

ABSTRACT

Title of dissertation: USING LINKED EMPLOYER-EMPLOYEE
DATA TO UNDERSTAND LABOR
MARKETS AND IMPROVE DATA
PRODUCTS

Gary Benedetto, Doctor of Philosophy, 2007

Dissertation directed by: Professor John C. Haltiwanger
Department of Economics

This thesis is comprised of three chapters. The first chapter (joint with John Haltiwanger, Julia Lane, and Kevin McKinney) explores a new way of capturing dynamics: following clusters of workers as they move across administrative entities. Information on firm dynamics is critical to understanding economic activity, yet fundamentally difficult to measure. The worker flow approach is shown to improve linkages across firms in longitudinal business databases. The approach also provides conceptual insights into the changing structure of businesses and employer-employee relationships. Many worker-cluster flows involve changes in industry – particularly movements into and out of personnel supply firms. Another finding, that a nontrivial fraction of firm entry is associated with such flows, suggests that a path for firm entry is a group of workers at an existing firm starting a new firm.

The second chapter makes use of linked employer-employee data from the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) Program and matches it to data on business acquisitions from the Federal Trade Commission to examine labor market outcomes of employees at firms undergoing mergers. Earnings and employment can be observed over time for workers at both the acquired firm and

the acquiring firm. The findings suggest that while wages tend to be about the same or higher for workers at these restructuring firms, turnover is significantly higher, and the costs of job-loss are large and long lasting.

The third chapter (joint with John Abowd and Martha Stinson) provides technical documentation for a project undertaken by the US Census Bureau, the Social Security Administration, and the Internal Revenue Service to explore a potential method of providing the public a valuable new dataset without compromising confidentiality. The underlying database was created by merging the respondents from the Census' own SIPP with administrative data on earnings and benefits from the IRS and SSA. The administrative variables combined with the detailed survey responses from the SIPP offer the potential to do interesting research especially in the areas of retirement, benefits, and lifetime earnings; however, they also add extensive new information for malicious data users to potentially reidentify SIPP respondents. This final chapter develops a cutting edge new technique for providing a micro-dataset that looks, in structure, just like the underlying confidential data. This "partially synthetic" database aims to preserve as many of the complex covariate relationships in the confidential data without posing any significant new risk to disclosure protection.

USING LINKED EMPLOYER-EMPLOYEE
DATA TO UNDERSTAND LABOR
MARKETS AND IMPROVE DATA
PRODUCTS

by

Gary Benedetto

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:

Professor John C. Haltiwanger, Chair/Advisor

Professor John M. Abowd, Advisor

Professor John Shea

Professor Seth Sanders

Professor John Horowitz

© Copyright by
Gary Benedetto
2007

DEDICATION

This thesis is dedicated to my children, Lucas and Mia Benedetto.

ACKNOWLEDGMENTS

I would like to thank the members of my dissertation committee, particularly Professors John C. Haltiwanger and John Abowd, for their support and advice on this thesis. The second chapter benefitted greatly from the comments received in both the microeconomics and macroeconomics brown bag seminars at the Department of Economics, University of Maryland. I would also like to thank everyone from the research staff of the Longitudinal Employer-Household Dynamics program of the United States Census Bureau for their helpful comments and encouragement.

My wife, Shannon Benedetto, has been an extremely strong source of support for me throughout my entire graduate career. I could not have had the persistence to finish this dissertation without her constant love and patience.

Contents

Contents	iv
List of Tables	vii
List of Figures	ix
1 Chapter	1
Using Worker Flows to Measure Firm Dynamics	1
1.1 Introduction	1
1.2 Background	3
1.2.1 Tracking Firm Births and Deaths	4
1.2.2 Changes in Firm Structure	5
1.2.3 Insourcing and Outsourcing	6
1.3 Data	7
1.4 Measuring Firm Dynamics	10
1.4.1 Classifying Clustered Worker Flows	10
1.4.2 Firm Entry and Exit	12
1.4.3 Identifying Firm Dynamics	13
1.5 Empirical Analysis of Clustered Worker Flows	16
1.5.1 Relative Frequency of Transitions	16
1.5.2 Using Industry to Shed Light on Firm Dynamics	17
1.6 Impact Analysis	21
1.7 External Validation Of the Worker Flow Approach	24
1.7.1 Successor/Predecessor Information from the QCEW Program	25
1.7.2 Match to the Census Business Register	27
1.8 Conclusion	27
2 Chapter	2
The Effects of Mergers on Workers' Earnings and Employment	39
2.1 Introduction	39
2.2 Background	40
2.3 Data	43
2.3.1 General Overview	43
2.3.2 Method of Identifying Firm Restructurings	46
2.3.3 Multiple Imputation of Missing Data: Reason for Separation .	49
2.3.4 Selecting the Population to Analyze	53
2.4 Empirical Model	54
2.4.1 General Strategy	54
2.4.2 Earnings Regression	55
2.4.3 Logistic Regression on Quits and Layoffs	57
2.4.4 Examining Earnings Losses of Quits and Layoffs	58
2.5 Results	61
2.6 Conclusion	65

3	Chapter	3
	Technical Description of the SIPP/SSA/IRS Public Use File Project	84
3.1	Executive Summary	84
3.1.1	Purpose and brief history	84
3.1.2	Structure of the inputs to the SIPP/SSA/IRS public use file	85
3.1.3	Completion of the missing data and synthesis of the confidentiality-protected data	88
3.1.4	Development of the weights	91
3.1.5	Analytical validity testing	93
3.1.6	Disclosure avoidance assessment	98
3.1.7	Using the SIPP/SSA/IRS-PUF	101
3.1.8	Next steps	101
3.2	Project Background	103
3.2.1	Purpose and brief history	103
3.2.2	Overview project description	104
3.3	Creation of the Gold Standard File	107
3.3.1	SIPP data	107
3.3.2	IRS/SSA earnings data	112
3.3.3	SSA data	114
3.3.4	Weight creation and use	114
3.4	Data Completion and Synthesis	116
3.4.1	General methodology	116
3.4.2	Bayesian Bootstrap	123
3.4.3	Sequential Regression Multivariate Imputation	126
3.4.4	Summary of synthetic data production	129
3.4.5	Modeling details	135
3.5	Weight Creation and Synthesis	158
3.5.1	Introduction and background	158
3.5.2	Summary of the weight creation process	161
3.5.3	Part A: Creation of poverty stratification variable for Census 2000 records	164
3.5.4	Poverty stratum assignment	166
3.5.5	Part B: Creation of stage-2 clusters for Census 2000 records	170
3.5.6	Part C: Creation of poverty stratification variable for SIPP records	174
3.5.7	Part D: Creation of stage-2 clusters for SIPP records	175
3.5.8	Part E: Matching SIPP individuals to Census 2000 records	177
3.5.9	Part F: Creation of preliminary weight	181
3.5.10	Part G: Creation of the final weight	182
3.5.11	Geography issues	183
3.5.12	Birth date issue	184
3.5.13	Overall evaluation of Gold Standard weight	184
3.5.14	Synthesizing the weight	185
3.5.15	Evaluation of the synthesized weight	187
3.6	Analytical Validity	189

3.6.1	Complete data estimation	190
3.6.2	Inference frameworks using multiple imputation	191
3.6.3	Application to the SIPP/SSA/IRS-PUF	197
3.6.4	Results	198
3.7	Assessing Disclosure Risk	209
3.7.1	Overview	209
3.7.2	Matching based on probabilistic record linking	210
3.7.3	Distance matching	216
3.8	Using Synthetic Data	220
3.9	Conclusion	223

References **266**

List of Tables

Table 1.1	29
Table 1.2	29
Table 1.3	30
Table 1.4	31
Table 1.5	32
Table 1.6	33
Table 1.7	35
Table 1.8	36
Table 2.1	67
Table 2.2	68
Table 2.3a	69
Table 2.3b	70
Table 2.4a	71
Table 2.4b	72
Table 2.5a	73
Table 2.5b	74
Table 2.6	75
Table 2.7	76
Table 2.8	77
Table 2.9	78
Table 3.1	225
Table 3.2	226
Table 3.3	227
Table 3.4	228
Table 3.5	229
Table 3.6	230
Table 3.7	231
Table 3.8	232
Table 3.9	233
Table 3.10	234
Table 3.11	235
Table 3.12	236
Table 3.13	237
Table 3.14	238
Table 3.15	239
Table 3.16	240
Table 3.17	241
Table 3.18	241
Table 3.19	242
Table 3.20	242
Table 3.21	243
Table 3.22	244
Table 3.23	247

Table 3.24	250
Table 3.25	251
Table 3.26	252
Table 3.27	253
Table 3.28	254
Table 3.29	255

List of Figures

Figure 1.1	37
Figure 1.2	37
Figure 1.3	38
Figure 2.1a	79
Figure 2.1b	79
Figure 2.2a	80
Figure 2.2b	80
Figure 2.3	81
Figure 2.4a	83
Figure 2.4b	83
Figure 3.1	256
Figure 3.2	257
Figure 3.3	258
Figure 3.4	259
Figure 3.5	260
Figure 3.6	261
Figure 3.7	262
Figure 3.8	263
Figure 3.9	264
Figure 3.10	265

Using Worker Flows to Measure Firm Dynamics^{*†}

1.1 Introduction

Information on firm dynamics is critical to understanding economic activity. This is particularly evident in the attention paid to firm deaths, mergers/acquisitions, and outsourcing in the popular press. In addition, recent research has shown that aggregate growth in the U.S. economy is closely linked to firm restructuring and reallocation activities, with resources being reallocated from less productive to more productive firms (Foster et al. 2001). The pace of the churning of jobs, workers, and firms underlying this ongoing reallocation is high and is an important factor for understanding worker and firm economic outcomes (Brown et al. 2006).

Given this importance, U.S. statistical agencies have improved the tracking of firm dynamics by devoting increased attention to the development of longitudinal business databases. However, developing the data infrastructure and new measures of business dynamics has posed serious measurement challenges – challenges that

^{*}This chapter is a reprint of "Using Worker Flows to Measure Firm Dynamics," Gary Benedetto, John Haltiwanger, Julia Lane, and Kevin McKinney. (benedetto et al, 2006).

[†]This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications, and is released to inform interested parties of ongoing research and to encourage discussion of work in progress. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grant SES-9978093 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging, and the Alfred P. Sloan Foundation. The views expressed herein are attributable only to the author(s) and do not represent the views of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau is preparing to support external researchers' use of these data; please contact U.S. Census Bureau, LEHD Program, FB 2138-3, 4700 Silver Hill Rd., Suitland, MD 20233, USA. We appreciate the useful comments of Katherine Abraham, Fredrik Andersson, and Jim Spletzer. John Abowd provided valuable guidance in structuring the approach. We also benefited from the comments of three unusually thoughtful referees.

have been exacerbated in recent years by the blurring of firm boundaries exemplified by outsourcing, insourcing, firm spin-offs, and breakouts. These changes in firm boundaries make it difficult to measure and interpret the expansion and contraction as well as entry and exit of firms.

This paper describes ways in which linked employer-employee datasets can be used to improve the measurement and interpretation of firm transitions. Our basic approach is novel in that we use information about the movement of worker-clusters across firms to develop a broad new set of linkages not typically present in longitudinal business data. It has long been argued that in order to truly understand the relationship between firms and workers, it is necessary to have universal, longitudinal data on firms, workers, and the match between the two (Lane et al. 1998, Hamermesh 1999). In this spirit, we take advantage of new linked employer-employee data, created by combining employer level information from the Bureau of Labor Statistics' (BLS) Quarterly Census of Employment and Wages (QCEW) program with state Unemployment Insurance (UI) worker records (Abowd et al. 2000). A few examples may help illustrate both the value of integrated worker firm data and our approach.

First, consider a firm that undergoes a change in administrative identifiers due to a change in ownership or legal form of organization. In many cases, though such an event may appear to be a firm birth and death, the activities, location, and, in particular, the workers remain largely unchanged. Although the BLS QCEW program has a record tracking system to capture such changes, the possibility remains that this will be recorded as the entry of one firm and the exit of a new one. The use of worker flows enables us to establish a link between the two firms.

Now consider examples of changes in the boundaries of firms. For instance, an existing firm might outsource a portion of its workforce to another firm – such as outsourcing a particular function (janitorial or accounting services) or outsourcing the production of an intermediate input. Another example occurs when a firm spins-off

or breaks out a subsidiary unit. These events represent a change in traditional firm boundaries and/or the employer-employee relationship that are particularly difficult to capture. The use of worker flows allows us to identify many of these “new” relationships and gain important insights into the prevalence of complex firm structures.

The paper proceeds as follows. Section 2 provides additional background motivation. Section 3 provides an overview of the data infrastructure at the Longitudinal Employer-Household Dynamics (LEHD) Program housed at the Bureau of the Census and describes how the data can be used to construct measures of the clustered flows of workers. Section 4 describes the clustered worker flow methodology we use to construct new measures of firm dynamics. Section 5 presents an analysis of the clustered flows of workers that quantifies the relative importance of different types of changes in firm structures and boundaries. We quantify the impact that transitions have on measures of firm dynamics in section 6. Section 7 presents a comparison of the firm linkages identified by our approach and those identified under the existing administrative data processing of predecessor/successor relationships. Section 8 presents concluding remarks.

1.2 Background

Many national statistical agencies such as the U.S. Census Bureau and the Bureau of Labor Statistics have developed longitudinal business databases that rely on links across firms to properly capture firm dynamics (see, for example, Doms and Bartelsman 2000, Faberman 2001, Jarmin and Miranda 2002, Carroll et al. 2002 or Clayton et al. 2003). The development of these databases faces three challenges: tracking firm births and deaths, capturing changes in firm structure, and identifying across-firm relationships such as insourcing and outsourcing, particularly the increased use of temporary-help businesses.

1.2.1 Tracking Firm Births and Deaths

Accurately identifying firm entry and exit is an important exercise. Although such events occur at the fringes of the economy, an accumulation of evidence suggests that this activity is disproportionately important in promoting economic change (Doms and Bartelsman 2000). The reallocation of jobs from exiting firms to entering firms contributes positively to productivity growth (Foster et al. 2001), and successful entering businesses grow at a much faster rate than do existing firms.

The two U.S. statistical agencies (Census Bureau and BLS) developing separate longitudinal business databases recognize the importance of accurately tracking firm births and deaths. Unfortunately, in addition to ownership changes, firms often change their identifiers for accounting convenience or to avoid administrative penalties, even when the factors of production are virtually identical in the “new” and “old” firm. The latter practice has become widespread enough in the Department of Labor’s database that it has acquired the term, “SUTA dumping,” and attracted the attention of regulators (see United States Department of Labor 2002)

In the case where a new firm inherits virtually all the factors of production from a recent firm death, little real structural change has taken place. Although it is clear that an ownership change is an economically significant event, indistinguishable from a pure job flow perspective to an administrative edit, both events should be treated equally. The act of transferring ownership in and of itself does not create new jobs; it is the effect on the future operation of the firm that should accrue to the new owners, whatever the source of the change in firm identifiers. A link between the two firms allows for the proper accounting of this event.

Spletzer (2000) found that the accurate measurement of the links between firms is important for series that estimate firm entry and exit as well as for series that estimate job creation and destruction (see Pivetz et al. 2001). To reduce the occurrence of spurious changes, both business databases use additional administrative and survey

information as well as geographic coding to link firm identifiers across time (see Pivetz et al. 2001, Jarmin and Miranda 2002, and Clayton et al. 2003). In principle, these links can be enhanced by using information on the clustered flow of workers across firms. This approach was first demonstrated on U.S. data by Pivetz and Chang (1998), but is in use internationally. In particular, Scandinavian and French statistical agencies have also begun implementing such approaches (e.g. Persson 1999).

1.2.2 Changes in Firm Structure

Accurately tracking mergers and acquisitions (as well as spin-off companies and breakouts) is important for a number of reasons. First, such events represent a substantial restructuring of economic activity for both the acquiring and the acquired firm. The acquiring firm changes its size and scope while the acquired firm often loses its corporate identity. Second, they account for a substantial portion of economic activity. In 1995, the value of mergers and acquisitions equaled 5% of GDP and were equivalent to 48% of non-residential gross investment (Andrade et al. 2001). In addition, Jovanovic and Rousseau (2002) note that mergers play an important reallocation role – particularly for capital. Indeed, in certain sectors such as health care (Gaynor and Haas-Wilson 1998) and financial services (Hunter et al. 2001), mergers and acquisitions are in many ways changing the very structure of the industry and the types of services produced.

Acs and Armington (1998) provide an extensive analysis of the measurement issues encountered when developing longitudinal links across businesses. They find that using administrative identifiers alone is insufficient to accurately track changes in firms' identities, and that the problem is not obviated by the use of survey based information, such as the Census Bureau's Company Ownership Survey.

Linked employer-employee data have the potential to identify changes in firm structure that are otherwise difficult to detect. In each case, whether the event is a

merger/acquisition or a spin-off/breakout, a significant cluster of workers moves from one administrative entity to another. We use entry and exit measures to differentiate between the two events: in the case of mergers, one of the entities will disappear; in the case of spin-offs, a new entity will appear. In the latter case, we are able to quantify the extent of all firm entry accounted for by such spin-offs.

1.2.3 Insourcing and Outsourcing

Firms often contract for services outside their core area of expertise. However, despite the interest of policy-makers in this phenomenon, there is little empirical evidence. The most frequently used approach is to measure the growth in the temporary-help services industry. The rapid growth of this industry clearly indicates a substantial change has taken place in the nature of the employment relationship. In particular, employment in temporary-help services grew five times as fast as overall non-farm employment between 1972 and 1997, an average annual growth rate of 11% (Estevo and Lach 1999, Autor 2003). By the 1990's this sector accounted for 20% of all employment growth.

