.0646	-0.0215	**out**	**no overlap**	0.0145	0.0129	1	0	63	55
Widowed	**-0.1975**	**-0.0827**	-0.2695	-0.1256	-0.1560	-0.0094	**out**	**overlap**	0.0435	0.0433	1	0	161	33
Got married	**0.1204**	**0.0163**	0.0571	0.1838	-0.0435	0.0761	**out**	**overlap**	0.0379	0.0356	1	0	59	44
Got divorced/widowed	**-0.0127**	**0.0072**	-0.0640	0.0386	-0.0490	0.0635	**in range**	**overlap**	0.0311	0.0341	1	0	453	335
Worked, t-2	**0.3299**	**0.3885**	0.3109	0.3488	0.3722	0.4048	**out**	**no overlap**	0.0115	0.0090	1	0	751	11
Worked, t-3	**0.2493**	**0.2692**	0.2285	0.2701	0.2535	0.2849	**out**	**overlap**	0.0124	0.0094	1	0	50	61
Number of years worked out of last 10	**0.0544**	**0.0512**	0.0505	0.0583	0.0476	0.0548	**in range**	**overlap**	0.0023	0.0022	1	0	42	39
Worked all of last 10 years	**0.3011**	**0.2873**	0.2849	0.3174	0.2733	0.3013	**in range**	**overlap**	0.0097	0.0085	1	0	59	641
Constant	**-0.2454**	**-0.1973**	-0.3574	-0.1335	-0.3233	-0.0712	**in range**	**overlap**	0.0676	0.0724	1	0	136	17
N	**545857**	**547974**												
Mean dependent variable	**0.944**	**0.942**												

Table 4-9: Probit model of employment at t for those not employed at t-1, 1990-1999
Comparison of Gold Standard to Completed File
Group: Non-immigrant male non-beneficiaries, 25+

	Coefficient		Confidence Interval				Gold Estimate in	CIs	Standard Error		Degrees of Freedom
	Gold	Completed	Gold		Completed		Completed CI	Overlap	Gold	Completed	Completed
Age spline 25 - 40	**-0.0246**	**-0.0174**	-0.0302	-0.0191	-0.0213	-0.0135	**out**	**overlap**	0.0034	0.0023	112.9
Age spline 41 - 55	**-0.0223**	**-0.0136**	-0.0280	-0.0166	-0.0181	-0.0091	**out**	**overlap**	0.0035	0.0027	143.4
Age spline 56+	**-0.0693**	**-0.0536**	-0.0832	-0.0553	-0.0674	-0.0398	**out**	**overlap**	0.0085	0.0082	32.8
Age62 - 64, indicator	**0.1705**	**0.1329**	0.0594	0.2816	0.0207	0.2450	**in range**	**overlap**	0.0675	0.0666	40.4
Age65-69, indicator	**0.1073**	**0.0597**	-0.0730	0.2876	-0.1061	0.2255	**in range**	**overlap**	0.1095	0.1006	533.1
Age 70 or older, indicator	**0.2573**	**0.2695**	0.0002	0.5144	0.0276	0.5114	**in range**	**overlap**	0.1562	0.1458	110.1
Cohort1931_40	**0.0307**	**0.0436**	-0.0726	0.1341	-0.0554	0.1426	**in range**	**overlap**	0.0628	0.0596	95.7
Cohort1941_50	**0.0022**	**0.0362**	-0.1200	0.1243	-0.0694	0.1418	**in range**	**overlap**	0.0742	0.0641	622.9
Cohort1951_60	**-0.0572**	**-0.0008**	-0.1938	0.0795	-0.1192	0.1176	**in range**	**overlap**	0.0830	0.0717	216.8
Cohort1961_70	**-0.1121**	**-0.0570**	-0.2589	0.0348	-0.1820	0.0680	**in range**	**overlap**	0.0892	0.0756	172.5
Cohort1971_80	**-0.1342**	**-0.0452**	-0.3113	0.0429	-0.2246	0.1343	**in range**	**overlap**	0.1076	0.1028	16.1
High School	**0.0806**	**-0.0379**	0.0453	0.1159	-0.0845	0.0086	**out**	**no overlap**	0.0214	0.0247	7.2
Some college	**0.1102**	**0.0279**	0.0719	0.1485	-0.0307	0.0865	**out**	**overlap**	0.0233	0.0302	6.1
College Degree	**0.1473**	**-0.0210**	0.0981	0.1965	-0.0808	0.0389	**out**	**no overlap**	0.0299	0.0321	7.9
Graduate Degree	**0.1723**	**0.0200**	0.1224	0.2222	-0.0374	0.0775	**out**	**no overlap**	0.0303	0.0315	9.4
Black	**-0.1111**	**0.0002**	-0.1478	-0.0744	-0.0340	0.0343	**out**	**no overlap**	0.0223	0.0196	16.6
Hispanic	**-0.0189**	**0.0041**	-0.0720	0.0341	-0.0681	0.0763	**in range**	**overlap**	0.0322	0.0375	6.3
Worked, t-2	**0.3474**	**0.3471**	0.3159	0.3788	0.3075	0.3868	**in range**	**overlap**	0.0191	0.0219	10.4
Worked, t-3	**0.1115**	**0.0949**	0.0772	0.1458	0.0569	0.1330	**in range**	**overlap**	0.0208	0.0216	14.2
Number of years worked out of last 10	**0.0916**	**0.1465**	0.0853	0.0979	0.1384	0.1547	**out**	**no overlap**	0.0038	0.0042	6.3
Constant	**-0.2190**	**-0.9782**	-0.4739	0.0359	-1.1723	-0.7840	**out**	**no overlap**	0.1548	0.1169	98.3
N	**37963**	**113632**									
Mean dependent variable	**0.266**	**0.152**									

Table 4-10: Probit model of employment t for those not employed at t-1, 1990-1999
Comparison of Gold Standard to Completed File
Group: Non-immigrant female non-beneficiaries, 25+

	Coefficient		Confidence Interval				Gold Estimate in	CIs	Standard Error		Degrees of Freedom
	Gold	Completed	Gold		Completed		Completed CI	Overlap	Gold	Completed	Completed
Age spline 25 - 40	**-0.0044**	**-0.0028**	-0.0076	-0.0012	-0.0062	0.0006	**in range**	**overlap**	0.0023	0.0018	49.0
Age spline 41 - 55	**-0.0267**	**-0.0233**	-0.0303	-0.0231	-0.0266	-0.0199	**out**	**overlap**	0.0028	0.0020	57.3
Age spline 56+	**-0.0580**	**-0.0479**	-0.0677	-0.0483	-0.0574	-0.0384	**out**	**overlap**	0.0053	0.0056	36.3
Age62 - 64, indicator	**0.1580**	**0.0812**	0.0684	0.2477	0.0073	0.1550	**out**	**overlap**	0.0381	0.0448	1109.3
Age65-69, indicator	**0.0660**	**-0.0144**	-0.0747	0.2067	-0.1452	0.1164	**in range**	**overlap**	0.0675	0.0784	62.3
Age 70 or older, indicator	**0.1309**	**0.0473**	-0.0574	0.3192	-0.1273	0.2220	**in range**	**overlap**	0.0940	0.1048	76.5
Cohort1931_40	**-0.0243**	**-0.0025**	-0.1092	0.0605	-0.0801	0.0751	**in range**	**overlap**	0.0494	0.0465	70.7
Cohort1941_50	**-0.0594**	**0.0009**	-0.1539	0.0351	-0.0828	0.0847	**in range**	**overlap**	0.0581	0.0506	104.2
Cohort1951_60	**-0.0950**	**0.0017**	-0.1971	0.0070	-0.1034	0.1068	**in range**	**overlap**	0.0656	0.0615	25.1
Cohort1961_70	**-0.0759**	**-0.0071**	-0.1822	0.0304	-0.1250	0.1108	**in range**	**overlap**	0.0689	0.0672	18.4
Cohort1971_80	**0.1246**	**0.0984**	-0.0002	0.2493	-0.0071	0.2039	**in range**	**overlap**	0.0860	0.0705	45.5
High School	**0.0691**	**-0.0100**	0.0468	0.0914	-0.0318	0.0119	**out**	**no overlap**	0.0226	0.0124	55.6
Some college	**0.1027**	**0.0469**	0.0784	0.1269	0.0227	0.0710	**out**	**no overlap**	0.0239	0.0143	34.3
College Degree	**-0.0298**	**-0.1147**	-0.0621	0.0024	-0.1514	-0.0781	**out**	**no overlap**	0.0302	0.0206	15.8
Graduate Degree	**0.1097**	**-0.0219**	0.0717	0.1477	-0.0616	0.0178	**out**	**no overlap**	0.0349	0.0225	22.4
Black	**0.0408**	**0.0187**	0.0170	0.0646	0.0000	0.0374	**out**	**overlap**	0.0247	0.0115	438.4
Hispanic	**0.0337**	**0.0324**	0.0024	0.0651	0.0073	0.0576	**in range**	**overlap**	0.0329	0.0154	491.0
Worked, t-2	**0.3506**	**0.3303**	0.3290	0.3722	0.3084	0.3522	**in range**	**overlap**	0.0128	0.0135	42.1
Worked, t-3	**0.1596**	**0.1302**	0.1378	0.1815	0.1060	0.1544	**out**	**overlap**	0.0129	0.0144	27.4
Number of years worked out of last 10	**0.0353**	**0.0906**	0.0318	0.0388	0.0870	0.0941	**out**	**no overlap**	0.0028	0.0020	43.7
Constant	**-0.7956**	**-1.2348**	-0.9561	-0.6352	-1.4145	-1.0552	**out**	**no overlap**	0.0738	0.0641	42.1
N	**104318**	**213675**									
Mean dependent variable	**0.202**	**0.140**									

Table 4-11: Probit model of employment at t for those not employed at t-1, 1990-1999
Completed versus Synthetic Files
Group: Non-immigrant male non-beneficiaries, 25+

	Coefficient		Confidence Interval				Completed Estimate in	CIs	Standard Error		Positive	Small	Degrees of Freedom	
	Completed	Synthetic	Completed		Synthetic		Synthetic CI	Overlap	Completed	Synthetic	Variance	DOF	Completed	Synthetic
Age spline 25 - 40	**-0.0187**	**-0.0275**	-0.0226	-0.0149	-0.0309	-0.0241	**out**	**no overlap**	0.0023	0.0021	1	0	199	228
Age spline 41 - 55	**-0.0141**	**-0.0179**	-0.0185	-0.0097	-0.0242	-0.0115	**in range**	**overlap**	0.0027	0.0035	1	0	202	11
Age spline 56+	**-0.0537**	**-0.0466**	-0.0676	-0.0399	-0.0569	-0.0362	**in range**	**overlap**	0.0082	0.0063	1	0	33	134
Age62 - 64, indicator	**0.1328**	**-0.0502**	0.0206	0.2449	-0.1382	0.0378	**out**	**overlap**	0.0666	0.0518	1	0	40	29
Age65-69, indicator	**0.0604**	**0.0264**	-0.1053	0.2262	-0.1438	0.1966	**in range**	**overlap**	0.1006	0.0997	1	0	536	25
Age 70 or older, indicator	**0.2697**	**0.1984**	0.0271	0.5123	-0.0020	0.3988	**in range**	**overlap**	0.1462	0.1215	1	0	106	308
Cohort1931_40	**0.0402**	**-0.0144**	-0.0592	0.1395	-0.0982	0.0694	**in range**	**overlap**	0.0598	0.0482	1	0	88	17
Cohort1941_50	**0.0309**	**-0.0663**	-0.0746	0.1364	-0.1693	0.0366	**in range**	**overlap**	0.0641	0.0582	1	0	652	13
Cohort1951_60	**-0.0024**	**-0.1416**	-0.1208	0.1160	-0.2797	-0.0035	**out**	**overlap**	0.0717	0.0750	1	0	218	9
Cohort1961_70	**-0.0508**	**-0.2406**	-0.1768	0.0753	-0.3990	-0.0822	**out**	**overlap**	0.0761	0.0842	1	0	147	7
Cohort1971_80	**-0.0299**	**-0.1648**	-0.2050	0.1452	-0.3355	0.0058	**in range**	**overlap**	0.1010	0.0926	1	0	18	9
High School	**-0.0386**	**0.0019**	-0.0860	0.0089	-0.0207	0.0244	**out**	**overlap**	0.0251	0.0135	1	0	7	66
Some college	**0.0269**	**-0.0007**	-0.0330	0.0868	-0.0467	0.0452	**in range**	**overlap**	0.0307	0.0241	1	0	6	7
College Degree	**-0.0200**	**-0.0401**	-0.0805	0.0406	-0.0914	0.0112	**in range**	**overlap**	0.0324	0.0272	1	0	8	7
Graduate Degree	**0.0209**	**-0.0173**	-0.0369	0.0787	-0.0547	0.0200	**out**	**overlap**	0.0317	0.0205	1	0	9	9
Black	**0.0072**	**0.0223**	-0.0268	0.0412	0.0002	0.0444	**in range**	**overlap**	0.0195	0.0133	1	0	17	90
Hispanic	**0.0053**	**0.0152**	-0.0676	0.0783	-0.0209	0.0513	**in range**	**overlap**	0.0378	0.0217	1	0	6	71
Married	**0.0550**	**0.0564**	0.0140	0.0960	0.0028	0.1099	**in range**	**overlap**	0.0217	0.0258	1	0	7	4
Divorced	**0.0986**	**0.0608**	0.0523	0.1450	0.0013	0.1202	**in range**	**overlap**	0.0270	0.0302	1	0	22	6
Widowed	**0.0626**	**0.0680**	-0.1688	0.2939	-0.2013	0.3374	**in range**	**overlap**	0.1343	0.1395	1	0	21	6
Got married	**0.1426**	**0.0196**	0.0425	0.2427	-0.1063	0.1455	**in range**	**overlap**	0.0599	0.0675	1	0	60	8
Got divorced/widowed	**-0.0056**	**0.0133**	-0.1314	0.1201	-0.0757	0.1022	**in range**	**overlap**	0.0728	0.0534	1	0	20	74
Worked, t-2	**0.3469**	**0.2620**	0.3072	0.3865	0.2352	0.2888	**out**	**no overlap**	0.0220	0.0158	1	0	10	32
Worked, t-3	**0.0949**	**0.1706**	0.0568	0.1331	0.1366	0.2046	**out**	**no overlap**	0.0217	0.0196	1	0	14	17
Number of years worked out of last 10	**0.1461**	**0.1287**	0.1382	0.1540	0.1224	0.1351	**out**	**no overlap**	0.0041	0.0035	1	0	7	9
Constant	**-0.9682**	**-0.4671**	-1.1640	-0.7725	-0.6808	-0.2533	**out**	**no overlap**	0.1178	0.1215	1	0	90	14
N	**113632**	**118717**												
Mean dependent variable	**0.152**	**0.144**												

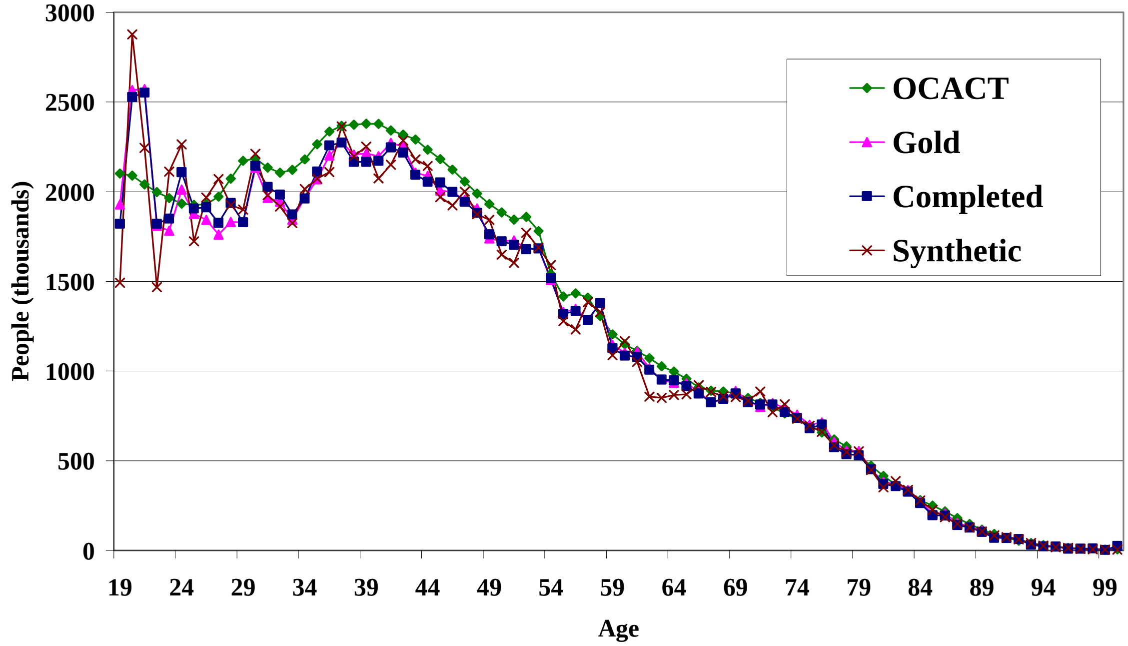
Appendix E

Figures for Chapter 5

Cross-Sectional and Longitudinal Comparisons of Completed and Synthetic Beta Files with Microsim and Gold

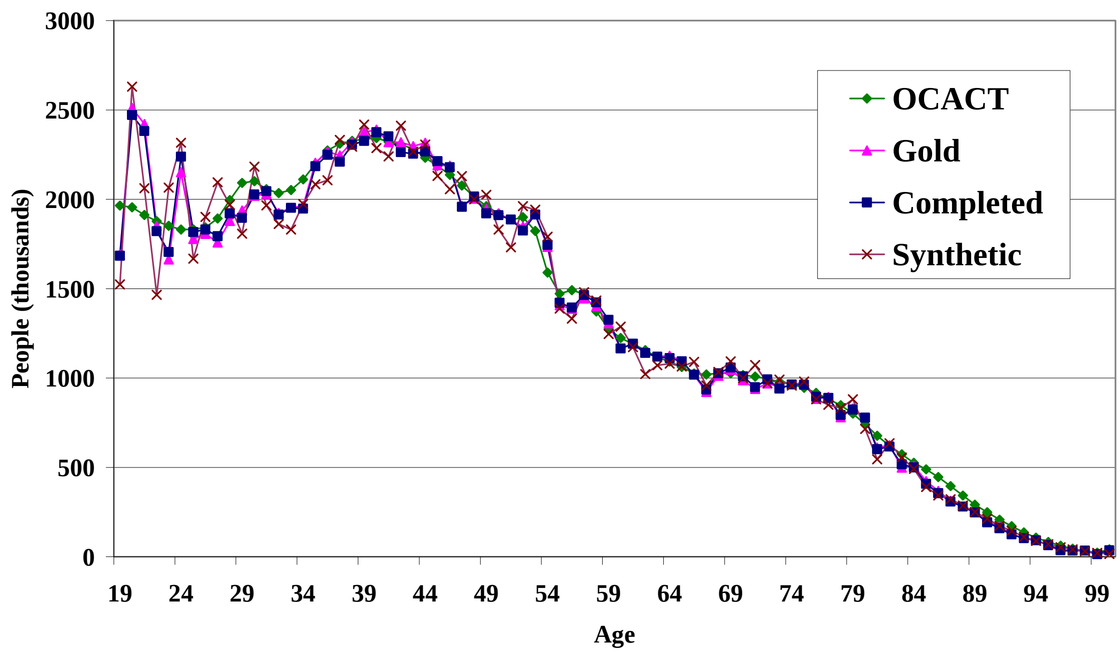
Karen Smith, Urban Institute

Figure 5-1. Male Population Totals 2000 by Age and Data Source



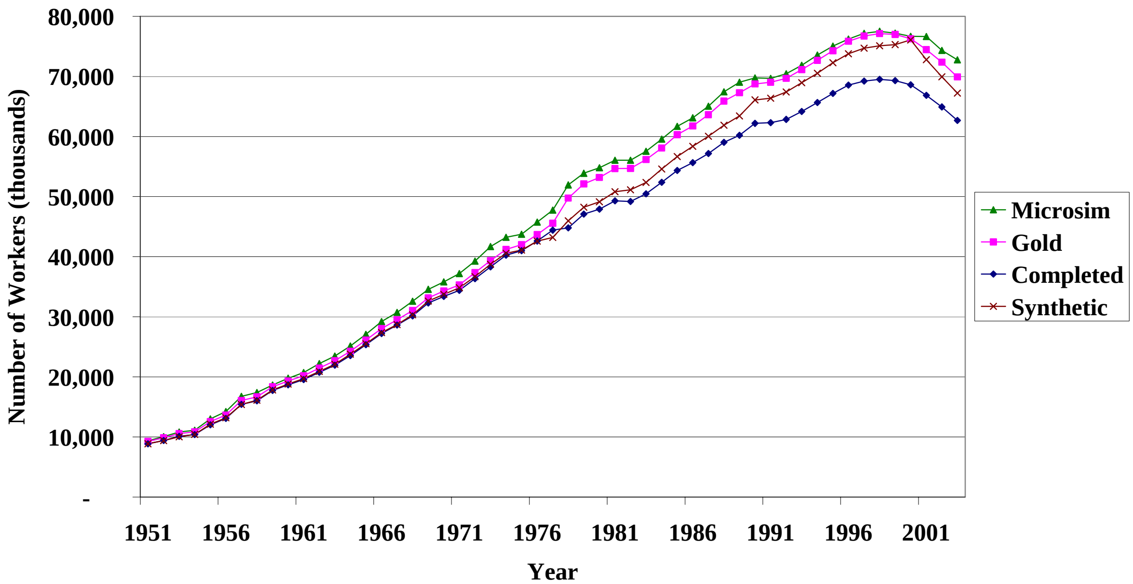
Source: OCACT 2008 and Urban Institute tabulations of the pooled Synthetic, pooled Completed, and Gold SIPP files.

Figure 5-2. Female Population Totals 2000 by Age and Data Source



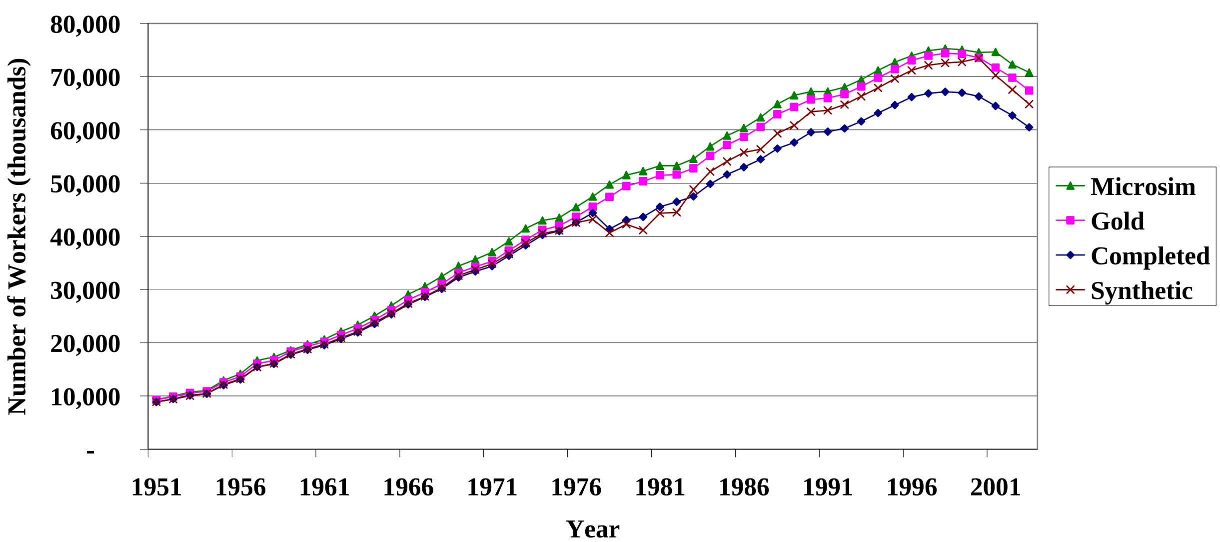
Source: OCACT 2008 and Urban Institute tabulations of the pooled Synthetic, pooled Completed, and Gold SIPP files.

Figure 5-3a. Number of Male Workers by Year and Data Source
Survivors to 2000



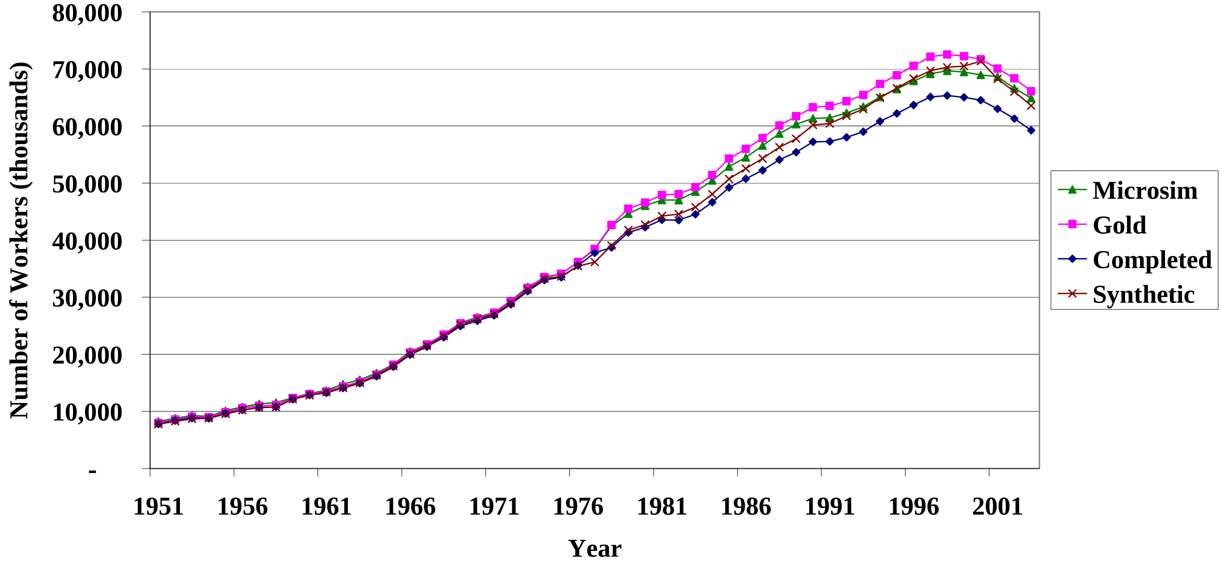
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings.

Figure 5-3b. Number of Male Workers in Covered Employment
by Year and Data Source
Survivors to 2000



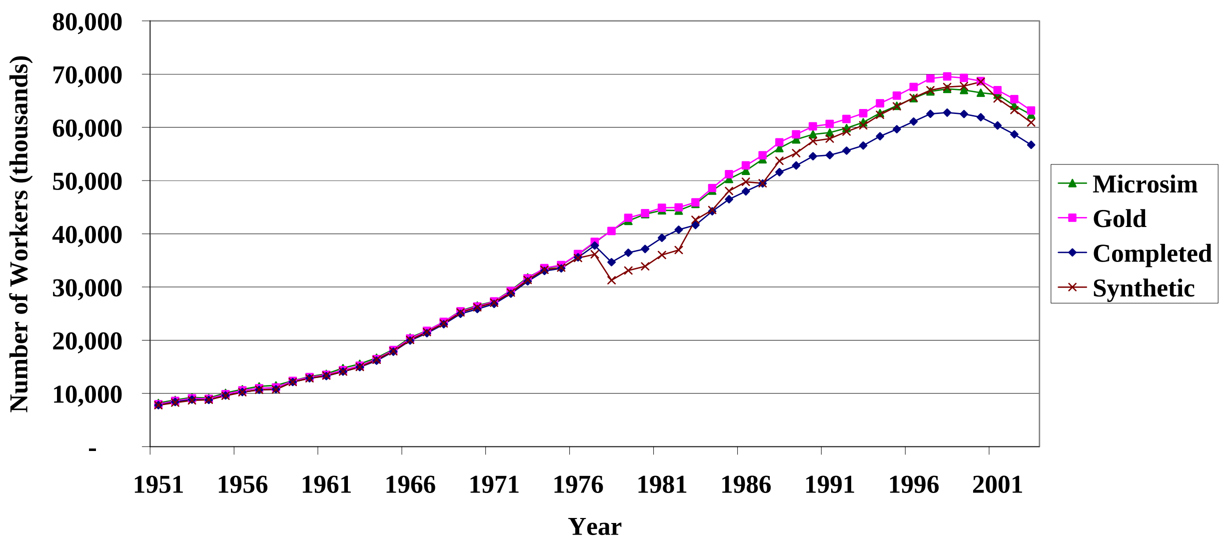
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual taxable earnings.

Figure 5-3c. Number of Female Workers by Year and Data Source
Survivors to 2000



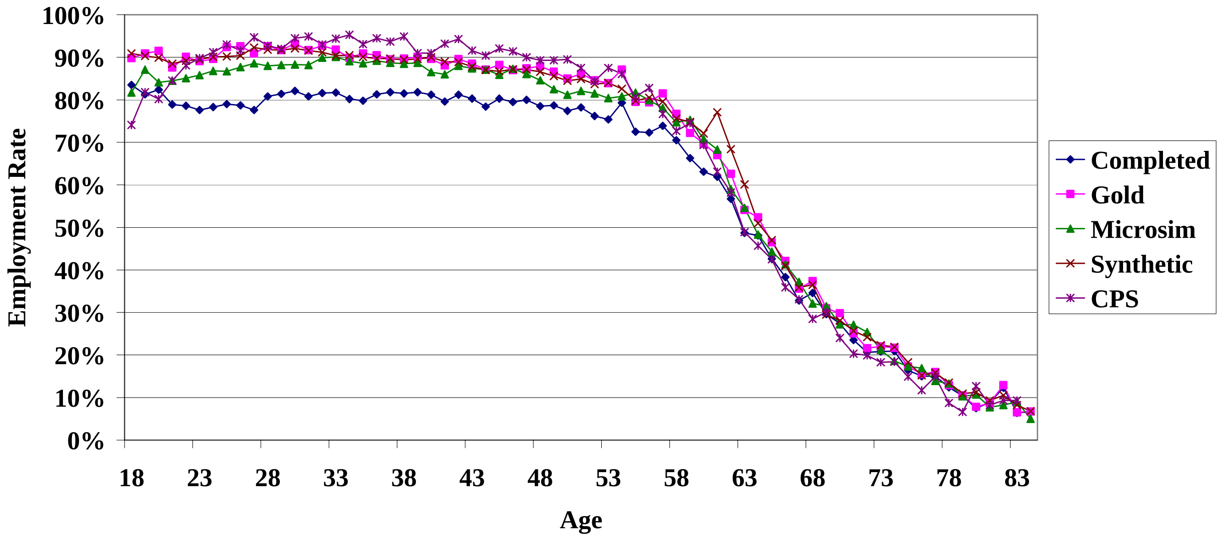
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings.

Figure 5-3d. Number of Female Workers in Covered Employment
by Year and Data Source
Survivors to 2000



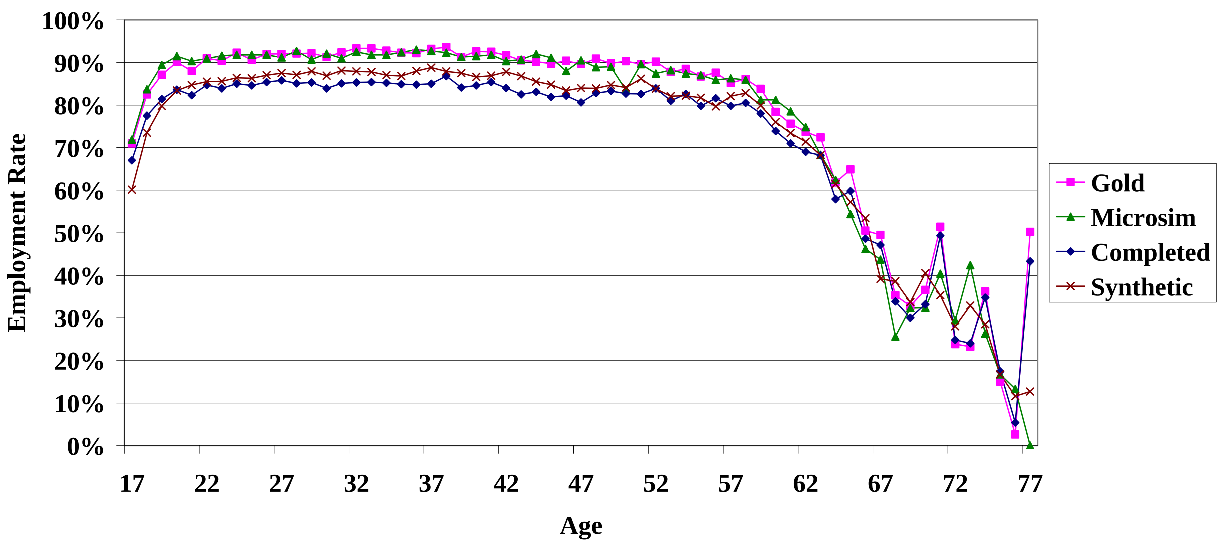
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual taxable earnings.

Figure 5-3e. Male Employment Rates 2000
by Age and Data Source
Survivors to 2000



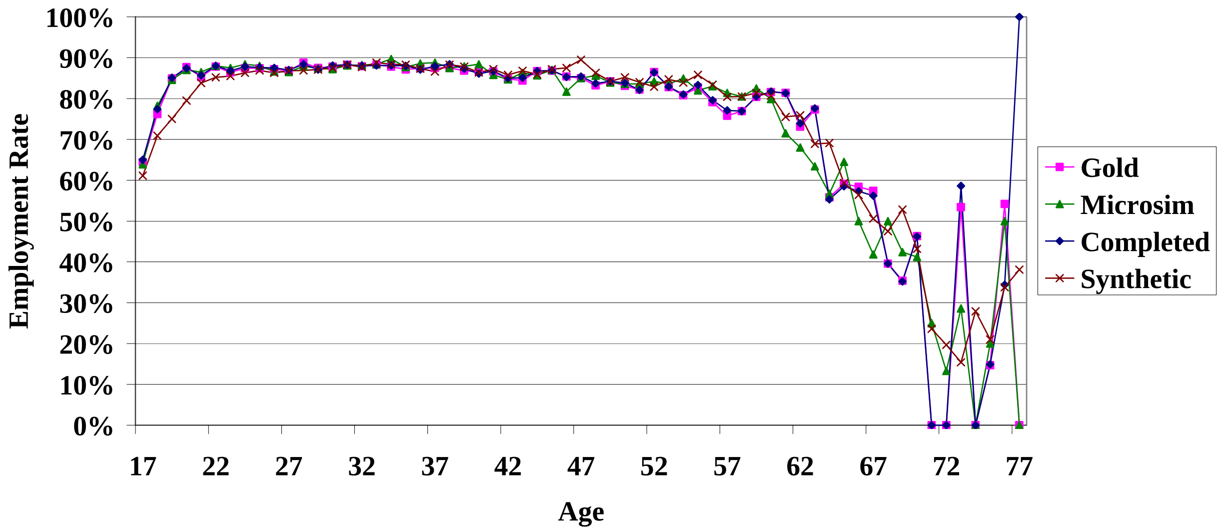
Source. Urban Institute tabulations of Microsim, 2000 March Current Population Survey, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings in 2000.

Figure 5-3f. Male Employment Rates 1979
by Age and Data Source
Survivors to 2000



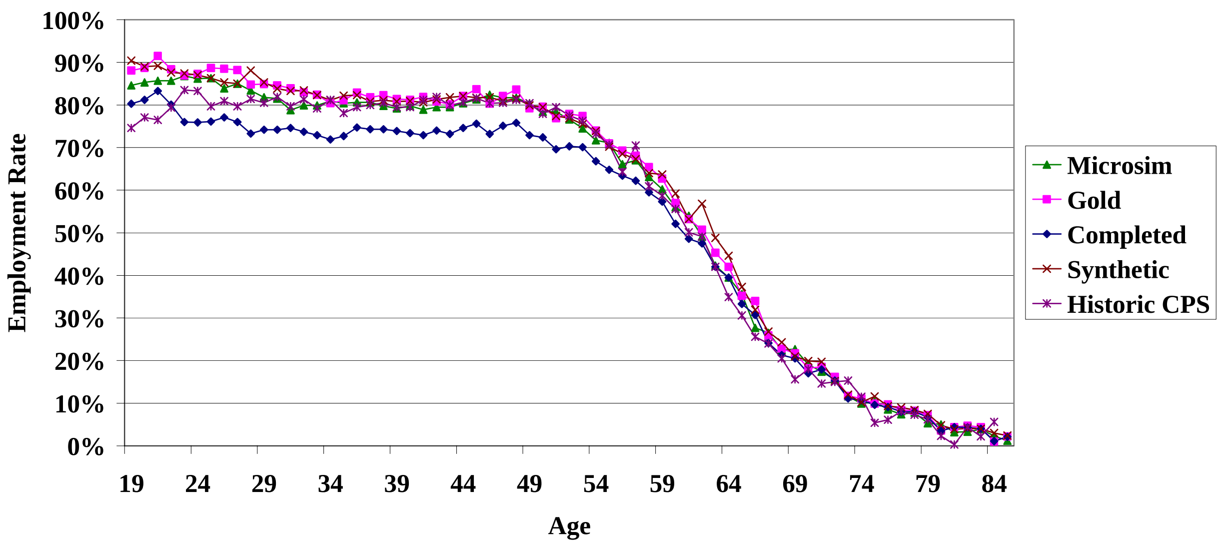
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings in 1979.