Given the large growth of the temporary-help industry, the usage of informal employment arrangements would appear to be a pervasive feature of the modern firm. However, the majority of the empirical evidence comes from worker-based surveys, such as the Contingent Worker Supplement to the Current Population Survey, which do a poor job capturing the distribution of informal employment across firms and industries. With the exception of a few small firm surveys, little is known about which businesses outsource employment, particularly since outsourcing may take forms other than the increased use of temporary or leased employee agencies. Houseman (1997) analyzed the data from one of these small surveys and found that 27% of the firms used on-call workers, 46% used agency temporaries, and 44% used contract workers, although this varied by size and industry.

Firms use workers in alternative work arrangements for a variety of reasons; cost effectiveness, flexibility, and the ability to screen workers prior to hiring are the most widely cited factors (Abraham and Taylor 1996). However, the empirical evidence suggests that while there are many reasons firms use alternative work arrangements, staffing needs, primarily short term, are the main source of demand for on-call workers and agency temporaries.

The converse of outsourcing, insourcing, is a relatively ignored dynamic in today's labor market. As we will see later in this paper, insourcing appears to be a direct by product of firms' increased outsourcing activity. The act of outsourcing or contracting for services may allow firms to evaluate the skills and employability of a substantial pool of workers. As market conditions or the size of the firm changes, firms often choose to bring in-house functions such as information technology or maintenance that were previously done by a contracting firm. The best candidates for this work are likely to be the employees of the contracting firm.

Linked employer-employee data can be used to help identify both insourcing and outsourcing. In particular, if firm A decides to outsource an administrative task to firm B, there is a strong incentive for firm B to employ at least some subset of workers that have accumulated experience at firm A. In such a situation, the event would likely be captured by documenting the flow of a cluster of workers moving from one continuing firm (firm A) to another continuing firm (firm B).

1.3 Data

We use a new linked employer-employee dataset available from the Census Bureau's Longitudinal Employer Household Dynamics (LEHD) program, (Abowd et al., 2000; <http://lehd.dsd.census.gov>). As noted above, these data are created by combining employer level information from the QCEW program with state UI worker records. Covered employers file regular earnings reports for each employee with pos-

itive earnings sometime during the quarter. From this information, quarterly employment and earnings histories are constructed for every person-firm combination in the data. QCEW data, the core of the BLS establishment database, are also filed by the employer and provide the industry, total employment, and location of every establishment on the 12th of the month.

The integrated LEHD master files have a number of key characteristics. Within a state, they are universal and longitudinal for both firms and workers, resulting in a very dense sample of about 96% of private wage and salary employment. Across states, coverage is very good; at the middle of 2006, the LEHD program consisted of 43 partner states, covering some 85% of US employment. For most states, the data series begin in the early 1990's and are updated on a quarterly basis (six months after the transaction date).

The data are beginning to be used to analyze many facets of employment dynamics. In work closest in spirit to this paper, Abowd and Vilhuber (2005) use them to examine the sensitivity of economic statistics to coding errors in personal identifiers. In other work, Davis et al. (2005) examine the dynamics of young and small businesses by matching the data to non-employer information.

The data have a number of drawbacks that are extensively documented in Abowd et al. (2000). One issue is that the UI wage records contain only a state employer identification number (SEIN), while the QCEW program reports data at the more disaggregated establishment level. Fortunately about 85% of all the SEINs in a given year and quarter have only one unit and of the 15% that have more than one unit, about 62% of those have the same 4-digit industry across all their establishments. For the small percentage of SEINs in our links that have multiple establishments with varying industries, the linking strategy is to attach the SEIN employment weighted modal industry to worker flows. More sophisticated methods of imputing a worker's industry will be used in future research. Another issue of importance for this paper is

the fact that the filing unit is a within-state administrative entity (identified by a state employer identification number, or SEIN) thus firms that operate in multiple states may appear as a single unit within a state. While an SEIN typically encompasses an entire firm, this is not a requirement for multiunits (about 30% of the employment). These firms are free to use multiple SEIN's and are allowed to group establishments within those SEIN's as best fits their corporate structure. The impact of this is discussed later in the paper.

In this paper, we analyze only a subset of the data currently available at LEHD. Specifically, we use the following 18 states over the period 1992 – 2001: CA, CO, FL, ID, IL, KS, MD, MN, MO, MT, NC, NJ, NM, OR, PA, TX, VA, and WV. Due to historical data availability issues, not all states are present in every year. Most states' data begins in 1992, except for MN (1994), NJ (1996), NM (1995), TX (1995), VA (1998), and WV (1997) Although additional states are part of the program, these 18 states were chosen based on data availability and processing constraints at the time we began our analysis. This selection rule should not bias our results if order of entry into the program is uncorrelated with state differences in clustered worker mobility patterns.

Using our 18 state sample, we search through each worker's employment history and create a new database containing 2,668,127,897 firm-to-firm, worker-cluster transitions (note that clustered job flows that move across boundaries are not captured, since all of our analysis is done within state) Each record in this new database is uniquely identified by the predecessor SEIN, successor SEIN, and the date at which the firm-to-firm transition occurred. The size of the predecessor and successor firm, industry, and the number of employees involved in the transition are attached to each record. To simplify our analysis, we ignore firm-to-firm worker transitions that take place over more than two quarters. This focus implies that our analysis does not capture the situation where a cluster of workers flow into nonemployment for more

than two quarters and subsequently find work together at a new firm, since to be included in our sample, a cluster of workers that leaves a firm in quarter q must begin employment at a new firm by the end of quarter $q+1$.

The vast bulk of the transitions are singletons (one worker moving from one firm to the next). These movements account for 98% of the records in our database, and just over 90% when weighted by cluster size. The frequency of the transitions by worker-cluster size is reported in Table 1.1.

While each record potentially represents a firm-to-firm relationship, it is reasonable to assume that the strength of this relationship is a function of the absolute magnitude of the flows between each firm. This implies that the vast majority of one worker “clusters” represent the normal dynamics of our labor market and therefore contain little information about a firm-level relationship. To make the analysis manageable and to focus in on those records most likely to reflect a decision made at the firm level, we analyze only worker-cluster transitions including five or more workers. We also exclude records for small predecessor firms (5 employees or less at the time of the transition). Although this cutoff is somewhat arbitrary as well, it is motivated by a desire to limit the impact of small firms, where the movement of a sizable proportion of total employment is a relatively frequent event. After the imposition of these restrictions, the resulting sample size of our analysis dataset is 4,557,451 firm-to-firm, worker-cluster transitions.

1.4 Measuring Firm Dynamics

1.4.1 Classifying Clustered Worker Flows

In order to simplify the analysis and presentation of our results, and in the absence of theoretical guidance, we create a set of classifying rules for the worker-cluster transitions. We first choose a relative threshold that captures the importance of the movement of a cluster of workers to the predecessor firm: the ratio of the number of

transitioning workers to total employment before the transition. The magnitude of this measure can be used to differentiate between different firm events; for example, when a firm dies, virtually all of the employees are likely to transition to a new firm as opposed to a spin-off where a much smaller fraction of the employees leave.

Our calculations of this ratio show that even though each transition involves a flow of at least five workers, the vast majority of the transitions are insignificant, in that they account for less than 10% of the predecessor firm's workforce. In Figure 1.1, we present the frequency distribution of transitions exceeding that proportion. The curve is generally U shaped, with most of the mass at the tails. For example, almost 16% of cluster transitions in scope for Figure 1.1 (greater than 10% of predecessor's employment) account for only 10 to 15% of the predecessor firm's work force. At the other end of the spectrum, a little over 14% of cluster transitions account for 95 to 100% of the predecessor's work force. Since Figure 1.1 shows a dramatic jump upwards in the relative frequency of transitions that contain at least 80% of the predecessor firm's employment, this appears to be a natural cutoff value and therefore defines our first condition.

Condition W1: Significant worker flow from predecessor firm -

80% or more of the predecessor's current employees
transition to the successor.

In order to establish a complementary rule for the impact of the flows into the receiving firm, we perform the same exercise for successor firms and display the results in Figure 1.2. This figure looks remarkably like Figure 1.1, with another distinct spike when the ratio exceeds more than 80% of the successor firm's workforce. Although the 80 to 95% region is not quite as large as before, over 18% of cluster transitions in scope for Figure 1.2 lie in the 95 to 100% range. The net effect is that approximately one-third of the transitions shown in Figure 1.2 contain over 80% of the successor firm's workforce, about the same as the results for the predecessor firms shown in

Figure 1.1.

As a result, we choose our second condition, namely, a significant flow of workers to a successor firm to be the following:

Condition W2: Significant worker flow to successor firm -
80% or more of the successor's employees
after transition came from the predecessor.

1.4.2 Firm Entry and Exit

The set of conditions outlined above enable us to identify significant flows of workers into and out of the firm. However, describing firm dynamics requires more than this, precisely because we are interested in separating out true births and deaths, mergers and acquisitions, as well as outsourcing, firm spin-offs, and breakouts. Many of these events are characterized by firm entry and exit and in both cases we would like to define conditions under which a firm has either ceased to be or has become economically viable.

The challenge with defining an exit is that while exits often occur over an extended period of time, statistical work requires some certainty about the exit date. As Pivetz et al. (2001) point out, it is difficult to pinpoint the exact date a firm shuts down production because in many instances the firm leaves a few staff in place to finalize the administrative details. In order to capture this, we somewhat arbitrarily choose a threshold of five workers (consistent with Census Bureau norms) and define the following condition:

Condition F1: The predecessor exits. This is defined to occur when

- i. the predecessor firm's employment drops below five in each of the two quarters after the transition, and
- ii. the average employment at the predecessor over the course of those two quarters is less than 10% of the predecessor's employment prior to the transition.

A similar challenge, also noted by Pivetz et al. (2001), is associated with identifying firm entry. Often firms will apply for an employer identification number and hire a small staff, but take additional time to become a full-fledged operation. As a result, we also make the following choice to define firm entry:

Condition F2: The successor is an entrant. This is defined to occur when

- i. the successor's employment is fewer than five workers in each of the two quarters prior to the transition, and
- ii. the average employment at the successor over the course of those two quarters is less than 10% of the successor's employment after the transition.

1.4.3 Identifying Firm Dynamics

In this section we pull together the two worker-cluster conditions and our two firm birth and death conditions to derive measures of firm dynamics for the following events: firm births and deaths, changes in firm structure, and across firm relationships such as insourcing and outsourcing. The heuristic discussion above suggests that the combination of the two worker-based and two firm-based conditions lend themselves to the following interpretations:

	Firm Condition	Worker Condition	Possible Interpretation
<i>Predecessor Category</i>			
1	F1 True: firm exit	W1 True: more than 80% of workers go to successor	ID change or merger/acquisition
2	F1 True: firm exit	W1 False: fewer than 80% of workers go to successor	Merger/acquisition or reason unclear
3	F1 False: firm continues	W1 True: more than 80% of workers go to successor	ID change or merger/acquisition
4	F1 False: firm continues	W1 False: fewer than 80% of workers go to successor	Merger/acquisition or reason unclear
<i>Successor Category</i>			
1	F2 True: firm entry	W2 True: more than 80% of workers come from predecessor	ID change or spin-off/breakout
2	F2 True: firm entry	W2 False: fewer than 80% of workers come from predecessor	Spin-off/Breakout or reason unclear
3	F2 False: firm continues	W2 True: more than 80% of workers come from predecessor	ID change or spin-off/breakout
4	F2 False: firm continues	W2 False: fewer than 80% of workers come from predecessor	Spin-off/Breakout or reason unclear

Taking this one step further, the predecessor and successor categories can be combined to summarize firm dynamics in a more detailed fashion as is done in Table 1.2. In part, the discussion of the classification below rely on the relative infrequency of event 3 compared to event 1 and the relative infrequency of event 2 compared to event 4

Approaching each of the measurement challenges in turn, the first is identifying true, versus spurious firm entry and exit. The categories identified in Table 1.2 suggest four sets of transition combinations that might capture such spurious events. Using

row and then column numbers to identify cells, the (1,1), (1,3), (3,1) and (3,3) cells may represent combinations of firm startup and shutdown events that simply reflect ID changes. For example, the combination of events described by the (1,1) cell is as follows:

1. The original administrative entity shut down
2. More than 80% of the workers in the predecessor firm moved to the successor firm
3. The successor firm was a new entrant
4. More than 80% of the workers in the successor firm came from the predecessor firm

This sequence of events strongly suggests that the factors of production in the two firms are virtually the same, and that the flows are the result of either an unlinked administrative edit or a change in ownership. The evidence is less strong for the (1,3) category, where the successor firm was already in existence, but is included because of the timing issue noted by Pivetz et al. (2001), and because it is possible that the few employees in the successor firm prior to the event form part of a shell corporation (United States Department of Labor 2002). It is, of course, possible to invent alternative scenarios, such as in the (1,3) cell whereby a startup firm is able to woo and attract large numbers of workers from another firm. Similar arguments can be made for the combination of events in the (3,1) and (3,3) categories..

Some of the combinations identified in Table 1.2 are consistent with the sequence of events that occur during a merger or acquisition. The (1, 2) cell, for example, identifies a firm shutdown combined with the move of over 80% of workers into a newly born firm, where those workers represent under 80% of the new firm's workforce. The (1,4) cell represents the same transition, albeit for a successor that is an existing firm. Similarly, the (3,2) and (3,4) cells identify a continuing firm that has more than 80% of its workers transitioning to a newly born firm (column 2) or a continuing firm

(column 4) respectively, also suggesting that either a merger or an acquisition took place.

The final task involves identifying insourcing and outsourcing relationships. Typically these types of arrangements involve peripheral firm functions such as payroll or human resources and are therefore not likely to make up more than 80% of either the predecessor or successor firm’s employment. Therefore, the best candidates for these types of transitions are likely to be found in the “reason unclear” cells. In each of these cells, a substantial cluster of workers (at least 5), but fewer than 80%, move from the predecessor to the successor firm and account for fewer than 80% of the workers at the successor firm. From the evidence presented in Table 1.1, clustered worker flows of this size are very rare events (under .2% of worker movements), and this suggests that the underlying transitions may be the result of an insourcing or outsourcing relationship between two employers. We explore this hypothesis in various ways in the remainder of the paper.

1.5 Empirical Analysis of Clustered Worker Flows

1.5.1 Relative Frequency of Transitions

In Table 1.3, we begin by documenting the relative frequency of the 16 worker flow classifications identified in Table 1.2. A brief analysis of Table 1.3 yields a number of interesting results. The most numerically frequent set of links, by an overwhelming margin, occurs in cell (4,4), where the transitioning cluster accounts for less than 80% of both the predecessor and successor firms’ total employment and where neither firm enters or exits. It is worth noting that although the “reason unclear” category dominates Table 1.4 in terms of the number of links, the size of the clusters tend to be relatively small. These results are quite robust: when we regenerated Table 1.4 for minimum flows of 8 and 10 workers, we found that the relative distributions remained essentially unchanged. Not surprisingly, however, the weakest link (cell

{4,4}) decreased in relative importance from the 73.49 reported in Table 1.4 to 72.79 for cluster size 8 and to 70.65 for cluster size 10.

One possible explanation for this phenomenon is outsourcing; another explanation is that the worker-clusters simply represent transfers between establishments under some larger, single corporate identity. Yet a third possibility is that within firm networking leads groups of workers to move to new opportunities together. We explore each of these possibilities later in the paper.

For events defined by either the exit of the predecessor or the entry of the successor firm, the majority of transitions are accounted for by clusters of workers that make up “over 80% of employment.” In contrast, the links that do not involve entry/exit are dominated by the “under 80%” condition. This suggests that clustered flows involving entry/exit are not associated with a fundamental change in firm activity and may well reflect a missed administrative edit or ownership change. This is particularly true for the (1,1) cell.

1.5.2 Using Industry to Shed Light on Firm Dynamics

The sixteen firm relationship classifications identified in Table 1.2 are vastly different in terms of their effect on a firm’s workforce. Some changes imply a large amount of structural change, while others have minimal impact. This suggests that certain types of links such as an ID change, where the factors of production are similar in both the successor and predecessor firm, likely involve two firms in the same industry. Other links, where relatively few workers have transitioned to a new firm, are more likely to be in a new industry.

In particular, outsourcing and insourcing are two activities that are likely to involve a change of industry. For example, if a firm outsources its information technology staff, the new employer is likely to be in the computer services business, which is almost by definition not the primary industry of the predecessor firm. The temporary-

help/personnel supply industry (Standard Industrial Classification 7363) is also an important source of both insourcing and outsourcing transitions. For example, as a firm grows they may choose to insource or permanently hire a cluster of workers from a temporary-help agency.

Table 1.4 reports the results of separating out the sixteen categories identified in Table 1.2 based on whether the link is within or across detailed industry (the four digit SIC industry code). We pay particular attention to the temporary-help/personnel supply industry (7363) and break the table into two panels: panel one for transitions not involving 7363 firms and panel two for transitions involving 7363 firms. In both panels, the percentage of each link category that moves across detailed industry is in bold.

On examining the first panel of Table 1.4, it is clear that the transitions we have identified as ID changes mostly occur within an industry. This finding gives further credence to the assertion that these are administrative edits or ownership changes. The worker flows involving 80% of either the successor's or the predecessor's employment (but not both) are much more likely to cross industry lines. This is particularly true when the break-out SEIN lives on after the link (or, in the case of mergers, the SEIN that absorbs the predecessor was already in existence prior to the link).

Several possible scenarios could account for such a result. For example, suppose firm A performs tasks that fall under industries I and II, but is recorded as an industry I firm. Then, firm A decides it would be more efficient to reorganize into two firms thus producing a breakout. If the resulting firms both have new identifiers, B and C, then two links would be found in the data, one of which is across-industry (I to II) and one of which is within-industry (I to I). However, if the new firm, which takes on the industry I tasks, keeps A as its identifier while the other firm, which takes on the industry II tasks, gets a new identifier, B, then only one link will be formed in the

data (A to B) which will be across-industry (I to II). Under this scenario, one would expect to see a higher percentage of across-industry links in breakouts where the predecessor ID lives on then in breakouts where the predecessor ID dies off (similarly for mergers).

It is also of interest that a large fraction of the “reason unclear” cases involve changes in industry. This finding suggests, perhaps not surprisingly, that outsourcing occurs across industry lines.

In addition, it is informative to note the importance of personnel supply companies, industry 7363. Although very few of the ID changes involve firms in industry 7363, a substantial fraction of the “reason unclear” transitions involve firms in industry 7363. Indeed, of the cells that account for less than 80% of both the predecessor’s and successor’s employment, about one-third of the transitions involve at least one firm in industry 7363.

How can this be interpreted? When the successor only is in industry 7363, it can be reasonably assumed that the predecessor is outsourcing a portion of its payroll to be managed by a personnel supply firm. When only the predecessor is in industry 7363, it is likely that the workers had been working at the successor firm prior to the link quarter as temporary-help workers and were eventually hired for permanent positions either due to their merits or due to an organizational decision by the successor firm. We also see a significant number of changes between two firms within industry 7363 (far more than within any other single industry).

The information provided by combining the upper and lower panels of Table 1.4 provides strong clues for the events underlying many of the “reason unclear” cases. To quantify the success of resolving such cases, consider cell (4,4) in the upper and lower panels of Table 1.4. The cumulative percentage in cell (4,4) from Table 1.3 is 73.49%. According to the first panel of Table 1.4, excluding flows to and from industry 7363, 38.60% of this total was the result of the transition of clusters of workers across

industries, which, as argued above, is likely to be outsourcing. The 14.92% number reported in the second panel is the proportion of the total that reflected transitions of workers to and from firms that were both within industry 7363, and the 22.50% number reflects the proportion of transitions in which at least either the destination or the origin firm was in industry 7363.

Thus, of the 73.49% of transitions accounted for by the (4,4) “reason unclear” cell, we estimate that between 37.42% and 76.02% are associated with outsourcing of some kind, while only the 23.98% that occur within the same industry remains fully unresolved. The larger estimate requires the assumption that all worker-clusters transitioning to a different industry not involving a 7363 firm (38.6%) retain some association with their previous employer. Since this assumption is quite strong, a more reasonable estimate likely lies somewhere between our upper and lower bounds.

The industry switching results suggest that it is of interest to explore in more detail the patterns of industry-to-industry changes between predecessor and successor firms. These changes can happen for a variety of reasons. One obvious possibility is that as businesses evolve, the focus of production may shift or become more specialized, especially after an ownership change. In the case of industrial reorganization, this may be due to branches performing the same tasks they have always performed, but now reporting separately (as a breakout or spin-off) or being absorbed into the reporting of another firm in a different industry (as a merger or an acquisition).

After examining the empirical patterns, most of the worker-clusters that cross industry lines do not stray far from the predecessor firm’s industry. For example, one common pattern is for a cluster of workers to move between 5812 (Eating Places) and 5813 (Drinking Places). Other common patterns include the following: 5311 (Department Stores) to/from 5411 (Grocery Stores); 6021 (Federal Reserve Banks) to/from 6022 (National Commercial Banks); 8011 (Offices and Clinics of Doctors of Medicine) to/from 8062 (General Medical and Surgical Hospitals); 0741 (Veterinary

Services for Livestock) to/from (Veterinary Services for Animal Specialties); 0781 (Landscape Counseling and Planning) to/from 0782 (Lawn and Garden Services). Links between 0761 (Farm Labor Contractors and Crew Leaders) and various other agriculture firms are conceptually somewhat different and are probably the result of outsourcing.

There are also other interesting patterns involving transitions of clusters of workers across very different industry combinations. For example, when we examine the transitions attributed to mergers/acquisitions, a common combination is 1711 (Plumbing, Heating, and Air-Conditioning) to 5812 (Eating Places) and 1731 (Electrical Work) to 5812 (Eating Places). One possible explanation for this phenomenon could be vertical integration; a large dining establishment might decide it would be more efficient to have its own fulltime maintenance staff.

1.6 Impact Analysis

The evidence presented above suggests that worker flows provide important information about firm dynamics. In this section, we document the impact of incorporating such information on establishment-based statistics relating to job creation, job destruction, firm entry, firm exit, and worker accessions and separations. We start by classifying each transition into one of the following categories: ID change, spin-off / breakout, acquisition/merger, and reason unclear. Then in a subsequent pass through the data we look for transitions involving a firm in industry 7363. Any link involving a firm in industry 7363 is reclassified in one of three ways: predecessor firm is in industry 7363 (reflecting insourcing), successor firm is in industry 7363 (reflecting outsourcing), or both firms are in industry 7363. Table 1.5 reports the relative frequency of each of the seven categories by year, both un-weighted and weighted by cluster size.

Several results are evident from an examination of Table 1.5. The first result,

as before, is the importance of the “reason unclear” category in describing clustered worker flows. Much of this activity, however, is likely to reflect some type of insourcing or outsourcing, especially given our interpretation of the industry change results from Table 1.4. The second result is that about one quarter of all clustered worker flows involves temporary-help firms. The third is the up-tick in ID changes during 1996 / 1997, a period during which a substantial reorganization of the QCEW data took place. Finally, the importance of weighting the cells by the size of the worker flow becomes readily apparent. Such weighting reduces the importance of the “reason unclear” category while substantially increasing the contribution of the first three columns. This pattern is not surprising in light of our previous discussion; the first three columns are more likely to involve the entry and exit of a complete business where the cluster size is inherently large.

Table 1.6 reports the impact of these links on six different labor market dynamics measures. The first two measures considered are worker accessions (accessions are defined as workers who are not employed with an SEIN in period $q-1$ but are employed with the SEIN in period q) and worker separations (worker employed with SEIN in period q but not in $q+1$). We also include measures of firm dynamics: firm entry (firm is counted as an entrant if SEIN had zero employment in period $q-1$ and positive employment in period q) and firm exit (SEIN had positive employment in period $q-1$ and zero employment in period q). Finally, we include measures of job flows: job creation (increase in employment from period $q-1$ to q for new firms or existing firms that increased employment) and job destruction (absolute magnitude of reduction in employment for firms that decreased employment or died, more detailed definitions for each of these measures are provided at <http://lehd.dsd.census.gov>).

A significant percentage of apparent worker flows (ranging from about 10% to 13%) are a result of clustered worker flows, due in a large part to the “reason unclear” category, and the results presented in Table 1.6 reflect only the impact of suppressing

clustered worker flows. The predecessor-successor links present in the QCEW data were not used. We explore the relationship between the QCEW predecessor-successor links and the UI worker-cluster links in the next section

The Quarterly Workforce Indicators (QWI) generated by the LEHD program, suppress the strongest links (columns 1 through 3) which account for anywhere from 2.3% to 3.1% of total worker flows in this sample. These columns are used to suppress flows because an ownership change or an administrative edit does not represent an actual gain or loss of jobs

Job creation and destruction are slightly more sensitive than worker flows to suppressions for columns 1-3, but are much less sensitive to the suppressions for columns 4-7. As before, the difference in results for 1997 illustrates the importance of taking these links into account when working with job flows. Row totals are not included for overall job creation/destruction because the non-linear nature of these variables precludes meaningful addition across link categories. The latter is a technical point but stems from the fact that job creation reflects expanding businesses and job destruction reflects contracting businesses. Thus, the suppression of a worker flow could change a business from being a contracting business to an expanding business, and the resulting movement across categories makes the calculation of row totals misleading

A little surprisingly, we even see job creation and destruction increase slightly after flow suppression for links between standard and 7363 firms (in columns 5-6 we see negative percentage differences implying increases after flow suppression). This is probably the result of high worker turnover at these firms. In this case, small net employment changes can become large net employment changes after flow suppression.

In order to understand this result, imagine firm A is an employment leasing firm that is observed to have 100 separations and 90 accessions in quarter q in the UI data before suppressions, but then it is found that firm B had 90 workers shifted from its own payroll to be managed by firm A in that quarter. Moreover, suppose in that

same period, firm B hired 50 additional workers onto its own payroll. Originally, job destruction at firm A would be 10 (100-90) and job destruction at firm B would be 40 (90-50). After suppression, however, job destruction at firm A would increase to 100 (a difference of 90) but only drop by 40 (from 40 to 0) at firm B. Moreover, firm B went from no job creation before suppression, to 50 net jobs created after suppression, while firm A's job creation remained at zero.

It is also worth noting that the aggregate change in job creation and destruction for the 7363 columns is very small, which suggests that the unusual cases are largely offset by the more intuitive cases where job creation or destruction decreases after flow suppression.

Also of great interest is the approximately 4% of firm entry and exit that is accounted for by the so-called "reason unclear" links. This finding suggests that the factors leading to firm entry are associated with the factors leading a cluster of workers at an existing firm to start a new firm. Given the important role of firm entry in economic growth, this finding raises a variety of interesting research questions about the impact of clustered flows of workers that we leave for future research. We do note, however, that this result is consistent with the finding in prior sections that many of the "reason unclear" cases involve switches in industry. Putting the two pieces together suggests a potentially important role for worker-clusters in firm entry. Along those lines, it suggests that at least some new firms may have a pre-history in that a group of the workers at the new firm have been coworkers at another firm in the past. Such links between existing firms and new firms raises a rich set of questions about firm dynamics and the factors that lead to business formation.

1.7 External Validation Of the Worker Flow Approach

We use two external sources of validation: the BLS QCEW data and the Census Business Register. In this section we compare the results of our worker flow approach

with those derived from administrative and survey data, respectively.

1.7.1 Successor/Predecessor Information from the QCEW Program

The QCEW program identifies predecessor/successor relationships in partnership with the states that assemble the data. When firms change ownership, the original firm is designated the predecessor, and the new firm is designated the successor. These “official” links reflect organizational (e.g., change of ownership) changes reported to the QCEW survey staff by the businesses involved. It is worth noting that while our approach is likely to capture many ownership changes, administrative edits for smaller employers (especially those with less than 5 employees) will likely be better captured by QCEW successor/predecessor links. Table 1.7 contains the results of comparing the worker flow approach with the successor/predecessor codes derived from the QCEW file. This table provides a broad overview of how the data match up, split into pre and post 1998 periods to reduce the impact of the 1997 change in the processing of the QCEW data: namely that states made a large effort to improve the reporting of employment and payroll by establishment for multi-unit firms. This resulted in many changes in administrative identifiers, both at the establishment and firm level. Predecessor / Successor links were created to capture these changes, but this type of widespread administrative change would likely distort our analysis and is therefore excluded.

Statistics are computed on both an un-weighted and flow-weighted basis, and describe whether the link is present only in the UI (clustered worker flow) or in both the UI and the QCEW. On an un-weighted basis, the vast majority (about 94%) of the UI wage record links are not found in the QCEW. On a weighted basis, this pattern still holds although the percentage drops to 74%. Almost all of the links categorized as “reason uncertain” or transitions to/from Personnel Supply Service firms occur in the set found only in the UI wage records. When the link is present in both data sources,

we find a very high percentage of links that appear to be ID changes, acquisitions, and spin-offs, and find very few links to temporary-help/personnel supply firms. Thus, much of what we have characterized as outsourcing is apparently not captured in the QCEW predecessor-successor links.

Looking at the first three categories of linkages, all of which result in the suppression of worker and job flow statistics, roughly 8% of the UI links on an un-weighted basis and about 18% of the UI links on an employment-weighted basis suppress flows that are not accounted for when using only the QCEW flags (either the QCEW links do not exist or they occur in a different quarter). Thus, there is non-trivial value-added from using worker flows even for the sole purpose of fixing missing links from the QCEW. The QWI flow statistics generated by LEHD also suppress flows resulting from links that agree exactly between the QCEW and UI-based links.

There is more information available when we consider the timing of the link and how this might differ in the UI wage records and the QCEW. Figure 1.3 shows the distribution of the difference in timing for those links found in both files.

The majority of this subset only disagrees by one quarter, but it appears that when there is a disagreement, the UI link tends to take place after the QCEW link. This makes sense since workers may still receive money (severance pay or bonuses) from an employer after their actual separation. The consequence of this is that workers will appear in the UI wage records as matched to the old employer ID after the employer has ceased reporting those workers in its QCEW employment counts.

To sum up, the QCEW links and the UI links overlap the most where they should, namely for ID changes, merger and acquisitions, and breakouts/spin-offs. The UI links and the QCEW links do not overlap much in the categories where the evidence suggests there is an economic change in the structure of the business or in the nature of the employer-employee relationship. In these latter cases, it is an open question as to whether worker and job flow statistics should be adjusted.

1.7.2 Match to the Census Business Register

One possible reason for the observed firm-to-firm transitions is that they reflect administrative changes or transfers within a broader firm structure, particularly since the SEIN may or may not directly correspond to an individual firm. We investigated this by matching the QCEW files to the Census Business Register. The Business Register tracks changes in ownership using parent/subsidiary information received from the IRS as well as from the Company Ownership Survey. We report these results in Table 1.8, which includes only the links in which both SEINs matched to the Business Register

An investigation of Table 1.8 demonstrates that about 85% of the links identified as ID changes and Mergers/Acquisitions do in fact reflect different firm relationships. The results are higher for Breakouts/Spin-offs where 95% reflect different firm relationships, implying in all three cases that the links represent mostly legitimate ownership changes. Also, about 5% of the reason uncertain category can be explained as transfers of employees within a larger corporate structure. Almost all of the temporary-help flows are across different economic entities.

1.8 Conclusion

Our new approach of following clustered flows of workers has uncovered a previously unknown set of facts about firm transitions. Our findings fall into two broad categories. First, we show that there are technical reasons to use a worker flow approach to improve linkages in longitudinal business databases. A small but important fraction of the worker-cluster, predecessor-successor links appear to “fix” problems in the administrative data for the purpose of generating job and worker flow statistics or other related measures of firm dynamics.

Second, we find that following clustered flows of workers provides important conceptual insights into the changing structure of businesses and the changing structure

of employer-employee relationships. In particular, we show that after abstracting from the ID changes, most of the worker-cluster flows involve changes in industry, and many involve movements into and out of personnel supply firms. Both of the latter reflect more than just an ID change and reflect some richer change in firm structure or employer-employee relationship. Having said this, depending on the question at hand, breaking out measures of worker and job flows along these dimensions is likely to be important. For example, a clustered flow of workers to an employee-leasing firm may not involve workers changing their production location or activities even though the workers involved have undergone an important change in the employer-employee relationship.

Another interesting facet of the clustered flows of workers is that a nontrivial fraction of firm entry is associated with these flows. This finding suggests that one of the paths for firm entry is a group of workers at an existing firm deciding to start a new firm. A variety of interesting questions immediately arise from this finding. One interesting possible line of inquiry is the relationship between clustered flows of workers and the transfer of knowledge. A related line of inquiry is whether a firm that is created as a result of a cluster of workers leaving another firm is more likely to survive. We leave such interesting questions for future research, but for now our findings suggest a whole new avenue for studying and analyzing the factors underlying firm entry.

Number in Cluster	Percent	Cumulative Percent	Percent	Cumulative Percent
	Un-weighted		Weighted by Cluster Size	
	1	98.07%	98.07%	90.74%
2	1.17%	99.24%	2.17%	92.91%
3	0.33%	99.57%	0.92%	93.82%
4	0.15%	99.72%	0.56%	94.38%
5	0.08%	99.80%	0.37%	94.75%
6	0.05%	99.85%	0.28%	95.02%
7	0.03%	99.88%	0.22%	95.25%
8	0.02%	99.90%	0.18%	95.43%
9	0.02%	99.92%	0.16%	95.59%
10	0.01%	99.93%	0.14%	95.72%
>10	0.07%	100.00%	4.28%	100.00%

Notes: A total of 2,668,127,897 firm-to-firm, worker-cluster transitions occurred over the sample period.

Link Description		Successor Category			
		1. 80% of Succ comes from Pred and Succ is entrant	2. Less than 80% of Succ comes from Pred and Succ is entrant	3. 80% of Succ comes from Pred and Succ was in existence	4. Less than 80% of Succ comes from Pred and Succ was in existence
Predecessor Category	1. 80% of Pred. Moves to Succ and Pred exits	ID change	Acquisition / Merger	ID change	Acquisition / Merger
	2. Less than 80% of Pred moves to Succ and Pred exits	Spin-off / Breakout	Reason unclear	Spin-off / Breakout	Reason unclear
	3. 80% of Pred moves to Succ and Pred lives on	ID change	Acquisition / Merger	ID change	Acquisition / Merger
	4. Less than 80% of Pred moves to Succ and Pred lives on	Spin-off / Breakout	Reason unclear	Spin-off / Breakout	Reason Unclear

Table 3: Relative Frequency of Successor/Predecessor Combinations						
		Successor Category				
		<i>1. 80% of Succ come from Pred and Succ is born</i>	<i>2. Less than 80% of Succ comes from Pred and Succ is born</i>	<i>3. 80% of Succ comes from Pred and Succ was in existence</i>	<i>4. Less than 80% of Succ comes from Pred and Succ was in existence</i>	Total
Predecessor Category	<i>1. 80% of Pred. Moves to Succ and Pred dies</i>	ID change 3.12 44.35 55.80	Acquisition / Merger 1.42 20.16 13.98	ID change 0.24 3.47 13.81	Acquisition / Merger 2.25 32.02 2.73	7.03
	<i>2. Less than 80% of Pred moves to Succ and Pred dies</i>	Spin-off / Breakout 0.98 10.23 17.53	Reason unclear 1.63 17.04 16.08	Spin-off / Breakout 0.29 3.00 16.28	Reason unclear 6.67 69.72 8.09	9.57
	<i>3. 80% of Pred moves to Succ and Pred lives on</i>	ID change 0.10 38.85 1.87	Acquisition / Merger 0.05 19.63 0.52	ID change 0.03 9.69 1.48	Acquisition / Merger 0.09 31.84 0.10	0.27
	<i>4. Less than 80% of Pred moves to Succ and Pred lives on</i>	Spin-off / Breakout 1.39 1.67 24.80	Reason unclear 7.04 8.47 69.42	Spin-off / Breakout 1.21 1.45 68.44	Reason unclear 73.49 88.41 89.08	83.13
	Total	5.59	10.14	1.77	82.50	100

Notes: The first element of each cell represents the proportion of all transitions; the second element represents the proportion of transitions in the row; the third element is the proportion of transition in the column. The total number of firm-to-firm cluster transitions is 4,557,451.