Figure 5-3g. Male Employment Rates 1975
by Age and Data Source
Survivors to 2000



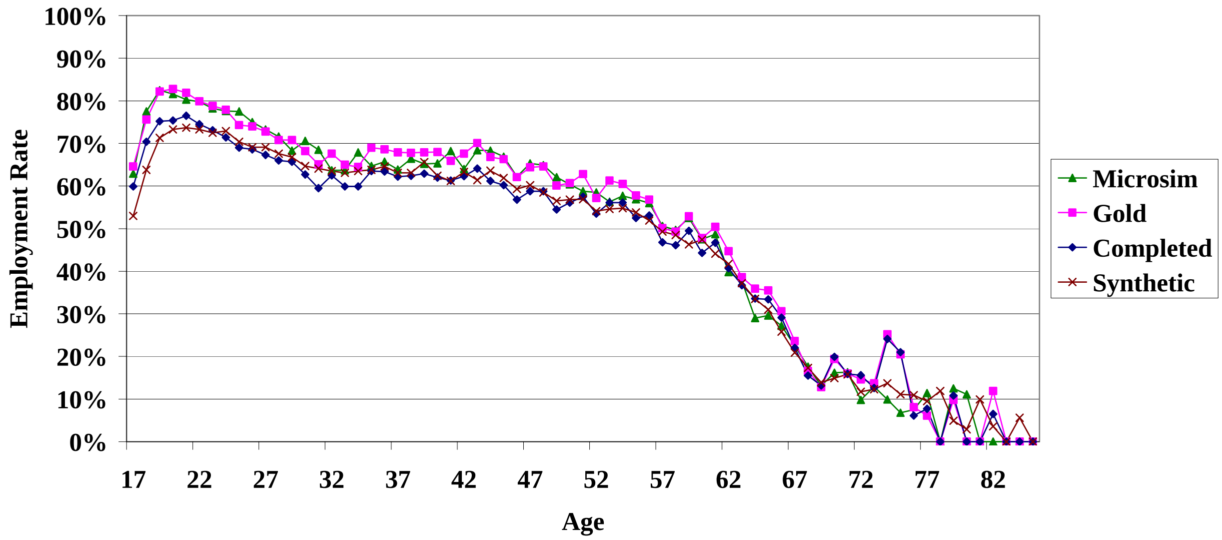
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings in 1975.

Figure 5-3h. Female Employment Rates 2000
by Age and Data Source
Survivors to 2000



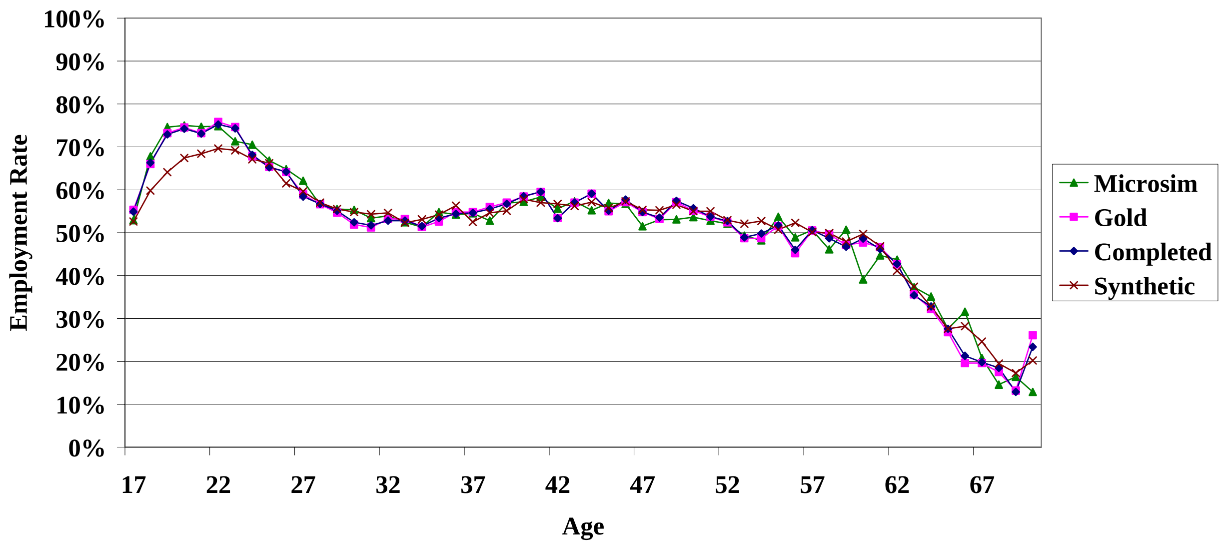
Source: Urban Institute tabulations of Microsim, 2000 March Current Population Survey, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings in 2000.

Figure 5-3i. Female Employment Rates 1979
by Age and Data Source
Survivors to 2000



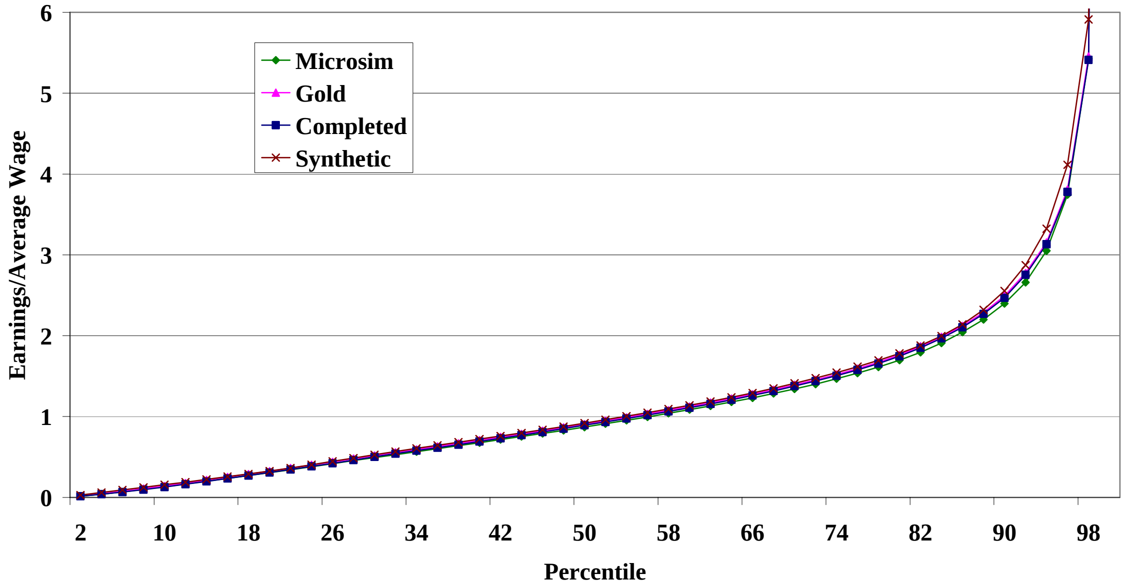
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings in 1979.

Figure 5-3j. Female Employment Rates 1975
by Age and Data Source
Survivors to 2000



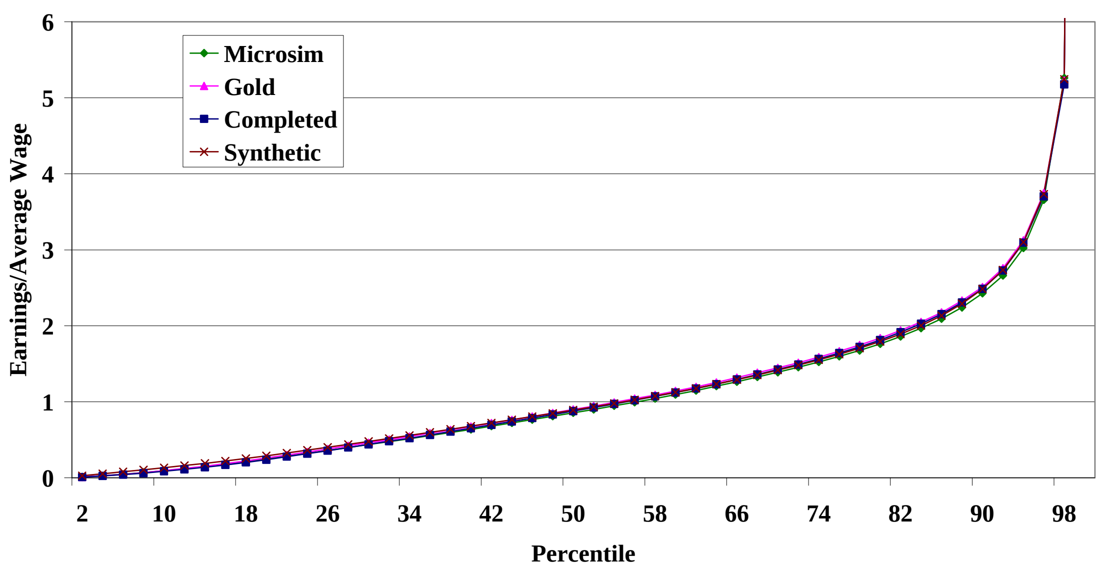
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Workers are defined as individuals with positive annual total earnings in 1975.

Figure 5-4a: Total Earnings / Average Wage of Male Workers in 2000
Survivors to 2000



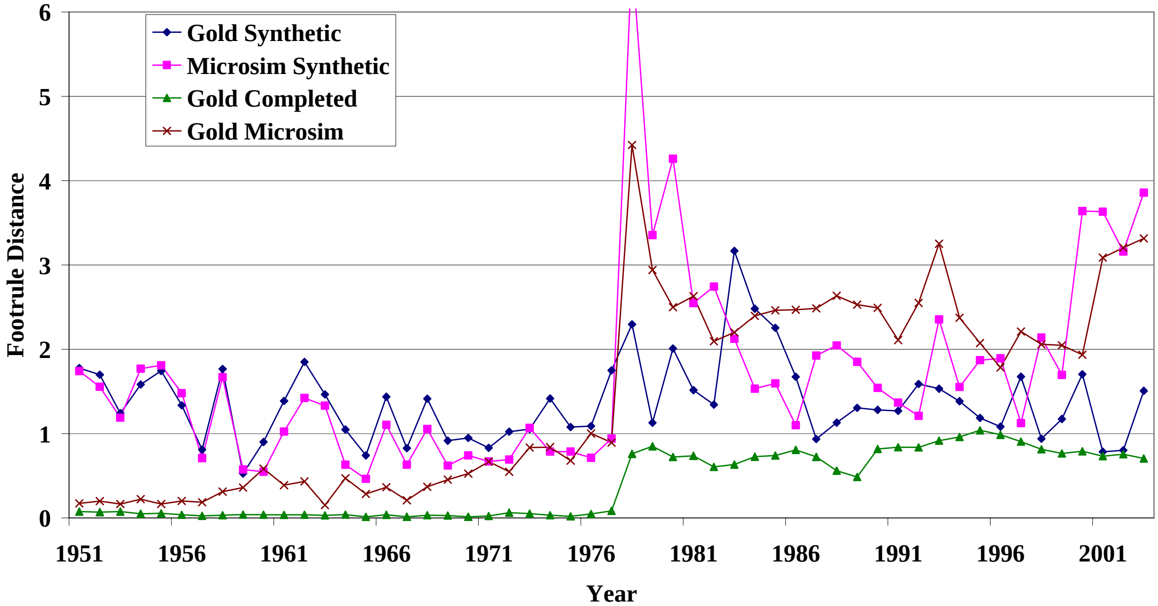
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive earnings in 2000.

Figure 5-4b: Total Earnings / Average Wage of Male Workers in 1995
Survivors to 2000



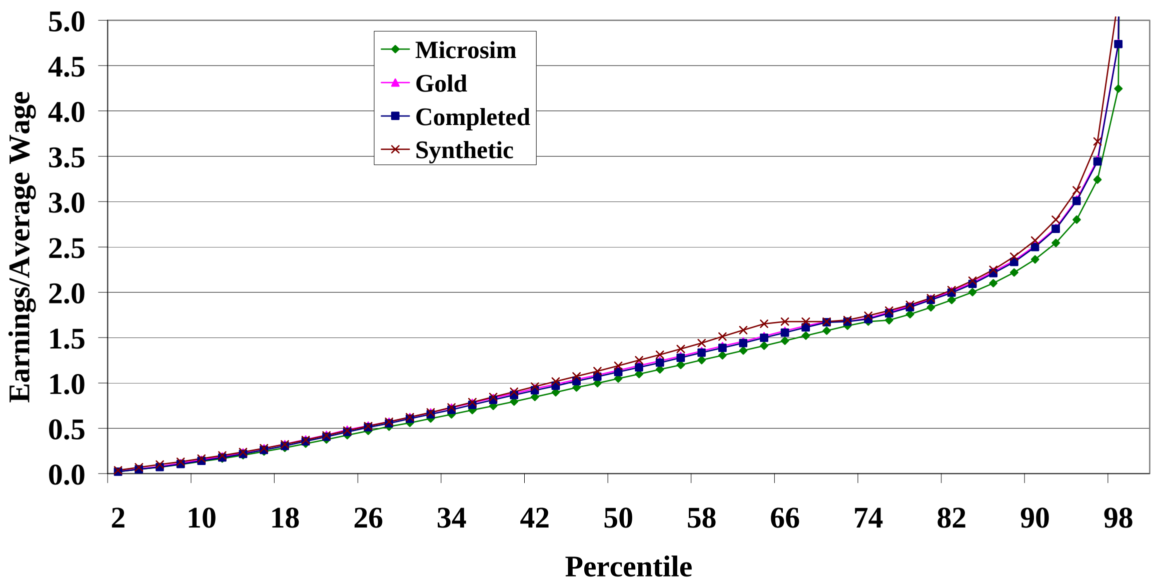
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive earnings in 1995.

Figure 5-5: Footrule Distance Total Earnings/Average Wage of Male Workers 1951 to 2003



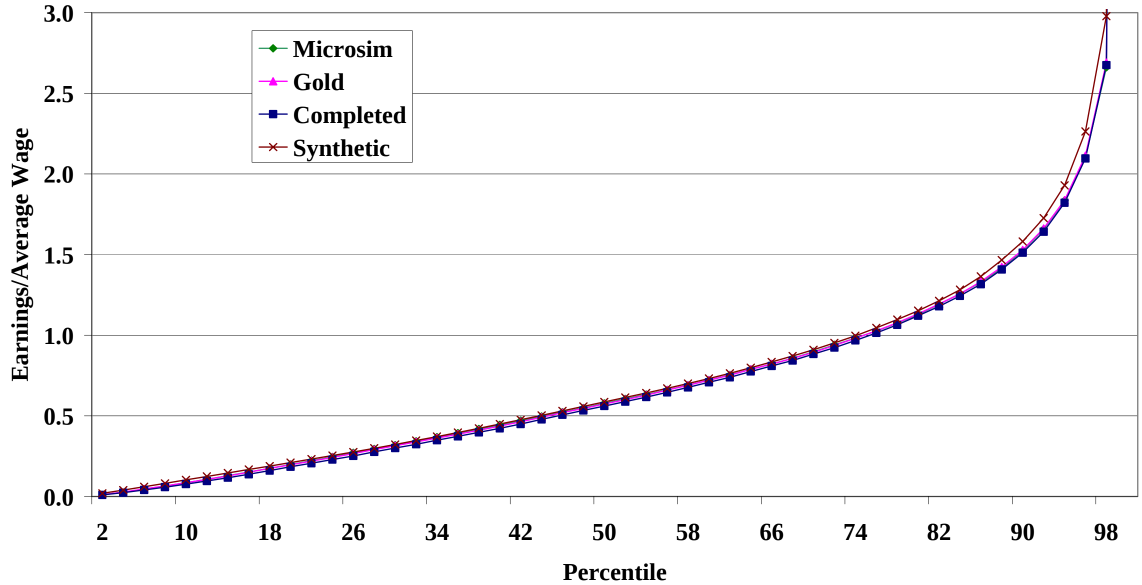
Source: Urban Institute footrule distance calculations of the distribution of total earnings of male workers in the Microsim, Gold, pooled Completed, and pooled Synthetic files. The distribution includes the sum of the absolute differences between the 1st and 98th percentiles. The data are limited to U.S. residents that survive to 2000.

Figure 5-6: Total Earnings / Average Wage of Male Workers in 1978 Survivors to 2000



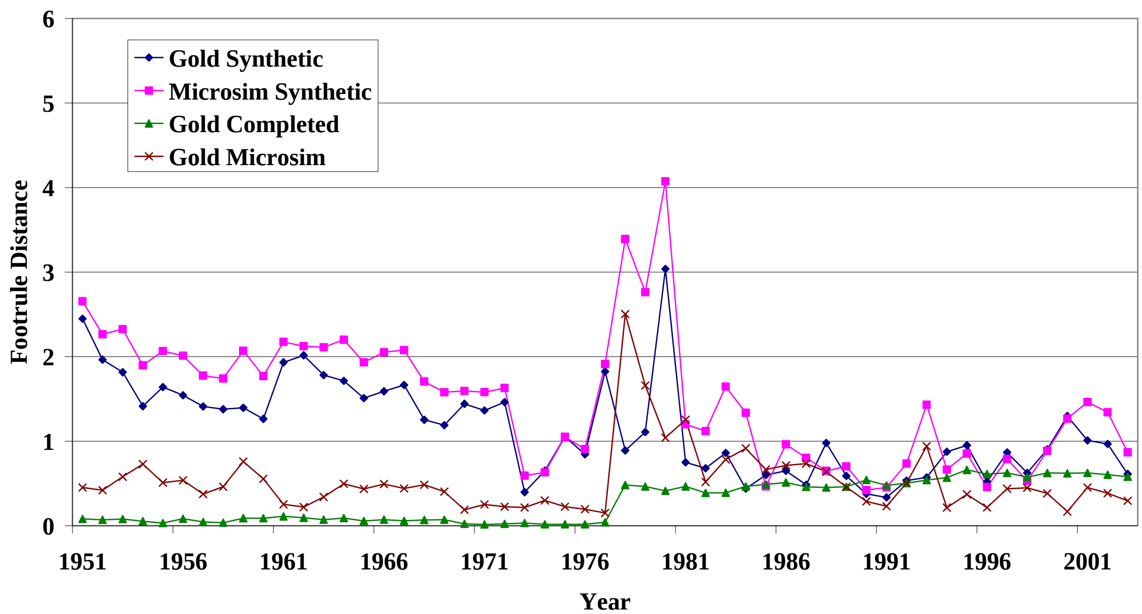
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive earnings in 1978.

Figure 5-7: Total Earnings / Average Wage of Female Workers in 2000
Survivors to 2000



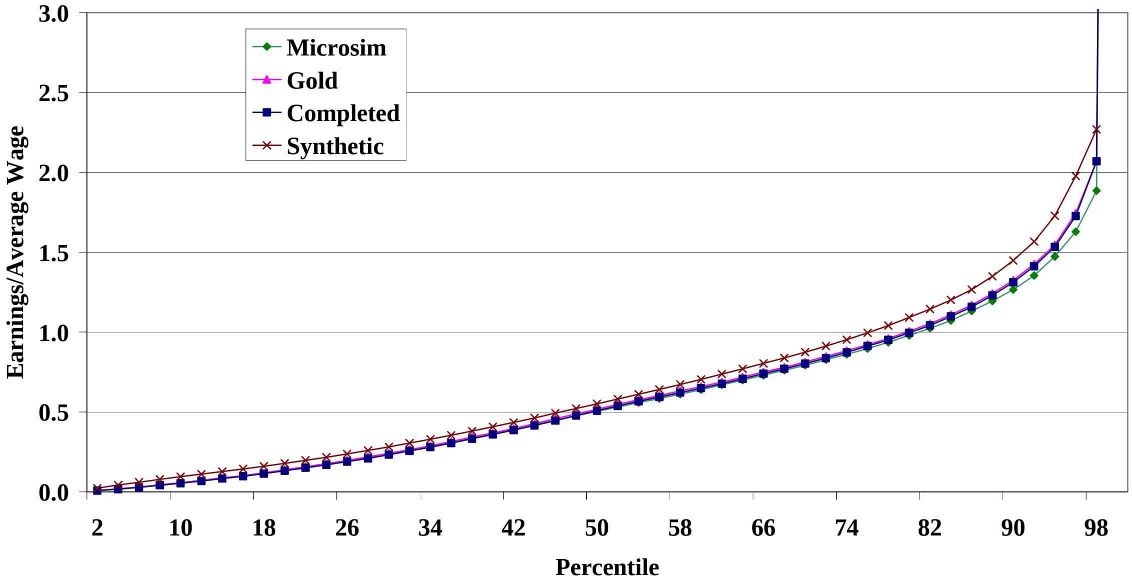
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive earnings in 2000.

Figure 5-8: Footrule Distance Total Earnings/Average Wage of Female Workers 1951 to 2003



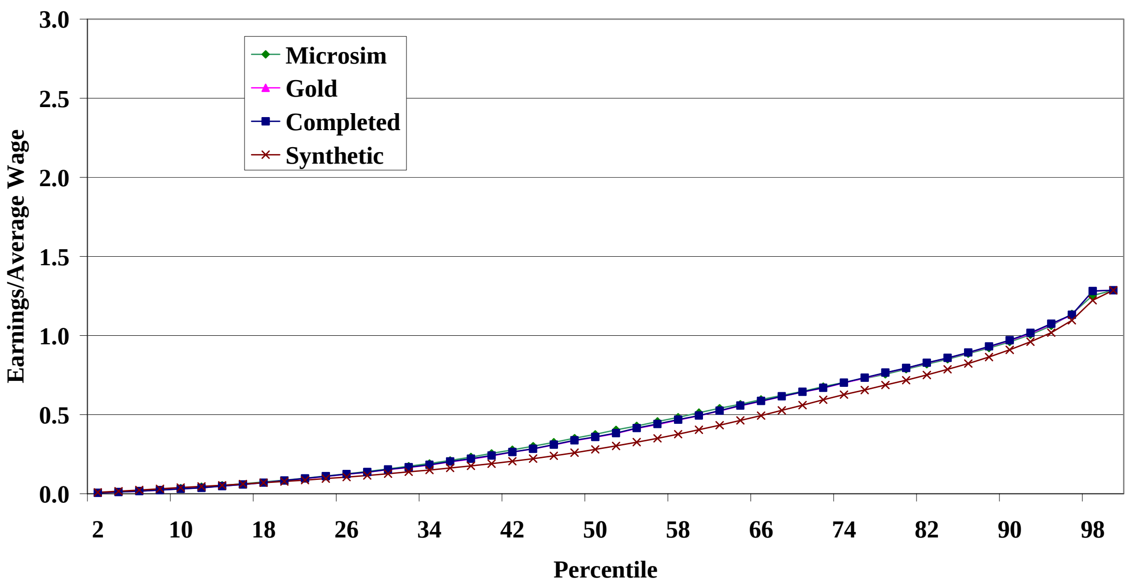
Source: Urban Institute footrule distance calculations of the distribution of total earnings of female workers in the Microsim, Gold, pooled Completed, and pooled Synthetic files. The distribution includes the sum of the absolute differences between the 1st and 98th percentiles. The data are limited to U.S. residents that survive to 2000.

Figure 5-9: Total Earnings / Average Wage of Female Workers in 1980 Survivors to 2000



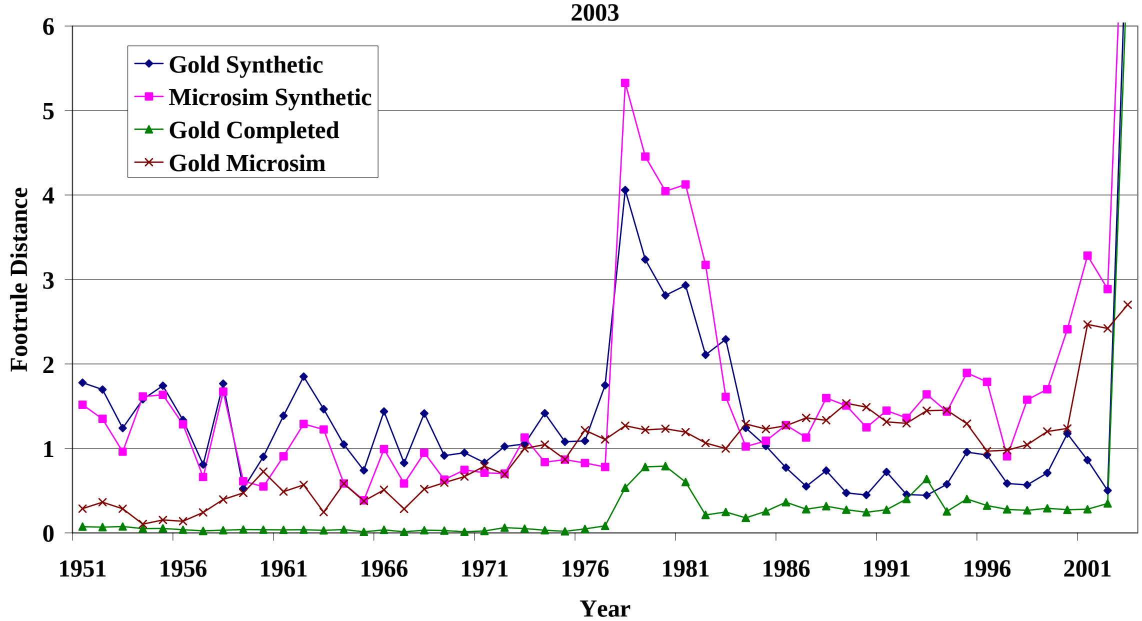
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive earnings in 1980.

Figure 5-10: Total Earnings / Average Wage of Female Workers 1951 Survivors to 2000



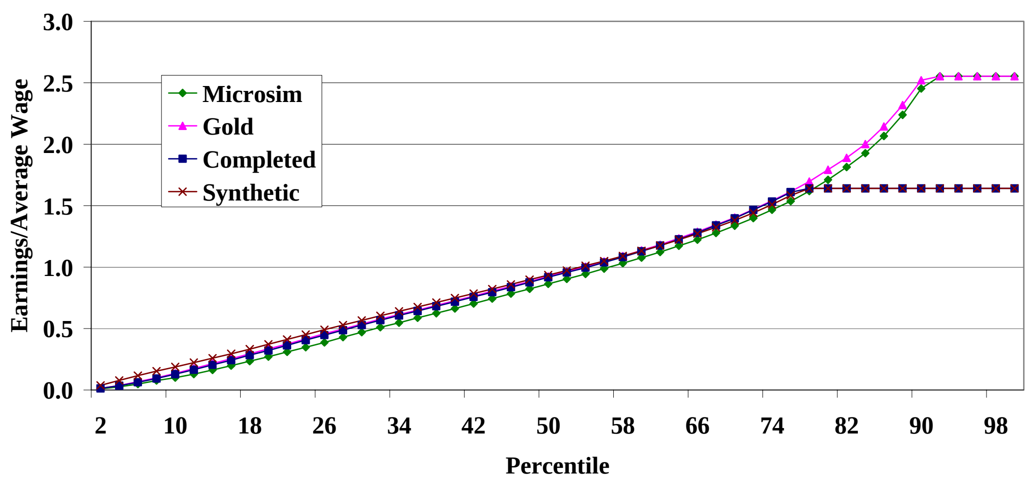
Source: Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive earnings in 1951.

Figure 5-11: Footrule Distance Taxable Earnings/Average Wage of Male Workers 1951 to 2003



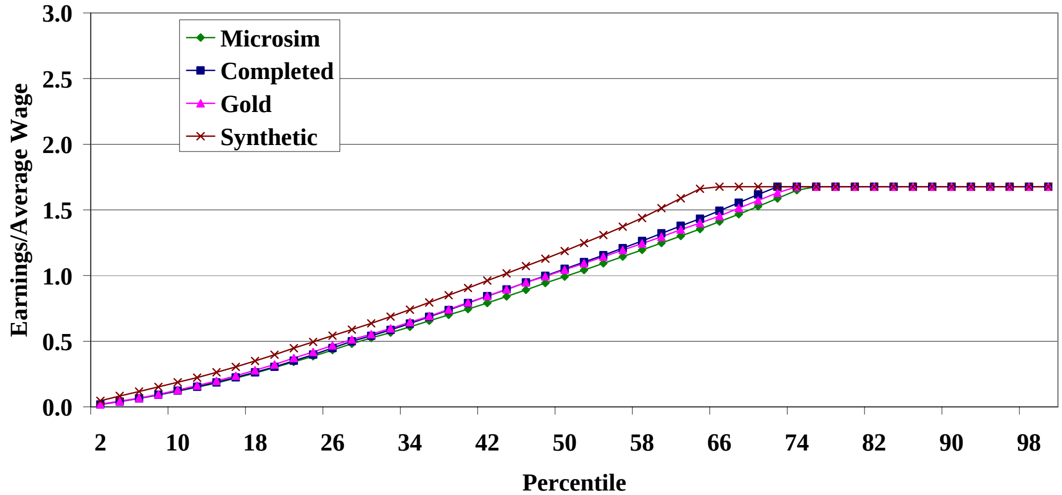
Source: Urban Institute footrule distance calculations of the distribution of taxable earnings of male workers in the Microsim, Gold, pooled Completed, and pooled Synthetic files. The distribution includes the sum of the absolute differences between the 1st and 98th percentiles. The data are limited to U.S. residents that survive to 2000.

Figure 5-12. Taxable Earnings / Average Wage of Male Workers in 2003
Survivors to 2003



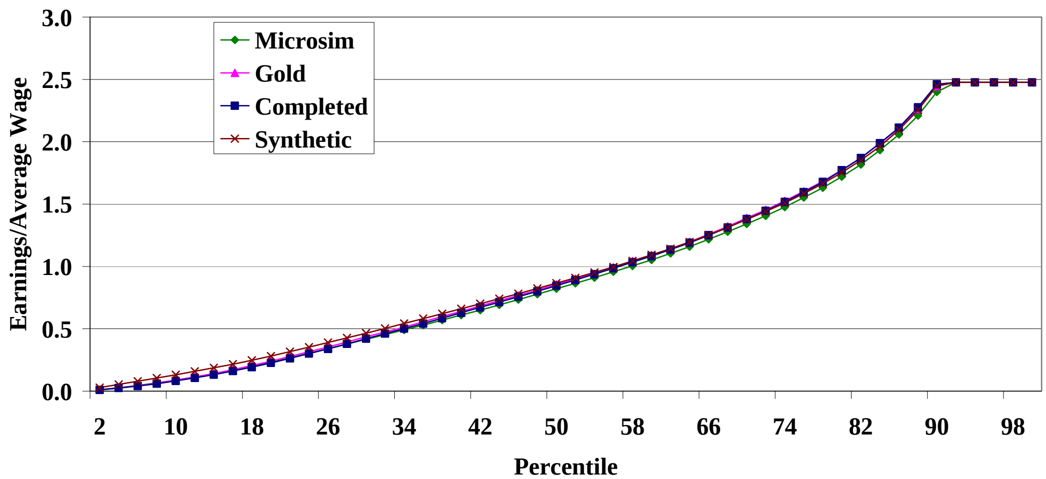
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2003 with positive taxable earnings in 2003.

Figure 5-13: Taxable Earnings / Average Wage of Male Workers in 1978
Survivors to 2000



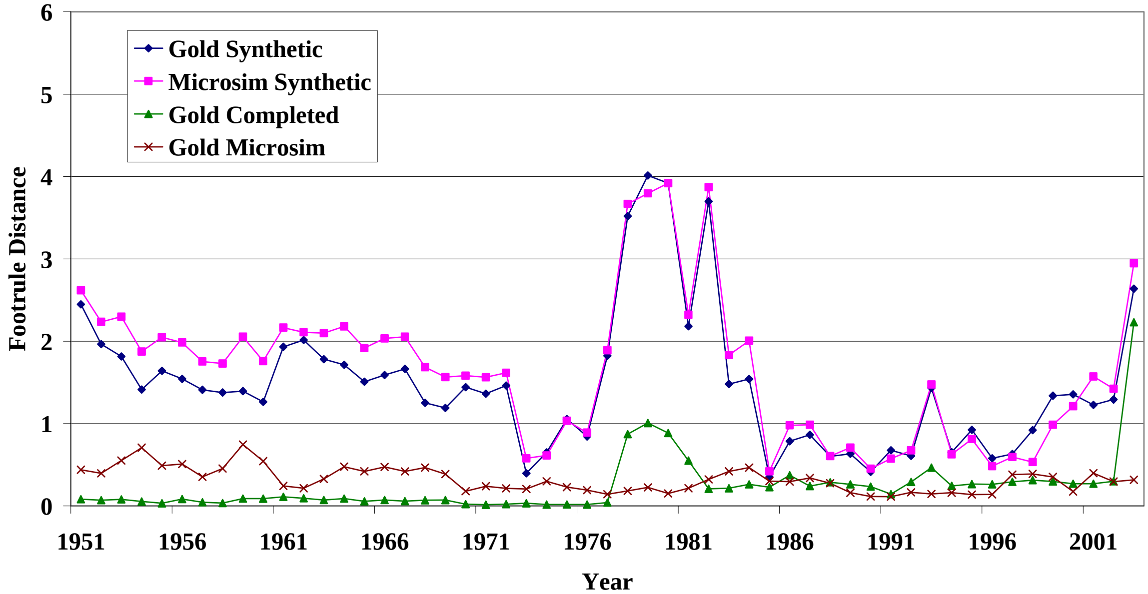
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with taxable earnings in 1978.

Figure 5-14: Taxable Earnings / Average Wage of Male Workers in 1995
Survivors to 2000



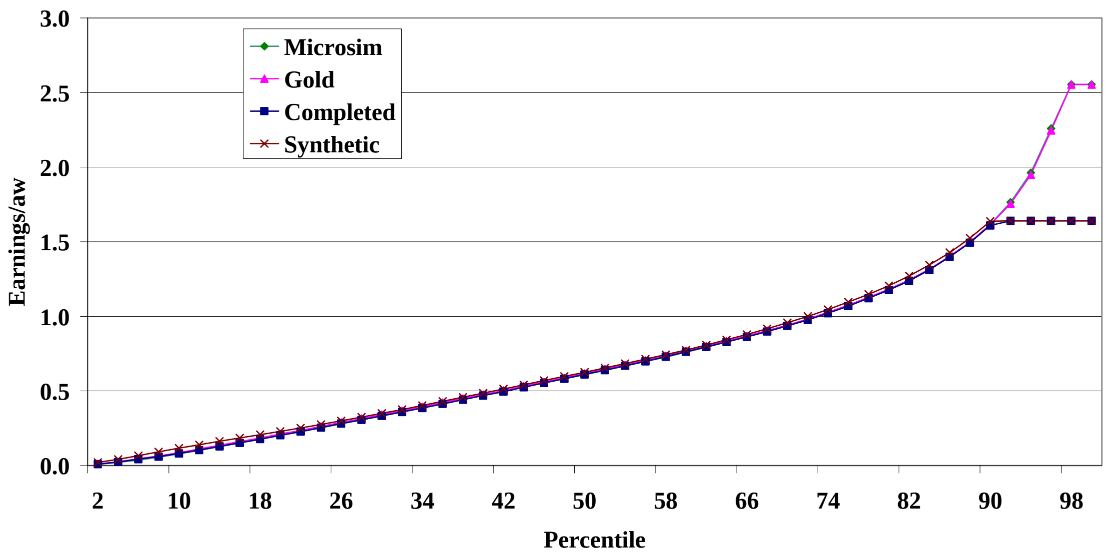
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive taxable earnings in 1995.

Figure 5-15: Footrule Distance Taxable Earnings/Average Wage of Female Workers 1951 to 2003



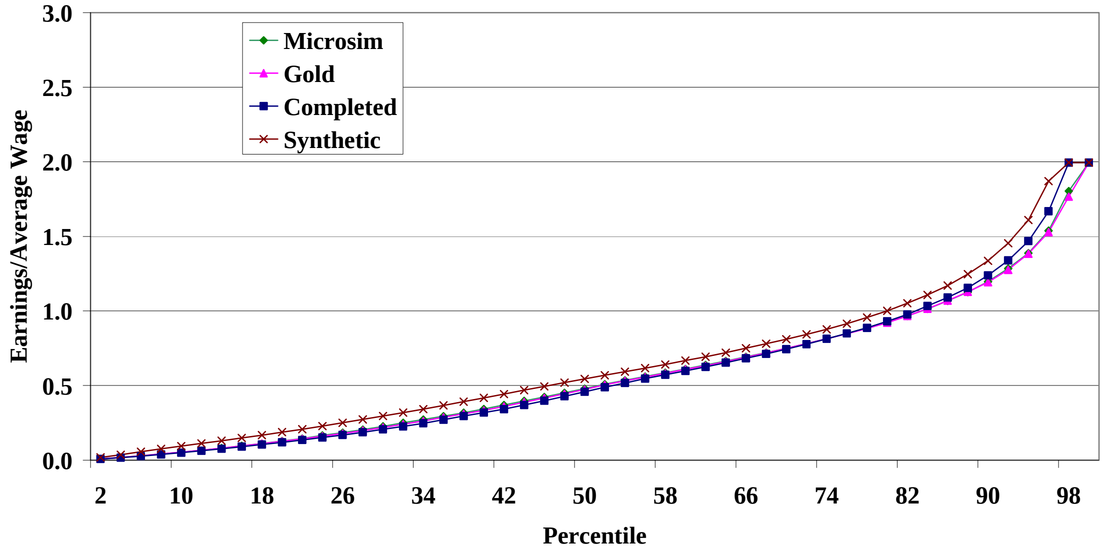
Source: Urban Institute footrule distance calculations of the distribution of taxable earnings of female workers in the Microsim, Gold, pooled Completed, and pooled Synthetic files. The distribution includes the sum of the absolute differences between the 1st and 98th percentiles. The data are limited to U.S. residents that survive to 2000.