Table 4: Panel 1:					
Successor/Predecessor Comparisons When Transitions do not Involve 7363 Firms					
		Successor Category			
		<i>1. 80% of Succ comes from Pred and Succ is born</i>	<i>2. Less than 80% of Succ comes from Pred and Succ is born</i>	<i>3. 80% of Succ comes from Pred and Succ was in existence</i>	<i>4. Less than 80% of Succ comes from Pred and Succ was in existence</i>
Predecessor Category	1. 80% of Pred. Moves to Succ and Pred dies	ID change 74.14 24.81	Acquisition / Merger 61.23 32.11	ID change 53.66 42.77	Acquisition / Merger 35.48 48.67
	2. Less than 80% of Pred moves to Succ and Pred dies	Spin-off / Breakout 59.63 35.19	Reason unclear 46.81 35.20	Spin-off / Breakout 34.54 53.72	Reason unclear 27.92 40.78
	3. 80% of Pred moves to Succ and Pred lives on	ID change 59.15 38.08	Acquisition / Merger 48.57 44.34	ID change 53.78 44.20	Acquisition / Merger 35.19 49.31
	4. Less than 80% of Pred moves to Succ and Pred lives on	Spin-off / Breakout 38.56 50.24	Reason unclear 28.12 40.34	Spin-off / Breakout 29.48 54.94	Reason unclear 23.98 38.60
Notes: The first element reflects the proportion of transitions that occurred within the same industry (4 digit SIC code); the second element the proportion that crossed industry lines. The numerator of each proportion reflects only transitions that did not involve firms in industry 7363, while the denominator includes all transitions for that cell in panel 1 and 2. This implies that the proportions in cell (i,j) across both panel 1 and 2 sum to 100.					

Table 4: Panel 2:					
Successor/Predecessor Comparisons When Transitions Involve 7363 Firms					
		<i>1. 80% of Succ comes from Pred and Succ is born</i>	<i>2. Less than 80% of Succ comes from Pred and Succ is born</i>	<i>3. 80% of Succ comes from Pred and Succ was in existence</i>	<i>4. Less than 80% of Succ comes from Pred and Succ was in existence</i>
Predecessor Category	1. 80% of Pred. Moves to Succ and Pred dies	ID change 0.40 0.66	Acquisition / Merger 0.93 5.73	ID change 0.61 2.95	Acquisition / Merger 0.68 15.17
	2. Less than 80% of Pred moves to Succ and Pred dies	Spin-off / Breakout 1.02 4.17	Reason unclear 6.53 11.46	Spin-off / Breakout 4.18 7.57	Reason unclear 10.81 20.49
	3. 80% of Pred moves to Succ and Pred lives on	ID change 1.15 1.61	Acquisition / Merger 1.66 5.43	ID change 0.42 1.60	Acquisition / Merger 1.38 14.12
	4. Less than 80% of Pred moves to Succ and Pred lives on	Spin-off / Breakout 1.27 9.93	Reason unclear 12.63 18.92	Spin-off / Breakout 7.11 8.47	Reason unclear 14.92 22.50
Notes: The first element reflects the proportion of transitions that occurred within the same industry (4 digit SIC code); the second element the proportion that crossed industry lines. The numerator of each proportion reflects only transitions that involve firms in industry 7363, while the denominator includes all transitions for that cell in panel 1 and 2. This implies that the proportions in cell (i,j) across both panel 1 and 2 sum to 100.					

Table 5: Successor/Predecessor Relationships by Year							
	ID Change	Merger / Acquisition	Spin-off / Breakout	Reason Unclear	Insourcing 7363 employees to regular payroll	Outsourcing regular employees to 7363 payroll	Transition between two 7363 firms
1993	3.76 12.90	3.62 8.94	4.73 9.15	62.71 49.56	8.42 5.45	6.86 4.42	9.90 9.58
1994	3.33 11.76	3.29 8.46	3.95 7.38	60.72 49.79	9.81 6.69	7.79 5.21	11.11 10.69
1995	3.36 12.48	3.41 8.60	3.54 7.03	58.28 45.98	10.64 7.79	8.51 5.91	12.27 12.22
1996	3.15 16.92	3.44 8.85	3.50 5.67	58.10 43.98	10.71 7.12	8.37 4.97	12.72 12.51
1997	4.81 18.31	3.52 8.82	3.00 5.48	55.65 41.99	11.25 7.43	8.79 5.52	12.97 12.47
1998	3.00 12.46	3.37 9.50	2.77 5.35	55.04 42.25	12.37 8.32	9.92 6.52	13.52 15.60
1999	2.69 11.76	2.93 8.02	2.59 5.46	54.30 42.50	12.30 8.58	10.73 7.23	14.47 16.45
2000	2.74 12.43	3.02 9.38	2.68 5.50	53.40 40.42	13.29 9.28	10.35 6.77	14.52 16.21
2001	3.37 14.35	3.15 8.27	3.41 7.05	56.73 41.02	11.35 7.80	9.19 7.49	12.80 14.02
Total	3.30 13.96	3.27 8.76	3.20 6.56	56.51 43.19	11.46 7.78	9.23 6.21	13.04 13.55

Notes: The elements in each cell reflect the proportion in each row and the columns in a row sum to 100. The first cell element is un-weighted; the second is weighted by the size of the flow.

Table 6: Effect of Successor/Predecessor Transitions on Selected Job-flow Statistics								
Percentage difference in job-flow statistics caused by suppression of worker flows due to UI successor/predecessor links								
	ID Change	Merge / Acquisition	Breakout / Spin-off	Reason Uncertain	Insourcing 7363 employees to regular payroll	Outsourcing regular employees to 7363 payroll	Transition between two 7363 firms	Total
WORKER ACCESSIONS								
1993	0.9%	0.7%	0.9%	5.8%	0.6%	0.5%	1.1%	10.4%
1994	0.8%	0.7%	0.8%	6.2%	0.8%	0.6%	1.3%	11.2%
1995	0.9%	0.7%	0.7%	6.0%	1.0%	0.8%	1.5%	11.6%
1996	1.5%	0.9%	0.7%	6.4%	1.0%	0.7%	1.7%	13.0%
1997	1.7%	0.8%	0.6%	5.5%	1.0%	0.7%	1.6%	11.9%
1998	1.1%	1.0%	0.6%	6.2%	1.2%	1.0%	2.1%	13.1%
1999	1.0%	0.8%	0.6%	6.0%	1.2%	1.0%	2.2%	12.9%
2000	1.1%	1.0%	0.6%	5.8%	1.4%	1.0%	2.3%	13.0%
2001	1.1%	0.7%	0.7%	5.5%	1.0%	1.0%	1.8%	11.7%
WORKER SEPARATIONS								
1993	1.0%	0.8%	1.0%	6.3%	0.7%	0.6%	1.2%	11.4%
1994	0.8%	0.7%	0.8%	6.1%	0.8%	0.6%	1.3%	11.2%
1995	0.9%	0.7%	0.7%	5.9%	1.0%	0.8%	1.5%	11.5%
1996	1.4%	0.8%	0.6%	6.0%	1.0%	0.7%	1.6%	12.1%
1997	1.9%	0.9%	0.7%	6.1%	1.1%	0.8%	1.8%	13.3%
1998	1.1%	1.0%	0.6%	6.3%	1.3%	1.0%	2.1%	13.3%
1999	1.0%	0.8%	0.6%	6.0%	1.2%	1.0%	2.2%	12.9%
2000	1.1%	1.0%	0.6%	5.9%	1.4%	1.0%	2.3%	13.3%
2001	1.0%	0.7%	0.7%	5.4%	1.0%	1.0%	1.7%	11.5%
FIRM ENTRY								
1993	0.9%	0.5%	0.9%	3.1%	0.3%	0.3%	0.3%	6.3%
1994	1.2%	0.6%	1.1%	3.7%	0.4%	0.3%	0.3%	7.4%
1995	1.2%	0.6%	1.1%	3.4%	0.4%	0.4%	0.5%	7.6%
1996	1.3%	0.7%	1.1%	4.1%	0.5%	0.5%	0.7%	9.0%
1997	1.9%	0.6%	0.9%	3.1%	0.4%	0.4%	0.6%	7.9%
1998	1.4%	0.7%	1.1%	3.8%	0.6%	0.7%	0.8%	9.3%
1999	1.2%	0.6%	1.1%	3.8%	0.6%	0.8%	0.8%	8.9%
2000	1.3%	0.6%	1.1%	3.9%	0.8%	0.7%	0.9%	9.4%
2001	1.4%	0.6%	1.2%	3.3%	0.6%	0.7%	0.7%	8.5%
FIRM EXIT								
1993	1.3%	1.5%	0.7%	3.7%	0.4%	0.6%	0.3%	8.5%
1994	1.2%	1.3%	0.6%	3.3%	0.4%	0.5%	0.3%	7.6%
1995	1.5%	1.7%	0.7%	3.8%	0.5%	0.6%	0.5%	9.3%
1996	1.1%	1.4%	0.6%	3.3%	0.6%	0.5%	0.5%	8.0%
1997	2.8%	2.0%	0.7%	4.1%	0.6%	0.7%	0.7%	11.4%
1998	1.6%	2.0%	0.6%	3.9%	1.1%	0.9%	0.8%	10.8%
1999	1.4%	1.8%	0.5%	4.3%	0.7%	0.9%	0.7%	10.3%
2000	1.4%	1.8%	0.5%	3.6%	0.8%	0.9%	0.7%	9.6%
2001	1.5%	1.5%	0.6%	3.3%	0.6%	0.8%	0.5%	8.8%

(Continued on next page)

Table 6 (continued)

ID	Merge / Change	Merge / Acquisition	Breakout / Spin-off	Reason Uncertain	Insourcing 7363 employees to regular payroll	Outsourcing regular employees to 7363 payroll	Transition between two 7363 firms	Total
JOB CREATION								
1993	1.8%	1.6%	1.0%	2.5%	-0.1%	0.0%	0.2%	-
1994	2.7%	2.2%	0.9%	3.4%	-0.3%	-0.1%	0.4%	-
1995	2.3%	1.9%	0.7%	3.3%	-0.3%	-0.4%	0.4%	-
1996	3.3%	2.3%	1.1%	3.9%	0.1%	-0.2%	0.6%	-
1997	5.3%	1.8%	0.8%	2.9%	-0.2%	-0.3%	0.5%	-
1998	3.4%	2.8%	1.0%	4.2%	-0.3%	-0.7%	0.5%	-
1999	3.1%	2.7%	1.3%	4.0%	-0.5%	-0.5%	0.9%	-
2000	3.4%	2.7%	1.2%	4.1%	-0.3%	-0.5%	0.8%	-
2001	3.8%	2.7%	1.3%	4.4%	0.3%	-0.1%	1.3%	-
JOB DESTRUCTION								
1993	2.8%	2.5%	1.4%	3.4%	-0.1%	0.1%	0.6%	-
1994	2.6%	2.0%	1.0%	3.1%	-0.1%	0.0%	0.4%	-
1995	2.8%	2.0%	1.0%	3.3%	0.1%	-0.2%	0.6%	-
1996	3.9%	1.9%	0.8%	3.0%	0.0%	-0.1%	0.7%	-
1997	6.3%	2.9%	1.3%	3.8%	-0.3%	-0.3%	0.7%	-
1998	3.7%	3.2%	1.3%	4.1%	-0.3%	-0.5%	1.1%	-
1999	3.1%	2.4%	1.3%	3.9%	-0.5%	-0.5%	0.8%	-
2000	3.7%	3.3%	1.3%	4.1%	-0.2%	-0.7%	0.9%	-
2001	3.2%	1.9%	1.3%	3.3%	0.0%	0.2%	0.8%	-

Notes: The total column is not presented for job creation and destruction due to conceptual issues associated with summing across the columns. See the text for a more detailed explanation.

Table 7: Clustered Worker Flow Links Compared with QCEW Links										
<i>ID Change</i>	<i>Merge / Acquisition</i>	<i>Breakout / Spin-off</i>	<i>Reason Uncertain</i>	<i>Insourcing 7363 employees to regular payroll</i>	<i>Outsourcing regular employees to 7363 payroll</i>	<i>Transition between two 7363 firms</i>	<i>Total</i>			
Prior to 1998										
<i>SEIN Pair Found Only in UI Links</i>	1.94 4.39	2.12 3.42	3.26 4.74	57.66 42.92	9.93 6.51	7.81 5.07	11.50 9.94	94.22 76.99		
<i>SEIN Pair Found in Both UI And ES202 Links and Agree on Quarter</i>	1.17 5.62	0.87 3.08	0.37 1.27	0.59 1.14	0.01 0.12	0.02 0.10	0.03 0.54	3.07 10.94		
<i>SEIN Pair Found in Both UI And ES202 Links but Disagree on Quarter</i>	0.71 4.85	0.55 2.45	0.33 1.22	1.01 1.67	0.02 0.12	0.02 0.10	0.06 0.54	2.71 10.94		
<i>Total</i>	3.83 14.86	3.55 8.95	3.96 7.33	59.26 45.73	9.96 6.75	7.85 5.31	11.60 11.07	100		
After 1998										
<i>SEIN Pair Found Only in UI Links</i>	1.11 2.20	1.72 2.84	2.23 2.62	53.16 37.89	12.05 8.02	9.75 6.31	13.67 13.34	93.71 73.22		
<i>SEIN Pair Found in Both UI And ES202 Links and Agree on Quarter</i>	0.93 5.61	0.63 2.85	0.33 1.61	0.46 1.15	0.02 0.25	0.04 0.22	0.04 1.18	2.46 12.87		
<i>SEIN Pair Found in Both UI And ES202 Links but Disagree on Quarter</i>	1.12 5.47	0.83 2.91	0.47 1.72	1.19 2.18	0.05 0.30	0.07 0.37	0.09 0.95	3.83 13.91		
<i>Total</i>	3.16 13.27	3.18 8.60	3.03 5.96	54.82 41.22	12.13 9.57	9.86 6.90	13.81 15.47	100		
Notes: The first element reflects the proportion of links in each cell as a proportion of all links; the second element reflects the proportion of total flows.										

Table 8: Transitions Type within and across Firm

	ID Change	Merge / Acquisition	Breakout / Spin-off	Reason Uncertain	Hiring 7363 employees to regular payroll	Outsourcing regular employees to 7363 payroll	Transition between two 7363 firms	Total
Different Firm on Census Business Register	1.40%	2.20%	2.36%	55.59%	11.67%	9.42%	13.85%	96.49%
Same Firm on Census Business Register	0.23%	0.40%	0.14%	2.58%	0.04%	0.03%	0.09%	3.51%
Elements of Cell: Percent of all links that matched to Business Register								

Figure 1

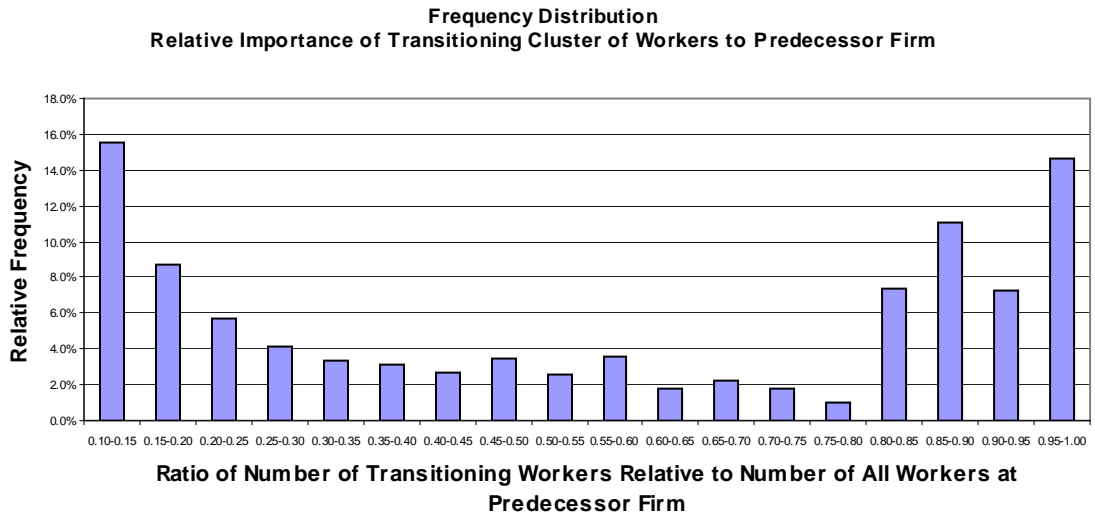


Figure 2

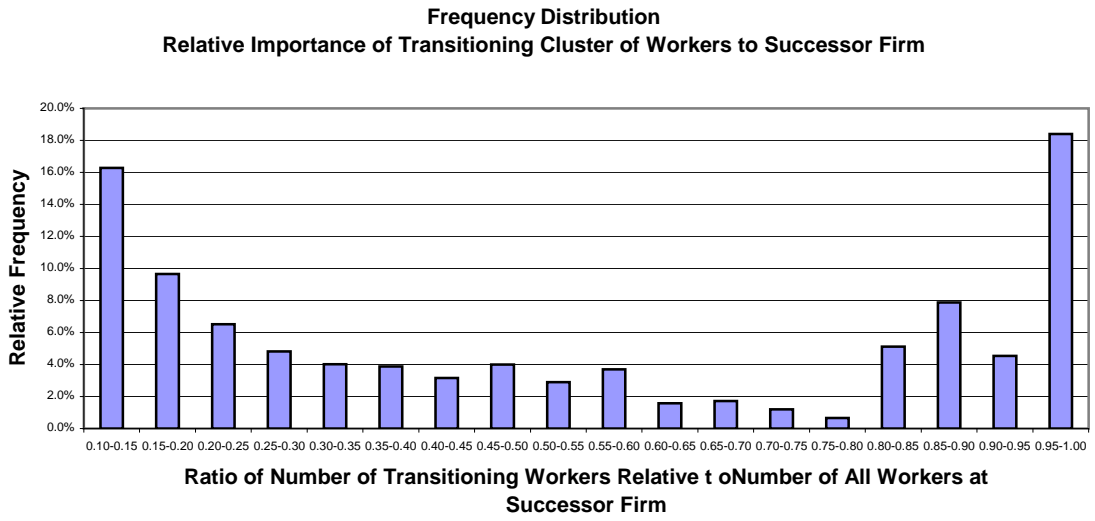
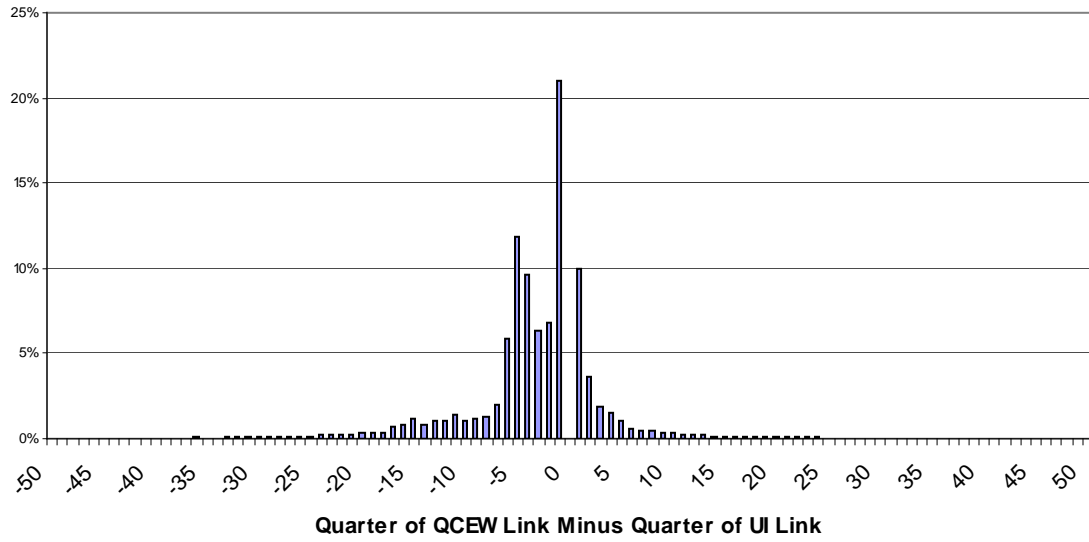


Figure 3

When QCEW and UI Agree on SEIN Pair But Not the Timing of the Link



The Effects of Mergers on Workers' Earnings and Employment*

2.1 Introduction

The costs of industrial reorganizations to experienced workers have been a concern in the debate over whether takeover activity should be restricted. However, these costs are tough to quantify, and the debate instead tends to focus on the implications of restructuring on productivity, and whether the gains for the firms in question come from tax avoidance and imperfect information or actual efficiency/productivity enhancements (Jensen 1988). The major concern, of course, is that such firm restructuring often goes hand-in-hand with significant worker turnover, and the long term earnings losses experienced by displaced workers has been well-documented in the labor economics literature. However, the displaced worker literature has typically focused on mass-layoff events which tend to be related to firm deaths, not mergers

*This document reports the results of research and analysis undertaken by the U.S. Census Bureau staff. This document is released to inform interested parties of research and to encourage discussion. This research is a part of the U.S. Census Bureau's Longitudinal Employer-Household Dynamics Program (LEHD), which is partially supported by the National Science Foundation Grants SES-9978093 and SES-0427889 to Cornell University (Cornell Institute for Social and Economic Research), the National Institute on Aging Grant R01 AG018854, and the Alfred P. Sloan Foundation. The views expressed on statistical, methodological, or technical issues are those of the author and not necessarily those of the U.S. Census Bureau, its program sponsors or data providers. Some or all of the data used in this paper are confidential data from the LEHD Program. The U.S. Census Bureau supports external researchers' use of these data through the Research Data Centers (see www.ces.census.gov). For other questions regarding the data, please contact Jeremy S. Wu, Program Manager, U.S. Census Bureau, LEHD Program, Demographic Surveys Division, FOB 3, Room 2138, 4700 Silver Hill Rd., Suitland, MD 20233, USA. (Jeremy.S.Wu@census.gov <http://lehd.dsd.census.gov>). I thank John Haltiwanger, John Abowd, Seth Sanders, John Shea, Simon Woodcock, Martha Stinson, and members of LEHD Program staff for helpful comments and suggestions.

and acquisitions. Moreover, it is an open question as to what the effects on earnings are for the workers caught in an industrial reorganization who do not lose their jobs. Similar to the methods used in recent displaced worker literature to identify mass layoffs, this paper uses linked employer-employee administrative data to help identify mergers and acquisitions and examine their long-term effects on earnings of workers at these firms.

The primary obstacles to the analysis of the impact of mergers on labor have been the lack of extensive, longitudinal data on employees and firms as well as the difficulty in identifying acquisitions in such data. This paper combines data from the Longitudinal Employer-Household Dynamics (LEHD) program of the U.S. Census Bureau with data from the Federal Trade Commission (FTC) of the Department of Commerce to overcome these issues. Workers at both the acquired and acquiring firms are observed over time and compared to workers at firms that do not experience a major restructuring in the same time period. The findings suggest that the wages of workers at restructuring firms are actually a little higher than their counterparts, but the turnover is significantly higher starting slightly before the reference period and persisting for a long time afterwards. The paper proceeds as follows: section 2.2 gives a little background to the discussion, section 2.3 describes the various data used in the analysis and how it was assembled, section 2.4 explains the methods of analyzing the data, section 2.5 reports the results, and section 2.6 offers some conclusions.

2.2 Background

Much of the study of the consequences of mergers and acquisitions focuses on productivity questions. In his paper, "Takeovers: Their Causes and Consequences," Michael Jensen (1988) summarizes many of the results in this literature on the value of the firm, shareholder behavior, and managerial incentives. He acknowledges that these "corporate control transactions and the restructurings that often accompany

them are frequently wrenching events in the lives of those linked to the involved organizations," but most of the analysis is directed at efficiency and productivity in the market. Because of the lack of good data on the workers at restructuring firms, the focus turns to concerns that the gains from acquisitions are illusory and based largely on tax incentives or short-term benefits. Jensen argues that the literature shows acquisitions and even the threat of takeover have real, positive benefits for the value of the firm and place heavy pressure on managers to maintain efficiency. However, he also contends that this same pressure is an incentive to form special interest groups supporting governmental restrictions on takeover activity.

Charles Brown and James Medoff (1987) use Michigan ES-202 data compiled by the Michigan Employment Security Commission to analyze mergers and acquisitions and their impact on labor. These data are quarterly data at the firm level, and they contain a field for identifying a predecessor or a successor firm in a given quarter. Brown and Medoff use this predecessor/successor information to identify acquisitions, but must use intuitive rules on overall firm employment counts to decide whether the workforce of the predecessor was acquired by the successor, in which case the event is deemed a merger. They find small negative changes in the average wage and slight increases in overall employment after a merger. However, because they only have firm-level employment and payroll, they cannot observe the compositional changes that may be driving the wage results.

Jagadeesh Gokhale, Erica Groshen, and David Neumark (1995) use smaller but more detailed survey data to explore how hostile takeovers affect implicit contracts, such as job security and steeper wage profiles, despite having little impact on current wages. Their data comes from the Community Salary Survey collected by the Federal Reserve Bank of Cleveland. The data covers select employers in the cities of Cleveland, Cincinnati, and Pittsburgh between 1980 and 1991. They identify mergers by linking employers who report ownership changes to hostile tender offers published by W.J.

Grimms and Co.'s *Mergestat Review* and the *Wall Street Journal Index*. They end up with a small set of eight hostile takeovers, but they have longitudinal information on compensation for workers in a large number of occupations at these firms. The results show that wage differentials increase after hostile takeovers. Moreover, they find that job security and returns to seniority decrease for the more senior workers.

Clearly, there is much more to be learned about the effects of mergers and acquisitions (not just hostile) on labor market outcomes for a more representative sample of the U.S. and for the entire workforce at these firms. Since much of the literature on this topic suffers from data limitations, this paper turns to the recent displaced worker literature for guidance on how to approach this problem with large, linked employer-employee data. Louis Jacobson, Robert LaLonde, and Daniel Sullivan (1993) use Pennsylvania Unemployment Insurance (UI) wage record data to identify mass layoffs and examine their effects on the long term earnings of workers. By observing large clusters of workers all separating from a firm in one quarter, they deduce that the worker-firm separations were not voluntary quits, and can then analyze the affected workers' earnings in a large window around the event. Because of the size of the sample they are able to study, they can estimate large regression equations with individual earnings components and a series of dummies for the quarter relative to a layoff. They find that the average worker caught in a mass layoff begins losing earnings a few quarters before the layoff, then takes a large hit at the time of the layoff, followed by some recovery but never achieving previous earnings levels.

Other papers since then have used similar data to advance the study of mass layoffs. Robert Schoeni and Michael Dardia (2000) use California administrative data, controlling for possible ownership changes when identifying mass layoffs and looking in more detail at the distribution of earnings losses. Paul Lengerhmann and Lars Vilhuber (2001) use the same LEHD administrative data that this paper uses to look at the distribution of human capital among the job leavers in the time period

leading up to the mass layoff. This paper will attempt to use some similar techniques from this branch of displaced worker literature to analyze labor market outcomes for workers caught in the middle of restructuring firms. There are certainly some similar questions to be asked since many of these restructurings go hand-in-hand with large clusters of job separations. On the other hand, there are also some new questions in that it is interesting to ask what happens to the workers who keep their jobs and if outcomes differ based on whether the worker started out at the acquired or the acquiring firm.

2.3 Data

2.3.1 General Overview

This paper makes use of administrative, linked employee-employer data put together by the LEHD program at the U.S. Census Bureau and combines it with public-use data on mergers and acquisitions provided by the FTC, and labor force data collected by the Census Bureau's Survey of Income and Program Participation (SIPP). The LEHD data provide the basis for the analysis of workers' labor market outcomes over several years, and the worker flows observed in these data are used to construct a set of candidate firm-pairs for possible merger/acquisition events. These candidate pairs are then matched to the FTC data to identify a set of clear-cut business acquisitions. Finally, responses from the SIPP on the reasons for job loss are used in the model estimation for multiply-imputing the missing data on the nature of worker-firm separations.

The LEHD program matches household and business data together using state level UI wage record data to create a comprehensive and unique resource for data analysis (Abowd et al., 2000). Every employer covered by the UI program reports earnings for each employee receiving positive earnings during the quarter (accounting for approximately 98% of employment in each state). The UI account numbers from

these data are then matched to the business data collected by the Quarterly Census of Employment and Wages (QCEW) program of the Bureau of Labor Statistics. The QCEW data provides information on industry, employment, and payroll for every establishment on the 12th of each month, as well as providing the establishments' employer identification numbers (EIN). Moreover, the micro-level data collected by the Census Bureau provides data on the workers, such as date of birth, race, and gender. Together, these data provide detailed information at the quarterly level on employment and earnings histories for every worker-firm pair.

The strengths of these data are that they are extensive and current offering an enormous sample size with rich variation. For most states the data series begins in the early 1990's and are updated on a quarterly basis (six months after the transaction date). As of the beginning of 2006, forty one states have partnered with the LEHD program, creating a longitudinal data set covering about 85% of US employment. This particular analysis uses data from 31 states accounting for approximately 69% of US employment and contains all the data for these states from the beginning of the LEHD sample through the year 2004.

There are also a number of drawbacks, which are extensively documented in Abowd et al 2000. The major weakness of using such a data set to examine labor market outcomes is that we do not know exactly why a worker leaves the sample (death, moves out of state, quits, etc) or why a worker appears at a different employer from one quarter to the next (quit and found new job, laid-off and found new job, same job but firm underwent some kind of administrative change). For firms, it is not clear when a firm ID appears/disappears from the sample whether the firm truly was born/died or whether there was some change in ownership, reporting, or coding.

The first external data set used to overcome some of these problems was the set of Early Termination Notices available on the FTC's website. The Hart-Scott-Rodino (HSR) Antitrust Improvements Act was instituted in 1976 in order to allow the federal

government to review mergers and acquisitions meeting certain criteria primarily regarding size of the companies involved and size of the transaction. As part of the act, the firms seeking permission to merge must wait thirty days before completing the transaction unless they file for and are granted "early termination" of this waiting period by the government. For the firms allowed to circumvent the waiting period, the FTC publishes the names of the acquiring and acquired companies and the date of early termination. The notices are available publicly on the FTC website covering acquisition activity from 1998 to the present. According to the FTC website, more than 95% of all HSR-reported transactions each year are approved during this initial waiting period. The remainder must go through a "second request," supplying the FTC with more data before potentially being cleared. Unfortunately, this means that the transactions reported are not a perfectly random sample of all mergers above the minimum size threshold specified in HSR; however, the good news is that the publicly available Early Termination notices contain the vast majority of acquisitions with which to attempt a match to the LEHD data.

The second external data set, used to gain some insight into the nature of worker/firm separations, was the 1996 SIPP panel. One of the labor force participation questions asked in this panel was for the reason an individual ceased working for his/her previous employer. There are 15 possible answers offered ranging from reasons such as retirement, health, or child care to layoff, quits, or new job opportunities. The individuals are interviewed in 12 waves spanning 4 years. With four rotation groups, the overall data begins in December 1995 and ends in February 2000, giving excellent overlap with the LEHD data used in this analysis. The raw internal SIPP files available to the LEHD program also contain business name information along with industry. Abowd and Stinson (2006) used these variables for a probabilistic match to the Census Bureau's Business Register to obtain the EIN which can be used in linking back to the administrative data. This match was of very high quality because

the candidates for the business name match were restricted to the names of employers ever seen matching to the SIPP respondent’s social security number in the UI wage records. As a result, the reported employer name from the SIPP file was compared to an extremely limited subset of employer names from the LEHD data which greatly reduces the risk of false matches. Moreover, this analysis on mergers only uses the observations where the reported date of job ending in the SIPP survey was within one quarter of the observed separation in the UI wage records, which essentially adds another matching variable into the mix to improve match quality.

2.3.2 Method of Identifying Firm Restructurings

Even though the LEHD data does not formally identify business restructurings, a great deal can be learned about such events by observing the flows of workers between firms. Benedetto et al. (2006) describe how the LEHD program flags large flows of workers between firms and offers a glimpse into how much can be learned about how modern firms organize themselves by examining the nature of these movements. A flow-based link between hypothetical firms A and B in quarter, t , is formed by finding all work patterns in the UI work histories that look like the following:

	$t - 1$	t	$t + 1$
Worker 1	1	1	0
Firm A			
Worker 1	0	1	1
Firm B			
Worker 2	1	1	0
Firm A			
Worker 2	0	0	1
Firm B			

If there are five or more such transitions, then Firm A is flagged as a potential predecessor of Firm B in the flow-based links. This cutoff is somewhat arbitrary, but it should eliminate most coincidental links, and still offers a very large set of potentially related firms.

If all the transitions for a given link look like Worker 2, then the assumption is that the workers were continuously employed and the change took place at the "quarter boundary." Otherwise, the transition is said to be "within quarter." The timing of the link may be an important clue into the nature of the relationship. A large cluster of workers all suddenly disappearing from one firm's records and appearing in another firm's records essentially overnight (as in the case of the quarter boundary links) is certainly a strong indication that this is not just coincidence. On the one hand, this would seem to suggest a simple record keeping change by the UI collection agencies. On the other hand, there is probably a large incentive to make acquisitions and real economic ownership changes official at the quarter boundary for ease of paperwork.

Benedetto et al. (2006) also categorized these links based on the relative size of the transitioning cluster to the predecessor and successor firms. Not surprisingly, clusters that were small percentages of the firms' employment dominated the links, but an interesting spike was found in the data for clusters greater than 80% of either the predecessor's prior employment or the successor's subsequent employment. A reasonable conclusion from this result is that most true ownership changes (as opposed to coincidental flows of workers between the same two firms) usually involve substantial percentages of at least one of the firms involved. While the links between firms created by this method offer strong evidence of firm restructurings, what exactly the nature of the relationship is between the linked firms remains an open question.