Figure 5-16. Taxable Earnings / Average Wage 2003 of Female Workers Survivors to 2003



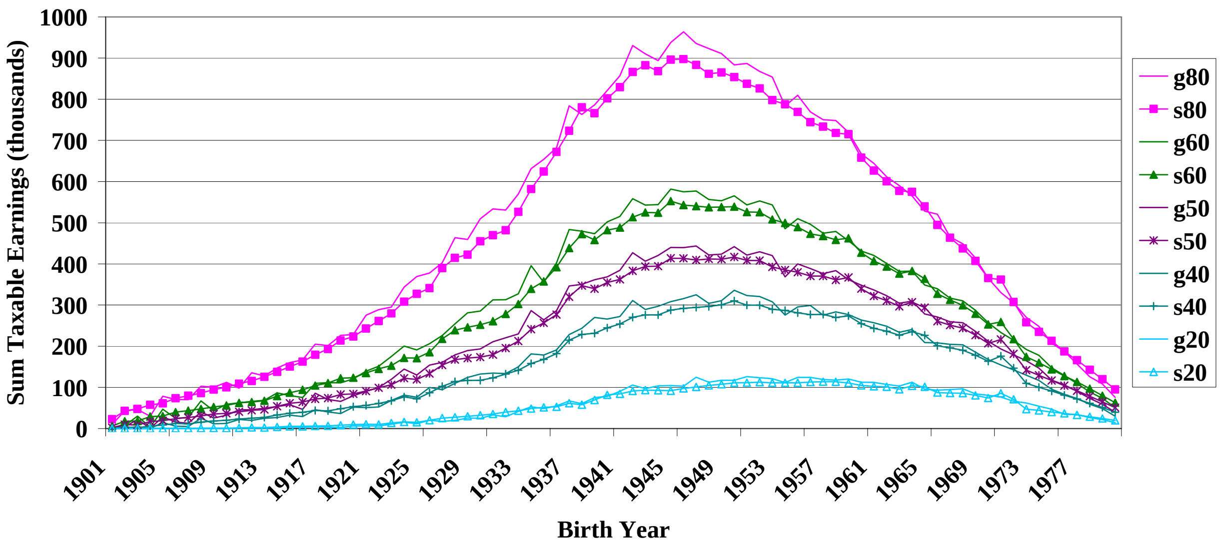
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2003 with positive taxable earnings in 2003.

Figure 5-17. Taxable Earnings / Average Wage of Female Workers in 1979
Survivors to 2000



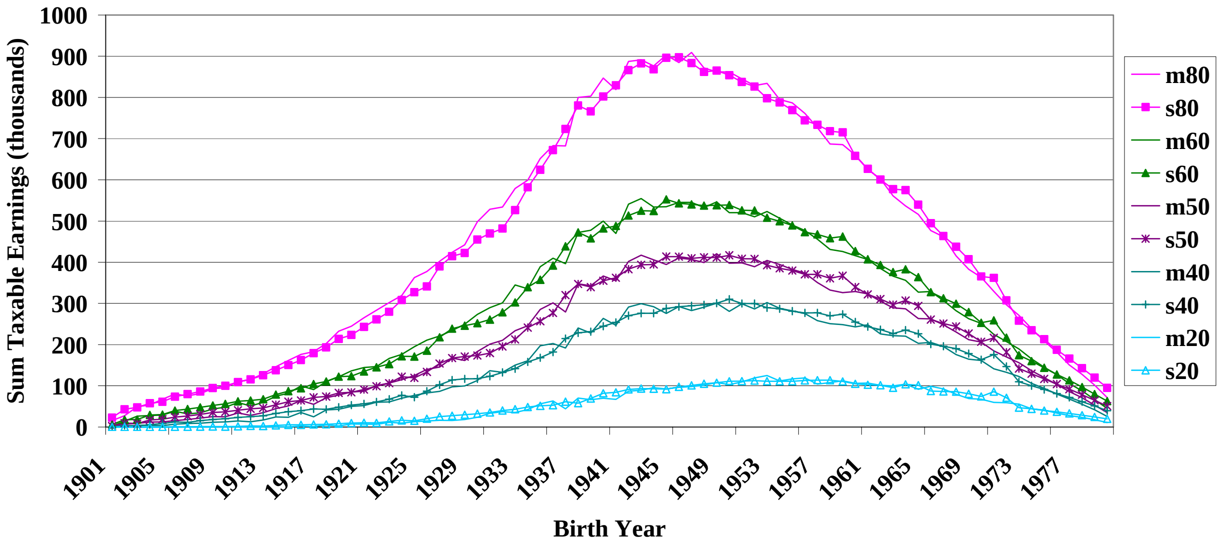
Source. Urban Institute tabulations of Microsim, pooled Synthetic, pooled Completed, and Gold SIPP files. Data are limited to U.S. resident survivors to 2000 with positive taxable earnings in 1979.

Figure 5-18a. Selected Percentiles of Sum of Nominal Taxable Earnings
(Age 16 - 65 between 1951-2003) by Birth Year
Synthetic Versus Gold Among Survivors to 2000



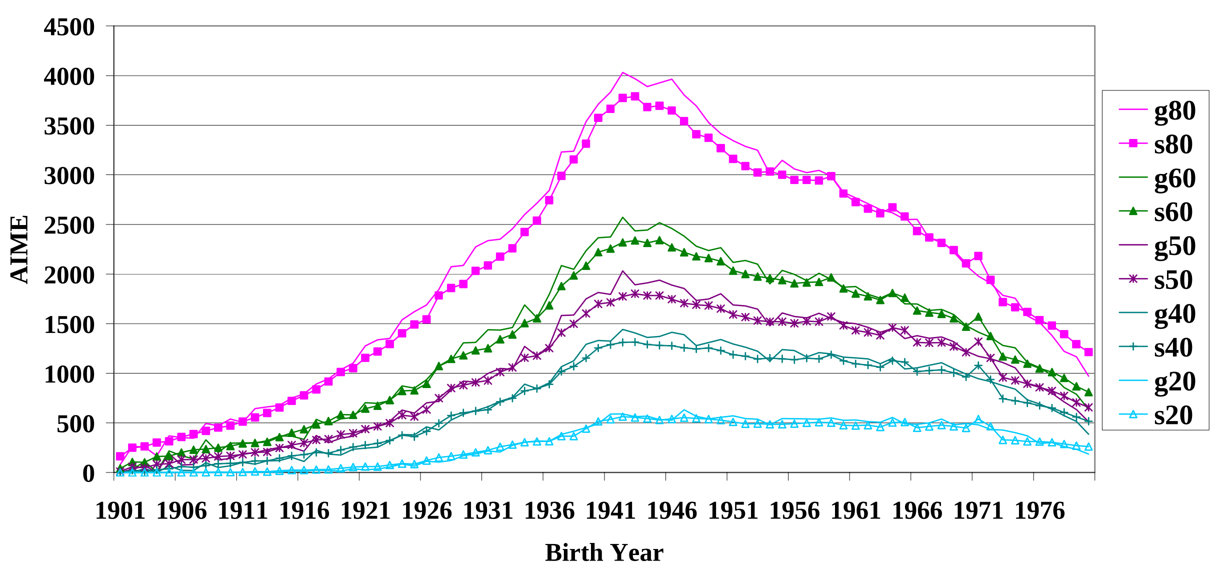
Source. Urban Institute tabulations of pooled Synthetic and Gold SIPP files. Data are limited to U.S. resident survivors to 2000. Total earnings is the sum of nominal taxable earnings from age 16 to 65 between 1951 and 2003. Prefixs *g* and *s* in the key indicate Gold and Synthetic file values respectively.

Figure 5-18b. Selected Percentiles of Sum of Nominal Taxable Earnings (Age 16 - 65 between 1951-2003) by Birth Year Synthetic Versus Microsim Among Survivors to 2000



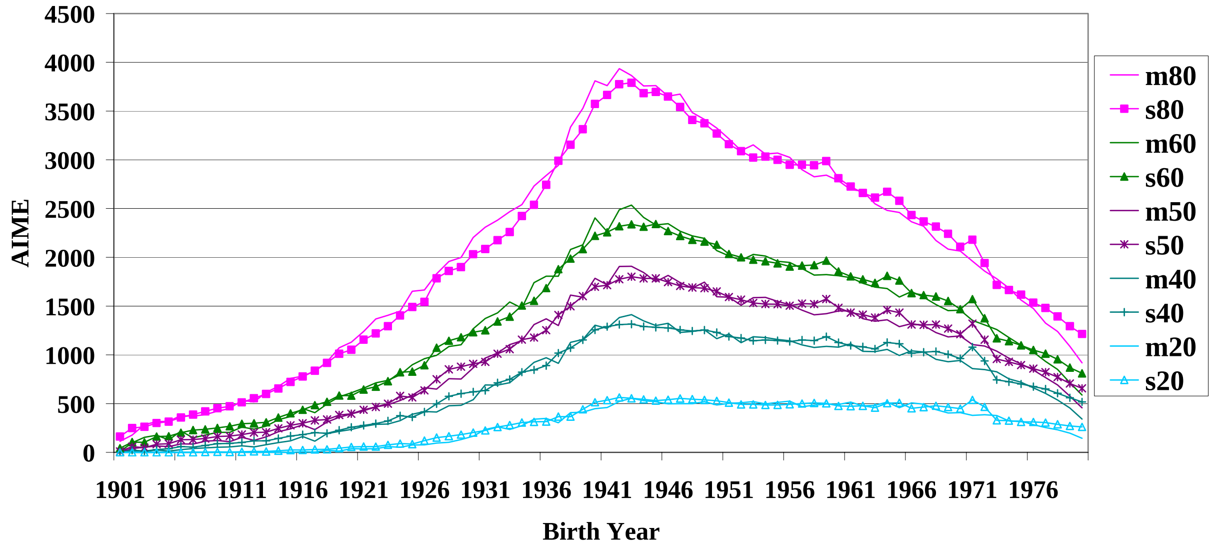
Source. Urban Institute tabulations of Microsim and pooled Synthetic SIPP files. Data are limited to U.S. resident survivors to 2000. Total earnings is the sum of nominal taxable earnings from age 16 to 65 between 1951 and 2003. Prefixs m and *s* in the key indicate Microsim and Synthetic file values respectively.

Figure 5-19a. Selected Percentiles of AIME by Birth Year Synthetic Versus Gold Among Survivors to 2000



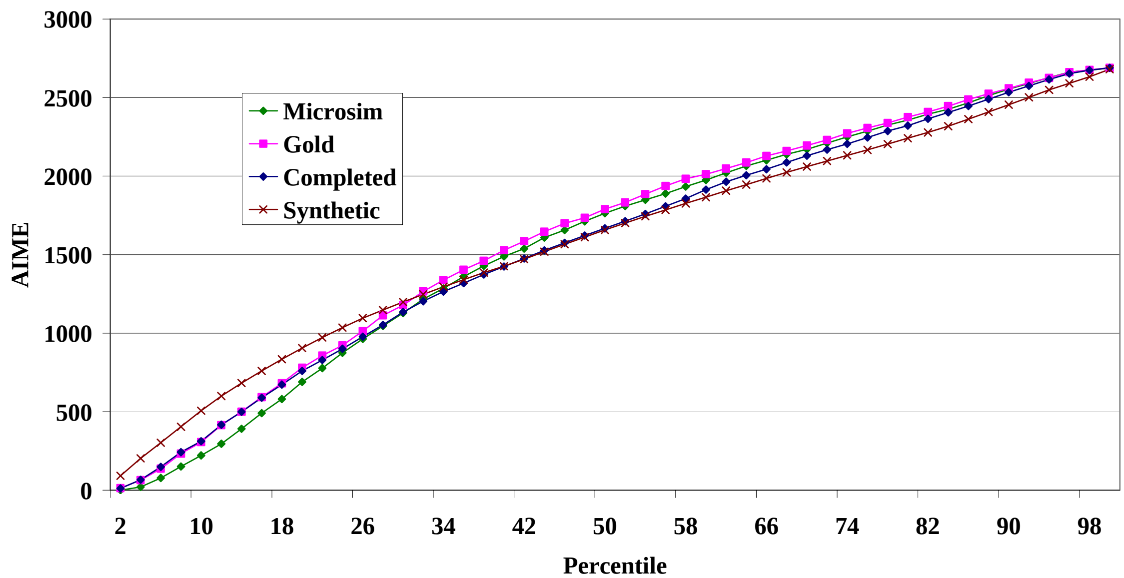
Source. Urban Institute tabulations of Gold and pooled Synthetic SIPP files. Data are limited to U.S. resident survivors to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003. Prefixs *g* and *s* in the key indicate Gold and Synthetic file values respectively.

Figure 5-19b. Selected Percentiles of AIME by Birth Year
Synthetic Versus Microsim Among Survivors to 2000



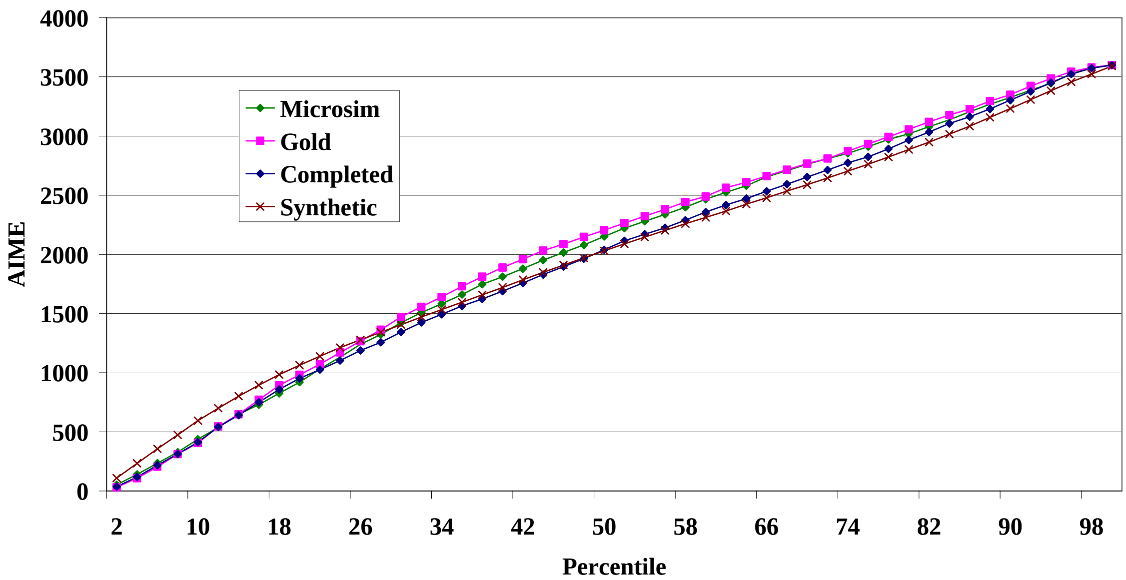
Source. Urban Institute tabulations of Microsim and pooled Synthetic SIPP files. Data are limited to U.S. resident survivors to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003. Prefixs *m* and *s* in the key indicate Microsim and Synthetic file values respectively.

Figure 5-20a. Distribution of AIME Men Born 1926 to 1930
Survivors to 2000



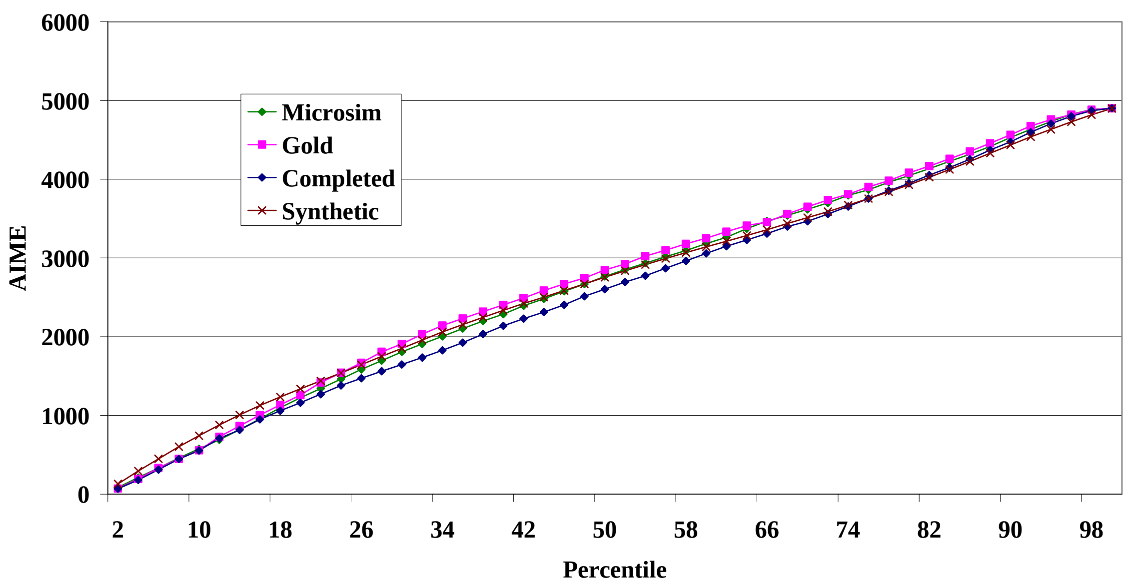
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1926 to 1930 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20b. Distribution of AIME Men Born 1931 to 1935
Survivors to 2000



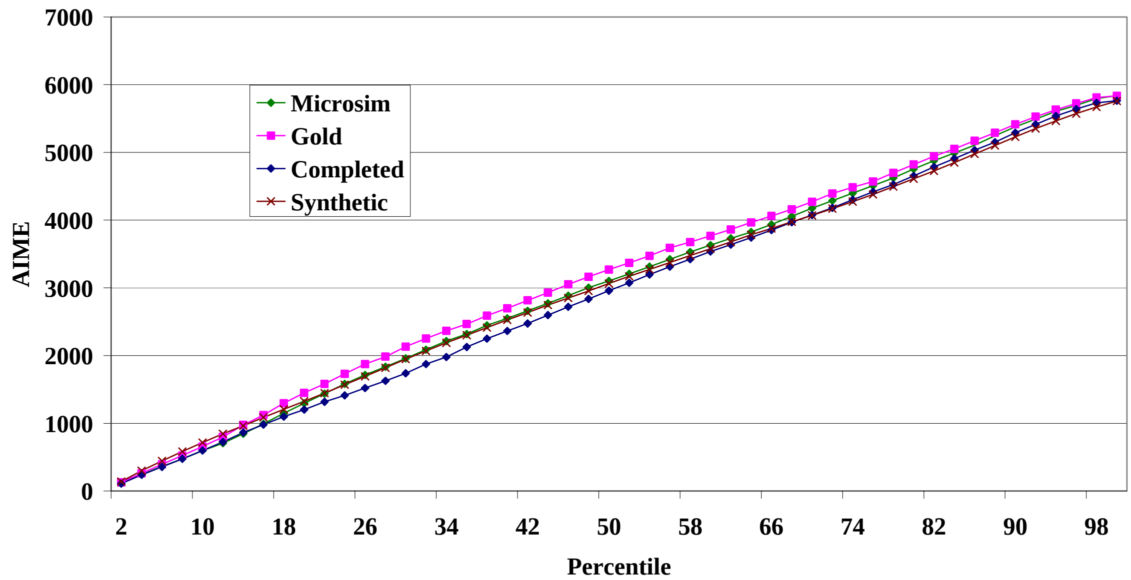
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1931 to 1935 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20c. Distribution of AIME Men Born 1936 to 1940
Survivors to 2000



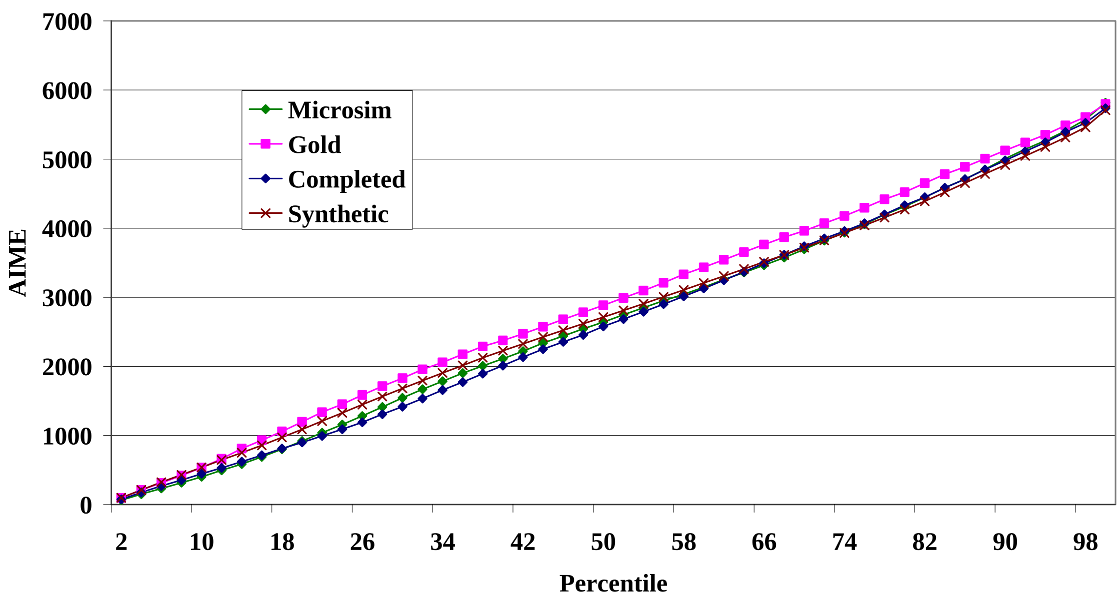
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1936 to 1940 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20d. Distribution of AIME Men Born 1941 to 1945
Survivors to 2000



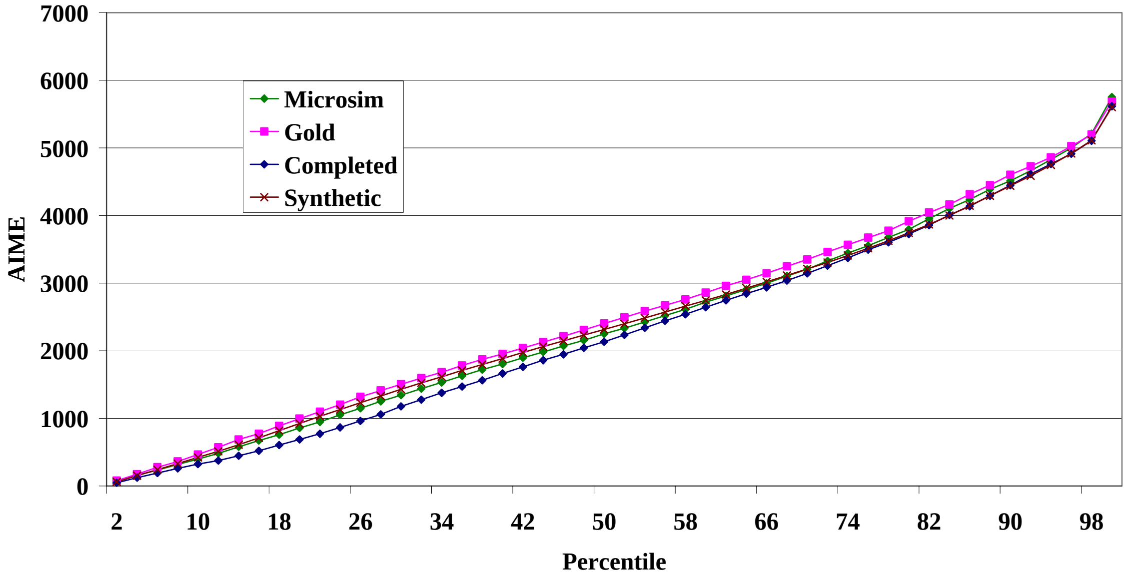
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1941 to 1945 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20e. Distribution of AIME Men Born 1946 to 1950
Survivors to 2000



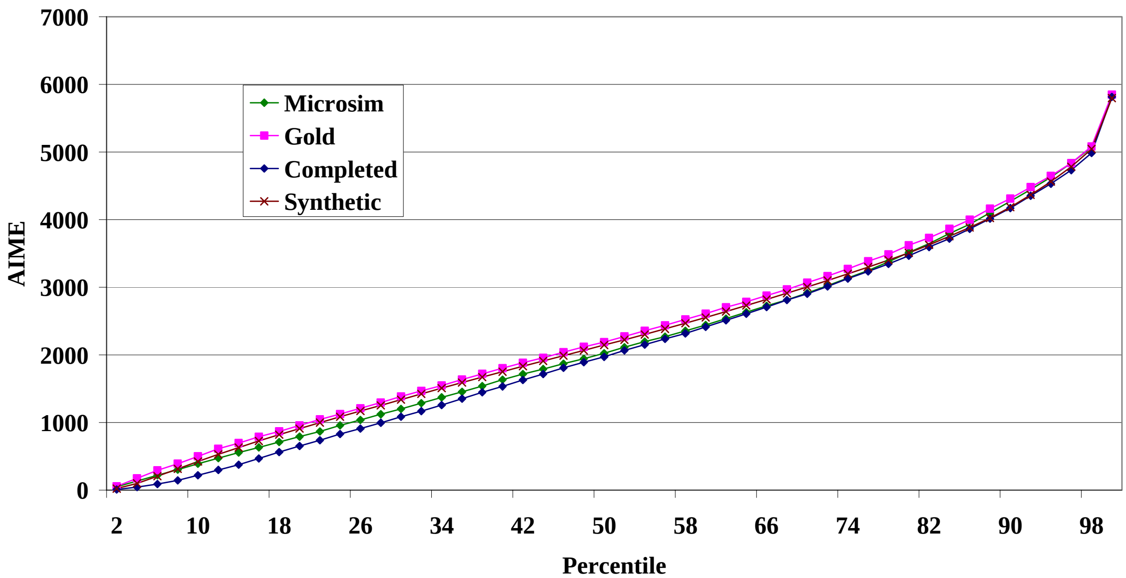
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1946 to 1950 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20f. Distribution of AIME Men Born 1951 to 1955
Survivors to 2000



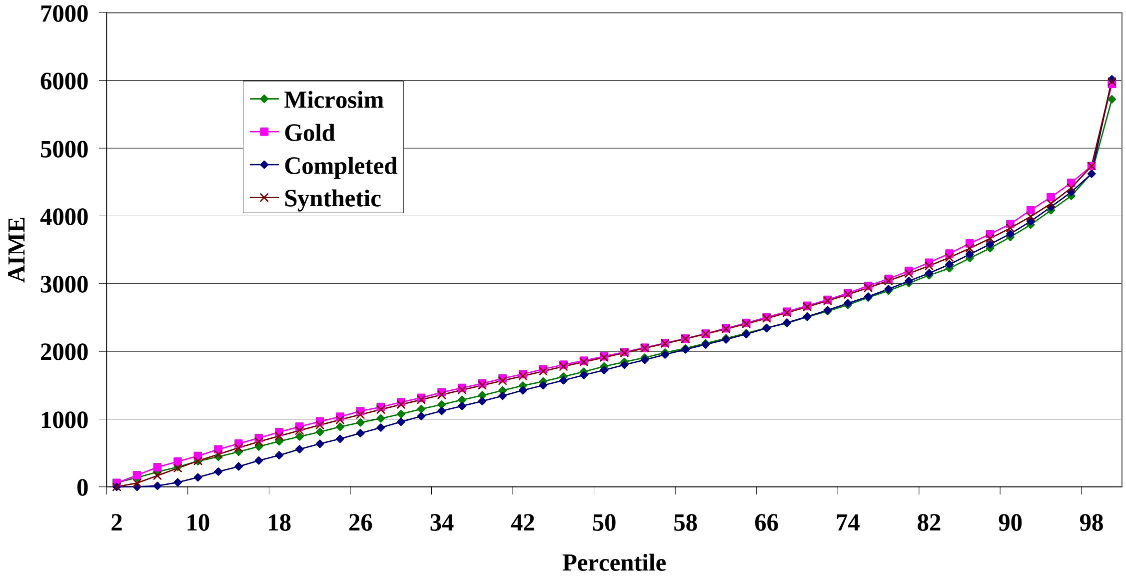
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1951 to 1955 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20g. Distribution of AIME Men Born 1956 to 1960
Survivors to 2000



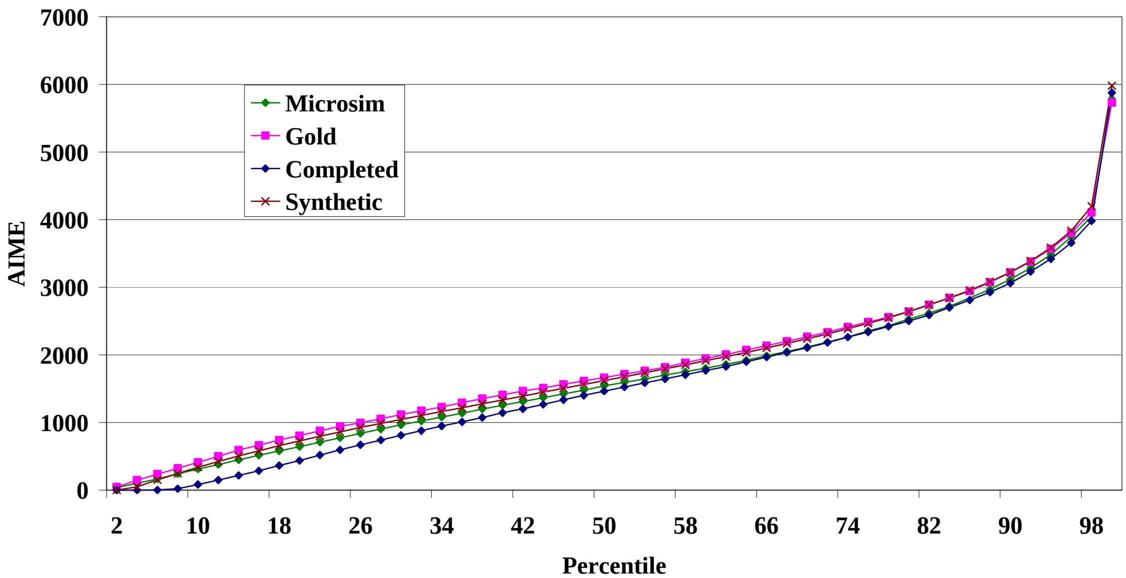
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1956 to 1960 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20h. Distribution of AIME Men Born 1961 to 1965
Survivors to 2000



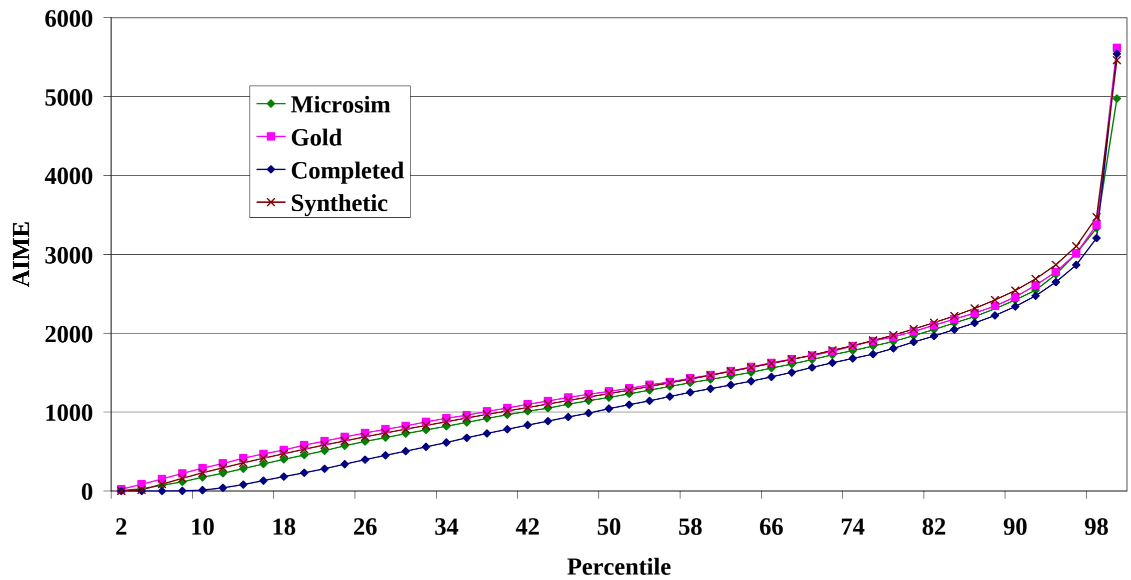
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1961 to 1965 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20i. Distribution of AIME Men Born 1966 to 1970
Survivors to 2000



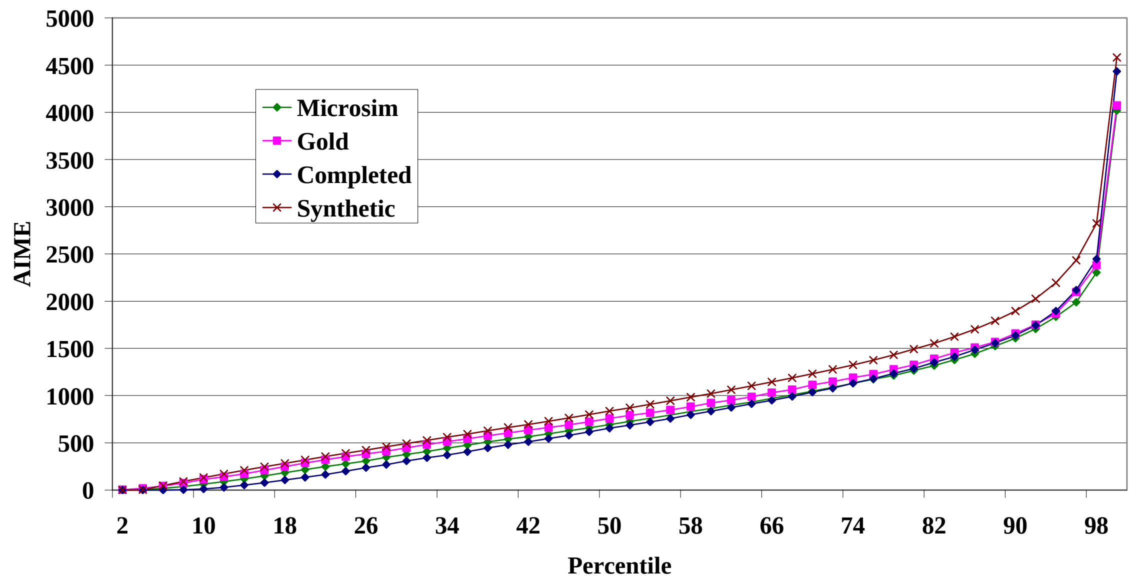
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1966 to 1970 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20j. Distribution of AIME Men Born 1971 to 1975
Survivors to 2000



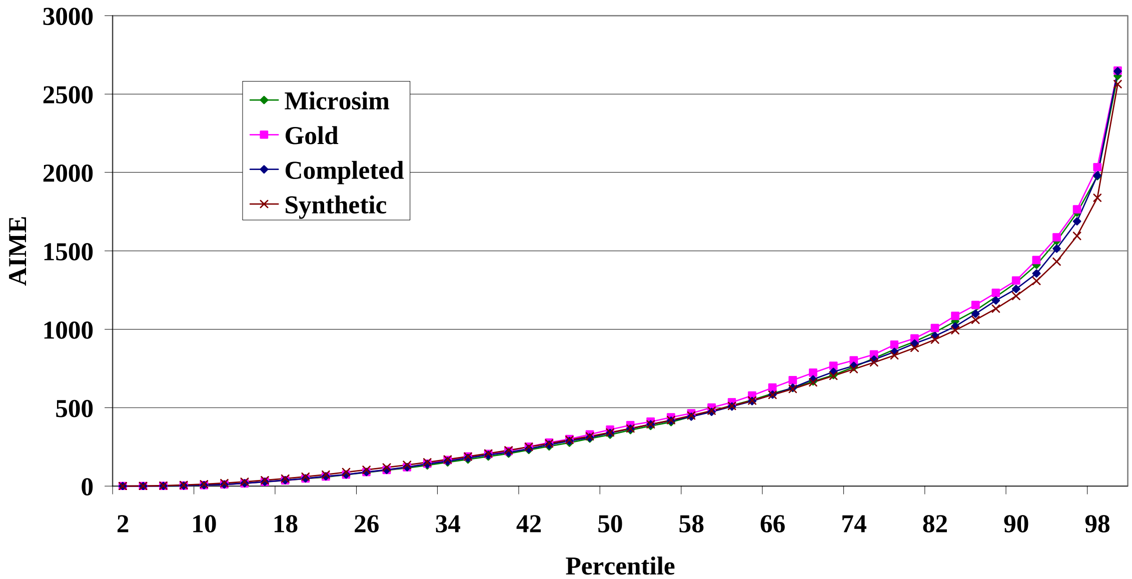
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1971 to 1975 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-20k. Distribution of AIME Men Born 1976 to 1980
Survivors to 2000