In order to make the jump from strong evidence to almost certain merger/acquisition, the set of candidate, flow-based links was matched with the set of Early Termination notices. This match is relatively difficult because the Early Termination notices

have only business names to identify the firms involved; however, the list of potential candidates has been severely reduced with the identification of the flow-based, predecessor/successor links. Generally, such a name matching exercise would be exceedingly difficult, but considering each record contains a pair of names (the acquired firm from the Early Termination notices compared to the origin firm in the worker-flow links and the acquiring firm from the Early Termination notices compared to the destination firm in the worker-flow links) and also the approximate time of the transaction (the date on the Early Termination notice compared to the quarter of the worker-flow), the probability of finding a decent set of matches was large. Moreover, the LEHD data offers two distinct sources for business name information (the QCEW and the Business Register), thus quadrupling the probability of finding at least one name match.

Minimum distance matching techniques were used with very low (ie very strict) reservation scores to insure that only the most convincing matches were used; after all, far more importance was placed on finding high quality matches than on finding a match for a large percentage of Early Termination notices. In the end 192 links were identified from the years 1998 through 2000, accounting for more than 1.5 million jobs (worker-firm matches). Table 2.1 shows how this set of matches compares to the overall set of flow-based, predecessor/successor links using categories defined with the 80% cutoff mentioned above and the timing of the link. The most striking difference, not surprisingly, is that the matched links had transitioning clusters accounting for at least 80% of one of the firms' employment much more frequently (58%) than the overall set (12%). Moreover, of those 58%, the vast majority involved more than 80% of the predecessor's employment but less than 80% of the successor's employment. Again, this makes perfect sense since one usually thinks of an acquisition as an already large company absorbing all of another company of comparable or lesser size. The timing of the link offers less, it seems, in distinguishing acquisitions from the rest of

the flow-based links. However, for the subset of links where the size of the transition is less than 80% of both the predecessor and successor, quarter boundary links are significantly more common in the matched set than in the overall set. Perhaps this implies that when the size of the link may not be significant enough to convince us that the flow is more than mere coincidence, the timing of the link can suggest a real event if the entire flow happens at the quarter boundary.

Unfortunately, there is no way of showing whether the matched links are a representative subset of the all the firms in the Early Termination notices since firm characteristics are only available for the matches. On the other hand, the quality of name matches between the two data sources should be unrelated to any economic variables of interest, so there should not be any systematic differences created by the process. The first of the two columns for the treatment group in table 2.3a shows some of the economic characteristics of the matches. The industry divisions outside of public administration and agriculture are represented fairly similarly to the general population of firms with perhaps a little more weight in manufacturing and wholesale trade and a little less weight in retail trade. The firm size distribution looks to be skewed slightly towards the larger categories which is to be expected for a population of merging firms.

2.3.3 Multiple Imputation of Missing Data: Reason for Separation

The final major obstacle to this analysis is the unknown reason for a separation of a worker from the employer. The most obvious problem with not knowing this information is that the effects of job loss on earnings should be allowed to differ between voluntary and involuntary separations. On the other hand, even if the reason is known, the difference between a quit and a layoff may not be all that striking, especially in a framework such as this where a worker might quit his/her job in anticipation of layoffs due to a major firm restructuring on the horizon. In such a

case, the "voluntary" separation may resemble more closely an involuntary separation since the choice was made due to firm-level events outside of the worker's control. As a result, this analysis does not distinguish between a quit and a layoff in estimating the risk of job loss around the time of the merger/acquisition event. Nevertheless, it is still important to distinguish between job loss due to firm-level events and separations due to shocks in the personal lives of the workers (e.g. health issues, death, child care problems, or even retirement). For these reasons, an effort was made to fill in this missing information in an unintrusive way.

Since the late 1970's, the theory and techniques for multiple imputation in order to fill missing data have been developed and refined (Rubin 1996). These methods offer an analytically useful set of completed data that allows the analyst to measure the noise introduced through imputation and properly take that into account in estimating statistics and their measures of uncertainty. Adapting Rubin's notation to this missing data problem, the data can be expressed as (Y, X) where Y is a variable with some missing values (in this case the reason for separation) and X is a set of complete covariates (ie no missing values). Y can be expressed as (Y_{obs}, Y_{mis}) where Y_{obs} represents the observed values of Y and Y_{mis} represents the missing values of Y . The inclusion indicator, I , is a structure equivalent in size to Y with elements equal to 1 where Y is non-missing and 0 otherwise. The database can then be expressed by the joint distribution, $p(X, Y, I, \theta)$, where θ are unknown parameters. In this case, the missing data mechanism is said to be missing at random if

$$p(I|Y, X) = p(I|Y_{obs}, X) \tag{1}$$

which is certainly a realistic assumption in this situation since being sampled by the SIPP should be entirely unrelated to the reason for job loss or even if job loss occurs.

Draws are taken from the posterior predictive distribution

$$p(\tilde{Y}|Y_{obs}, X) = \int p(\tilde{Y}|X, \theta)p(\theta|Y_{obs}, X)d\theta \quad (2)$$

to produce L multiply-imputed completed data files (Y^ℓ, X) where $Y^\ell = (Y_{obs}, \tilde{Y}^\ell)$ for $\ell = 1, \dots, L$. The resulting L data files are individually referred to as implicates.

One of the huge advantages of this data completion method is the ease with which statistical inference can be performed on the completed data. For a given estimand Q , the analyst calculates the estimator, q , and its variance estimator, u , on implicate, ℓ , exactly as it would be done on a complete data set. Doing this for every implicate gives $q^{(\ell)}$ and $u^{(\ell)}$ for $\ell = 1, \dots, L$. From these, the following can be calculated:

$$\bar{q}_L = \sum_{\ell=1}^L q^{(\ell)} / L \quad (3)$$

$$b_L = \sum_{\ell=1}^L (q^{(\ell)} - \bar{q}_L)^2 / (L - 1) \quad (4)$$

$$\bar{u}_L = \sum_{\ell=1}^L u^{(\ell)} / L \quad (5)$$

$$T_L = (1 + 1/L)b_L + \bar{u}_L \quad (6)$$

$$\nu_L = (L - 1)(1 + \bar{u}_L / ((1 + 1/L)b_L))^2 \quad (7)$$

Using \bar{q}_L as the estimator for Q , and T_L as the estimate of the variance of \bar{q}_L , inferences can then be based on a t-distribution with degrees of freedom, ν_L . (Rubin 1987).

The merged SIPP-LEHD data described earlier offers a large amount of useful information as a basis for estimating the joint distribution of the "reason for separation" variable with administrative variables. When the SIPP data was matched to the set of UI wage record histories, 13,245 records were found where the person

and EIN matched and the date of separation in the SIPP was within one quarter of the observed separation in the administrative data. Using only the variables in the LEHD data available to both the matches (now with non-missing reason for separation) and the non-matches (the records to receive multiple imputations of reason for separation), a large number of stratification variables were used to break down the data into detailed sub-domains. In other words, workers with similar demographic characteristics and similar work histories (e.g. tenure and length of unemployment after separation) were grouped together. Next, the Bayes' bootstrap described by Rubin (1981) was used to sample from the posterior predictive distribution in each of these sub-domains and produce ten (in the notation above, $L = 10$) draws of the imputed reason for work.

Besides the nice features of this imputation for statistical analysis, there are strong reasons for optimism that the multiple imputation, Bayes' bootstrap method can provide quality imputes of this variable. Table 2.2 shows how the "reason for separation" variable (grouped into three categories: [1] quit and stay in the labor force, [2] lay-off, fire, or discharge, [3] exit the labor force) relates to some of the covariates in the LEHD data that are available for everyone in the sample. While there are not significant differences between the results for quits and layoffs, there are large differences for those leaving the labor force, which is the group that this paper wants to separate out anyway. The people exiting the labor force are more frequently females most likely leaving work for child care reasons. Moreover, the average age of exiters is significantly higher indicating that many from this group are retiring. The most striking difference, however, is that the number of quarters without a job after separation is dramatically higher for those exiting the labor force. For these reasons, it seems reasonable to think the imputation can do a good job of distinguishing labor force exits from quits and layoffs in the sample. The results of the imputation are summarized in the last three rows of Table 2.3b. The fact that labor force exits are

relatively less frequent in the treatment group is encouraging since turnover is much higher for these firms, and one would not expect workers to grow older faster, have more children, or get sick at a significantly greater rate just because their employers undergo corporate restructurings.

2.3.4 Selecting the Population to Analyze

This analysis uses a five year window around the quarter of a restructuring event to examine its effects on earnings and employment. The firms identified to have acquired another firm or to have been acquired by another firm form the basis of the treatment group. Every employee observed to have worked three full quarters at these firms inside the half of the five year window leading up to the event is included in the treatment group, so the sample is composed of workers who have a non-trivial attachment to the relevant firms prior to the acquisition. This group can be divided into workers who originated at the acquired firm (Type A) and those who originated at the acquiring firm (Type B). The industry of the firms involved is also an important factor to take into consideration when examining labor market outcomes. One might expect that the workers at both firms get affected differently when the acquiring firm purchases another firm that performs similar tasks (within industry) than when the acquiring firm obtains a new set of tasks with the acquired firm (across industry). As a result, the treatment group is further divided into workers who are involved in an acquisition within 4-digit SIC (Type I) and those who are involved in an acquisition across distinct 4-digit SICs (Type II). The analysis compares the outcomes of these four types of workers (AI, AII, BI, and BII) to a set of controls who are workers at firms that do not undergo a merger/acquisition in a given period of time.

With the enormous size of the LEHD data, one of the toughest challenges is reducing the control data to a size that makes estimation less computationally burdensome. Given that multiple imputation is already being used to address the missing data on

the reason for job loss, independent random samples were drawn to form the control groups for each of the ten implicates to reduce the impact of a single draw on the estimates. All firm-year-quarters belonging to firms that were not identified to have undergone a restructuring were treated as potential controls for this analysis. Of course, not every possible merger/acquisition was identified in forming the treatment group, so it remains possible for one of the firms from this control set to in fact be involved in a restructuring. However, given the relatively small set of restructurings in comparison to the universe of firm-year-quarters, this probability is very small and will be ignored.

For each of the ten implicates, a random sample of firm-year-quarters was drawn from the overall distribution of controls weighted so as to mimic the features of the firm-year-quarters in the treatment group. The year-quarters drawn in the control sample mark the timing of the hypothetical restructuring around which a similar five year window will be examined. A set of stratification variables including state, SIC division, a seven category firm size class variable, year, and quarter were used to form the weights. Once these firm-year-quarters were selected, the set of workers forming the control group was assembled in the same fashion as the treatment group. The results of this method can be seen by comparing the control and treatment columns of Tables 2.3a and 2.3b.

2.4 Empirical Model

2.4.1 General Strategy

The overall approach was to analyze labor market outcomes of interest sequentially. First, a wage regression was estimated for workers in the sample just using quarters in which they had positive full quarter earnings at a treatment or control firm. From this regression, wage profiles of employees can be compared between Type AI, Type AII, Type BI, Type BII, and control workers. This regression can also pro-

vide estimates for an individual earnings component to be used in later models. Next, a logit regression was used to compare the probabilities of job loss for workers of the various types. Finally, the earnings losses were estimated for those who did lose their jobs with a regression similar to those found in the displaced worker literature. In the end, the pieces can be put together to tell a fairly detailed story of what happens to employees caught in the midst of major firm reorganizations.

2.4.2 Earnings Regression

The first model estimated was an earnings regression restricted to observations where workers were observed with full quarter earnings in order to simulate a wage rate. The dependent variable, w_{ijt} , is the log of the full quarter earnings. A worker is said to have full quarter earnings in period, t , at firm, j , if he/she has positive wages at firm j in periods $t-1$, t , and $t+1$. The natural assumption is that this wage record pattern implies continuous employment during quarter, t ; therefore, the earnings in that quarter can be thought of as a quarterly wage rate. This wage rate is regressed on a set of time varying person-firm characteristics, X_{ijt} , an individual component, θ_i , a set of dummies referring to any existing separation in the near future, and dummies to identify the merger effects for workers at both the acquiring and acquired firms.

$$\begin{aligned}
w_{ijt} = & X_{ijt}\beta + DI'_{iJ(i,t)}\alpha + DJAI'_{J(i,t)}\gamma^{AI} + DJAII'_{J(i,t)}\gamma^{AII} \\
& + DJBI'_{J(i,t)}\gamma^{BI} + DJBII'_{J(i,t)}\gamma^{BII} + \theta_i + \varepsilon_{ijt}
\end{aligned} \tag{8}$$

$$\begin{aligned}
DI'_{ij}\alpha &= \sum_{0 \leq \tau \leq M_s} DI'_{ij}\alpha_\tau \\
\text{where } DI'_i &= \begin{cases} 1 & \text{if worker } i \text{ separates from firm } j \text{ in period } t + \tau \\ 0 & \text{otherwise} \end{cases} \\
DJAI'_{ij}\gamma^{AI} &= \sum_{-M_r \leq \tau \leq M_r} DJAI'_{ij}\gamma_\tau^{AI} \\
\text{where } DJAI'_{ij} &= \begin{cases} 1 & \text{if firm } j \text{ is acquired in period } t + \tau \text{ within SIC4} \\ 0 & \text{otherwise} \end{cases} \\
DJAII'_{ij}\gamma^{AII} &= \sum_{-M_r \leq \tau \leq M_r} DJAII'_{ij}\gamma_\tau^{AII} \\
\text{where } DJAII'_{ij} &= \begin{cases} 1 & \text{if firm } j \text{ is acquired in period } t + \tau \text{ across SIC4} \\ 0 & \text{otherwise} \end{cases} \\
DJBI'_{ij}\gamma^{BI} &= \sum_{-M_r \leq \tau \leq M_r} DJBI'_{ij}\gamma_\tau^{BI} \\
\text{where } DJBI'_{ij} &= \begin{cases} 1 & \text{if firm } j \text{ acquires another firm} \\ & \text{in period } t + \tau \text{ within SIC4} \\ 0 & \text{otherwise} \end{cases} \\
DJBII'_{ij}\gamma^{BII} &= \sum_{-M_r \leq \tau \leq M_r} DJBII'_{ij}\gamma_\tau^{BII} \\
\text{where } DJBII'_{ij} &= \begin{cases} 1 & \text{if firm } j \text{ acquires another firm} \\ & \text{in period } t + \tau \text{ across SIC4} \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

$J(h, s)$ identifies firm of worker h at time s .

The model is generalized from the OLS case by allowing ε_{ijt} to be an AR(1) process for every individual (ie $\varepsilon_{ijt} = \rho\varepsilon_{ijt-1} + v_{ijt}$ where v_{ijt} is white noise, $E(\varepsilon_{ijt}\varepsilon_{iks}) = 0$ for all s and all $k \neq j$, and $E(\varepsilon_{ijt}\varepsilon_{hjs}) = 0$ for all s and all $h \neq i$). The time-varying person-firm variables include an estimate of the firm wage component, year dummies, age, and observed experience over the course of the sample. The firm wage component was separately estimated on the full sample using techniques pioneered by

Abowd, Kramarz, and Margolis (1999) and later applied to the LEHD data by Abowd, Haltiwanger, Lane, and Sandusky (2001) and Abowd, Lengermann, and McKinney (2002). The separately measured firm wage component was used because there is not enough variation in this sample to jointly estimate individual and firm wage components, but the previously estimated firm effect should offer a good measure to control for high wage firms in the regression. The tenure variable is potentially limited by the lower bound of the dates in the sample; however, any unobserved initial tenure should be soaked into the individual wage component, θ_i .

The results of this regression should offer some more insight into the wage questions explored by Brown and Medoff (1987). While they concluded wages decreased only slightly at merging firms, they were only able to observe a firm-level average quarterly earnings and acknowledged there could be unobserved compositional effects biasing the results. With this more detailed sample of data, these compositional changes can be controlled for, and the question of what happens to wages of workers from the acquiring and acquired firms around the time of a merger can be answered with more precision.

2.4.3 Logistic Regression on Quits and Layoffs

The second piece of the puzzle is the question of how these restructurings affect the probability of job loss. The logit model was used to regress whether a worker separated from his/her employer in a given period ($m_{ijt} = 1$ if worker i separates from firm j in quarter t and $m_{ijt} = 0$ otherwise) on a set of time-varying worker-firm characteristics, X_{ijt}^ℓ , and the same merger-effect dummies from the initial earnings

regression.

$$\begin{aligned} \text{let } \tilde{m}_{ijt} = & X_{ijt}^{\ell} \beta^{\ell} + DJAI'_{J(i,t)} \varphi^{AI} + DJAII'_{J(i,t)} \varphi^{AII} \\ & + DJBI'_{J(i,t)} \varphi^{BI} + DJBII'_{J(i,t)} \varphi^{BII} \end{aligned} \quad (9)$$

$$\Pr(\text{worker } i \text{ quits or is laid-off from firm } j \text{ at time } t) = \frac{1}{1+e^{-\tilde{m}_{ijt}}}$$

The worker-firm characteristics include all the characteristics from the previous regression as well as sex and race dummies and the estimates of θ_i fully interacted with dummies for type A and B workers and type I and II acquisitions.

Even though previous literature has found small changes in wages and employment at restructuring firms, that is more a consequence of the typical productivity gains from corporate takeovers, and does not reflect the cost of potentially higher turnover rates. Certainly, the gains from more efficient management must be weighed against the costs of job loss, especially if the job losers during mergers face long term earnings losses similar to those caught in mass layoffs. Once again, the detail of the data allows for distinguishing the risks of job loss between workers starting at the acquired firms and those originally employed at the acquiring firm. Intuition suggests that the workers from the acquiring firms would be better matches with the organizational structure of the new merged firm and, in turn, face lower risk of job loss than the acquired workers.

2.4.4 Examining Earnings Losses of Quits and Layoffs

Finally, the consequences of job loss for workers at the restructuring firms were examined to see if the results from the displaced worker literature apply to the job losers in this sample. Another earnings regression was run, but this time the log of total earnings, y_{ijt} , was regressed on a set of time-varying worker-firm characteristics,

X_{ijt}^q , and a set of indicators for future job loss. The model can be written as:

$$y_{ijt} = X_{ijt}^q \beta^q + DS'_{iJ(i,t)} \delta + \varepsilon_{ijt} \quad (10)$$

$$DS'_{ij} \delta = \sum_{-M_r \leq \tau \leq M_r} DS_{ij}^\tau \delta_\tau$$

$$\text{where } DS_i^\tau = \begin{cases} 1 & \text{if worker } i \text{ loses job at firm } j \text{ in period } t + \tau \\ 0 & \text{otherwise} \end{cases}$$

$J(h, s)$ identifies original firm in sample of worker h at time s .

The major challenge with this regression was how to treat quarters of zero earnings. Clearly, some employees who lose their job will experience some quarters of no employment after the job loss. Ideally, those quarters would influence the parameter estimates properly showing the cost of job loss on future earnings. However, in a log earnings regression those observations would be dropped. Moreover, some of the zero earnings observed in this data will in fact be workers who obtained jobs in states outside of this sample. As a result, this regression was run several times with different strategies on handling the zero earnings observations and different sample restrictions in an attempt to get an upper and lower bound on the potential earnings losses faced by job losers in this sample.

In the first regression, zero earnings observations were recoded to \$1 prior to taking the log, following the strategy of Kenneth Couch and Dean Lillard (1998) in their paper, "Sample Selection Rules and the Intergenerational Correlation of Earnings." As with the previous regressions, all workers at the sampled firms during the time of the event (or hypothetical event in the case of the control set) with at least three full quarters of earnings at some point in the first half of the five year window around the event were kept. The outcome of this regression can be thought of as a lower bound since clearly some of the job losers would have gotten jobs in states outside of the current LEHD sample (or jobs not covered by the UI program) and show up in this

regression as false zeros. On the other hand, this sample does account for most of the US labor force, so it is reasonable to think that there are not too many false zeros. In an attempt to get some idea of the impact of this problem, another regression was run restricting the analysis to 15 of the original 31 states from this sample, accounting for approximately 37% of US employment. The total earnings measure used for this second regression was recalculated by summing up earnings only over these same 15 states. Presumably this regression should overstate earnings losses of job losers even more than the first regression, but how much more might give some insight into the size of this bias.

A third regression was run using the same sample as the first regression, but all observations with zero earnings were dropped from the data matrix. The resulting parameter estimates should provide an upper bound to the earnings losses experienced by the job losers in this sample, since obviously some of the zeros reflect true unemployment spells. Moreover, the log transform in this regression is more natural since there is no spike in the data at zero earnings, and no need for any recoding.

Finally, one last regression was run using a similar sample selection rule to the one used by Jacobson, LaLonde, and Sullivan (1993). In this regression, only individuals with earnings in at least one quarter of every year of the sample were kept. Zero earnings quarters were again recoded to \$1 before the log transformation. This strategy should prevent many of the false zeros from entering the regression; therefore, it should provide a more conservative estimate of earnings losses than the first regression. Nevertheless, it is still possible that some of the zero earnings quarters could still be false for very mobile workers or workers residing near state boundaries, so it is not clear on which side of the truth this estimate should lie.

2.5 Results

To examine the results of the wage and earnings regressions, the expected values of the dependent variable with and without the treatment effect were compared. Expressing this mathematically, if $\ln(z)$ is the dependent variable, and it is regressed on a set of variables, W , and a treatment indicator, d , then one can calculate the expected value of $\ln(z)$ given W for either value of the treatment dummy as:

$$E(\ln(z)|W, d) = \hat{\vartheta}W + \hat{\kappa}d \quad (11)$$

where $\hat{\vartheta}$ and $\hat{\kappa}$ are the estimated regression coefficients. Transforming these expected values back to their natural scale and taking the ratio gives the following:

$$\frac{\exp(E(\ln(z)|W, d = 1))}{\exp(E(\ln(z)|W, d = 0))} = \exp(\hat{\kappa}) \quad (12)$$

The interpretation of this ratio is that the expected value of z with the treatment in its natural scale is $\exp(\hat{\kappa})$ times greater than the expected value of z without the treatment in its natural scale.

Applying this strategy to the first wage regression gives a time-series of these ratios, $\exp(\gamma_{\tau}^{type})$, where the treatment is to be at a firm that underwent a merger τ periods ago. Figure 2.1a plots $\exp(\gamma_{\tau}^{type})$ for $type = AI$ (workers at the acquired firm of a transition within SIC4) and for $type = BI$ (workers at the acquiring firm of a transition within SIC4), and τ ranges from 9 quarters before the reference period to 8 quarters after the reference period along the x-axis. Wages for both treatment groups are higher than the controls. The workers who start out at the acquiring firm for intra-industry transitions have the highest wages by a large margin; although, their wages also seem to be the least stable relative to the controls as the ratio jumps around quite a bit over time. Figure 2.1b plots the same ratio for $type = AII$ (workers

at the acquired firm of a transition across SIC4) and for *type = BII* (workers at the acquiring firm of a transition across SIC4). For these inter-industry transitions, wages are again highest for workers who started at the acquiring firm, but the wage gap closes slightly over time presumably as the workers from the acquired firm that are the best matches for the new management survive and the lower quality matches lose their jobs. Since the wage regression controls for the firm wage component, these ratios reflect the wage differentials all else being equal. One might expect that the acquiring firms are fundamentally different from the acquired firms. To get a sense of this difference with regards to wage, one can look at the average firm wage component at the sample of acquiring and acquired firms:

Firm Wage Component	Sample Mean	Standard Error
Acquired Firms	0.262	0.302
Acquiring Firms	0.304	0.253

The acquiring firms, not surprisingly, have a higher firm wage component on average, but the difference actually is not very large and not very statistically significant. Therefore, the ratios plotted in figures 2.1a and 2.1b only slightly understate the differential in wages between type A and B workers. The full set of parameter estimates and their measures of uncertainty can be observed in tables 2.4a and 2.4b.

Figures 2.2a and 2.2b plot the odds ratios calculated from coefficients on the treatment dummies in the job-loss, logistic regression where job-loss is defined as having an imputed reason for separation to be a layoff or quit but remain in the labor force. The data are plotted from 5 quarters prior to the transition to 10 quarters after the event, because the minimum tenure restriction prohibits any job-losers in the first year of the sample window to enter the regression. Turnover is generally higher for workers at the restructuring firms, but there are dramatic differences in the patterns of these odds ratios over time depending on the type of worker and

the type of restructuring. In the case of intra-industry transitions (Figure 2.2a), the odds of job-loss for a worker who started at the acquired firm are about 5 times higher than the controls from a year and a half before the transition to half a year before the transition. The odds then climb to a peak of nearly 10 times that of the controls right at the time of the restructuring, linger there for a few quarters after the event, and then drop to about twice that of the controls for the remainder of the sample window. Meanwhile the odds of job-loss for those who started at the acquiring firm is not significantly different from the controls for much of the sample window. However, two quarters after the restructuring these odds climb to about twice that of the controls for the rest of the sample period. Since the transition is within industry, the workers who are retained from the acquired firm are probably good substitutes for the incumbent workers at the acquiring firm in general which might explain the rise in instability for type B workers after the acquisition.

For inter-industry acquisitions, the odds ratios tell a similar story with slightly worse outcomes for type A workers and slightly better outcomes for type B workers. The workers from the acquired firm start out similar to their counterparts in the intra-industry acquisitions, but then peak at a significantly higher odds ratio during the quarter of the restructuring. The odds of these workers losing their jobs is almost 20 times that of the controls at the time of the sale. This seems a little strange considering the workers at the acquiring firm are probably not close substitutes in general for those at the acquired firm; however, it does fit the model of a successful firm purchasing an inefficient, struggling firm and putting its own managerial stamp on the entity. Moreover, the workers at the purchasing firms are not significantly different from the controls throughout the entire window in terms of the risk of job loss. As a result, it seems these acquisitions are cases of vertical or horizontal integration where the incumbent staff at the purchasing firm is essentially unaffected while the acquired businesses face substantial reorganizations and the inevitable worker turnover. Tables

2.5a and 2.5b give the coefficient estimates from the job-loss logit and their levels of uncertainty and significance.

For the sake of robustness, the logistic regression was run again defining job-loss as any separation. Since the ratio of imputed quits and layoffs to all separations is generally higher for the treatment group, intuition implies that the odds ratios associated with the relative quarter indicators should be closer to 1. After all, while one would expect slightly higher rates of retirement at restructuring firms, one would not expect this effect to be as large as the effect on layoffs and quits. However, as shown in figures 2.4a and 2.4b, the logistic regression run with all separations produces much larger estimates of the risk of job-loss for workers at the acquired firm. This result will require more investigation down the road.

Using the ratios of transformed expected earnings, $\exp(\delta_\tau)$, calculated from the results of the final earnings regressions, figure 2.3 verifies that the earnings losses for job losers (as they have been defined in this analysis) are similar to the losses found in the displaced worker literature. For all four regressions, there are some earnings loss prior to job loss followed by a severe drop in earnings directly after the separation. The different strategies for handling zero earnings quarters do, however, result in vastly different estimates of the size of the earnings drop immediately after separation and of the speed and extent of earnings recovery. Not surprisingly, when the zeros are dropped from the regression, the earnings hit at the time of the separation is not nearly as large as it is in the other three regressions. The first two regressions also show nearly the same results with the 15 state curve only slightly lower than the 31 state curve, implying that the false zero problem is not very large. When the sample was restricted to those with some earnings in every year, the earnings drop at the time of the separation is essentially just as large as the lower bound, but the recovery afterwards is much steeper. This curve is more in line with previous estimates from the displaced worker literature, although all four curves essentially have a similar

pattern with varying magnitudes. Tables 2.6-2.9 give the coefficient estimates and their significance for the four earnings regressions.

2.6 Conclusion

The results of this analysis answer three basic questions regarding the labor market outcomes for the average worker at a firm involved in a major corporate acquisition where at least part of the workforce of the acquired firm is merged with the acquiring firm. Wages are similar at the acquired firm to those at non-restructuring firms, and they are significantly higher at the acquiring firm. Despite generally higher wages, however, acquisitions also imply significantly higher risk of job loss. Job insecurity is especially pronounced for workers starting out at the acquired firm, but even workers at the acquiring firm face larger risks of losing their jobs when their company purchases another company within the same industry. Not surprisingly, the costs to overall earnings for the workers who lose their jobs around the time of such a corporate restructuring follow a similar pattern to what have been consistently shown in the displaced worker literature. The earnings of job losers dip before the separation, plummet immediately after separation, and only partially recover from the main earnings hit in the first couple of years after separation.

There is still much that can be done with this data set in the study of mergers and their effect on labor. Certainly it would be interesting and feasible to expand this research in much the same way that Dardia and Schoeni (2000) and Lengermann and Villhuber (2001) expanded on the displaced worker literature. For instance, looking at the distribution of wages and earnings around the time of these restructurings, as well as the distribution of human capital for the various types of workers who stay and leave from these firms would be a natural progression. Moreover, many of the techniques used in this paper to build the data set could be improved upon. The matching techniques used to link on the Early Termination notices could be expanded with

the increasing availability of high quality probabilistic and distance-based matching software. The weights used to draw the control set could be constructed with more detail on geographical location and finer industry information. The imputation of "reason for job separation" might be improved by using information on firm growth rates as those have been shown to be highly correlated with turnover (Davis et al., 2006). As a robustness measure, all the analysis should also be performed using the separately estimated person wage component as well as the separately estimated firm wage component. Also, since LEHD is rapidly expanding, more states should soon be able to be incorporated into the analysis resulting in better match rates to the Early Termination notices and more accurate measures of total earnings.

Table 1

	Size of Transition Relative to Predecessor	Size of Transition Relative to Successor	All Links		Matched Links	
			Percentage of Column	Percentage of Cell That Are Quarterly Boundary Links	Percentage of Column	Percentage of Cell That Are Quarterly Boundary Links
Less Than 80% of Predecessor's Employment Moves To Successor	Less Than 80% of Predecessor's Employment Comes From Predecessor	More Than 80% of Successor's Employment Comes From Predecessor	88%	9%	42%	33%
	More Than 80% of Predecessor's Employment Moves To Successor	Less Than 80% of Successor's Employment Comes From Predecessor	4%	33%	6%	25%
More Than 80% of Predecessor's Employment Moves To Successor	Less Than 80% of Predecessor's Employment Comes From Predecessor	More Than 80% of Successor's Employment Comes From Predecessor	4%	47%	41%	54%
	More Than 80% of Predecessor's Employment Comes From Predecessor	Less Than 80% of Successor's Employment Comes From Predecessor	4%	63%	10%	60%

Table 2

Reason for Separation	Training Data for Multiple Imputation of "Reason for Separation"		
	Variable	Mean	Standard Deviation
Layoff	Male	0.55	0.50
	Age	35.18	12.81
	Length of Unemployment	3.47	7.83
Quit	Male	0.47	0.50
	Age	30.46	11.86
	Length of Unemployment	3.72	8.41
Exit Labor Force	Male	0.31	0.46
	Age	41.78	16.44
	Length of Unemployment	10.10	12.15

Table 3a

Firm-Year-Quarter level data	Control Group			Treatment Group			Matched SIPP-LEHD sample		
	Percent	Between-Implicate Standard Deviation	Overall Standard Deviation	Percent	Between-Implicate Standard Deviation	Overall Standard Deviation	Percent	Between-Implicate Standard Deviation	Overall Standard Deviation
Year									
1995							0.29	0.000	0.046
1996							25.87	0.000	0.381
1997							25.47	0.000	0.379
1998	32.27	32.467	7.341	30.57	0.000	2.604	23.94	0.000	0.371
1999	21.15	4.734	4.397	21.02	0.000	2.303	23.62	0.000	0.369
2000	46.59	24.994	6.953	48.41	0.000	2.825	0.82	0.000	0.078
Size Class									
[0,5)							5.60	0.000	0.206
[5,20)	14.00	10.833	4.685	13.06	0.000	1.904	16.27	0.000	0.330
[20,50)	15.98	5.785	4.207	17.20	0.000	2.133	13.70	0.000	0.308
[50,100)	12.74	8.643	4.340	13.38	0.000	1.924	11.10	0.000	0.281
[100,250)	22.98	25.830	6.568	22.93	0.000	2.376	13.59	0.000	0.307
[250,500)	7.18	3.367	3.053	7.32	0.000	1.473	9.50	0.000	0.263
>500	27.13	18.235	6.059	26.11	0.000	2.483	30.24	0.000	0.411
SIC division									
A	1.17	0.361	1.174	1.59	0.000	0.708	5.80	0.000	0.203
B	8.05	9.856	4.130	8.28	0.000	1.558	0.44	0.000	0.057
C	13.84	6.922	4.206	11.15	0.000	1.779	4.95	0.000	0.189
D	9.36	9.577	4.202	9.24	0.000	1.637	10.49	0.000	0.266
E	16.18	20.185	5.783	16.24	0.000	2.085	4.21	0.000	0.174
F	13.47	5.826	4.025	12.74	0.000	1.885	4.39	0.000	0.178
G	8.87	14.732	4.795	10.19	0.000	1.710	29.87	0.000	0.398
H	29.05	24.442	6.641	30.57	0.000	2.604	4.11	0.000	0.173
I							34.17	0.000	0.412
J							1.56	0.000	0.108

Table 3b

	Control Group			Treatment Group			Matched SIPP-LEHD sample		
	Percent	Between-Implicate Standard Deviation	Overall Standard Deviation	Percent	Between-Implicate Standard Deviation	Overall Standard Deviation	Percent	Between-Implicate Standard Deviation	Overall Standard Deviation
Worker-Firm level data									
Gender									
F	49.03	26.023	5.351	48.79	0.000	0.040	53.66	0.000	0.433
M	50.97	26.023	5.351	51.21	0.000	0.040	46.34	0.000	0.433
Race									
White	60.19	25.278	5.274	63.88	0.000	0.038	77.77	0.000	0.361
Black	16.96	8.553	3.068	14.77	0.000	0.028	8.64	0.000	0.244
Hispanic	12.38	15.044	4.069	10.57	0.000	0.024	8.24	0.000	0.239
Other	10.48	4.613	2.253	10.78	0.000	0.025	5.36	0.000	0.196
Age									
<18	4.30	2.679	1.717	4.39	0.000	0.016	0.68	0.000	0.071
[18,25)	25.55	13.807	3.898	24.22	0.000	0.034	25.65	0.000	0.379
[25,35)	26.23	1.407	1.247	26.19	0.000	0.035	30.32	0.000	0.399
[35,45)	21.66	4.998	2.346	22.12	0.000	0.033	20.58	0.000	0.351
[45,55)	14.38	3.274	1.899	14.31	0.000	0.028	13.36	0.000	0.296
[55,65)	6.00	0.547	0.777	6.57	0.000	0.020	6.35	0.000	0.212
>65	1.87	0.048	0.231	2.21	0.000	0.012	3.07	0.000	0.150
Reason for Separation									
1. Layoff	11.96	0.835	0.960	14.50	0.002	0.058	16.53	0.000	0.323
2. Quit	56.84	8.363	3.035	57.28	0.002	0.058	73.20	0.000	0.385
3. Exit Labor Force	31.19	13.198	3.811	28.22	0.001	0.044	10.27	0.000	0.264