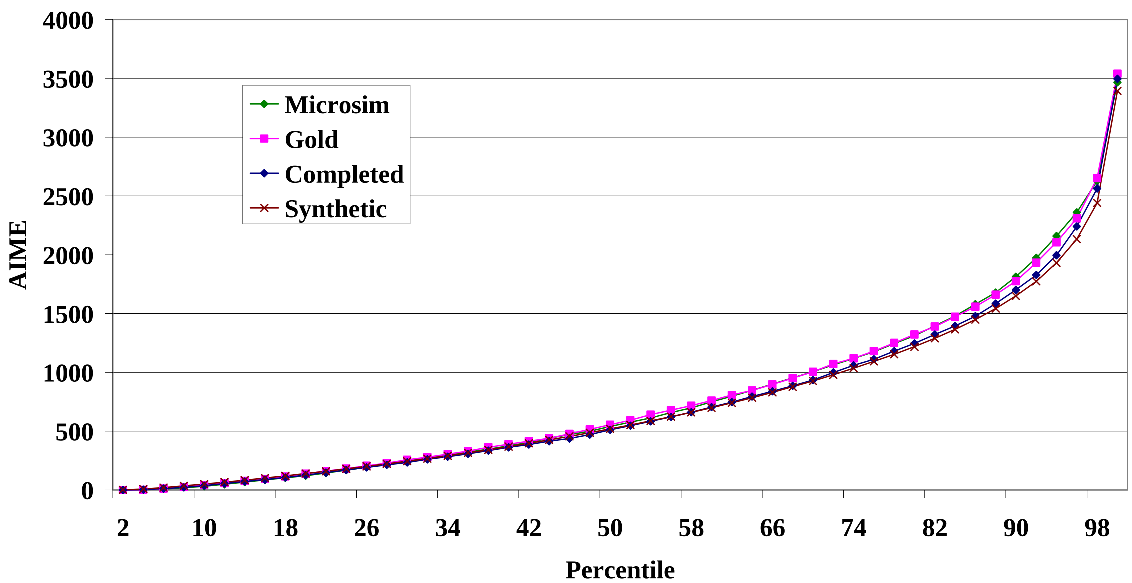
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1976 to 1980 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21a. Distribution of AIME Females Born 1926 to 1930 Survivors to 2000



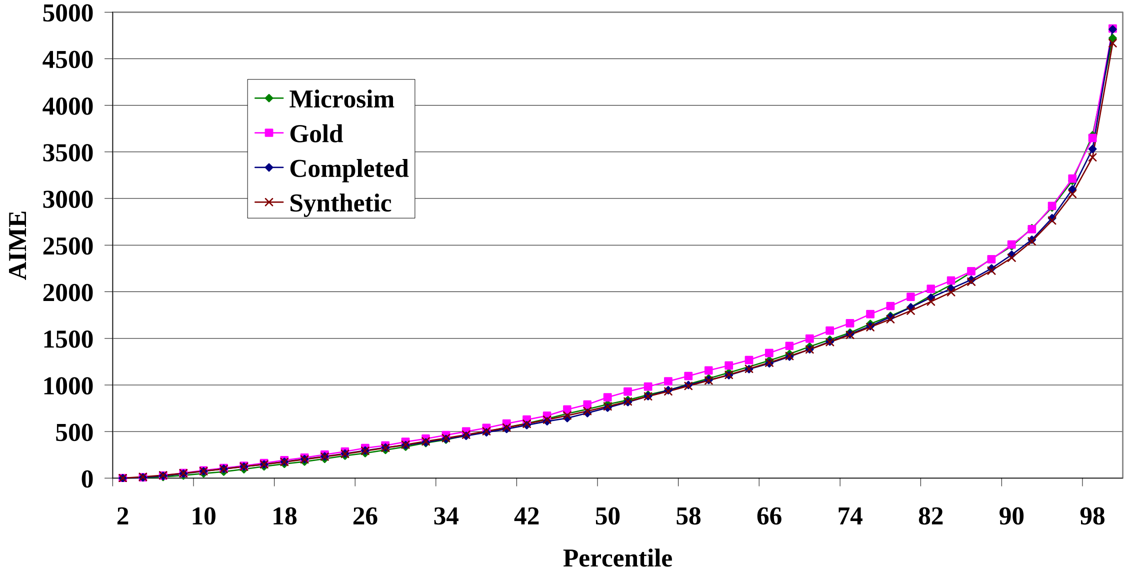
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1926 to 1930 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21b. Distribution of AIME Females Born 1931 to 1935 Survivors to 2000



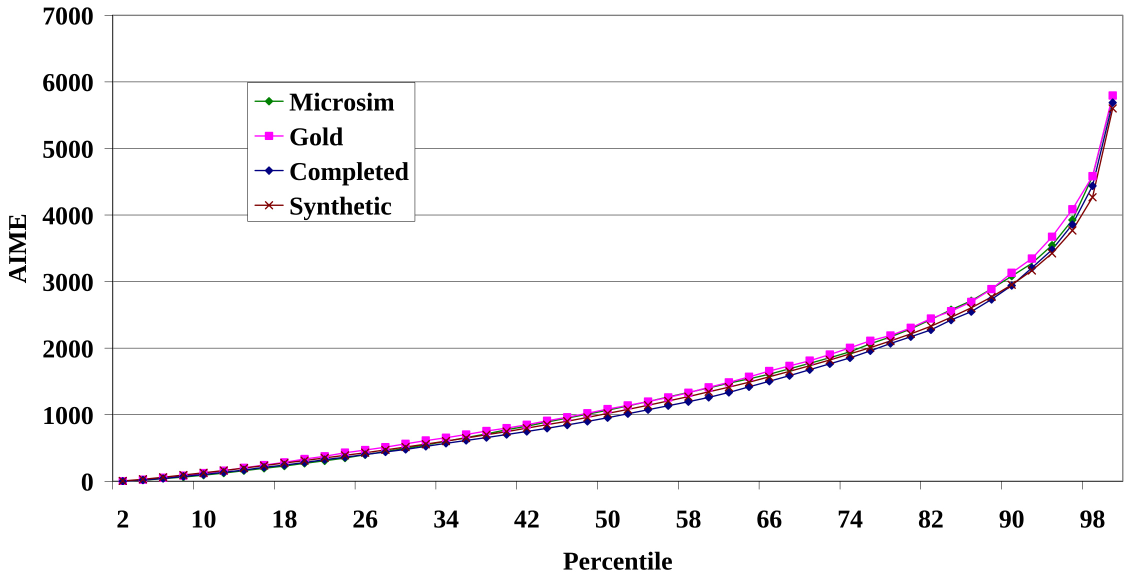
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1931 to 1935 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21c. Distribution of AIME Females Born 1936 to 1940
Survivors to 2000



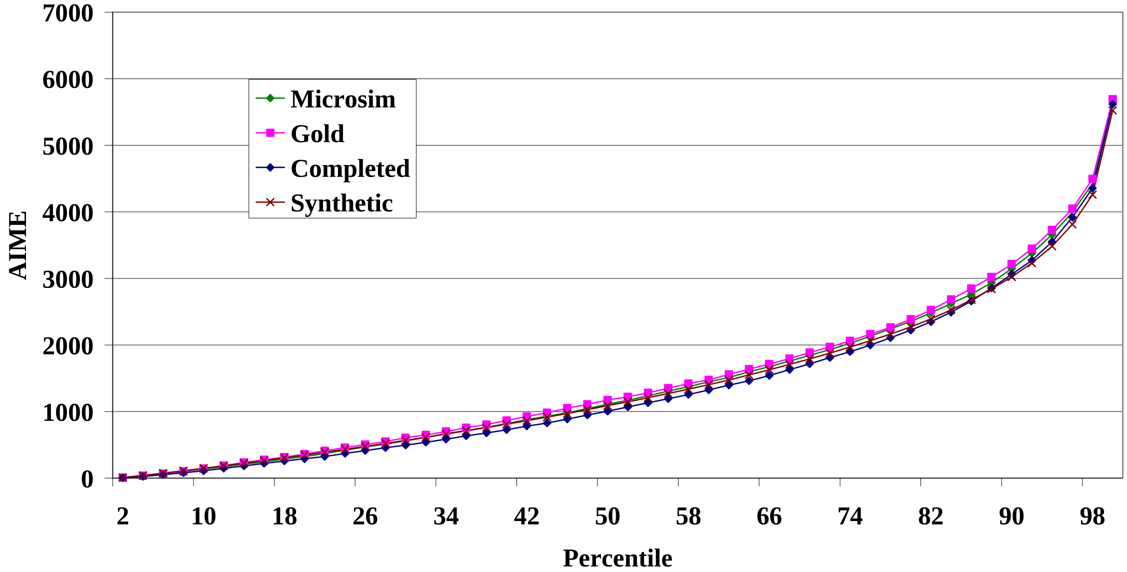
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1936 to 1940 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21d. Distribution of AIME Females Born 1941 to 1945
Survivors to 2000



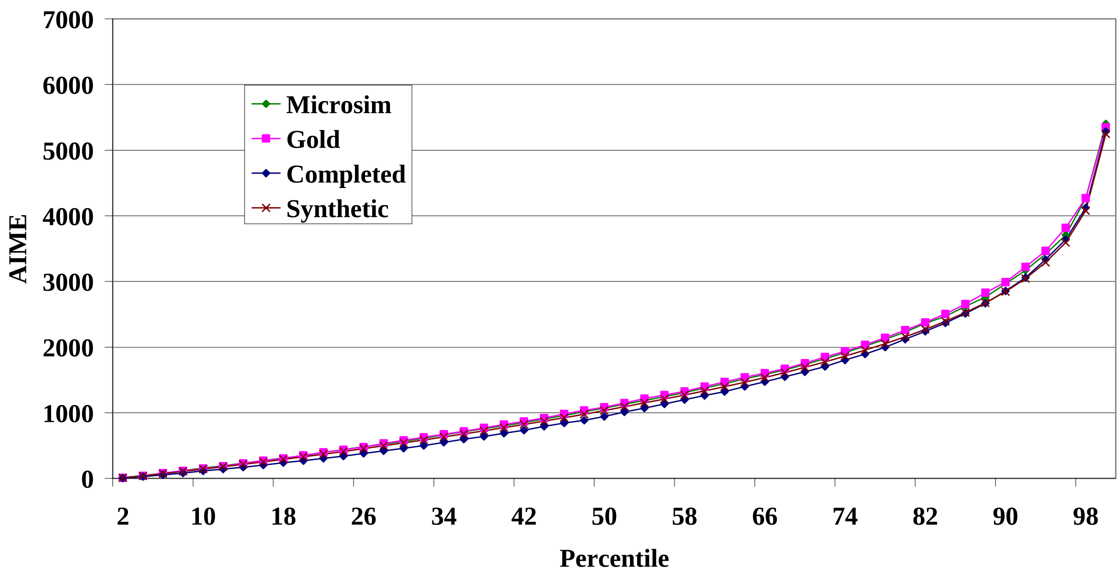
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1941 to 1945 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21e. Distribution of AIME Females Born 1946 to 1950
Survivors to 2000



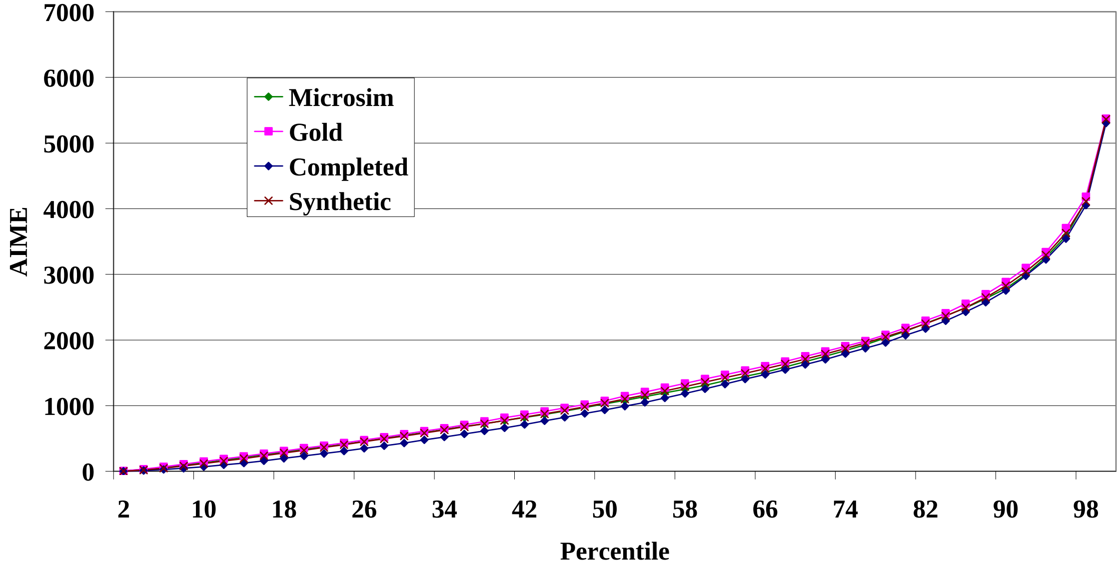
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1946 to 1950 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21f. Distribution of AIME Females Born 1951 to 1955
Survivors to 2000



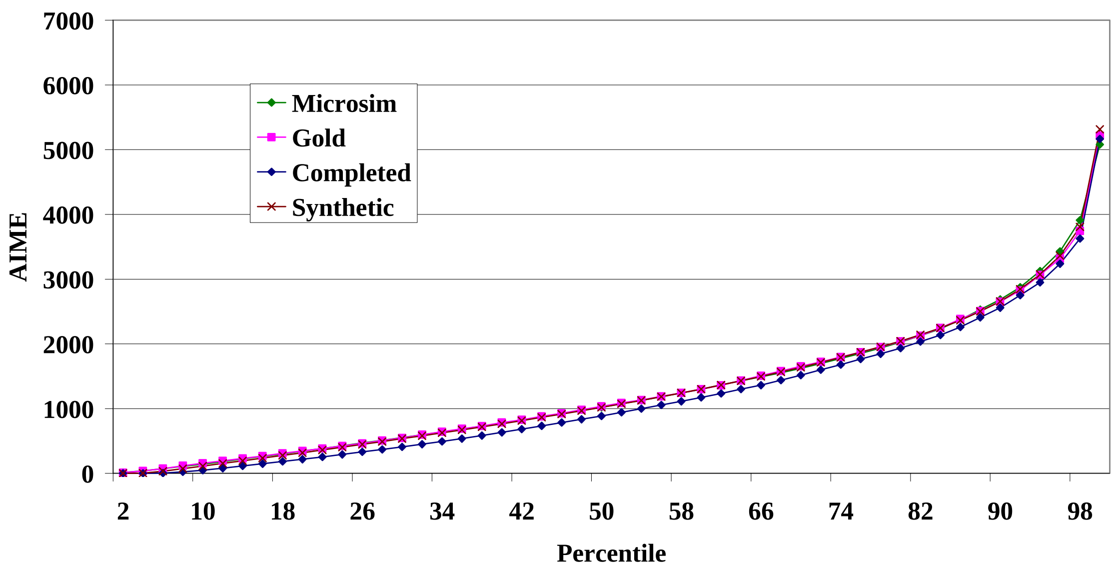
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1951 to 1955 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21g. Distribution of AIME Females Born 1956 to 1960
Survivors to 2000



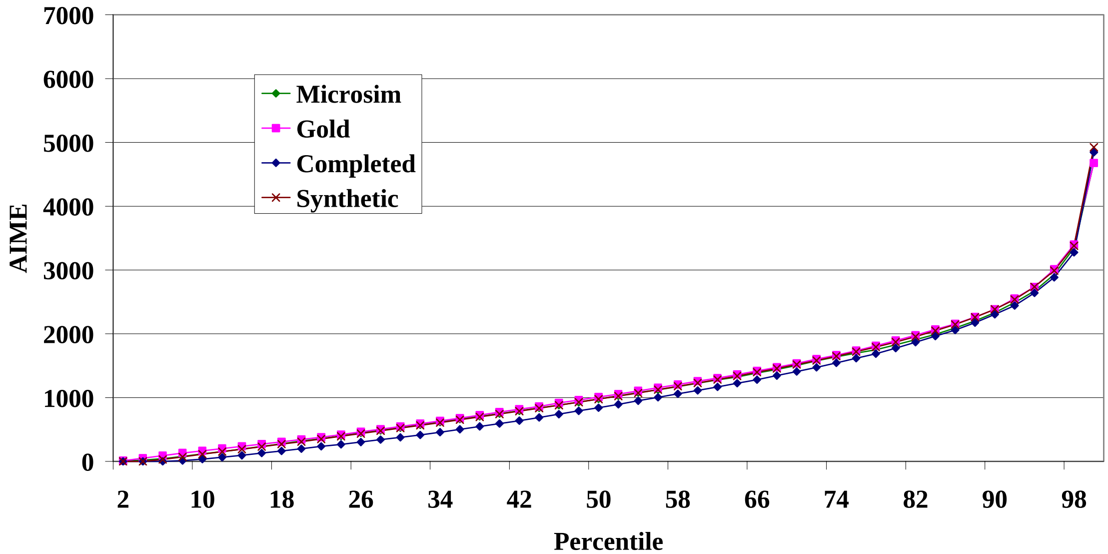
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1956 to 1960 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21h. Distribution of AIME Females Born 1961 to 1965
Survivors to 2000



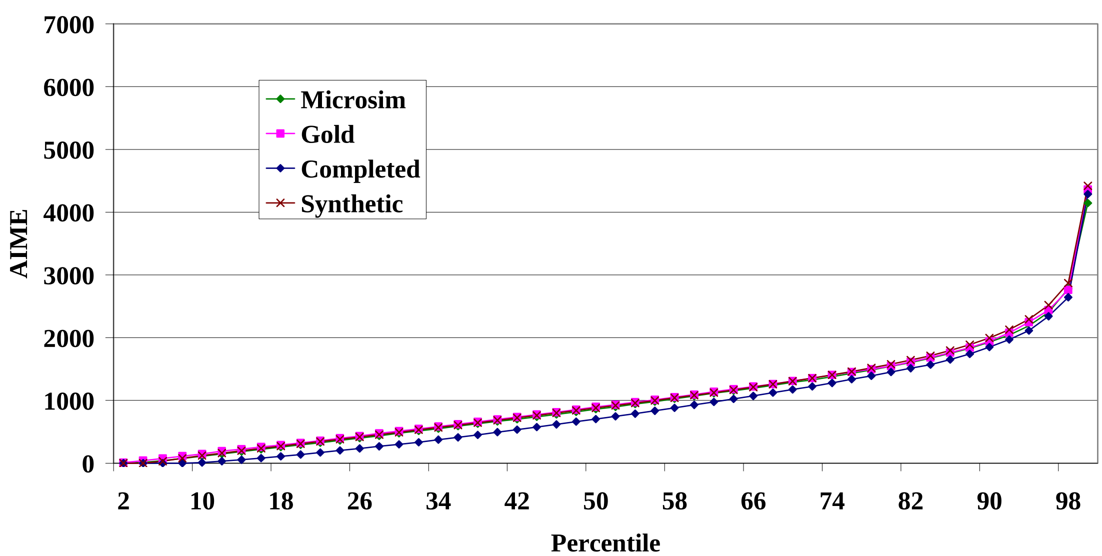
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1961 to 1965 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21i. Distribution of AIME Females Born 1966 to 1970
Survivors to 2000



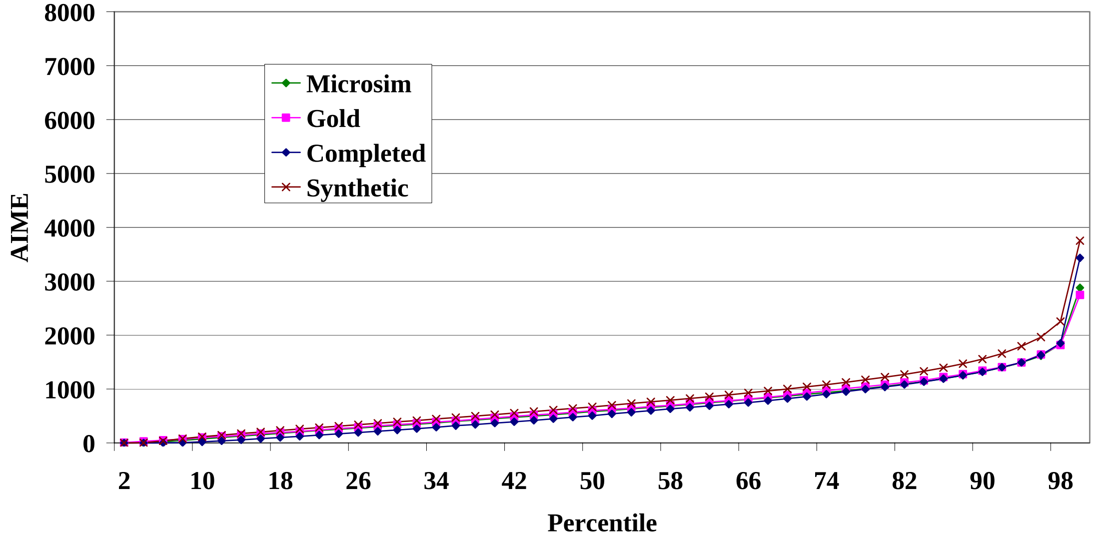
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1966 to 1970 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21j. Distribution of AIME Females Born 1971 to 1975
Survivors to 2000



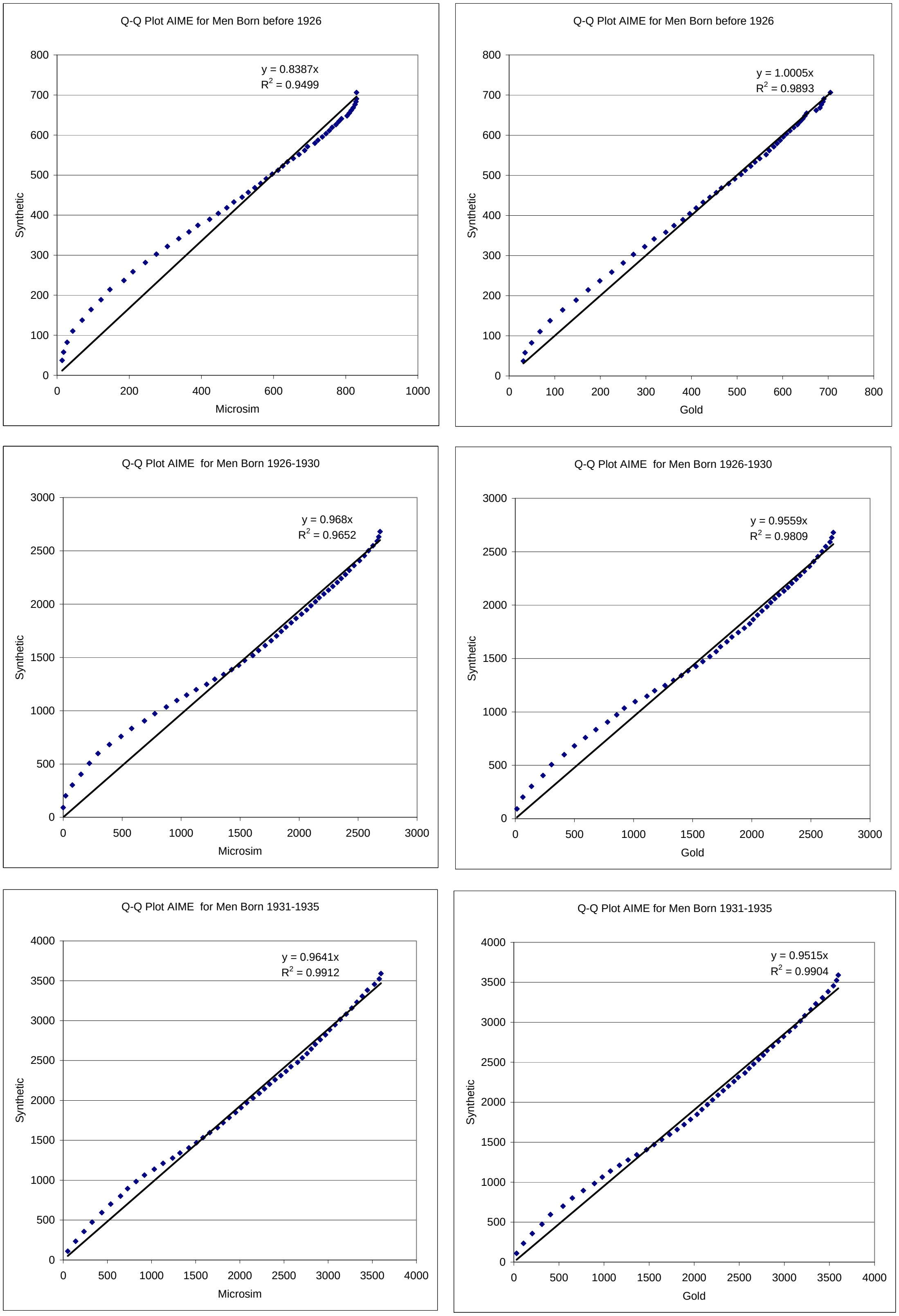
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1971 to 1975 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-21k. Distribution of AIME Females Born 1976 to 1980 Survivors to 2000



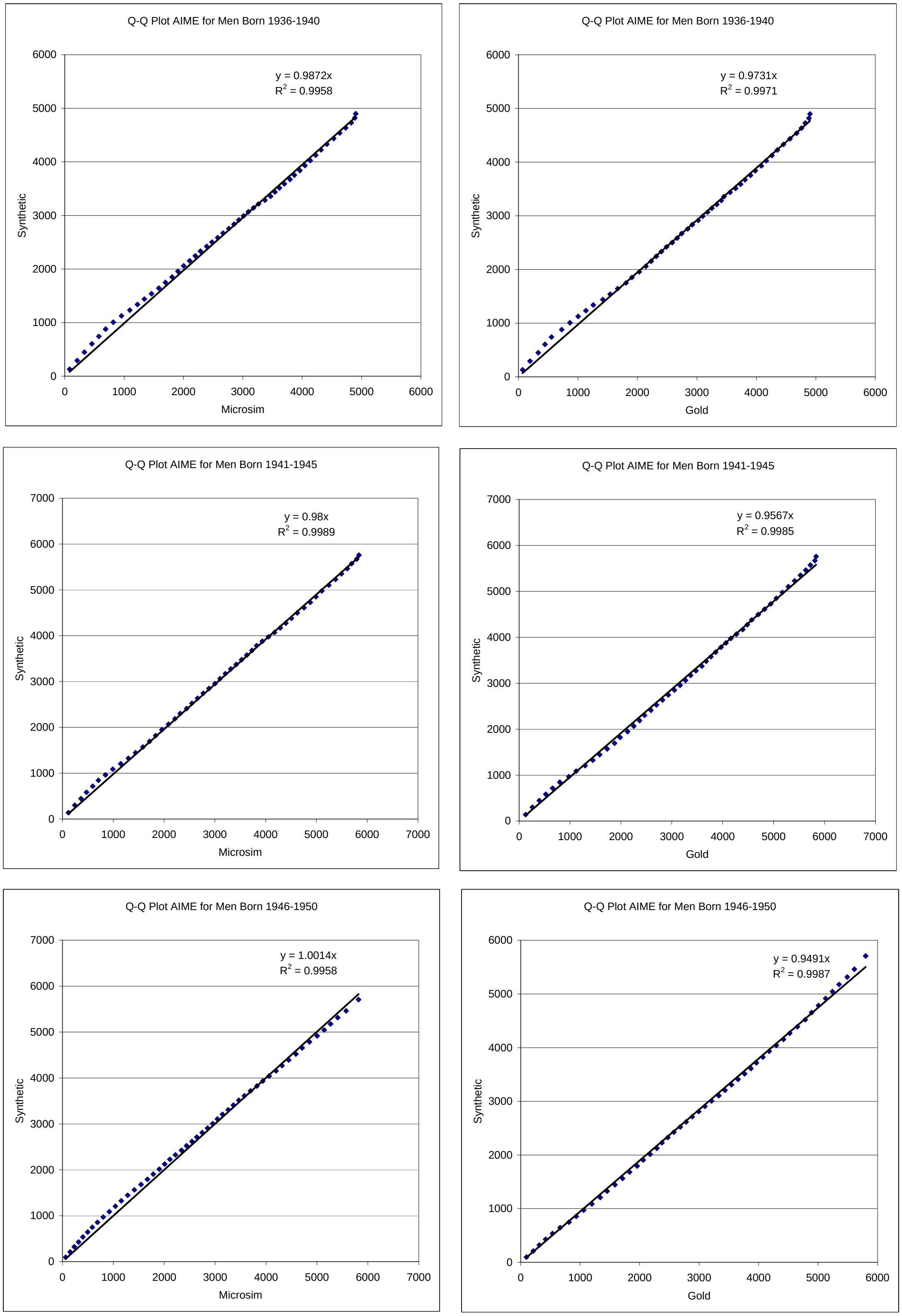
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1976 to 1980 that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-22a: Q-Q Plots for AIMEs for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



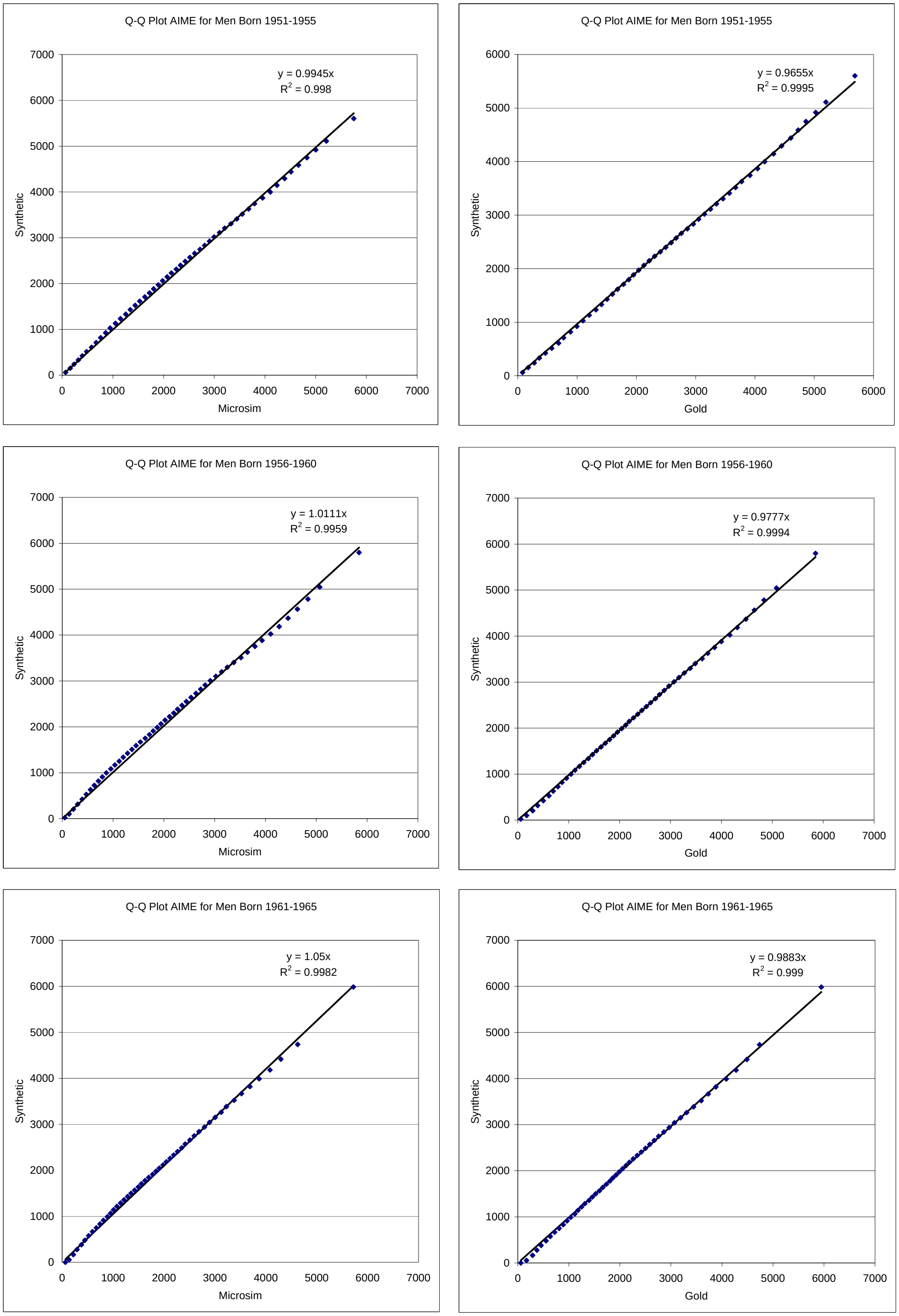
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-22a: Q-Q Plots for AIMEs for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



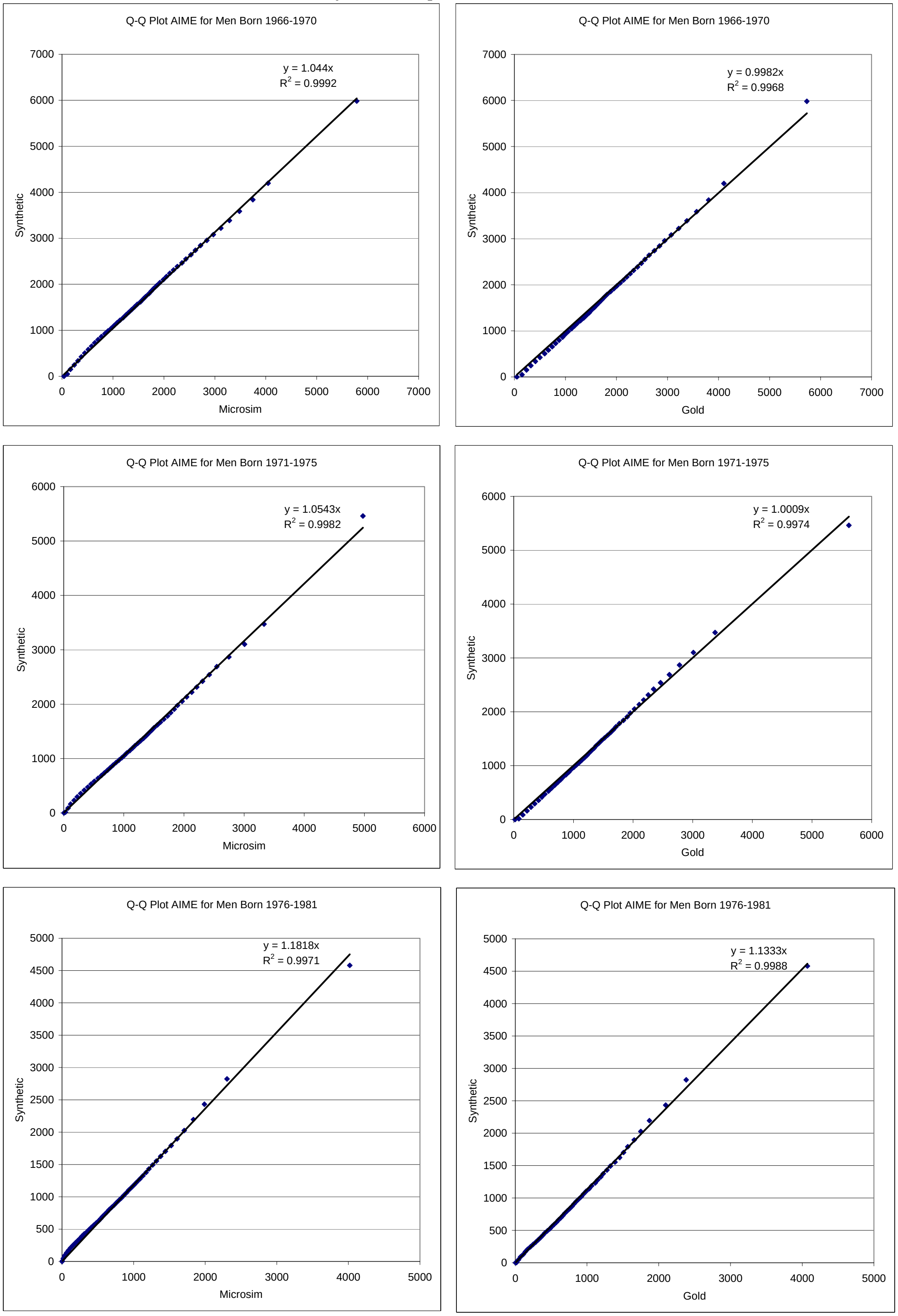
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-22a: Q-Q Plots for AIMEs for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



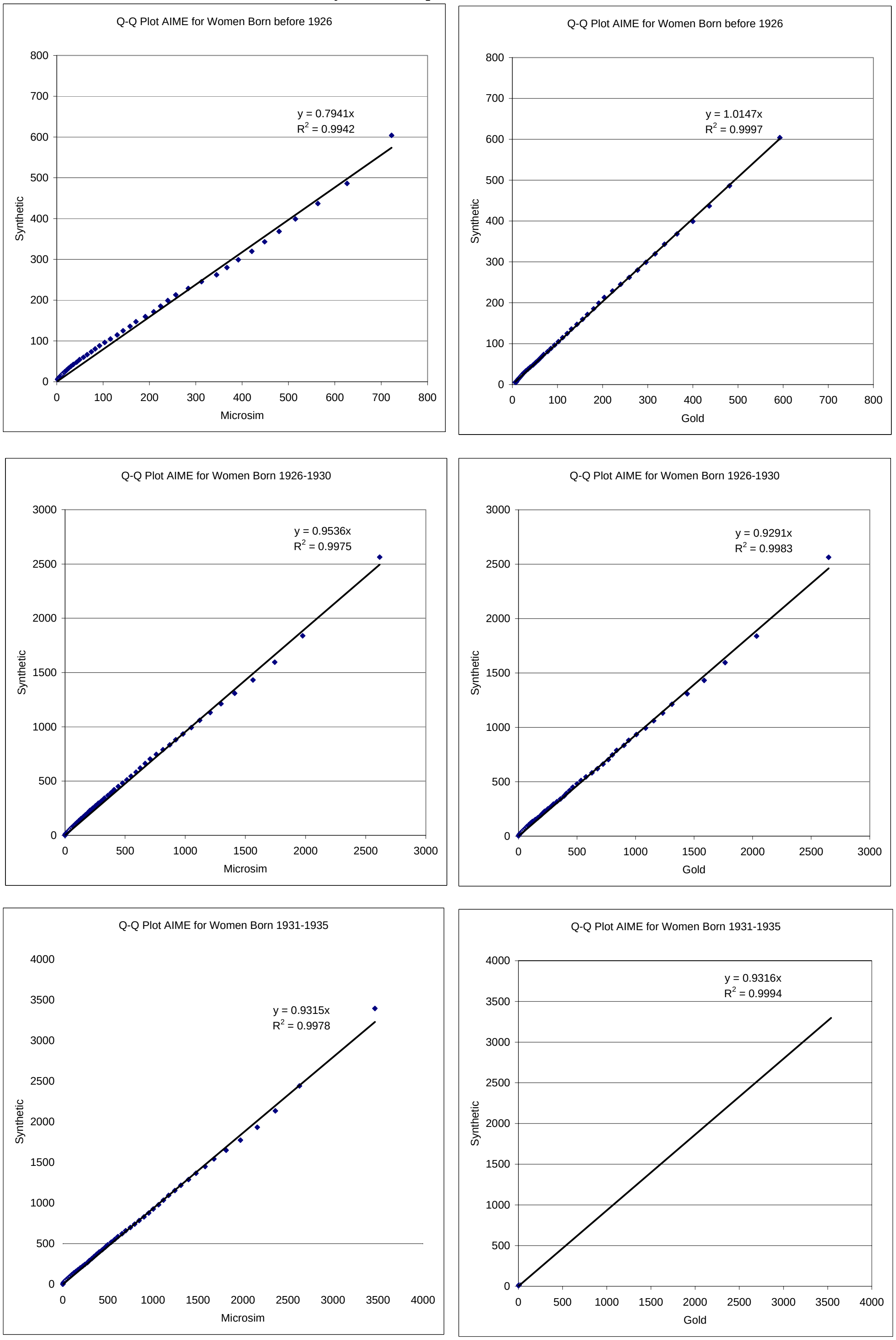
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-22a: Q-Q Plots for AIMEs for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



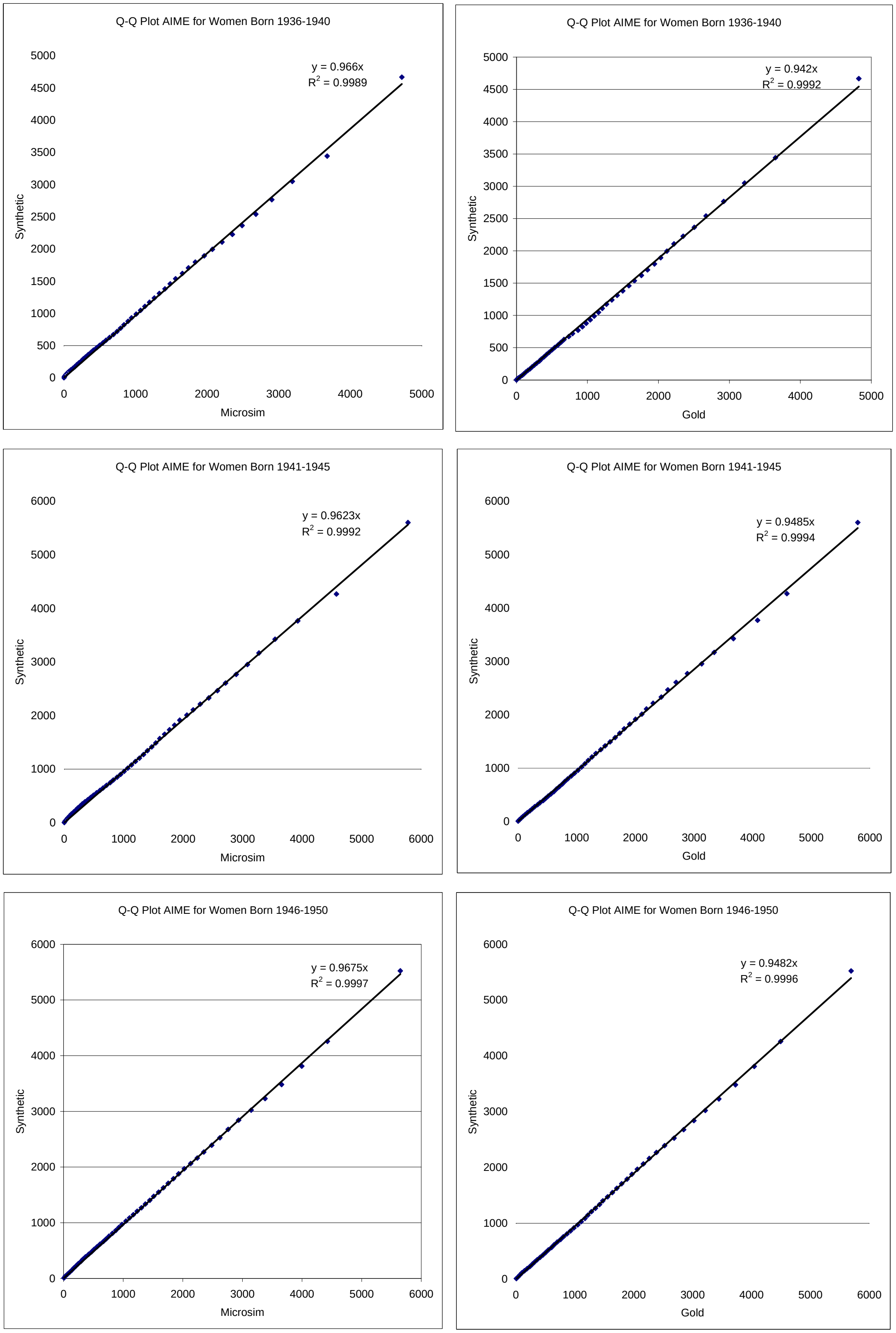
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-22b: Q-Q Plots for AIMEs for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold



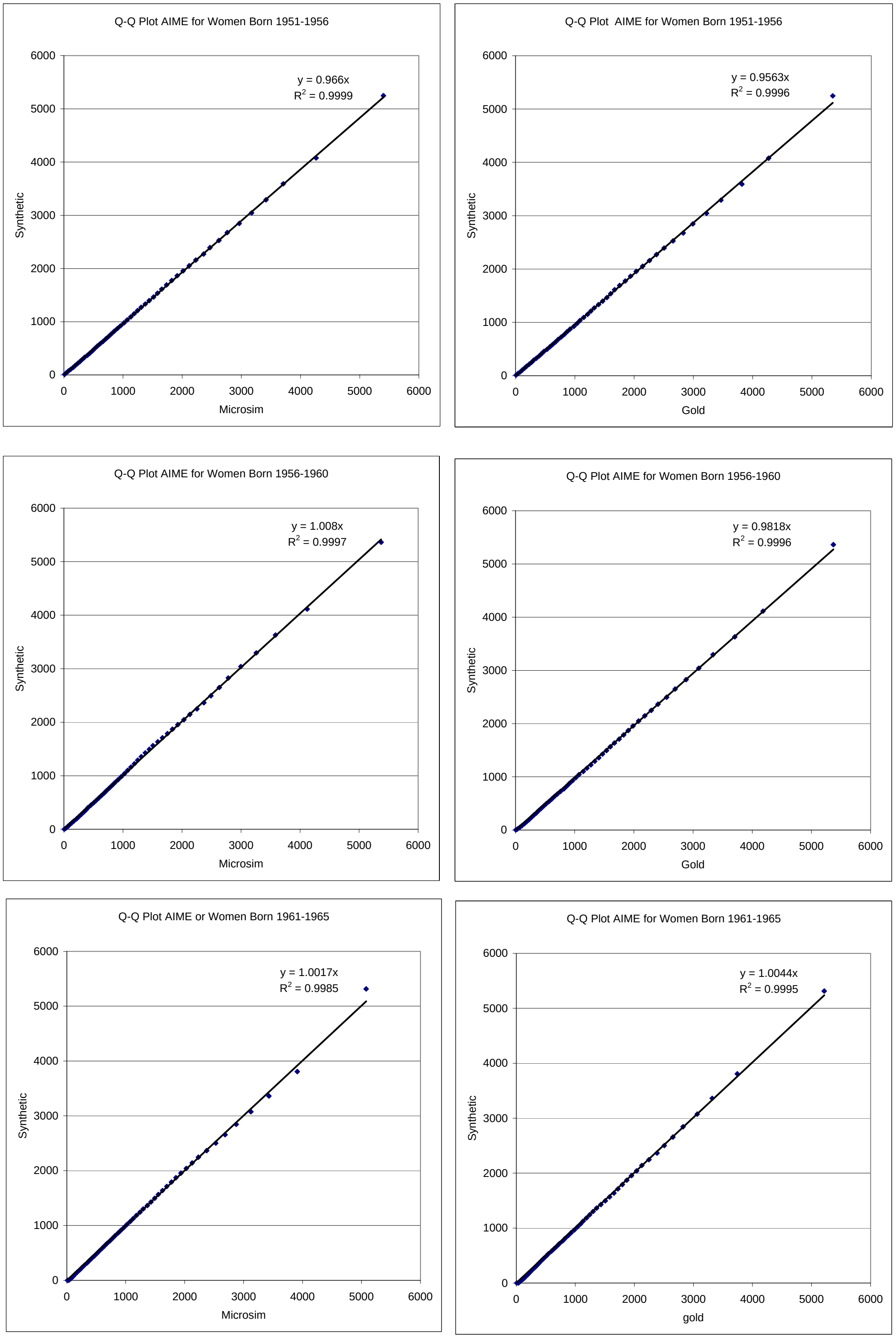
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-22b: Q-Q Plots for AIMEs for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold



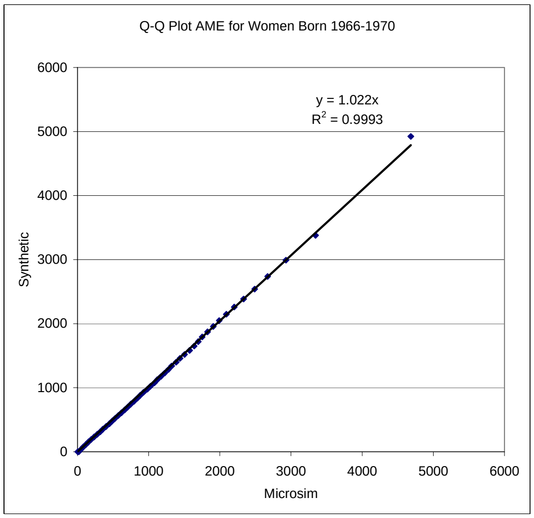
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

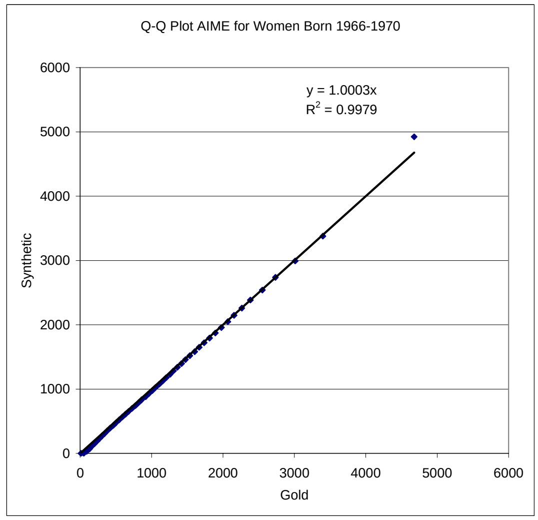
Figure 5-22b: Q-Q Plots for AIMEs for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold

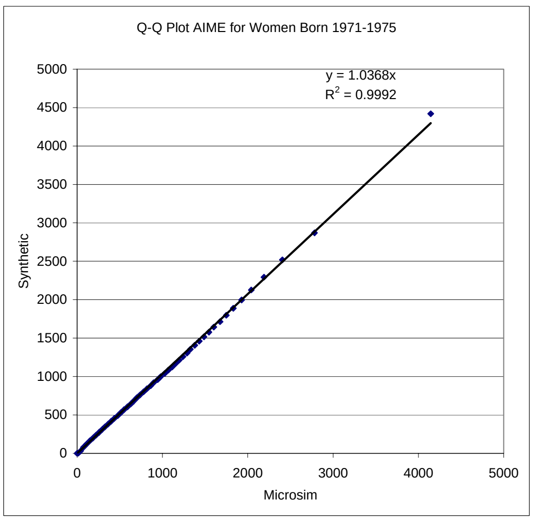


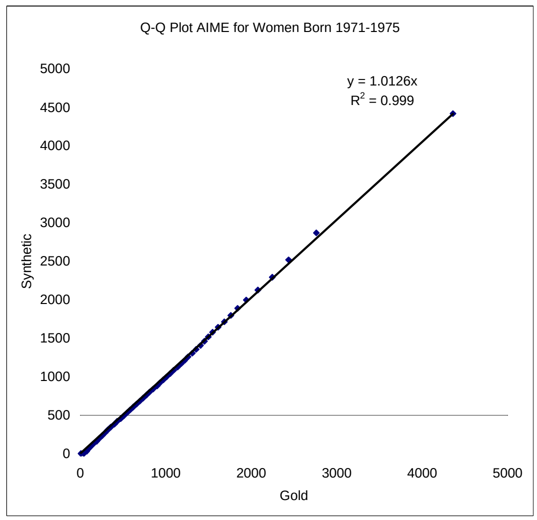
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

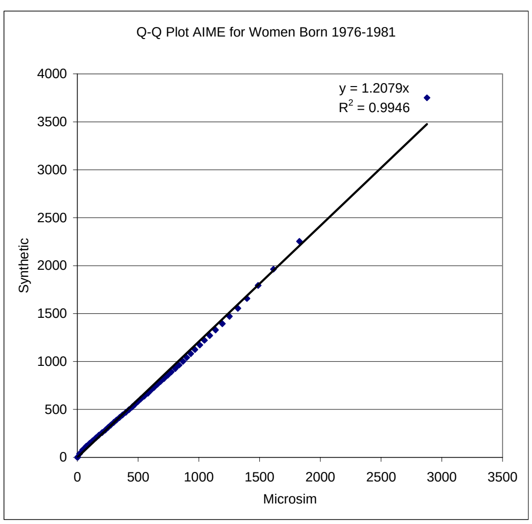
Figure 5-22b: Q-Q Plots for AIMEs for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold

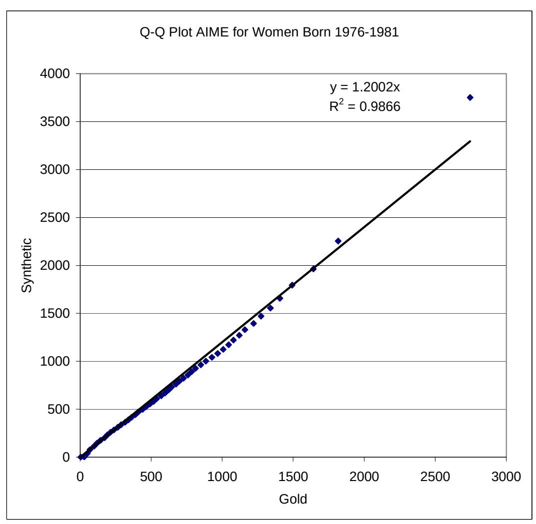






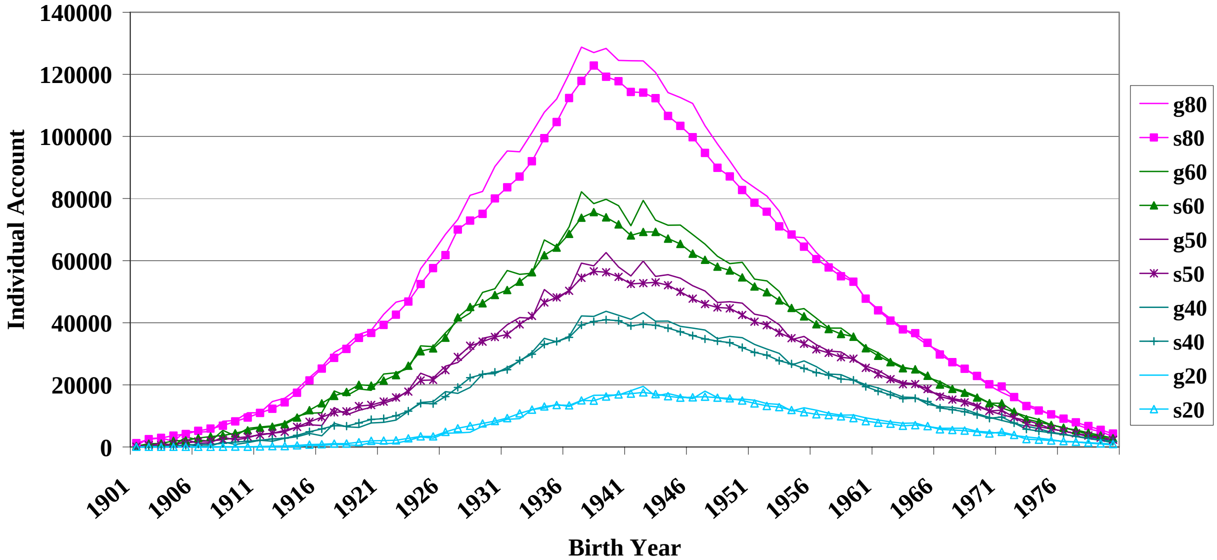






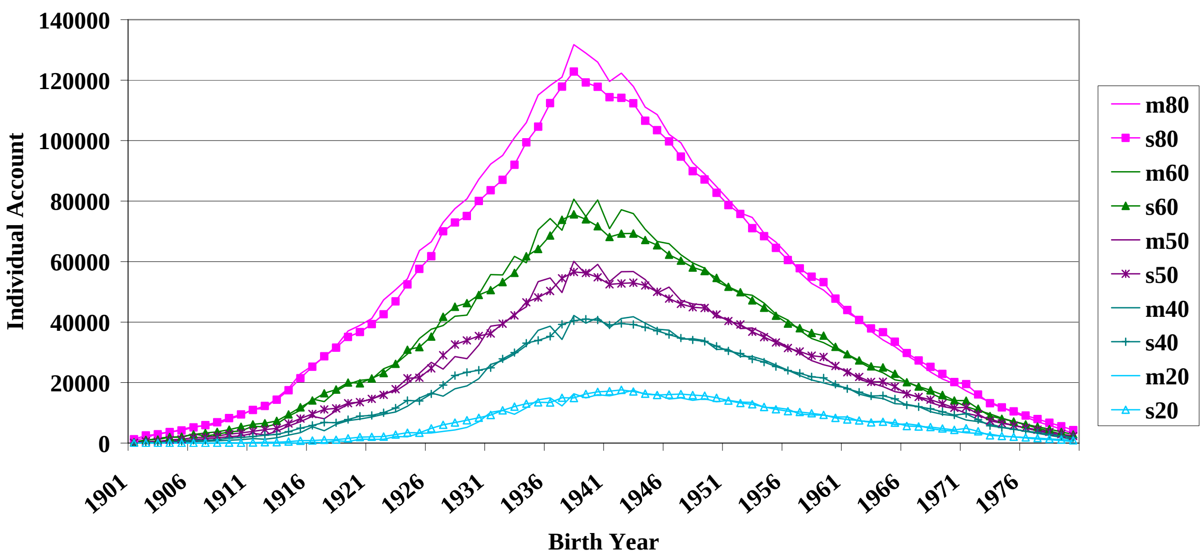
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born in the identified birth years that survive to 2000. AIME is the average of the top 35 years of wage-indexed taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-23a. Selected Percentiles of Individual Account by Birth Year
Synthetic Versus Gold Among Survivors to 2000



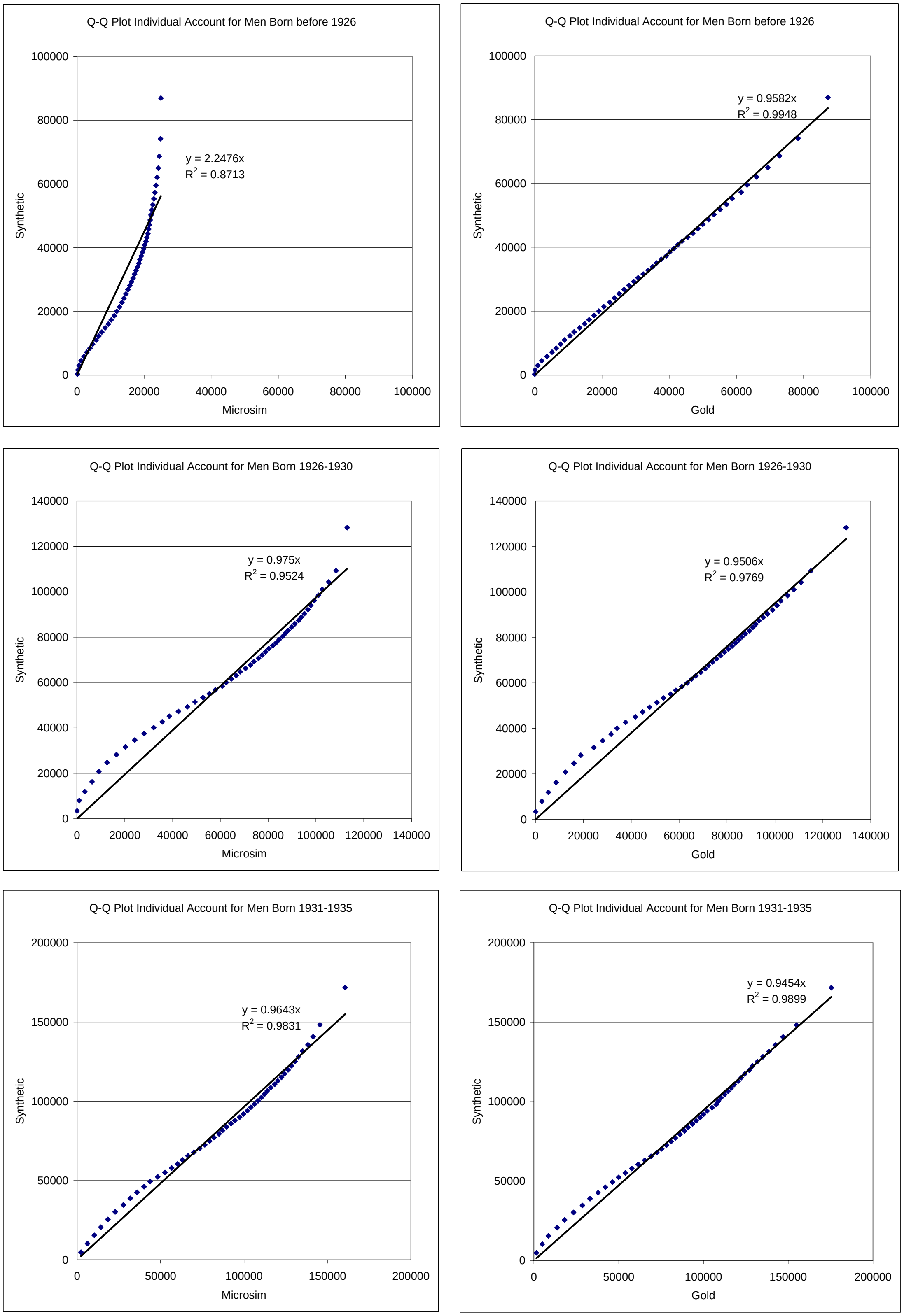
Source. Urban Institute tabulations of Gold, and pooled Synthetic SIPP files. Data are limited to U.S. residents that survive to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return. Prefixs *g* and *s* in the key indicate Gold and Synthetic file values respectively.

Figure 5-23b. Selected Percentiles of Individual Account by Birth Year
Synthetic Versus Microsim Among Survivors to 2000



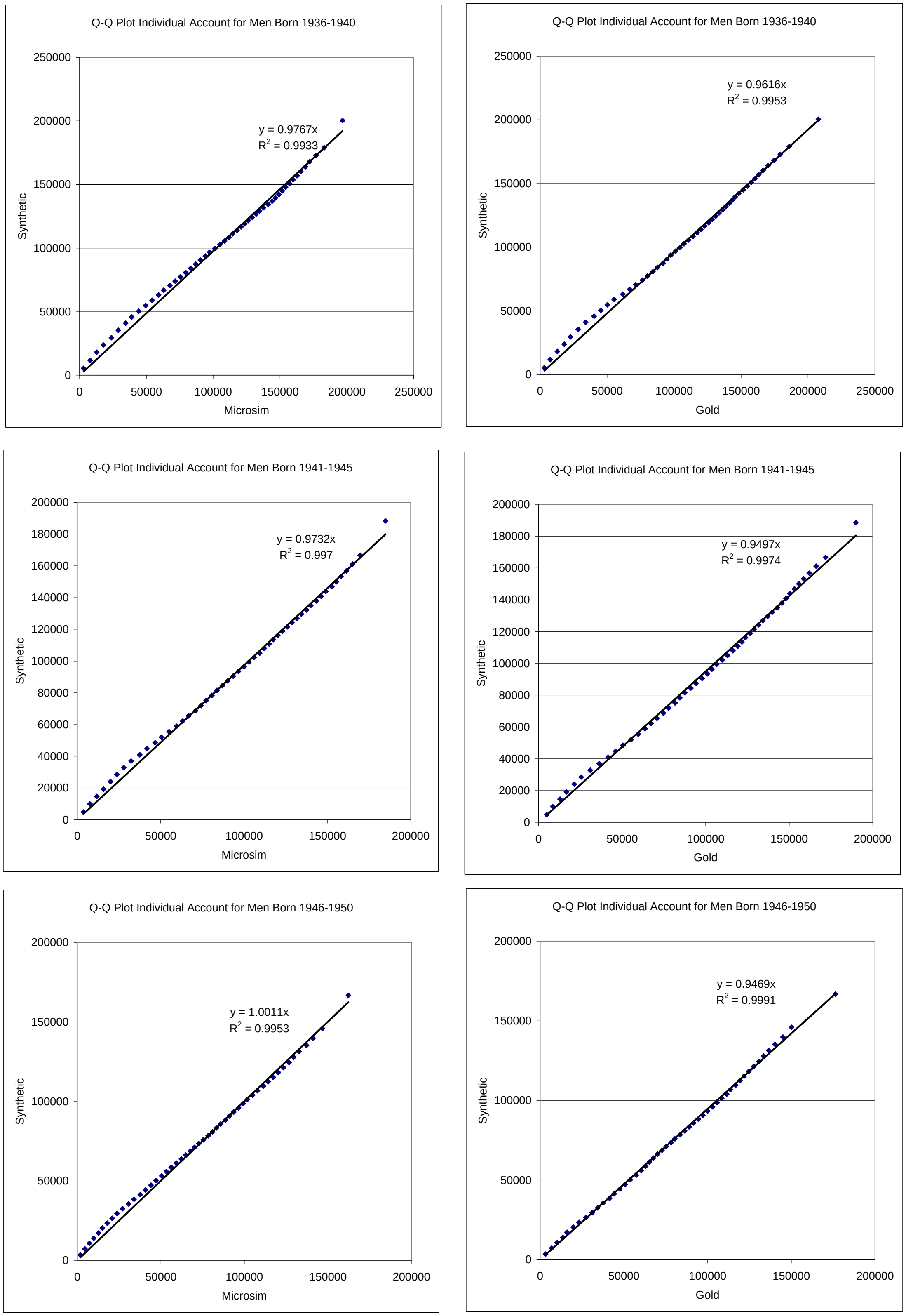
Source. Urban Institute tabulations of Microsim, and pooled Synthetic SIPP files. Data are limited to U.S. residents that survive to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return. Prefixs *m* and *s* in the key indicate Microsim and Synthetic file values respectively.

Figure 5-24a: Q-Q Plots for Individual Account Balances for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



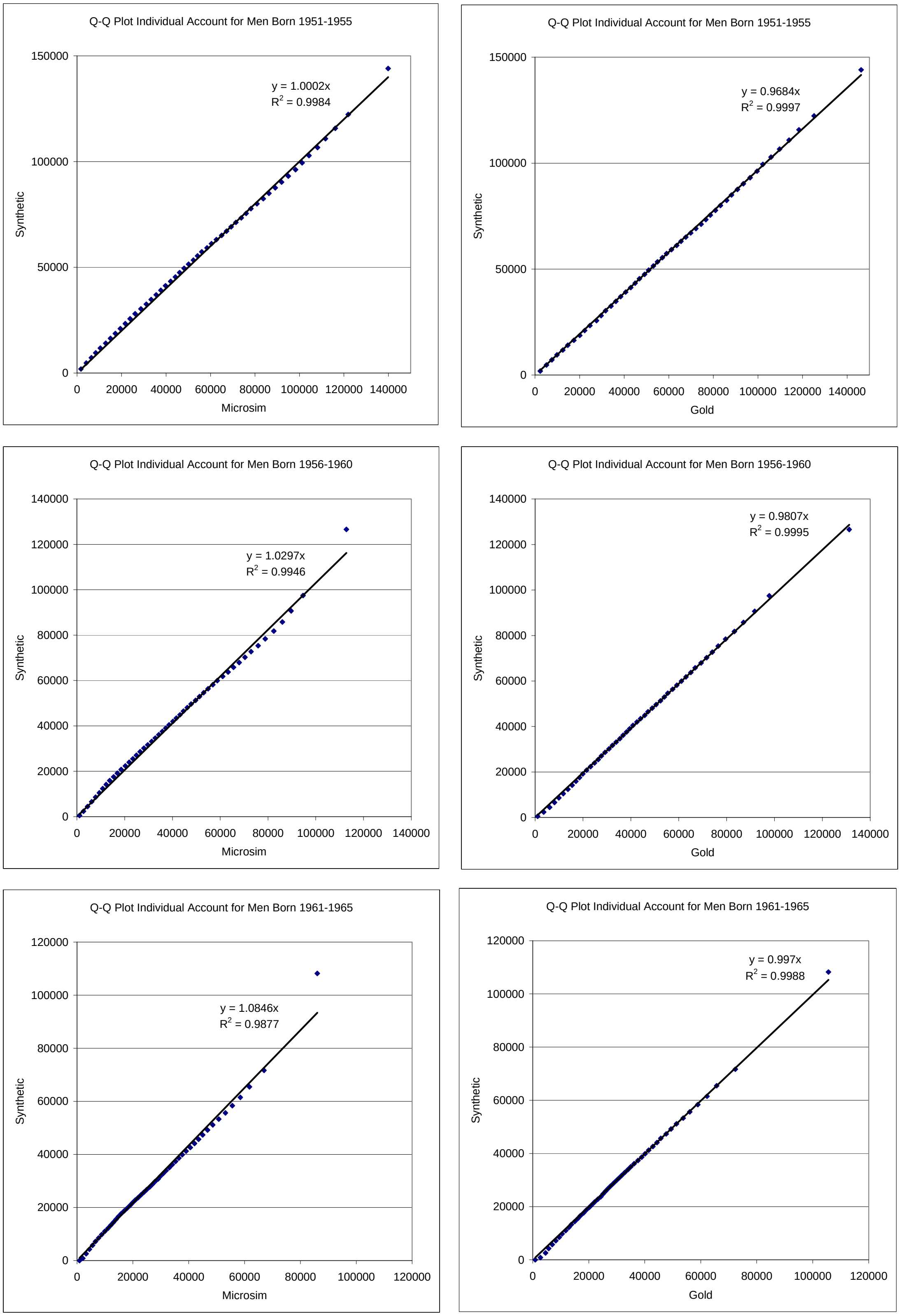
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident male survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

Figure 5-24a: Q-Q Plots for Individual Account Balances for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



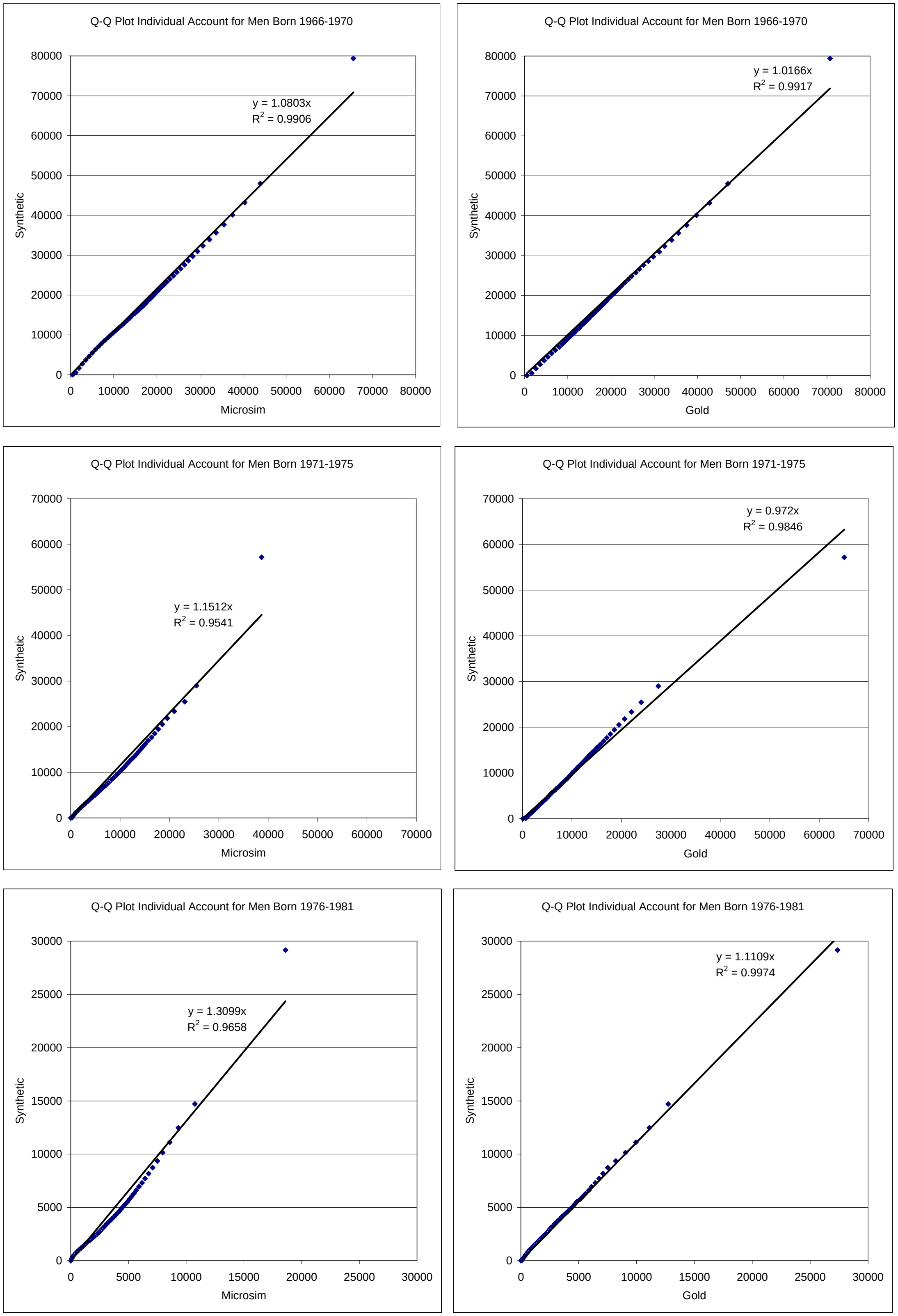
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident male survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

**Figure 5-24a: Q-Q Plots for Individual Account Balances for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold**



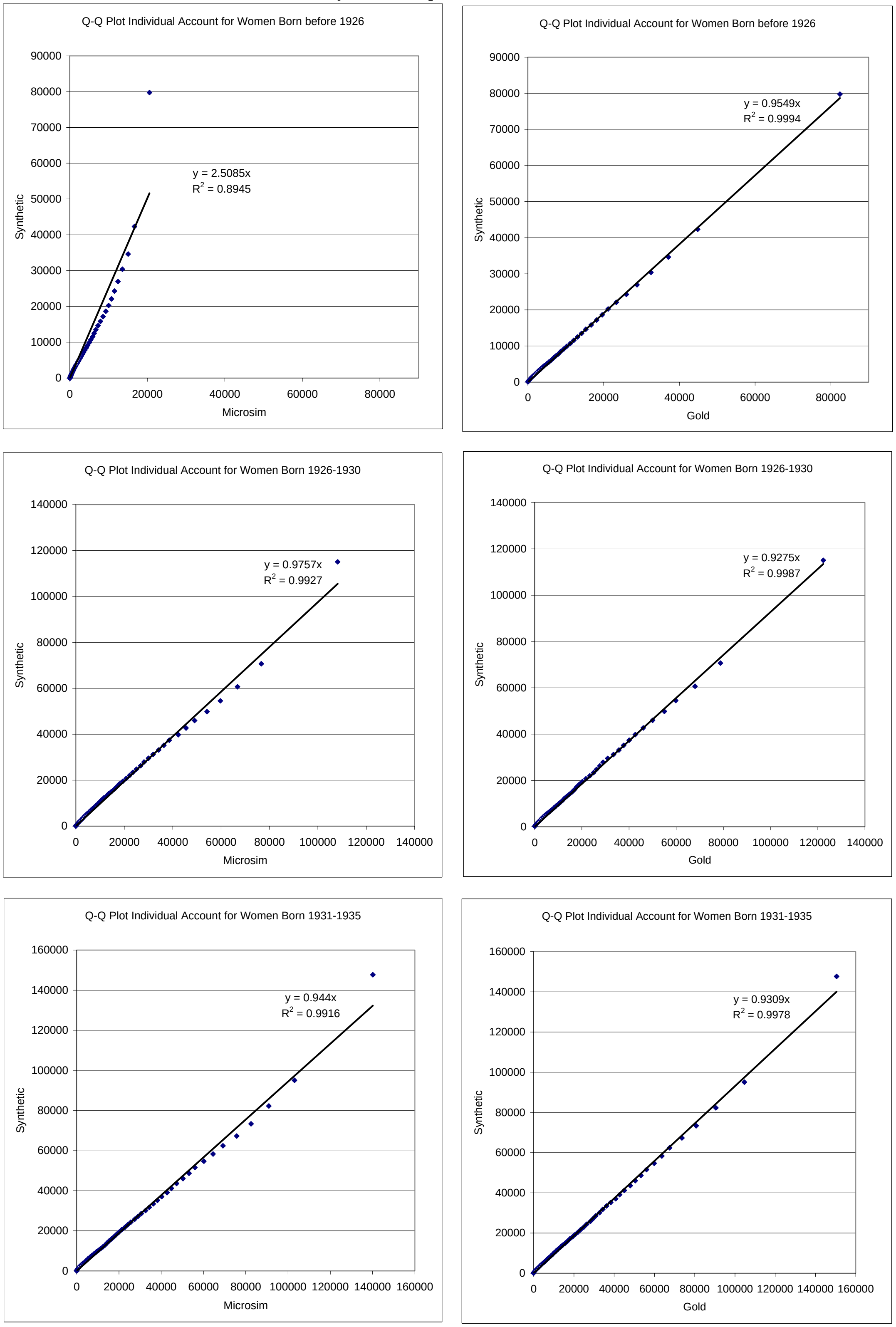
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident male survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

Figure 5-24a: Q-Q Plots for Individual Account Balances for Men by Birth Year
Pooled Synthetic compared to Microsim and Gold



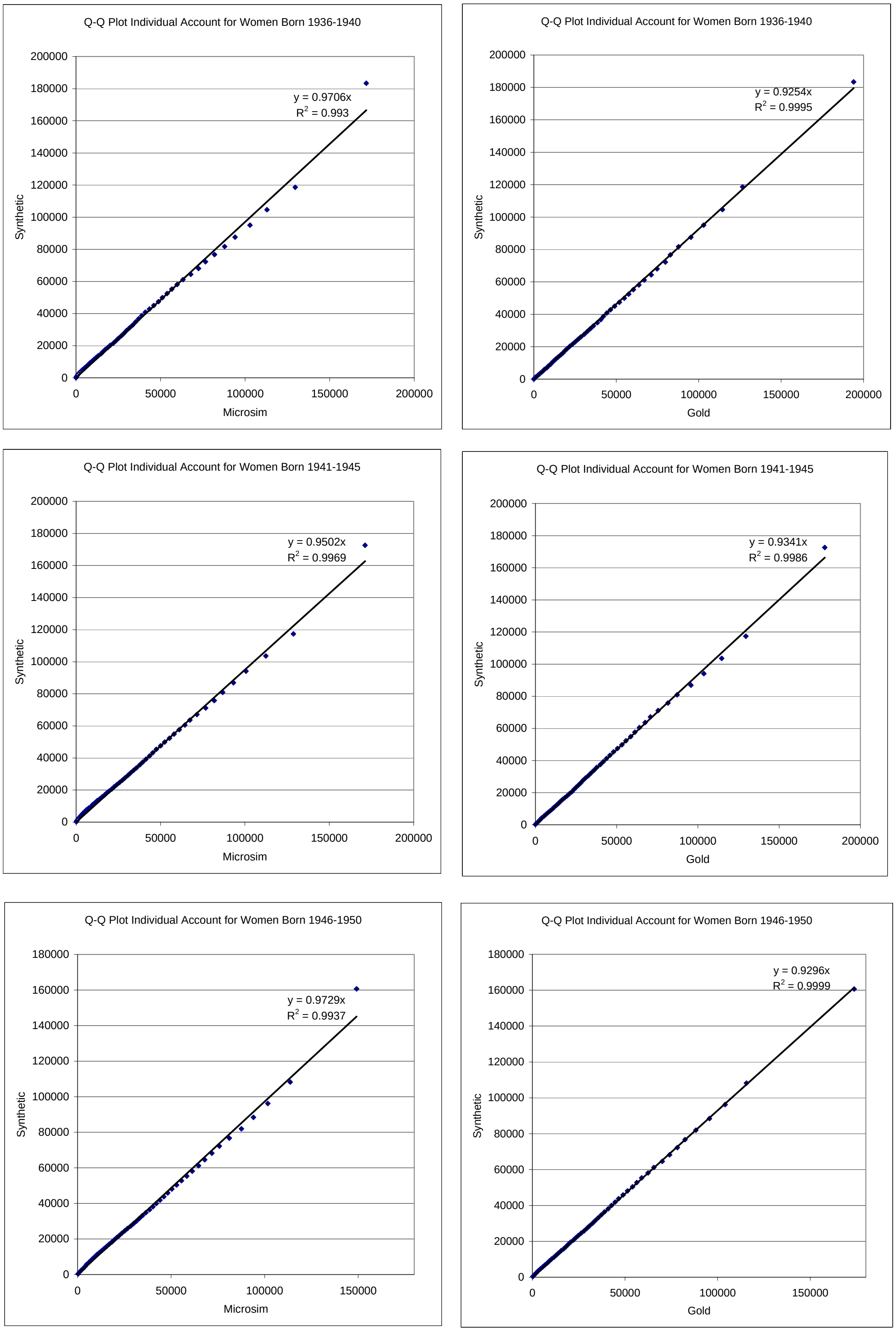
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident male survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

Figure 5-24b: Q-Q Plots for Individual Account Balances for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold



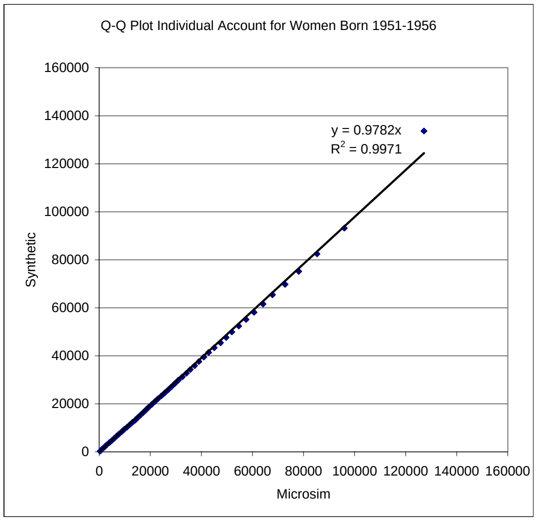
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident female survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

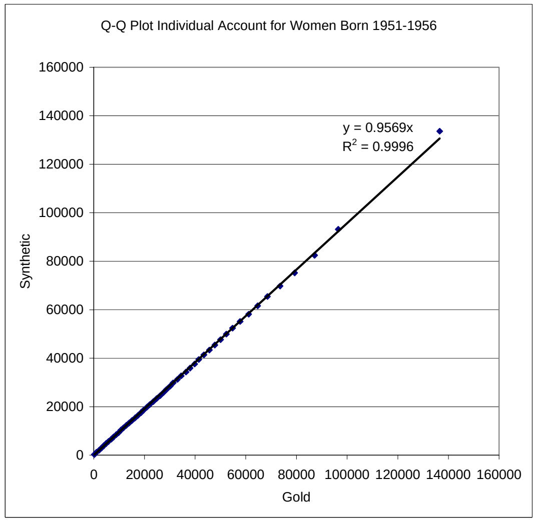
Figure 5-24b: Q-Q Plots for Individual Account Balances for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold

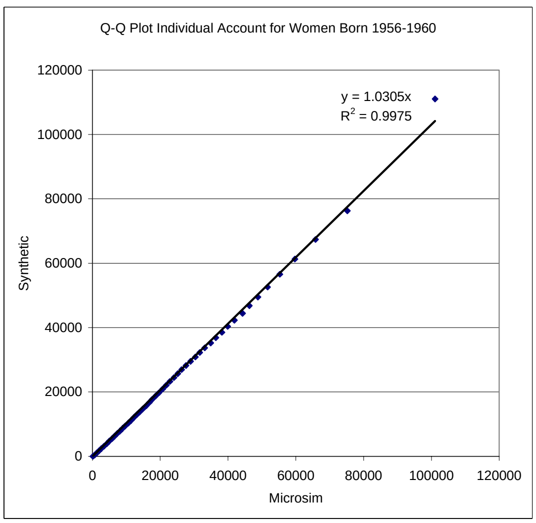


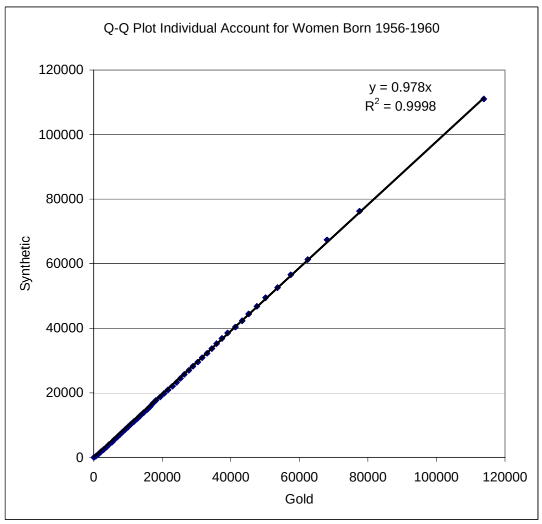
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident female survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

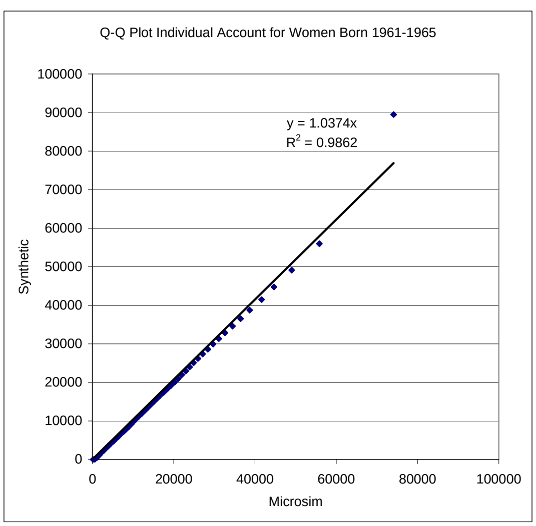
Pooled Synthetic compared to Microsim and Gold

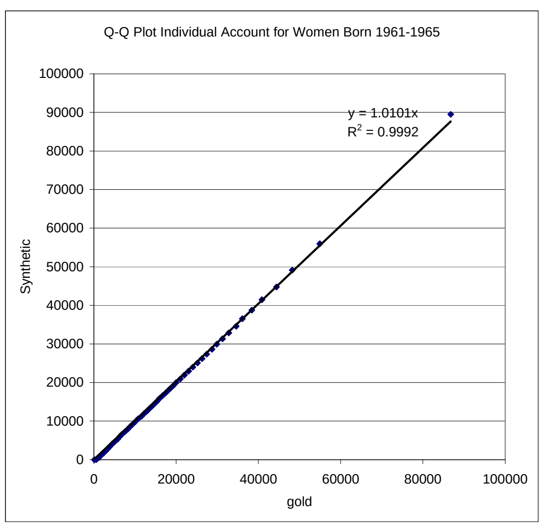






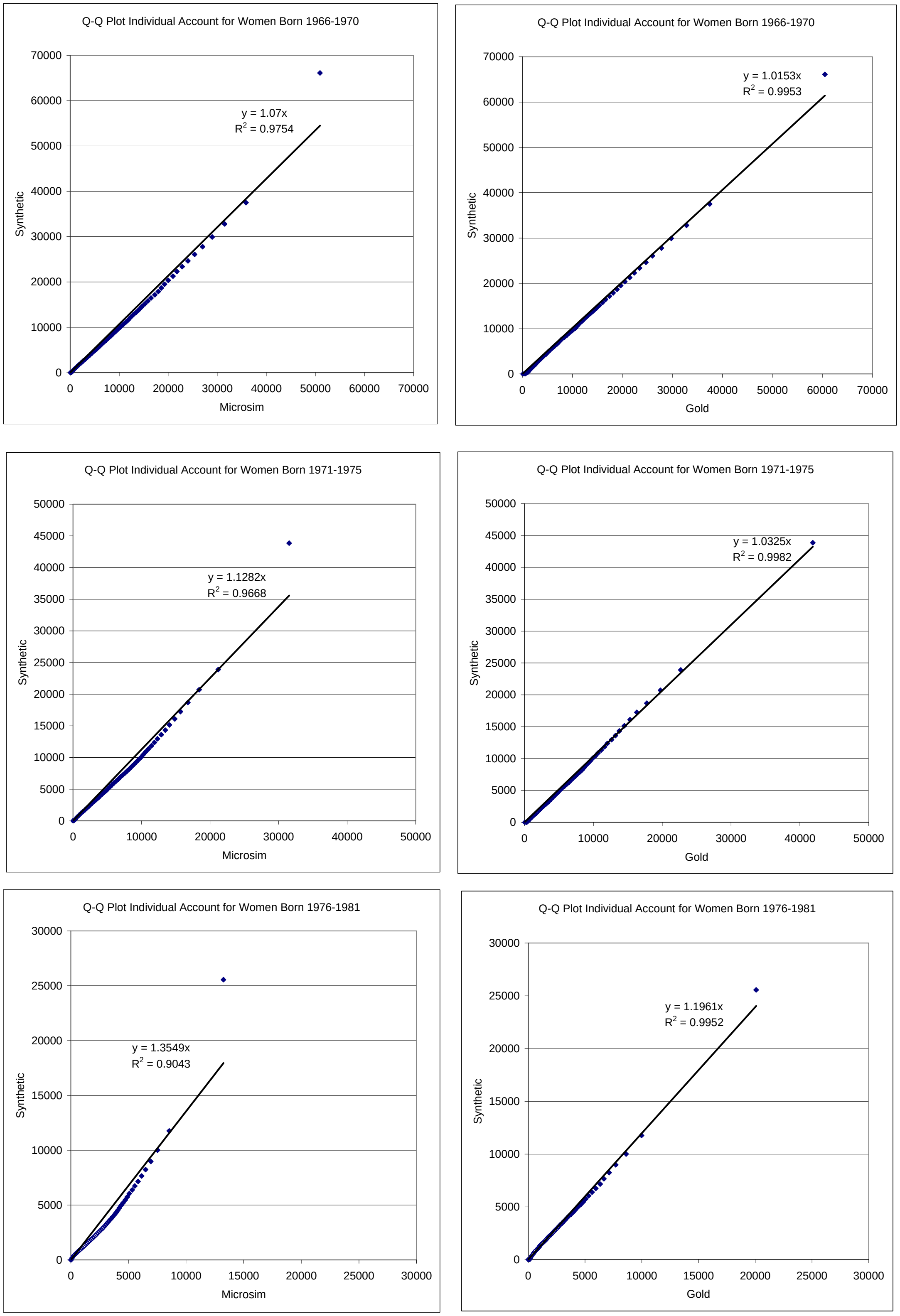






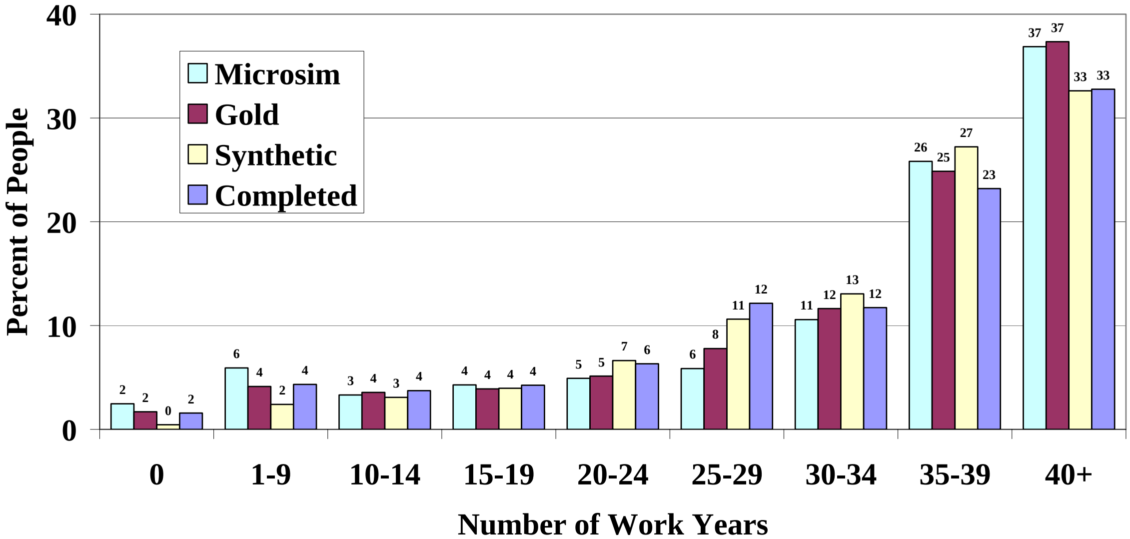
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident female survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

Figure 5-24b: Q-Q Plots for Individual Account Balances for Women by Birth Year
Pooled Synthetic compared to Microsim and Gold



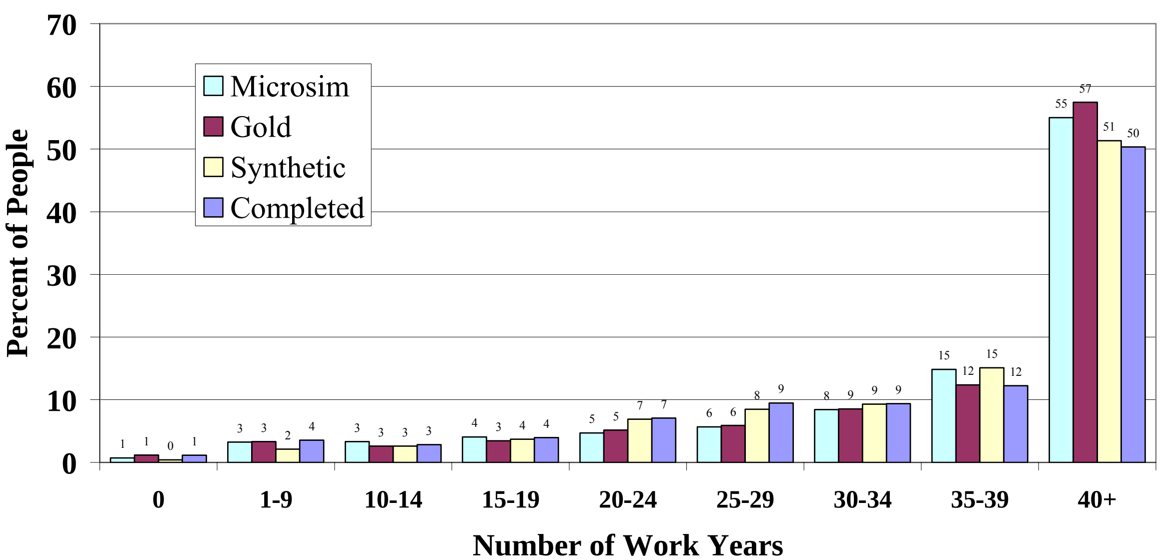
Source: Urban Institute tabulations of Pooled Synthetic and Microsim data. Data are limited to U.S. resident female survivors to 2000. Individual accounts are based on the investment value of a 4 percent contribution of taxable wages from age 16 to 62 between 1951 and 2003. Balances earn a 3.3 percent annual real rate of return.

Figure 5-25a. Distribution of Total Covered Work Years for Men Born 1926 - 1930, Survivors to 2000



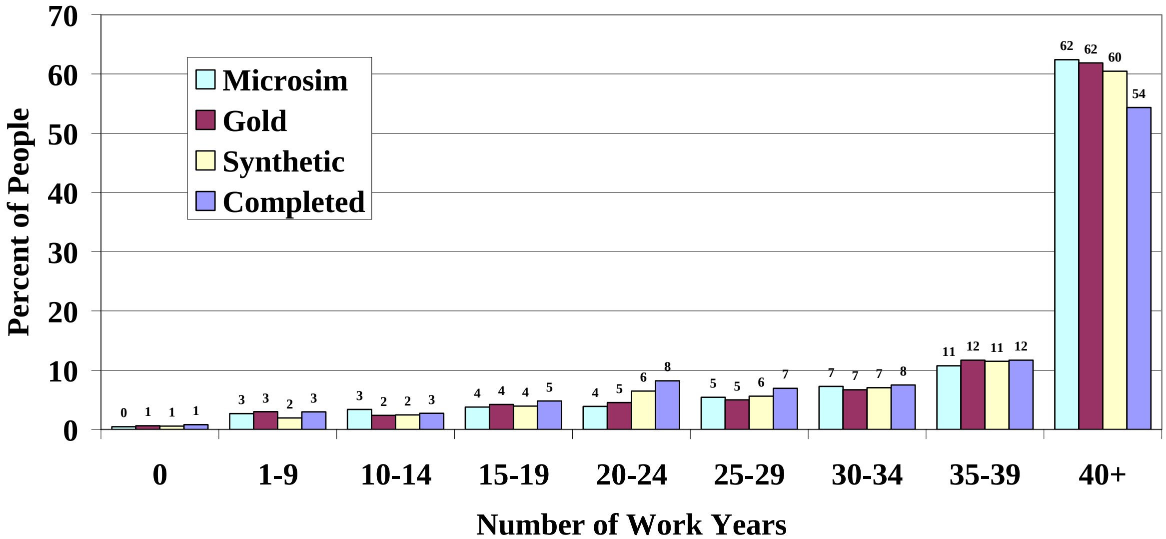
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1926 to 1930 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25b. Distribution of Total Covered Work Years for Men Born 1931 - 1935, Survivors to 2000



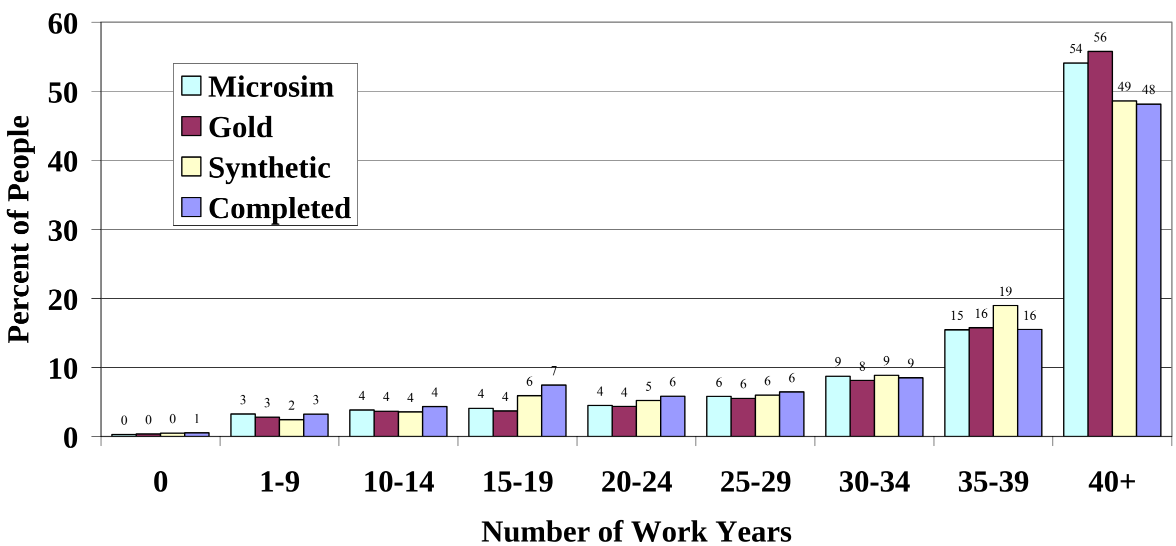
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1931 to 1935 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25c. Distribution of Total Covered Work Years for Men Born 1936 - 1940, Survivors to 2000



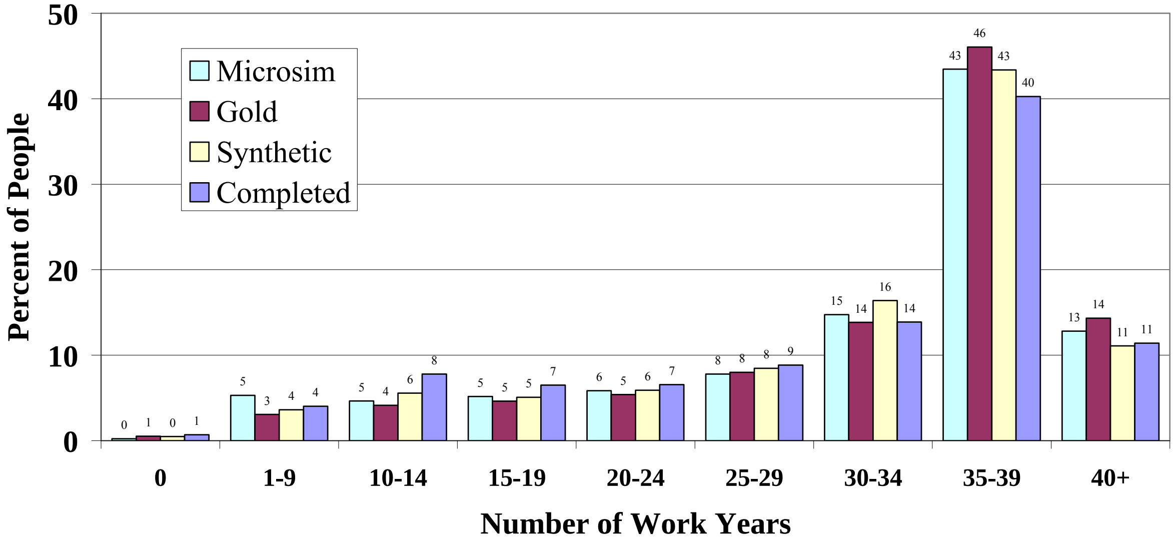
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1936 to 1940 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25d. Distribution of Total Covered Work Years for Men Born 1941 - 1945, Survivors to 2000



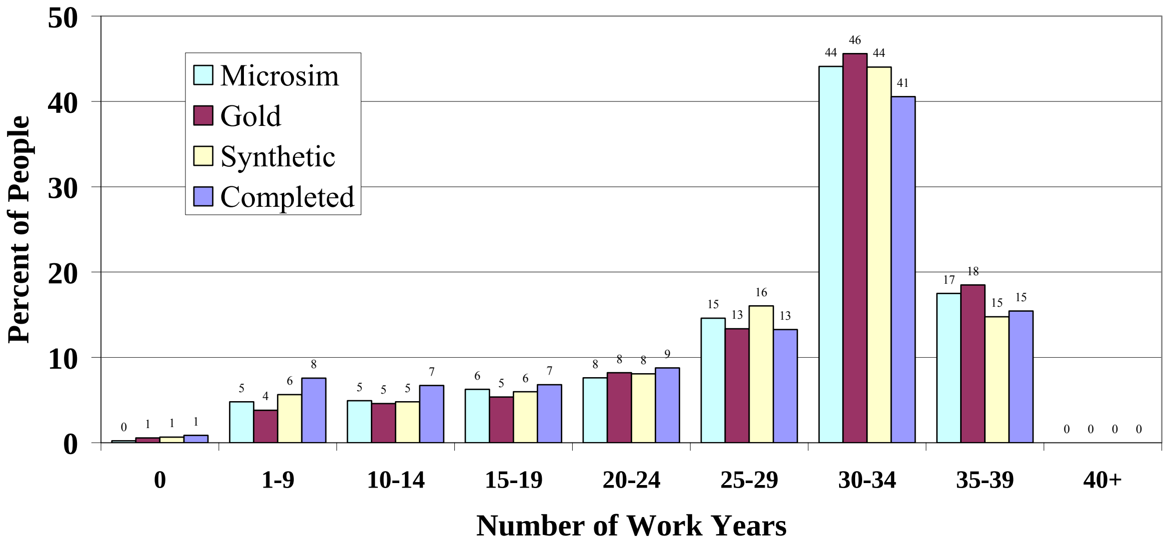
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1941 to 1945 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25e. Distribution of Total Covered Work Years for Men Born 1946 - 1950, Survivors to 2000



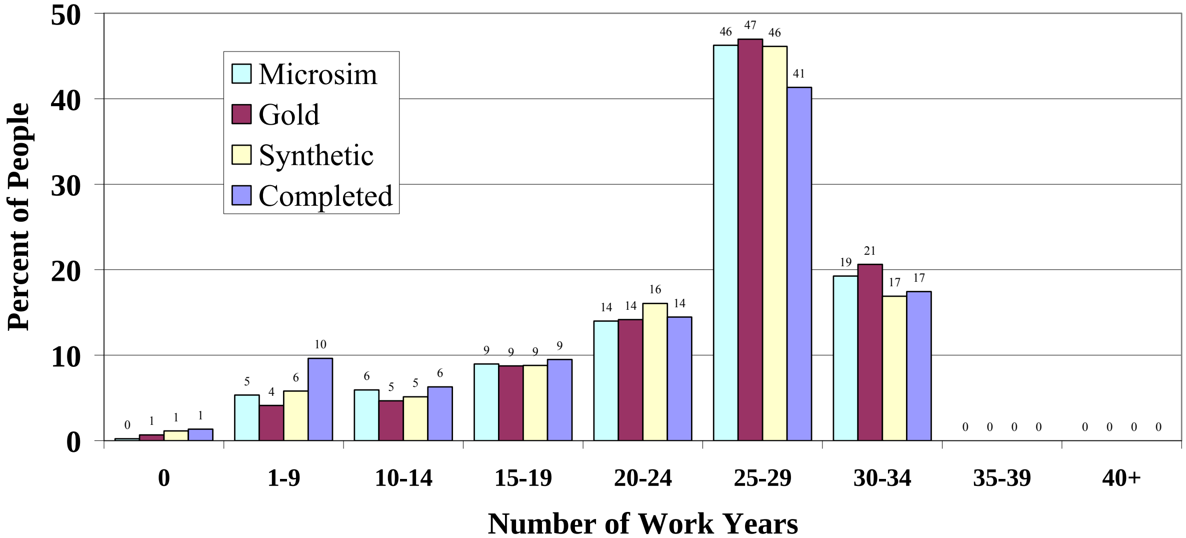
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1946 to 1950 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25f. Distribution of Total Covered Work Years for Men Born 1951 - 1955, Survivors to 2000



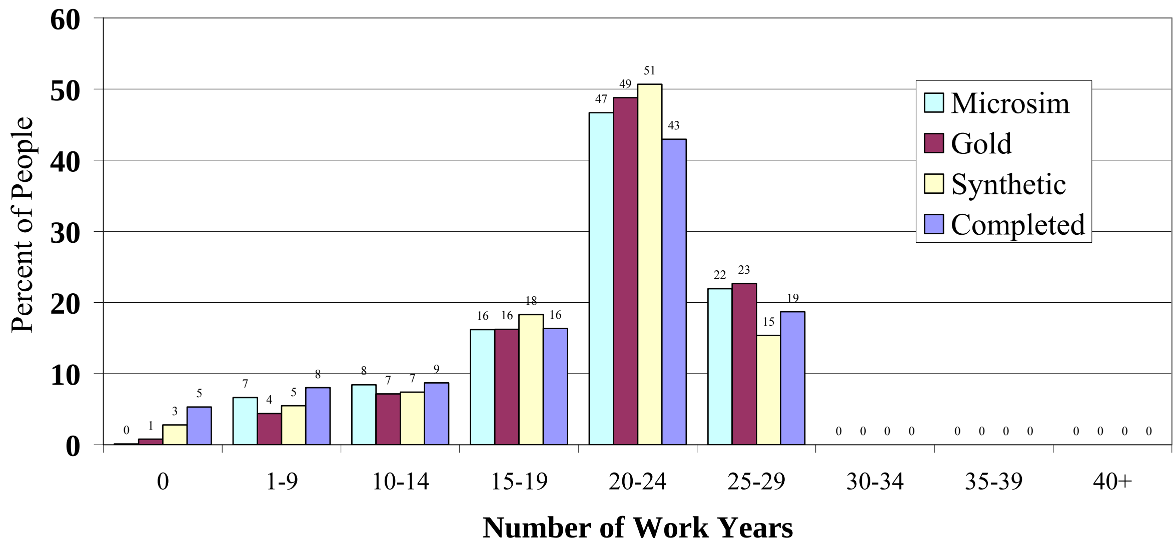
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1951 to 1955 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25g. Distribution of Total Covered Work Years for Men Born 1956 - 1960, Survivors to 2000



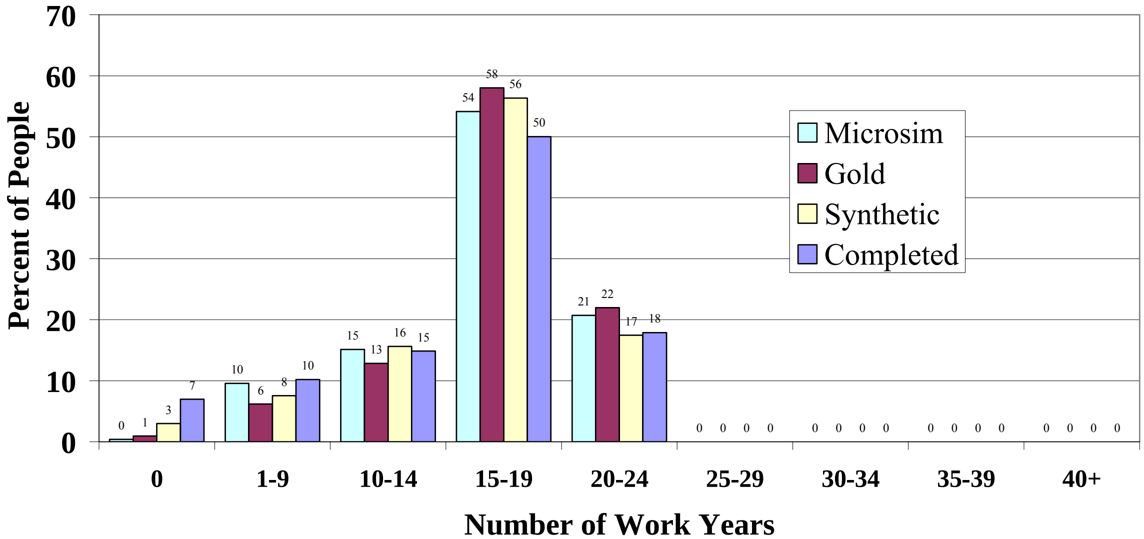
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1956 to 1960 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25h. Distribution of Total Covered Work Years for Men Born 1961 - 1965, Survivors to 2000



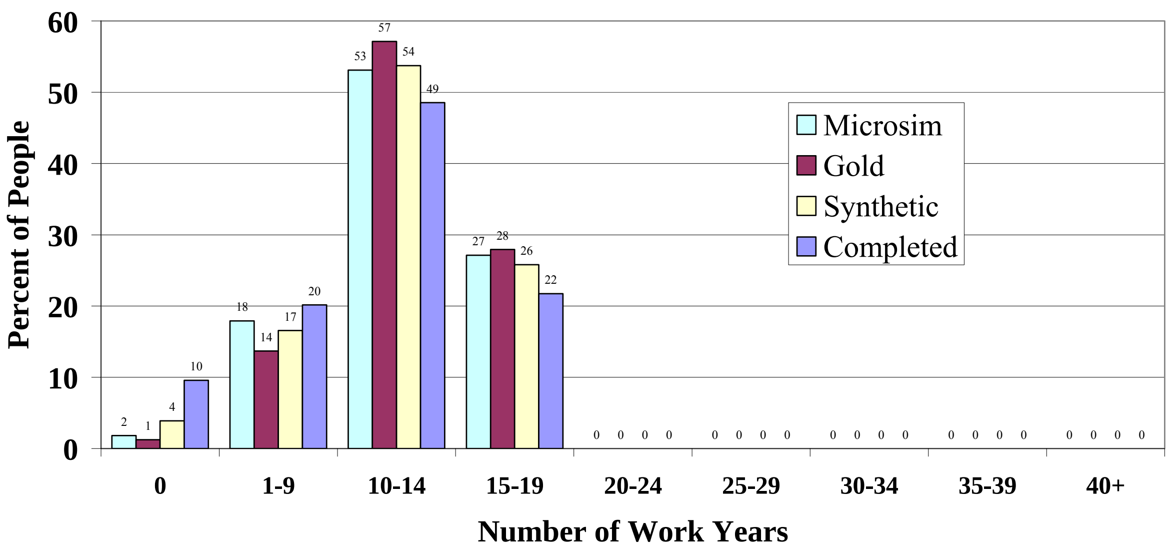
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1961 to 1965 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25i. Distribution of Total Covered Work Years for Men Born 1966 - 1970, Survivors to 2000



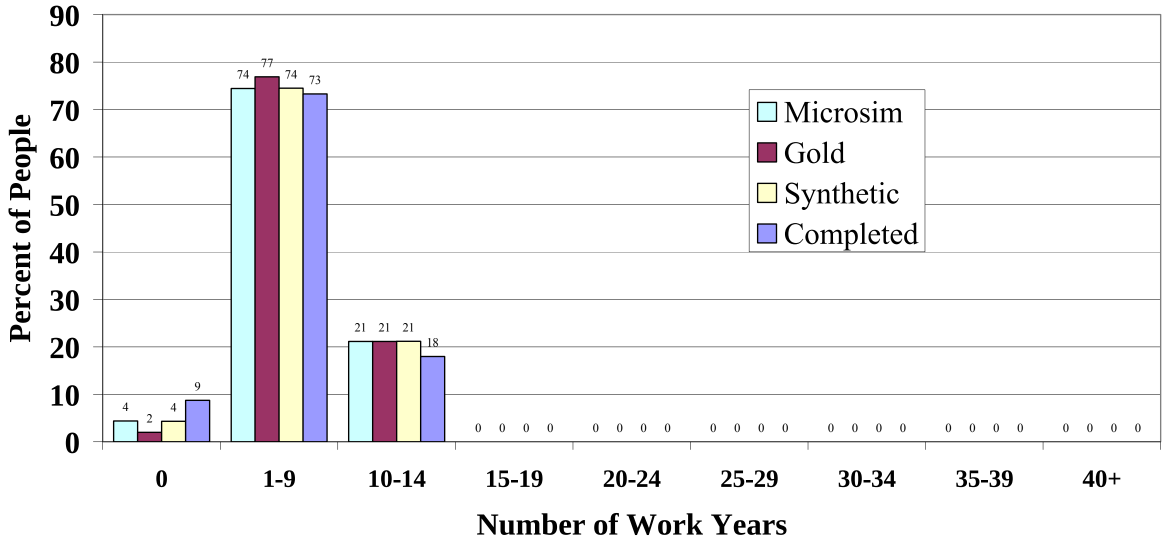
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1966 to 1970 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25j. Distribution of Total Covered Work Years for Men Born 1971 - 1975, Survivors to 2000



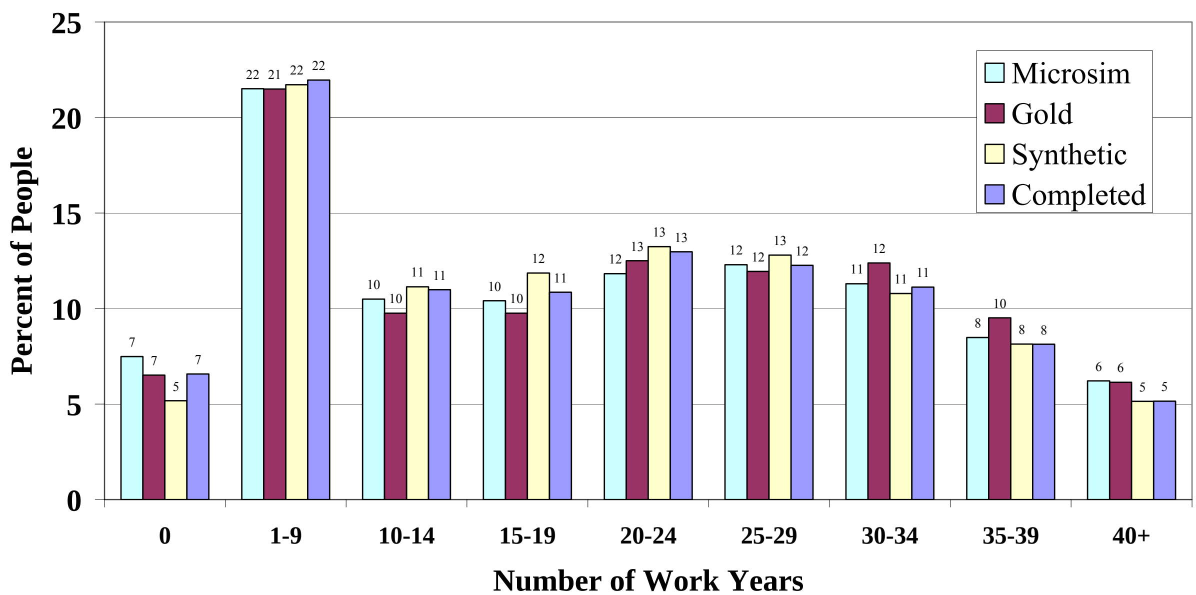
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1971 to 1975 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-25k. Distribution of Total Covered Work Years for Men Born 1976 - 1981, Survivors to 2000



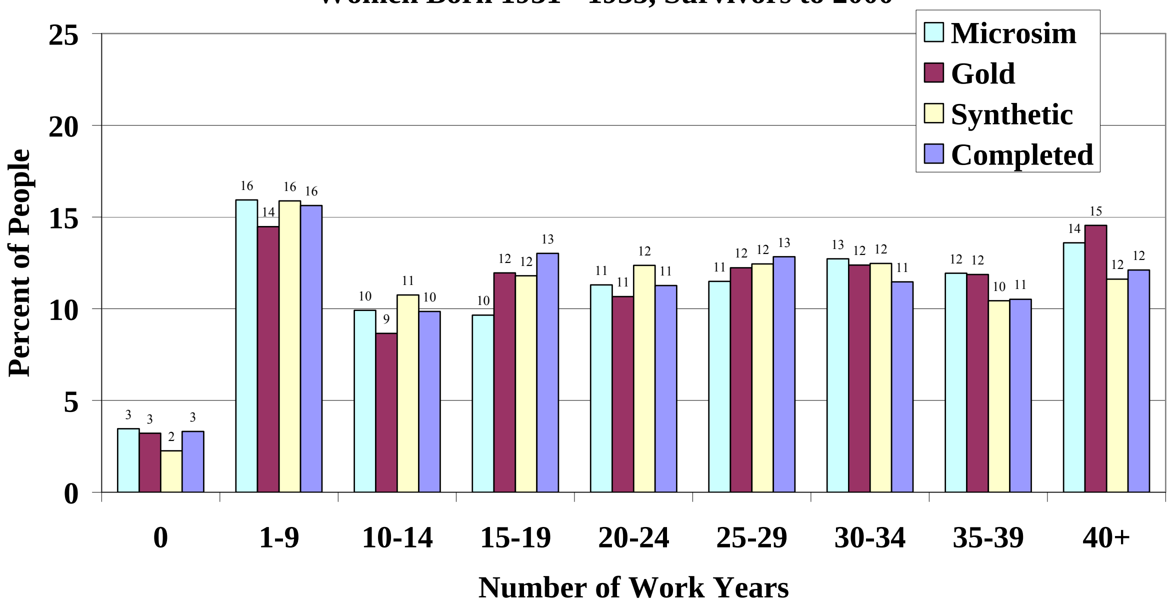
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident men born from 1976 to 1981 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26a. Distribution of Total Covered Work Years for Women Born 1926 - 1930, Survivors to 2000



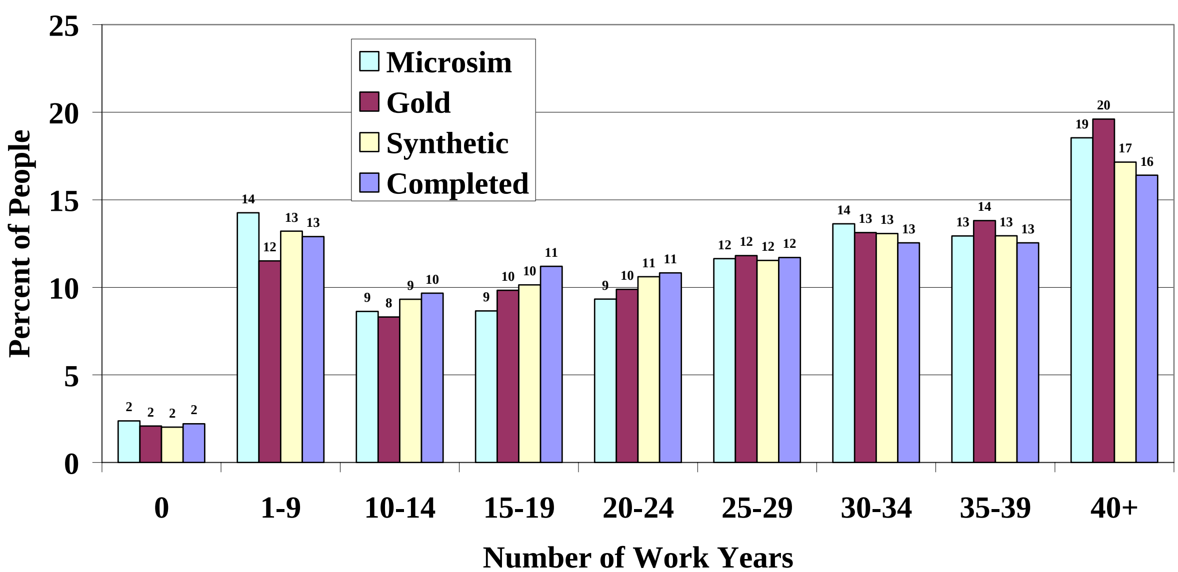
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1926 to 1930 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26b. Distribution of Total Covered Work Years for Women Born 1931 - 1935, Survivors to 2000



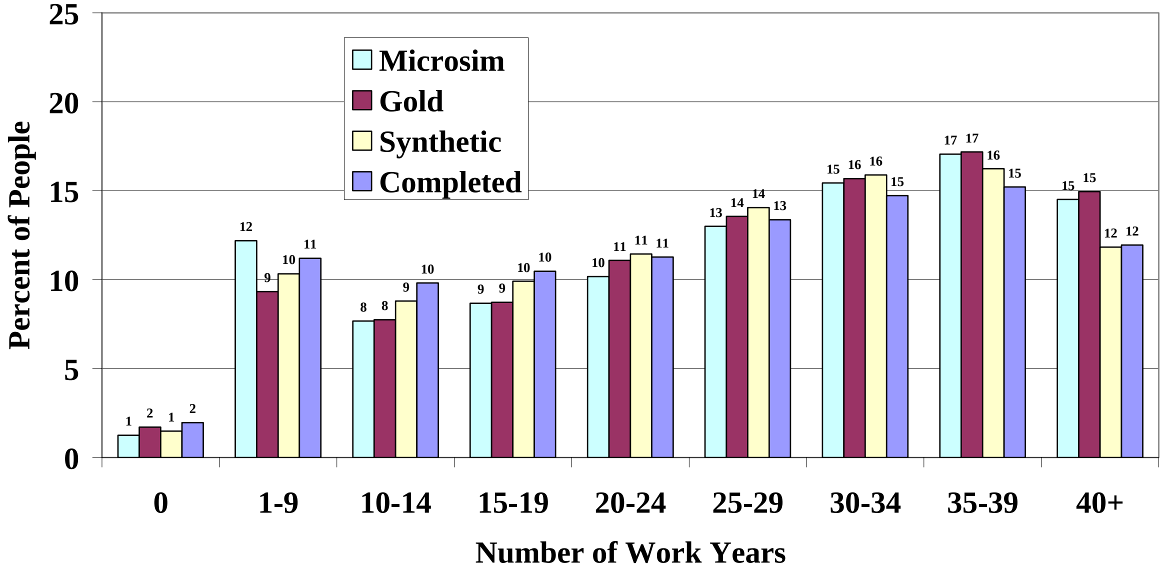
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1931 to 1935 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26c. Distribution of Total Covered Work Years for Women Born 1936 - 1940, Survivors to 2000



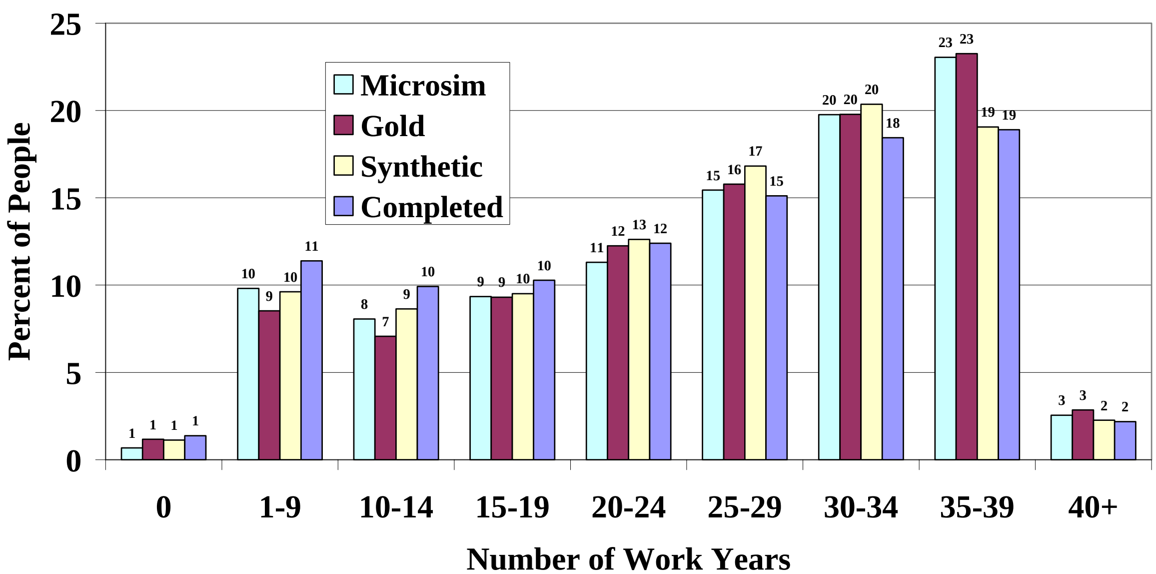
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1936 to 1940 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26d. Distribution of Total Covered Work Years for Women Born 1941 - 1945, Survivors to 2000



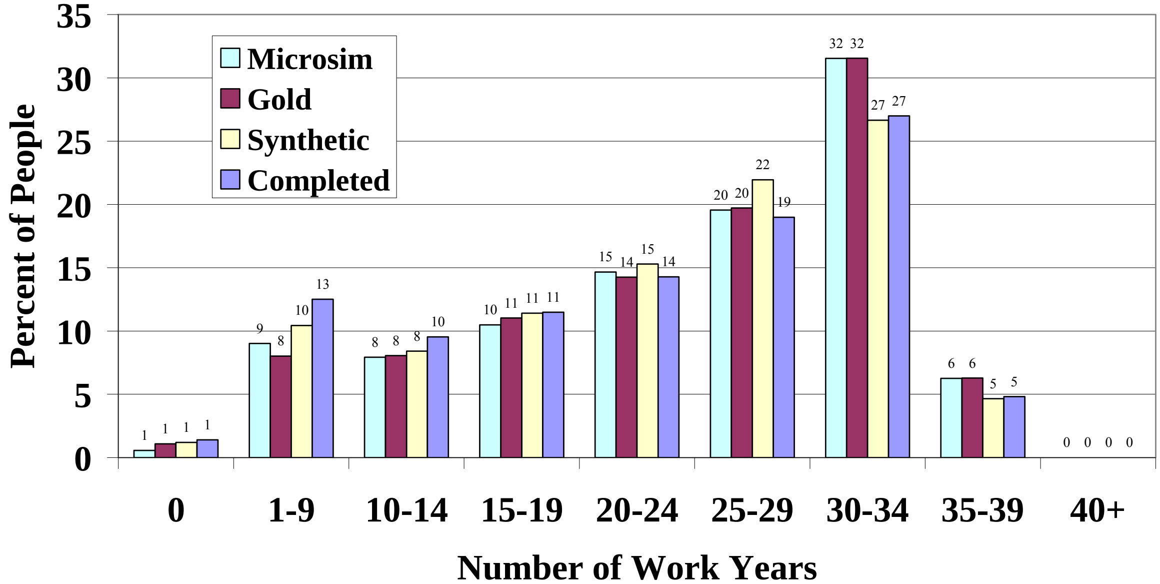
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1941 to 1945 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26e. Distribution of Total Covered Work Years for Women Born 1946 - 1950, Survivors to 2000



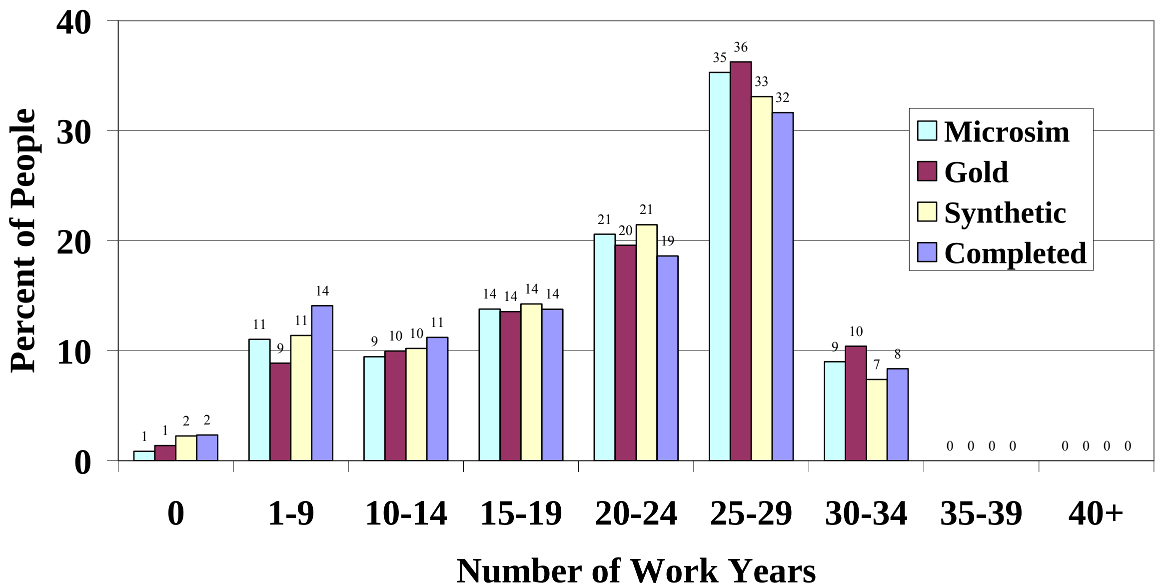
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1946 to 1950 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26f. Distribution of Total Covered Work Years for Women Born 1951 - 1955, Survivors to 2000



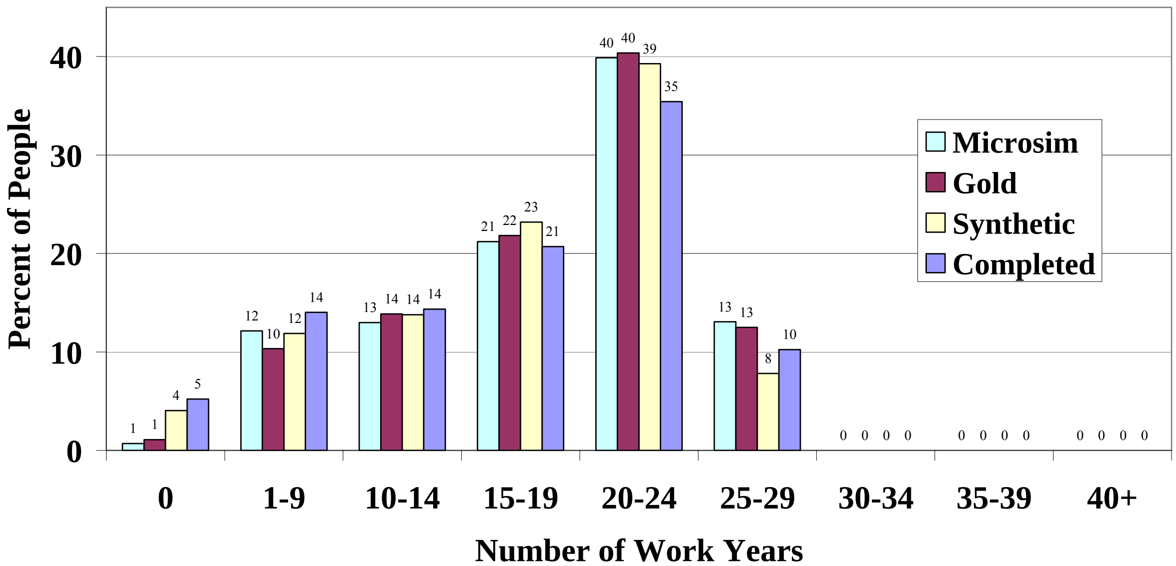
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1951 to 1955 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26g. Distribution of Total Covered Work Years for Women Born 1956 - 1960, Survivors to 2000



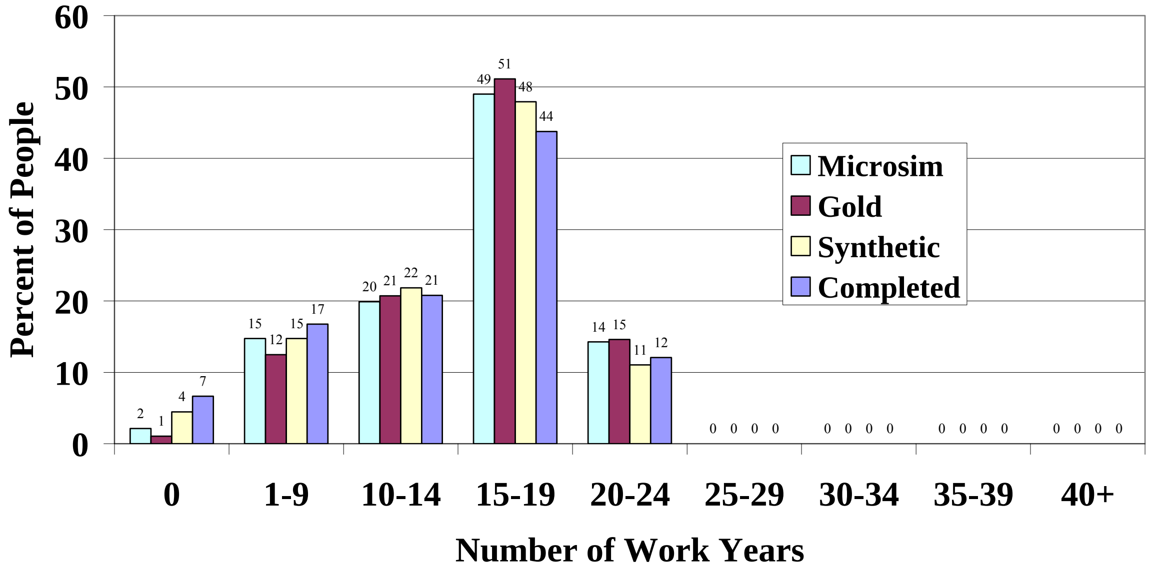
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1956 to 1960 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26h. Distribution of Total Covered Work Years for Women Born 1961 - 1965, Survivors to 2000



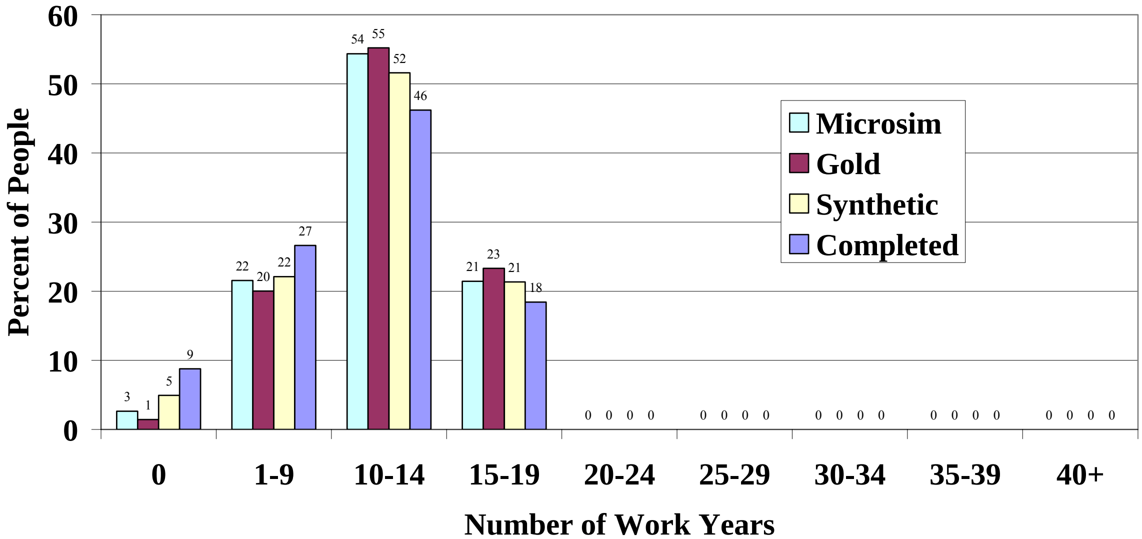
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1961 to 1965 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26i. Distribution of Total Covered Work Years for Women Born 1966 - 1970, Survivors to 2000



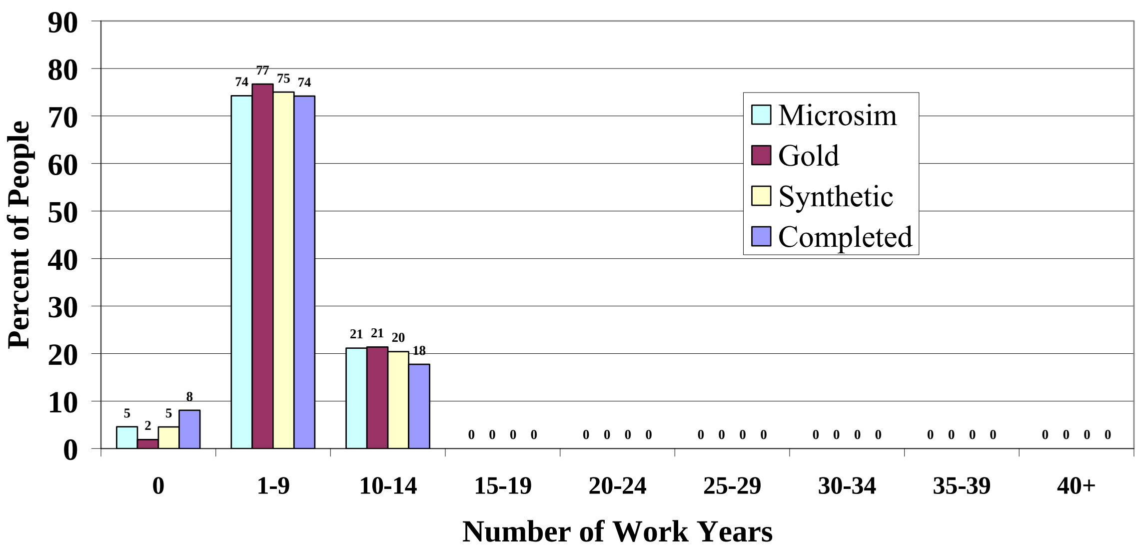
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1966 to 1970 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26j. Distribution of Total Covered Work Years for Women Born 1971 - 1975, Survivors to 2000



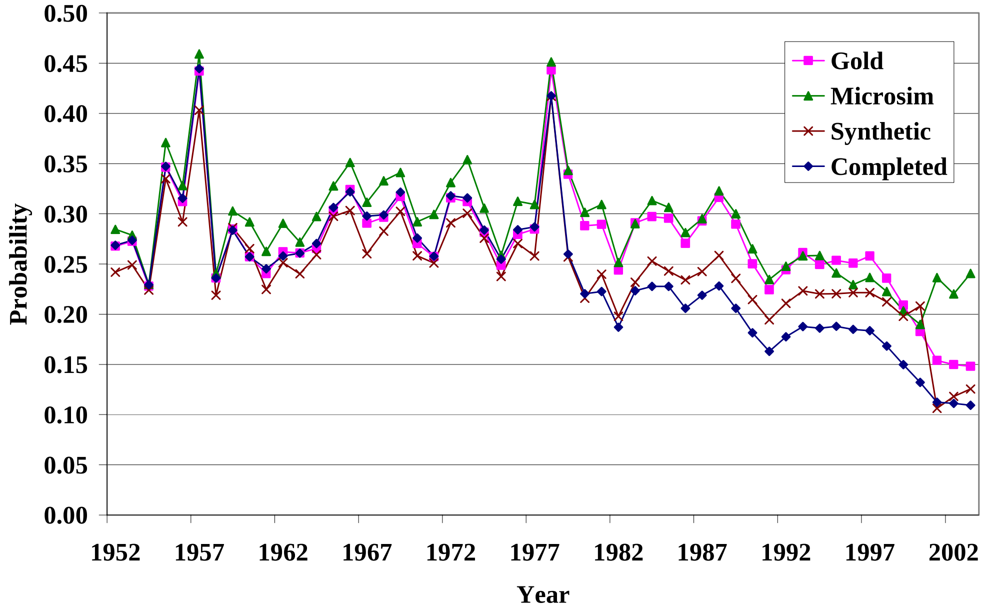
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1971 to 1975 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-26k. Distribution of Total Covered Work Years for Women Born 1976 - 1981, Survivors to 2000



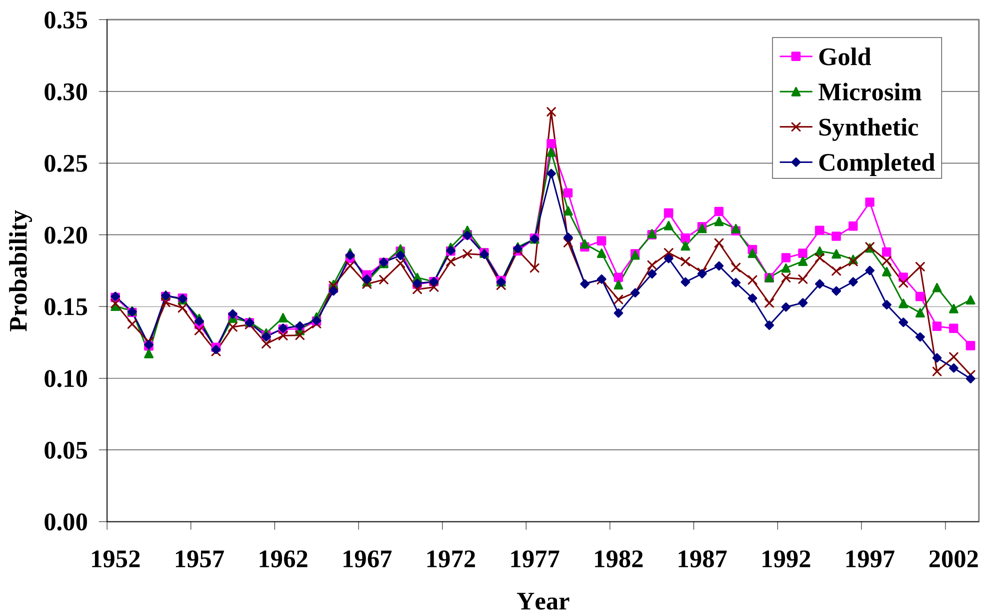
Source. Urban Institute tabulations of Microsim, Gold, pooled Completed, and pooled Synthetic SIPP files. Data are limited to U.S. resident women born from 1976 to 1981 that survive to 2000. Total covered work years is the total number of years with positive taxable earnings from age 16 to 62 between 1951 and 2003.

Figure 5-27a. Male Employment Probability Given NOT Worked Last Year 1952-2003



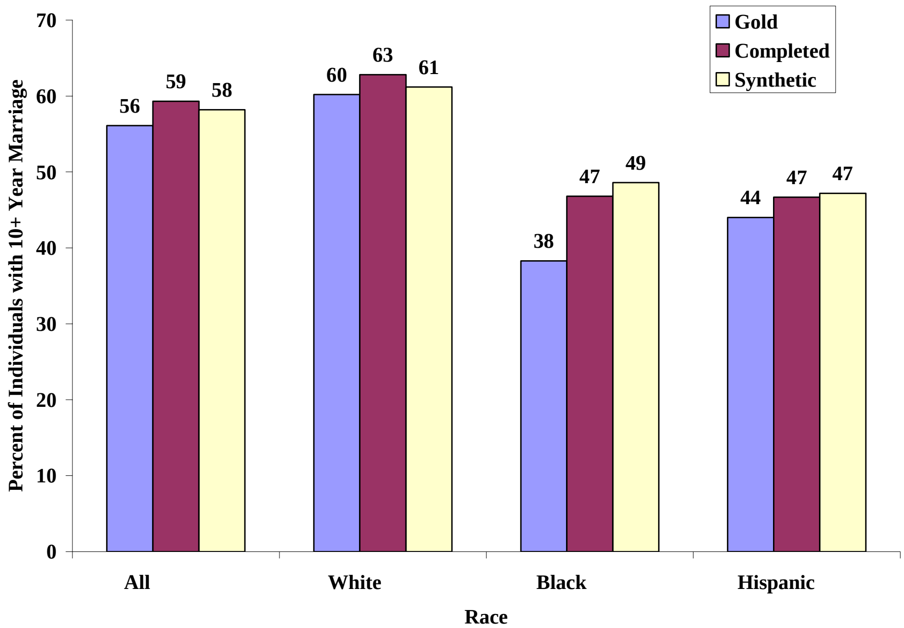
Source: Urban Institute tabulations of Gold, Synthetic (implicate 1, replicate 2), Completed (implicate 1), and Microsim. Person-year data file limited to U.S. resident male survivors to 2000 with no total earnings in the prior year.

Figure 5-27b. Female Employment Probability Given NOT Work Last Year



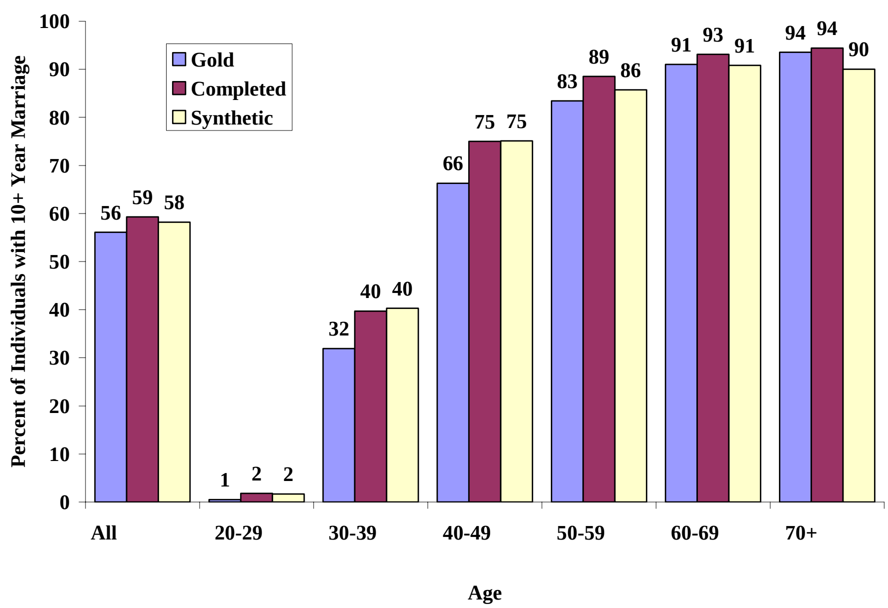
Source: Urban Institute tabulations of Gold, Synthetic (implicate 1, replicate 2), Completed (implicate 1), and Microsim. Person-year data file limited to U.S. resident female survivors to 2000 with no total earnings in the prior year.

Figure 5-28a. Percent of Individuals with Ten or More Years of Marriage by Race and Data Source



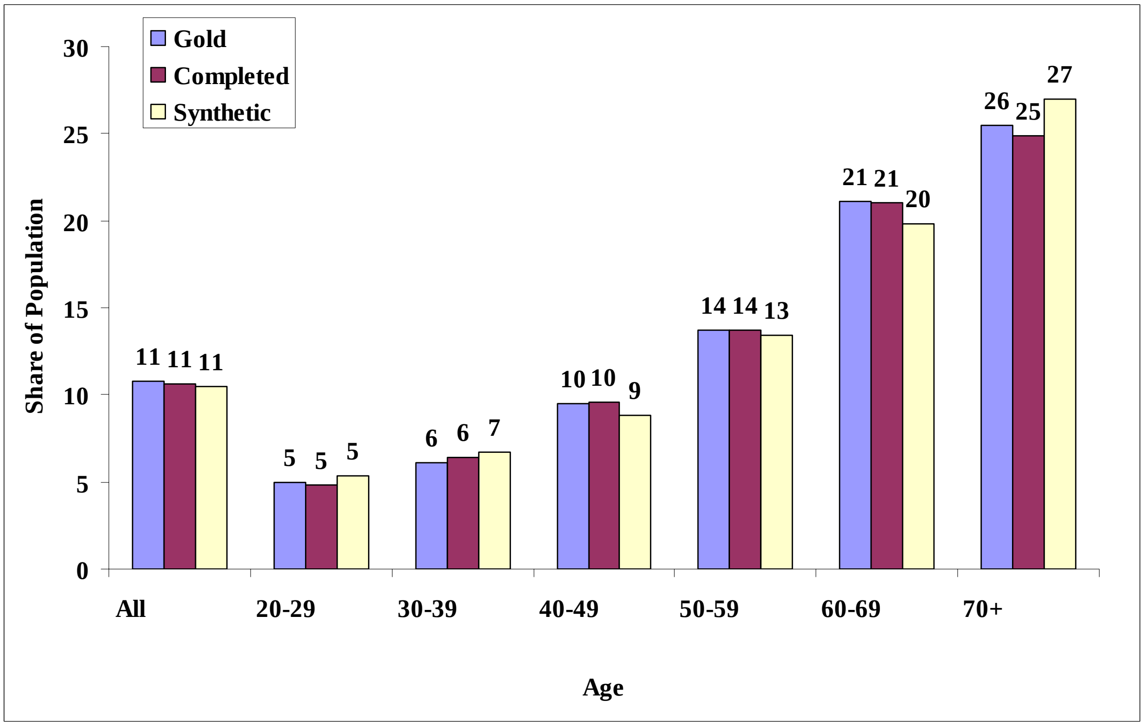
Source: Urban Institute tabulations from the Gold, Completed 1, and Synthetic 1.1 files. Data reported in Table 4.

Figure 5-28b. Percent of Individuals with Ten or More Years of Marriage by Age and Data Source



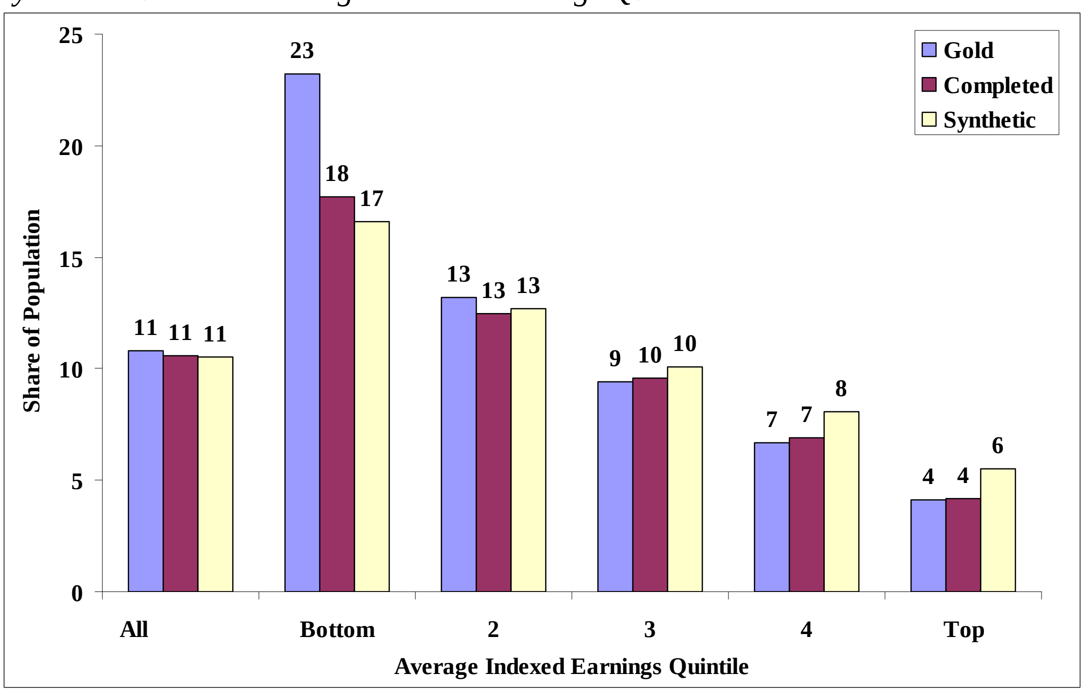
Source: Urban Institute tabulations of the Gold, Completed implicate 1, and Synthetic 1.1 (implicate 1, replicate 1) files. Data reported in Appendix E Table 5-8.

Figure 5-29. Share of Individuals Reporting a Health Disability Which Limits the Scope of Work, by Data Source and Age



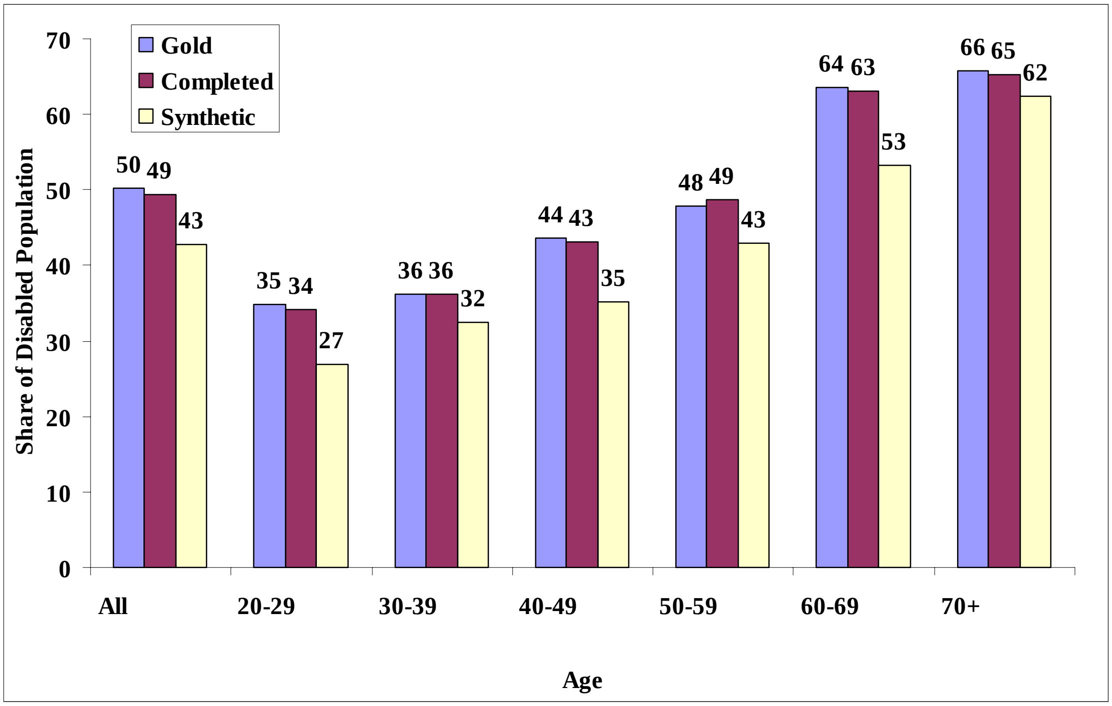
Source: Urban Institute tabulations of pooled Synthetic, pooled Completed, and Gold SIPP files.

Figure 5-30. Share of Individuals Reporting a Health Disability Which Limits the Scope of Work, by Data Source and Average Indexed Earnings Quintile



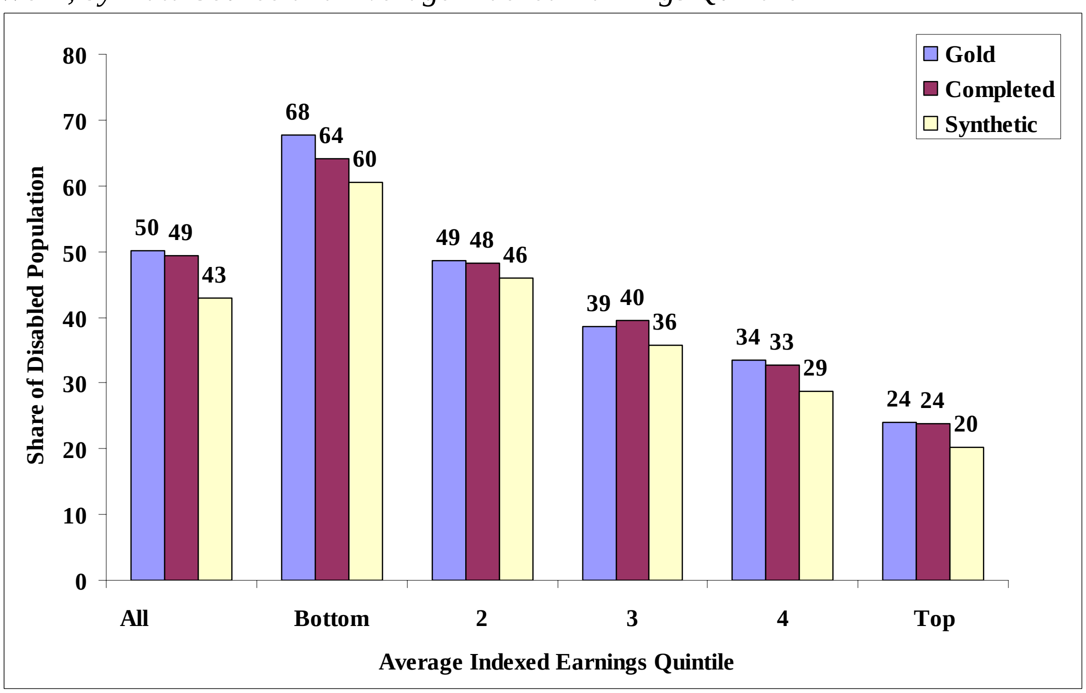
Source: Urban Institute tabulations of pooled Synthetic, pooled Completed, and Gold SIPP files.

Figure 5-31. Share of Disabled Population Reporting a Health Disability that Prevents All Work, by Data Source and Age



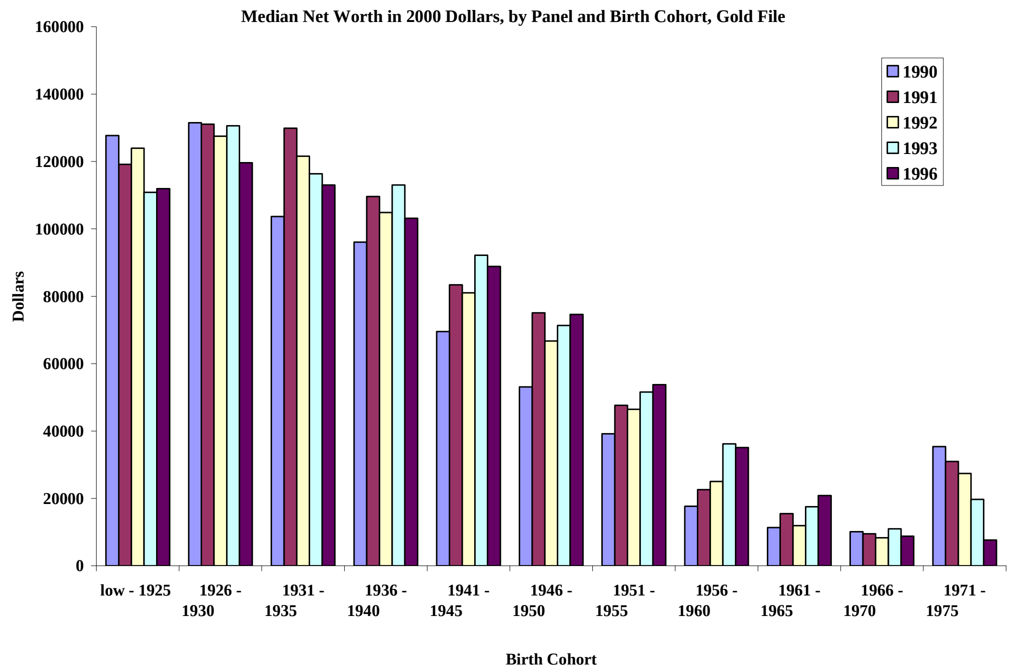
Source: Urban Institute tabulations of pooled Synthetic, pooled Completed, and Gold SIPP files.

Figure 5-32. Share of Disabled Population Reporting a Health Disability that Prevents All Work, by Data Source and Average Indexed Earnings Quintile



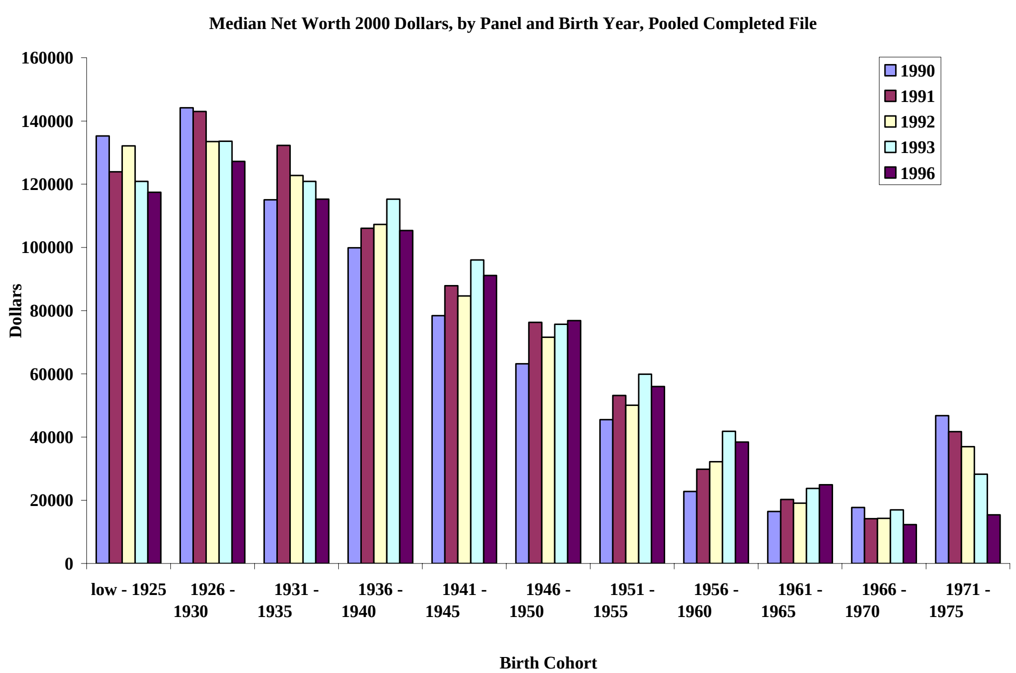
Source: Urban Institute tabulations of pooled Synthetic, pooled Completed, and Gold SIPP files.

Figure 5-33a. Median Net Worth in 2000 Dollars by Panel and Birth Cohort, Gold File



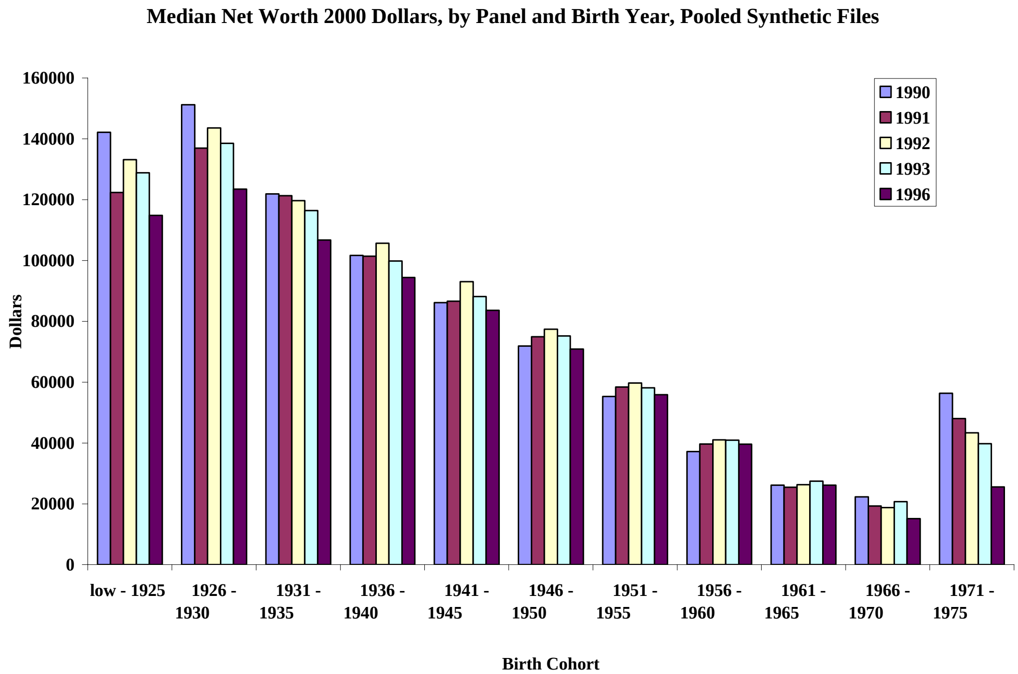
Source: Urban Institute tabulations Gold SIPP files.

Figure 5-33b. Median Net Worth in 2000 Dollars by Panel and Birth Cohort, Pooled Completed Files



Source: Urban Institute tabulations of pooled Completed file.

Figure 5-33c. Median Net Worth in 2000 Dollars by Panel and Birth Cohort, Average of Synthetic Files



Source: Urban Institute tabulations of pooled Synthetic with panel merged from the Gold file.

Table 5-4a. Number and Share of Non-missing Observations Reporting Any Health Insurance Coverage by Year, Data Source, and SIPP Panel Year

	Year									
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
SIPP Panel	**Number of Non-missing Observations**									
	Gold									
1990	38,944	37,341	39,417	0	0	0	0	0	0	0
1991	21,675	24,818	24,165	25,489	0	0	0	0	0	0
1992	0	29,558	34,527	33,243	32,734	0	0	0	0	0
1993	0	0	29,855	34,786	32,737	32,310	0	0	0	0
1996	0	0	0	0	0	18,456	62,436	54,247	50,069	49,039
All	**60,619**	**91,717**	**127,964**	**93,518**	**65,471**	**50,766**	**62,436**	**54,247**	**50,069**	**49,039**
	Completed									
1990	42,239	45,403	47,870	0	0	0	0	0	0	0
1991	22,476	29,595	30,911	29,097	0	0	0	0	0	0
1992	0	30,595	41,300	38,907	37,527	0	0	0	0	0
1993	0	0	31,265	38,615	37,426	35,670	0	0	0	0
1996	0	0	0	0	0	18,801	69,246	70,879	71,853	72,250
All	**64,715**	**105,593**	**151,345**	**106,619**	**74,953**	**54,472**	**69,246**	**70,879**	**71,853**	**72,250**
	Synthetic									
1990	42,925	46,689	47,561	15,326	2,636	1,171	3	107	108	381
1991	21,564	29,070	31,665	13,185	5,092	1,100	5	100	93	271
1992	3,334	29,723	42,374	40,276	31,260	8,510	4	168	160	312
1993	5	4,238	32,663	40,151	39,552	27,956	2	212	197	392
1996	2	5	2,159	1,473	2,221	19,609	71,400	73,026	73,452	73,465
All	**67,830**	**109,725**	**156,422**	**110,411**	**80,761**	**58,346**	**71,414**	**73,613**	**74,010**	**74,821**
Data Source	**Share of Observations with Non-missing Health Insurance Coverage**									
Gold	0.23	0.35	0.49	0.35	0.25	0.19	0.24	0.21	0.19	0.19
Completed	0.25	0.40	0.57	0.40	0.28	0.21	0.26	0.27	0.27	0.27
Synthetic	0.25	0.40	0.57	0.40	0.28	0.21	0.26	0.27	0.27	0.27

Source: Urban Institute tabulations of the Synthetic version 1.1 and pooled Completed files.

Notes: Table includes only surviving, U.S. residents out of 263,793 unweighted observations. Pooled completed counts are divided by 4.

Table 5-4b. Any Health Insurance Coverage Rates by Year, Data Source, and SIPP Panel

	Year									
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
SIPP Panel					**Gold**					
1990	0.86	0.85	0.82	.	.	.	.	.	.	.
1991	0.78	0.86	0.86	0.82	.	.	.	.	.	.
1992	.	0.77	0.84	0.84	0.83	.	.	.	.	.
1993	.	.	0.75	0.84	0.84	0.84	.	.	.	.
1996	.	.	.	.	.	0.73	0.80	0.81	0.81	0.82
All	**0.83**	**0.83**	**0.82**	**0.83**	**0.83**	**0.78**	**0.80**	**0.81**	**0.81**	**0.82**
					Completed					
1990	0.86	0.85	0.81	.	.	.	.	.	.	.
1991	0.78	0.84	0.84	0.81	.	.	.	.	.	.
1992	.	0.77	0.83	0.83	0.83	.	.	.	.	.
1993	.	.	0.75	0.83	0.83	0.83	.	.	.	.
1996	.	.	.	.	.	0.72	0.79	0.80	0.80	0.80
All	**0.83**	**0.82**	**0.81**	**0.82**	**0.83**	**0.77**	**0.79**	**0.80**	**0.80**	**0.80**
					Synthetic					
1990	0.85	0.85	0.81	0.81	0.78	0.74	0.38	0.48	0.69	0.82
1991	0.79	0.86	0.84	0.81	0.82	0.78	0.60	0.59	0.65	0.78
1992	0.67	0.79	0.83	0.83	0.83	0.81	1.00	0.62	0.67	0.77
1993	1.00	0.64	0.76	0.83	0.83	0.83	0.63	0.55	0.64	0.80
1996	1.00	1.00	0.68	0.78	0.77	0.74	0.80	0.81	0.82	0.82
All	**0.83**	**0.82**	**0.81**	**0.83**	**0.83**	**0.79**	**0.80**	**0.81**	**0.82**	**0.82**

Source: Urban Institute tabulations of the Synthetic version 1.1 and pooled Completed files

Notes: Dots reflect cells with no non-missing data.

Table 5-5a. Number and Share of Non-missing Observations Reporting Employer Health Insurance Coverage by Year, Data Source, and SIPP Panel Year

	Year									
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
SIPP	**Number of Non-missing Observations**									
Panel	**Gold**									
1990	47,622	45,524	41,825	0	0	0	0	0	0	0
1991	0	30,505	28,978	26,910	0	0	0	0	0	0
1992	0	0	41,751	40,143	37,795	0	0	0	0	0
1993	0	0	0	41,916	40,224	37,225	0	0	0	0
1996	0	0	0	0	0	0	76,422	71,158	63,834	58,794
All	**47,622**	**76,029**	**112,554**	**108,969**	**78,019**	**37,225**	**76,422**	**71,158**	**63,834**	**58,794**
	Completed									
1990	51,692	51,331	50,893	50,426	49,952	49,476	48,923	48,396	47,893	47,361
1991	33,205	33,046	32,786	32,497	32,193	31,902	31,580	31,256	30,928	30,555
1992	46,189	46,177	45,957	45,617	45,198	44,789	44,344	43,898	43,459	42,989
1993	46,162	46,147	46,124	45,855	45,433	45,049	44,588	44,139	43,679	43,211
1996	86,275	86,245	86,185	86,108	86,015	85,937	85,504	84,816	84,143	83,330
All	**263,523**	**262,946**	**261,944**	**260,502**	**258,791**	**257,153**	**254,938**	**252,503**	**250,103**	**247,446**
	Synthetic									
1990	51,918	51,654	51,361	51,050	50,763	50,460	50,123	49,798	49,470	49,116
1991	33,209	33,134	32,977	32,799	32,623	32,422	32,242	32,062	31,836	31,628
1992	46,197	46,154	46,039	45,832	45,593	45,332	45,060	44,789	44,518	44,207
1993	46,171	46,142	46,087	45,930	45,691	45,449	45,184	44,906	44,664	44,359
1996	86,297	86,251	86,178	86,079	85,975	85,843	85,573	85,175	84,735	84,249
All	**263,792**	**263,335**	**262,642**	**261,690**	**260,645**	**259,506**	**258,182**	**256,730**	**255,223**	**253,559**
Data										
Source	**Share of Observations with Non-missing Health Insurance Coverage**									
Gold	0.18	0.29	0.43	0.41	0.30	0.14	0.29	0.27	0.24	0.22
Completed	1.00	1.00	0.99	0.99	0.98	0.97	0.97	0.96	0.95	0.94
Synthetic	1.00	1.00	0.99	0.99	0.98	0.97	0.97	0.96	0.95	0.94

Source: Urban Institute tabulations of the Synthetic version 1.1 and pooled Completed files.

Notes: Table includes only surviving, U.S. residents out of 263,793 unweighted observations. Pooled completed counts are divided by 4.

Table 5-5b. Employer-Provided Health Insurance Coverage Rates by Year, Data Source, and SIPP Panel

	Year									
	1990	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
SIPP Panel					**Gold**					
1990	0.45	0.45	0.43	.	.	.	.	.	.	.
1991	.	0.43	0.43	0.41	.	.	.	.	.	.
1992	.	.	0.41	0.42	0.43	.	.	.	.	.
1993	.	.	.	0.40	0.41	0.43	.	.	.	.
1996	.	.	.	.	.	.	0.38	0.40	0.42	0.42
All	**0.45**	**0.44**	**0.42**	**0.41**	**0.42**	**0.43**	**0.38**	**0.40**	**0.42**	**0.42**
					Completed					
1990	0.37	0.39	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1991	0.27	0.38	0.39	0.36	0.00	0.00	0.00	0.00	0.00	0.00
1992	0.00	0.25	0.37	0.35	0.36	0.00	0.00	0.00	0.00	0.00
1993	0.00	0.00	0.25	0.34	0.34	0.34	0.00	0.00	0.00	0.00
1996	0.00	0.00	0.00	0.00	0.00	0.09	0.30	0.33	0.34	0.35
All	**0.08**	**0.12**	**0.17**	**0.12**	**0.09**	**0.09**	**0.15**	**0.17**	**0.17**	**0.18**
					Synthetic					
1990	0.37	0.40	0.37	0.12	0.02	0.01	0.00	0.00	0.00	0.00
1991	0.23	0.37	0.40	0.16	0.06	0.01	0.00	0.00	0.00	0.00
1992	0.02	0.19	0.37	0.36	0.29	0.08	0.00	0.00	0.00	0.00
1993	0.00	0.02	0.27	0.35	0.36	0.26	0.00	0.00	0.00	0.00
1996	0.00	0.00	0.01	0.01	0.01	0.11	0.32	0.35	0.36	0.37
All	**0.10**	**0.15**	**0.23**	**0.16**	**0.12**	**0.10**	**0.12**	**0.13**	**0.13**	**0.14**

Source: Urban Institute tabulations of the Synthetic version 1.1 and pooled Completed files

Notes: Dots reflect cells with no non-missing data.

Table 5-6. Number and Percent of Immigrants Reporting Earnings Prior to Entering the United States by Data Source, Weighted and Unweighted

	Number of Immigrants		**Number of Immigrants with Earnings Prior to Immigration**		**Percent of Immigrants with Earnings Prior to Immigration**	
Data Source	Weighted (thousands)	Unweighted	Weighted (thousands)	Unweighted	Weighted	Unweighted
Gold	16,711	20,371	1,201	1,437	7.2	7.1
Completed						
V1	24,863	28,564	4,607	5,816	18.5	20.4
V2	25,296	29,287	4,687	5,937	18.5	20.3
V3	24,576	28,126	4,872	5,954	19.8	21.2
V4	24,486	28,148	4,786	6,019	19.5	21.4
Synthetic						
1.1	23,951	31,043	7,610	10,486	33.8	33.8
1.2	24,034	30,846	7,387	10,209	33.1	33.1
1.3	23,395	30,318	7,269	10,047	33.1	33.1
1.4	23,823	30,840	7,476	10,370	33.6	33.6
2.1	21,521	27,377	7,099	9,727	35.5	35.5
2.2	20,950	26,997	6,887	9,576	35.5	35.5
2.3	23,782	30,467	8,189	11,291	37.1	37.1
2.4	22,318	28,631	7,460	10,364	36.2	36.2
3.1	21,990	27,985	7,162	9,874	35.3	35.3
3.2	22,714	29,291	7,340	10,172	34.7	34.7
3.3	23,119	29,761	7,553	10,508	35.3	35.3
3.4	22,255	28,679	7,356	10,176	35.5	35.5
4.1	22,628	28,872	7,332	10,162	35.2	35.2
4.2	22,982	30,019	7,564	10,618	35.4	35.4
4.3	22,299	28,475	7,307	10,147	35.6	35.6
4.4	22,809	29,149	7,397	10,255	35.2	35.2

Source: Urban Institute tabulations of the Synthetic, Completed, and Gold Files.

Notes: The four Completed files are identified by the implicate number V1 to V4. The Synthetic 16 files are identified by the implicate (first digit) and replicate (decimal digit).

Table 5-7. Share of Individuals with at Least One Marriage Lasting Ten or More Years by Data Source and Demographic Characteristics

	Gold	Completed				Synthetic															
		c1	c2	c3	c4	1.1	1.2	1.3	1.4	2.1	2.2	2.3	2.4	3.1	3.2	3.3	3.4	4.1	4.2	4.3	4.4
All	56.1	59.3	59.3	59.3	59.3	58.2	57.5	58.3	57.6	57.9	58.1	58.1	57.8	58.2	57.6	57.8	57.6	58.0	57.9	57.8	57.7
Age																					
20-29	0.5	1.8	1.6	1.6	1.6	1.7	1.8	1.7	1.8	2.0	1.6	1.8	1.5	1.7	1.9	1.7	1.6	1.9	1.7	1.6	1.6
30-39	31.9	39.7	39.7	39.6	39.8	40.3	38.6	40.6	39.5	39.7	40.3	40.0	39.7	40.0	39.6	40.0	39.4	39.6	40.2	39.3	40.1
40-49	66.3	75.0	75.2	75.1	75.2	75.1	75.1	75.7	73.7	75.0	75.1	74.8	74.9	75.5	74.6	75.1	74.7	75.1	74.9	74.6	74.3
50-59	83.4	88.5	88.4	88.5	88.4	85.7	84.8	85.9	85.1	85.3	85.5	85.3	84.8	86.0	85.0	84.7	84.7	85.9	84.8	85.1	84.9
60-69	91.0	93.1	93.3	93.0	93.1	90.8	89.6	90.6	89.8	89.8	89.8	90.0	89.9	90.7	89.6	90.3	89.5	90.7	89.7	90.5	89.9
70+	93.5	94.4	94.2	94.2	94.1	90.0	89.4	89.9	89.6	89.5	89.2	89.7	89.6	89.9	89.6	90.0	89.1	89.3	89.4	89.8	89.4
Sex																					
Female	57.7	62.6	62.5	62.5	62.6	60.9	59.6	60.6	59.9	60.9	60.8	60.4	60.5	60.8	60.2	60.2	59.9	60.7	60.3	60.4	60.5
Male	54.3	55.6	55.7	55.7	55.7	55.3	55.3	55.7	55.0	54.5	55.0	55.6	54.8	55.3	54.6	55.1	55.1	55.1	55.3	55.1	54.6
Education																					
Less than High School	52.6	54.5	54.6	54.5	54.5	51.5	51.1	52.0	50.6	51.3	51.9	52.1	52.2	53.8	52.8	52.7	52.2	52.7	52.3	52.4	51.9
HS Graduate	59.2	62.2	62.1	62.2	62.2	62.5	60.8	61.6	61.9	61.3	61.3	61.6	61.0	61.0	60.3	60.8	60.4	61.8	61.2	61.2	61.2
Some College	50.2	54.9	54.8	54.8	55.0	53.8	53.0	53.9	52.1	53.5	52.8	52.7	52.9	53.4	53.6	53.0	53.5	52.8	52.6	52.5	52.7
College Graduate	62.7	64.9	65.1	65.1	65.1	64.3	65.1	65.6	65.0	64.4	65.6	64.9	64.7	64.0	63.3	64.2	63.8	64.0	64.7	64.4	63.9
Race																					
White	60.2	62.8	62.8	62.8	62.9	61.2	60.7	61.4	60.9	61.0	61.1	61.3	60.9	61.3	60.7	61.0	60.7	61.2	61.1	60.9	60.9
Black	38.3	46.8	46.8	46.3	46.4	48.6	47.0	48.5	46.4	47.4	47.8	47.1	47.5	47.7	47.9	46.6	47.6	47.5	46.6	47.7	46.1
Hispanic	44.0	46.7	46.7	46.7	46.7	47.2	46.1	46.7	45.6	46.4	46.7	46.6	46.5	46.7	45.8	46.2	46.4	46.3	46.6	46.1	46.3

Source: Urban Institute tabulations of Synthetic, Completed, and Gold files.

Notes: The four Completed files are identified by the implicate number c1 to c4. The Synthetic 16 files are identified by the implicate (first digit) and replicate (decimal digit).

Table 5-8. Average Annual Poverty Thresholds by Data Source and Year, Before and After Adjustments

	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Gold										
Before Adjustment	$134,147	$138,619	$142,392	$146,600	$149,192	$152,449	$161,832	$164,737	$166,038	$168,175
After Adjustment	$11,179	$11,552	$11,866	$12,217	$12,433	$12,704	$13,486	$13,728	$13,836	$14,015
Completed										
Before Adjustment	$129,678	$135,912	$139,175	$143,604	$146,660	$149,438	$155,213	$154,949	$158,996	$162,718
After Adjustment	$10,806	$11,326	$11,598	$11,967	$12,222	$12,453	$12,934	$12,912	$13,250	$13,560
Synthetic										
Before Adjustment	$954	$1,018	$1,006	$1,034	$1,072	$1,127	$1,115	$1,102	$1,104	$1,189
After Adjustment	$11,444	$12,221	$12,074	$12,413	$12,858	$13,520	$13,375	$13,229	$13,253	$14,270

Source: Urban Institute tabulations of the Synthetic v1.1, Completed v1, and Gold Files.

Notes: Synthetic adjustment multiplied poverty threshold by 12. Completed and Gold adjustment divided poverty threshold by 12.

Table 5-9. Median Total Net Worth (in 2000 dollars) and Ratio of Completed to Gold and Synthetic to Gold by Panel and Birth Year for the Gold File

	Panel					
	1990	**1991**	**1992**	**1993**	**1996**	**All**
Birth Year	**Gold**					
low - 1925	127,701	119,170	123,936	110,849	111,946	117,434
1926 - 1930	131,494	131,086	127,511	130,604	119,629	126,214
1931 - 1935	103,678	129,895	121,553	116,336	113,044	116,322
1936 - 1940	96,092	109,636	104,869	113,044	103,166	104,264
1941 - 1945	69,540	83,419	81,035	92,191	88,899	84,509
1946 - 1950	53,104	75,077	66,735	71,338	74,631	69,540
1951 - 1955	39,195	47,668	46,476	51,583	53,778	48,859
1956 - 1960	17,701	22,642	25,026	36,218	35,120	28,601
1961 - 1965	11,379	15,492	11,917	17,560	20,853	16,437
1966 - 1970	10,115	9,534	8,342	10,975	8,780	9,534
1971 - 1975	35,402	30,984	27,409	19,755	7,683	15,365
1976 - 1981	77,127	44,093	27,409	31,828	35,120	34,023
	Ratio of Completed to Gold					
low - 1925	1.06	1.04	1.07	1.09	1.05	1.05
1926 - 1930	1.10	1.09	1.05	1.02	1.06	1.05
1931 - 1935	1.11	1.02	1.01	1.04	1.02	1.02
1936 - 1940	1.04	0.97	1.02	1.02	1.02	1.02
1941 - 1945	1.13	1.05	1.04	1.04	1.02	**1.05**
1946 - 1950	1.19	1.02	1.07	1.06	1.03	1.06
1951 - 1955	1.16	1.12	1.08	1.16	1.04	1.10
1956 - 1960	1.29	1.32	1.29	1.15	1.09	1.21
1961 - 1965	1.44	1.31	1.60	1.35	1.19	1.31
1966 - 1970	1.75	1.49	1.71	1.54	1.40	1.50
1971 - 1975	1.32	1.35	1.35	1.43	2.00	1.63
1976 - 1981	0.43	1.14	1.48	1.24	1.17	1.20
	Ratio of Synthetic to Gold					
low - 1925	0.94	1.07	1.17	1.36	0.87	1.07
1926 - 1930	1.07	0.89	1.14	1.00	1.03	1.05
1931 - 1935	1.14	0.76	1.08	0.72	1.01	0.94
1936 - 1940	1.09	0.73	1.00	0.64	0.97	0.88
1941 - 1945	1.18	0.72	1.16	0.58	0.89	**0.88**
1946 - 1950	1.34	0.71	1.15	0.57	0.89	0.90
1951 - 1955	1.39	0.73	1.24	0.53	0.98	0.94
1956 - 1960	1.88	0.90	1.41	0.51	1.03	1.02
1961 - 1965	1.89	0.94	1.52	0.61	1.01	1.10
1966 - 1970	1.46	1.00	1.42	0.68	1.19	1.19
1971 - 1975	0.49	1.00	0.90	1.54	2.42	1.42
1976 - 1981	0.28	0.90	1.09	1.43	1.53	1.41

Source: Urban Institute tabulations of the Gold, pooled Completed, and pooled Synthetic SIPP files.

Table 5-10. Correlation of SIPP Self-Reported and Administrative Total Earnings by Year and Data Source

Data Source	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	All
	Year										
Gold	0.70	0.70	0.71	0.70	0.68	0.64	0.68	0.67	0.66	0.62	0.68
Completed											
V1	0.70	0.69	0.70	0.69	0.68	0.64	0.67	0.69	0.67	0.65	0.68
V2	0.69	0.70	0.70	0.70	0.68	0.64	0.68	0.68	0.66	0.64	0.68
V3	0.68	0.69	0.70	0.70	0.68	0.64	0.68	0.68	0.67	0.65	0.68
V4	0.68	0.70	0.71	0.70	0.68	0.64	0.68	0.68	0.67	0.64	0.68
Synthetic											
1.1	0.68	0.67	0.70	0.67	0.66	0.61	0.65	0.65	0.64	0.61	0.65
1.2	0.69	0.69	0.70	0.66	0.64	0.62	0.62	0.63	0.63	0.60	0.65
1.3	0.69	0.68	0.69	0.66	0.63	0.63	0.65	0.63	0.63	0.60	0.65
1.4	0.70	0.70	0.70	0.66	0.65	0.62	0.64	0.64	0.63	0.61	0.66
2.1	0.69	0.70	0.71	0.69	0.68	0.65	0.67	0.66	0.64	0.61	0.67
2.2	0.70	0.70	0.71	0.69	0.68	0.64	0.68	0.66	0.66	0.61	0.67
2.3	0.69	0.70	0.71	0.70	0.68	0.63	0.67	0.65	0.66	0.61	0.67
2.4	0.70	0.70	0.70	0.70	0.69	0.64	0.68	0.66	0.65	0.61	0.67
3.1	0.70	0.69	0.70	0.69	0.67	0.64	0.66	0.65	0.62	0.59	0.66
3.2	0.70	0.70	0.71	0.67	0.66	0.63	0.67	0.65	0.63	0.59	0.66
3.3	0.69	0.70	0.71	0.69	0.66	0.62	0.66	0.65	0.63	0.58	0.66
3.4	0.70	0.70	0.71	0.70	0.67	0.63	0.66	0.65	0.63	0.58	0.66
4.1	0.70	0.71	0.71	0.70	0.69	0.64	0.67	0.66	0.64	0.60	0.67
4.2	0.69	0.69	0.71	0.68	0.68	0.63	0.67	0.66	0.64	0.60	0.67
4.3	0.70	0.70	0.71	0.69	0.67	0.63	0.66	0.65	0.65	0.61	0.67
4.4	0.69	0.70	0.71	0.67	0.67	0.64	0.67	0.66	0.65	0.60	0.67

Source: Urban Institute tabulations of the Synthetic, Completed, and Gold Files.

Notes: Table includes weighted Pearson Correlation Coefficients of SIPP and Administrative total earnings among non-missing observations. SIPP earnings are based on totearn1990-totearn1999. Administrative earnings are the sum of defer_der_fica_{year}, nondefer_der_fica_{year}, defer_der_nonfica_{year}, nondefer_der_nonfica_{year}. To reduce the impact of outliers, earnings are capped at 3 times the economy-wide average earnings.

Notes: The four Completed files are identified by the implicate number V1 to V4. The Synthetic 16 files are identified by the implicate (first digit) and replicate (decimal digit).