Table 4a: Wage Regression

Time-varying Worker-Firm Characteristics			
Variable Description	Parameter Estimate	Variable Description	Parameter Estimate
Age	0.0201 0.0412	Firm Wage Component	0.355 0.0261**
0.01*(Age squared)	0.00835 0.0291	Tenure	0.0264 0.00763**
0.001*(Age cubed)	-0.00609 0.00215**	0.01*(Tenure squared)	-0.0391 0.00832**
Male*Age	0.0716 0.0101**	0.001*(Tenure cubed)	0.0033 0.00179**
Male*0.01*(Age squared)	-0.15 0.0274**	1996 Indicator	0.0298 0.0381
Male*0.001*(Age cubed)	0.00902 0.00233**	1997 Indicator	0.0321 0.0299
Quarter 1 Indicator	-0.00628 0.0279	1998 Indicator	0.0177 0.0214
Quarter 2 Indicator	-0.0455 0.0136**	1999 Indicator	-0.0201 0.00597**
Quarter 3 Indicator	-0.0318 0.0118**	2001 Indicator	0.001 0.016
Male*Quarter 1 Indicator	0.0444 0.00783**	2002 Indicator	0.0119 0.0102
Male*Quarter 2 Indicator	0.00198 0.00431	Individual Wage Components	
Male*Quarter 3 Indicator	0.0316 0.0063**	Average	7.51
		Standard Deviation	1.62
		AR(1) coefficient	0.321

Table 4b: Wage Regression

Indicators for Quarter Relative to Restructuring Event				
Relative Quarter	Within SIC4		Across SIC4	
	Type A	Type B	Type A	Type B
-9	0.0562 0.025**	0.0778 0.0259**	0.0173 0.026	-0.044 0.0287
-8	0.122 0.0271**	0.521 0.0223**	0.0141 0.0264	0.0652 0.0279**
-7	0.0423 0.0219**	0.146 0.0231**	-0.0238 0.0239	0.048 0.0242**
-6	0.0418 0.0245*	0.223 0.0184**	0.00297 0.0263	0.102 0.0278**
-5	0.0245 0.0231	0.227 0.0214**	-0.028 0.0248	0.00751 0.0312
-4	0.096 0.0256**	0.379 0.0246**	-0.0217 0.0259	0.105 0.0303**
-3	-0.00168 0.0214	0.271 0.0183**	-0.0809 0.0235**	0.103 0.0255**
-2	0.0275 0.0236	0.33 0.0194**	-0.00861 0.0251	0.122 0.0274**
-1	0.0063 0.0211	0.205 0.0193**	-0.0757 0.02**	0.039 0.0249
0	0.055 0.0236**	0.393 0.0193**	-0.0319 0.0227	0.128 0.0237**
1	0.0257 0.0195	0.284 0.015**	-0.0947 0.0174**	0.0627 0.0208**
2	0.032 0.0224	0.14 0.0129**	-0.0302 0.0209	0.076 0.0247**
3	-0.00093 0.0208	0.179 0.0157**	-0.0562 0.0171**	0.0275 0.0271
4	0.0779 0.022**	0.518 0.0204**	0.032 0.0203	0.0505 0.025**
5	0.0373 0.0164**	0.184 0.0174**	0.0181 0.018	0.0341 0.0197
6	0.0275 0.0183	0.0693 0.0127**	-0.0154 0.0191	0.0765 0.0165**
7	0.000654 0.0147	0.114 0.0154**	-0.00711 0.0111	-0.0196 0.00834**
8	0.0442 0.0133**	0.395 0.0103**	-0.0404 0.0128**	0.0725 0.00588**

Time-Varying Worker-Firm Characteristics in Job Loss Regression			
Variable Description	Parameter Estimate	Variable Description	Parameter Estimate
Intercept	-5.7 0.174**	Female*Black	0.155 0.0858*
Male	6.4 3.5**	Female*Hispanic	-0.0667 0.0572
Firm Wage Component	-0.331 0.257	Female*Age	-0.117 0.0307**
1995 Indicator	-15 160	Female*0.01*(Age Squared)	0.196 0.0729**
1996 Indicator	-15 51	Female*0.001*(Age Cubed)	-0.0107 0.00518**
1997 Indicator	-1.1 0.303**	Female*Tenure	0.33 0.0367**
1998 Indicator	-1.1 0.213**	Female*0.01*(Tenure Squared)	-1.5 0.206**
1999 Indicator	-0.315 0.156**	Female*0.001*(Tenure Cubed)	0.194 0.0336**
2001 Indicator	-0.023 0.185	Female*Theta	2.5 1.3**
2002 Indicator	-0.16 0.444	Female*0.01*(Theta Squared)	-36 17**
2003 Indicator	-0.41 0.336	Female*0.001*(Theta Cubed)	16 7.1**
		Female*Quarter 1 Indicator	-0.386 0.174**
		Female*Quarter 2 Indicator	-0.229 0.229
		Female*Quarter 3 Indicator	-0.229 0.155*
		Male*Black	0.265 0.0616**
		Male*Hispanic	0.0576 0.076
		Male*Age	-0.166 0.0322**
		Male*0.01*(Age Squared)	0.285 0.0678**
		Male*0.001*(Age Cubed)	-0.0148 0.0046**
		Male*Tenure	0.363 0.0477**
		Male*0.01*(Tenure Squared)	-1.6 0.261**
		Male*0.001*(Tenure Cubed)	0.206 0.0402**
		Male*Theta	-0.0162 0.359
		Male*0.01*(Theta Squared)	-5.2 5.5
		Male*0.001*(Theta Cubed)	4.7 3.7
		Male*Quarter 1 Indicator	-0.332 0.183*
		Male*Quarter 2 Indicator	-0.225 0.131*
		Male*Quarter 3 Indicator	-0.155 0.144

Theta=Person Wage Component

Indicators for Quarter Relative to Restructuring Event				
Relative Quarter	Type A	Type B	Type A	Type B
-5	1.6 0.226**	0.363 0.218*	2 0.2**	0.306 0.214*
-4	1.9 0.218**	0.611 0.258**	1.7 0.195**	0.339 0.214*
-3	1.7 0.205**	0.183 0.259	1.6 0.2**	0.237 0.166*
-2	1.6 0.227**	0.399 0.205**	1.7 0.168**	0.452 0.194**
-1	1.8 0.155**	0.174 0.257	2.1 0.151**	0.195 0.166
0	2.1 0.192**	0.196 0.214	2.9 0.163**	0.251 0.173*
1	2.2 0.163**	0.356 0.229*	2.6 0.16**	0.294 0.149**
2	0.934 0.268**	0.117 0.182	1.6 0.234**	0.24 0.199
3	0.686 0.256**	0.274 0.266	2 0.212**	0.107 0.206
4	1.1 0.274**	0.606 0.213**	0.711 0.229**	0.369 0.279
5	0.793 0.212**	0.373 0.291	0.416 0.272*	0.761 0.229**
6	-0.0943 0.309	0.329 0.235*	0.354 0.333	-0.272 0.338
7	0.55 0.355*	0.368 0.283	0.579 0.38*	-0.102 0.473
8	0.854 0.365**	0.641 0.509	1.3 0.349**	-0.29 0.501
9	0.465 0.39	-0.505 0.506	1 0.49**	-0.195 0.456
10	-15 300	-18 170	-15 220	-15 68

Table 6: Full sample with zeros recoded to \$1 (upper bound of earnings losses)

Time-Varying Worker-Firm Characteristics					
Variable Description	Parameter Estimate	Variable Description	Parameter Estimate	Variable Description	Parameter Estimate
Intercept	6 0.0803**	Firm Wage Component	1.2 0.155**	Male	-2.6 1.7*
1994 Indicator	-0.537 0.0739**	Female*Age	0.164 0.026**	Male*Age	0.219 0.0232**
1995 Indicator	-0.412 0.0846**	Female*	-0.232 0.0682**	Male*	-0.312 0.0519**
1996 Indicator	-0.57 0.188**	Female*(Age Squared)	0.00497 0.00519	Male*(Age Squared)	0.00775 0.00357**
1997 Indicator	-0.206 0.0355**	Female*(Age Cubed)	-0.137 0.0346**	Male*(Age Cubed)	-0.261 0.0307**
1998 Indicator	-0.0652 0.0256**	Female*Black	-0.095 0.0362**	Male*Black	-0.154 0.0335**
1999 Indicator	0.0484 0.06	Female*Hispanic	-0.145 0.375	Male*Hispanic	0.559 0.837
2001 Indicator	-0.0661 0.0344**	Female*Theta	-1.7 4.7	Male*Theta	-9.3 12
2002 Indicator	-0.217 0.055**	Female*(Theta Squared)	3.8 2.1*	Male*(Theta Squared)	6.2 4.9
2003 Indicator	-0.807 0.255**	Female*(Theta Cubed)	0.162 0.0803**	Male*(Theta Cubed)	0.115 0.054**
2004 Indicator	-5.5 0.771**	Quarter 1 Indicator	0.0604 0.0855	Quarter 1 Indicator	0.00387 0.0694
		Female*	0.146	Male*	0.116
		Quarter 2 Indicator	0.0834*	Quarter 2 Indicator	0.0666*
		Female*		Male*	
		Quarter 3 Indicator		Quarter 3 Indicator	

Table 6 (continued)					
Indicators for Quarter Relative to Job Loss					
Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate
-10	0.216 0.0616**	-3	0.0356 0.0393	4	-2 0.19**
-9	-0.00588 0.06	-2	-0.0046 0.0409	5	-1.9 0.261**
-8	-0.0903 0.0459**	-1	-0.0482 0.0375	6	-1.8 0.319**
-7	-0.0842 0.0384**	0	-0.0886 0.0472**	7	-1.9 0.311**
-6	-0.0213 0.0401	1	-2 0.177**	8	-1.9 0.306**
-5	0.0718 0.0366**	2	-1.9 0.193**	9	-1.8 0.227**
-4	0.104 0.0379**	3	-2.1 0.186**	10	-1.8 0.224**

Table 7: 15 state sample with zeros recoded to \$1

Time-Varying Worker-Firm Characteristics					
Variable Description	Parameter Estimate	Variable Description	Parameter Estimate	Variable Description	Parameter Estimate
Intercept	5.9 0.116*	Firm Wage Component	1.2 0.194**	Male	-2.5 1.8*
1994 Indicator	-0.624 0.213**	Female*Age	0.149 0.0284**	Male*Age	0.214 0.0236**
1995 Indicator	-0.543 0.146**	Female* 0.01*(Age Squared)	-0.207 0.0753**	Male* 0.01*(Age Squared)	-0.321 0.0504**
1996 Indicator	-0.371 0.218	Female* 0.001*(Age Cubed)	0.00449 0.00588	Male* 0.001*(Age Cubed)	0.0102 0.0032**
1997 Indicator	-0.253 0.0501**	Female*Black	-0.0864 0.0384**	Male*Black	-0.206 0.0422**
1998 Indicator	-0.0772 0.0387**	Female*Hispanic	-0.0943 0.0348**	Male*Hispanic	-0.126 0.0374**
1999 Indicator	0.0839 0.101	Female*Theta	0.0487 0.712	Male*Theta	0.666 1.1
2001 Indicator	-0.0723 0.0356**	Female* 0.01*(Theta Squared)	-4.9 8.5	Male* 0.01*(Theta Squared)	-11 14
2002 Indicator	-0.181 0.0371**	Female* 0.001*(Theta Cubed)	5.5 4.1	Male* 0.001*(Theta Cubed)	6.8 5
2003 Indicator	-0.593 0.272**	Female* Quarter 1 Indicator	0.174 0.116	Male* Quarter 1 Indicator	0.144 0.0743**
2004 Indicator	-6.1 0.564**	Female* Quarter 2 Indicator Female* Quarter 3 Indicator	0.0639 0.133 0.122 0.123	Male* Quarter 2 Indicator Male* Quarter 3 Indicator	0.0315 0.0922 0.116 0.0923
Indicators for Quarter Relative to Job Loss					
Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate
-10	0.0161 0.0653	-3	0.012 0.14	4	-2.5 0.348**
-9	-0.0796 0.0517	-2	-0.174 0.0444**	5	-2.3 0.35**
-8	-0.172 0.0542**	-1	-0.117 0.0388**	6	-2.3 0.394**
-7	-0.0874 0.144	0	-0.226 0.048**	7	-2.2 0.331**
-6	-0.185 0.0591**	1	-1.8 0.266**	8	-2.1 0.267**
-5	0.018 0.039	2	-1.9 0.222**	9	-2.1 0.416**
-4	-0.0122 0.0548	3	-2.5 0.391**	10	-2.1 0.288**

Table 8: Zeros dropped (lower bound of earnings losses)

Time-Varying Worker-Firm Characteristics					
Variable Description	Parameter Estimate	Variable Description	Parameter Estimate	Variable Description	Parameter Estimate
Intercept	5.3 0.0117**	Firm Wage Component	1.2 0.0765**	Male	-0.881 1.2
1994 Indicator	-0.568 0.0439**	Female*Age	0.155 0.0129**	Male*Age	0.169 0.00923**
1995 Indicator	-0.412 0.0498**	Female*	-0.255 0.0339**	Male*	-0.249 0.0253**
1996 Indicator	-0.325 0.0275**	Female* 0.01*(Age Squared)	0.0113 0.00247**	Male* 0.01*(Age Squared)	0.00865 0.00222**
1997 Indicator	-0.246 0.0135**	Female*Black	-0.135 0.0249**	Male*Black	-0.268 0.0209**
1998 Indicator	-0.151 0.00685**	Female*Hispanic	-0.104 0.0142**	Male*Hispanic	-0.181 0.023**
1999 Indicator	-0.08 0.00514**	Female*Theta	0.53 0.448	Male*Theta	0.646 0.866
2001 Indicator	0.0206 0.013	Female* 0.01*(Theta Squared)	-12 5.4**	Male* 0.01*(Theta Squared)	-11 12
2002 Indicator	0.0225 0.0124	Female* 0.001*(Theta Cubed)	8.3 1.9**	Male* 0.001*(Theta Cubed)	7 5
2003 Indicator	0.0284 0.0376	Female* Quarter 1 Indicator	0.0357 0.0117**	Male* Quarter 1 Indicator	0.0256 0.0137**
2004 Indicator	-0.0523 0.121	Female* Quarter 2 Indicator Female* Quarter 3 Indicator	-0.0261 0.0146 0.0146 0.0149	Male* Quarter 2 Indicator Male* Quarter 3 Indicator	-0.0231 0.0135 0.0397 0.00945**
Indicators for Quarter Relative to Job Loss					
Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate
-10	-0.0396 0.0284	-3	-0.0928 0.0243**	4	-0.276 0.029**
-9	0.0121 0.0313	-2	-0.13 0.019**	5	-0.206 0.03**
-8	-0.0391 0.0239	-1	-0.103 0.0255**	6	-0.238 0.0453**
-7	-0.0327 0.0287	0	-0.327 0.03**	7	-0.169 0.0551**
-6	-0.0871 0.0229**	1	-0.287 0.0333**	8	-0.221 0.0526**
-5	-0.0494 0.0263**	2	-0.369 0.0281**	9	-0.172 0.06**
-4	-0.062 0.0195**	3	-0.213 0.0444**	10	-0.179 0.0611**

Table 9: JLS restriction (only individuals with positive earnings in the sample every year)

Time-Varying Worker-Firm Characteristics					
Variable Description	Parameter Estimate	Variable Description	Parameter Estimate	Variable Description	Parameter Estimate
Intercept	6.3 0.102	Firm Wage Component	0.909 0.153**	Male	-1.7 2.3
1994 Indicator	-0.619 0.275**	Female*Age	0.131 0.0317**	Male*Age	0.175 0.0241**
1995 Indicator	-0.467 0.189**	Female* 0.01*(Age Squared)	-0.2 0.0755**	Male* 0.01*(Age Squared)	-0.264 0.0611**
1996 Indicator	-0.7 0.403	Female* 0.001*(Age Cubed)	0.00679 0.00569	Male* 0.001*(Age Cubed)	0.00898 0.00483**
1997 Indicator	-0.296 0.0654**	Female*Black	-0.123 0.0377**	Male*Black	-0.211 0.0411**
1998 Indicator	-0.157 0.0359**	Female*Hispanic	-0.0913 0.0324**	Male*Hispanic	-0.139 0.0293**
1999 Indicator	-0.111 0.0873	Female*Theta	0.0638 1.1	Male*Theta	0.268 1.4
2001 Indicator	0.0236 0.0248	Female* 0.01*(Theta Squared)	-5.7 14	Male* 0.01*(Theta Squared)	-4.5 19
2002 Indicator	-0.0258 0.055	Female* 0.001*(Theta Cubed)	6.7 4.9	Male* 0.001*(Theta Cubed)	4.9 8.3
2003 Indicator	-0.00812 0.0949	Female* Quarter 1 Indicator	0.0852 0.102	Male* Quarter 1 Indicator	0.0496 0.0497
2004 Indicator	-4.4 1**	Female* Quarter 2 Indicator	0.00728 0.0962	Male* Quarter 2 Indicator	-0.0237 0.0661
		Female* Quarter 3 Indicator	0.0829 0.093	Male* Quarter 3 Indicator	0.0829 0.0632
Indicators for Quarter Relative to Job Loss					
Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate	Relative Quarter	Parameter Estimate
-10	0.0159 0.111	-3	0.0148 0.105	4	-0.941 0.343**
-9	-0.0902 0.119	-2	-0.113 0.0413**	5	-0.769 0.269**
-8	-0.215 0.105**	-1	-0.0705 0.0523	6	-0.603 0.11**
-7	-0.121 0.137	0	-0.222 0.0856**	7	-0.446 0.19**
-6	-0.141 0.0697**	1	-1.6 0.32**	8	-0.731 0.121**
-5	0.0145 0.0587	2	-1.3 0.249**	9	-0.909 0.26**
-4	-0.0588 0.0505	3	-0.903 0.129**	10	-1.2 0.301**

Figure 2.1a

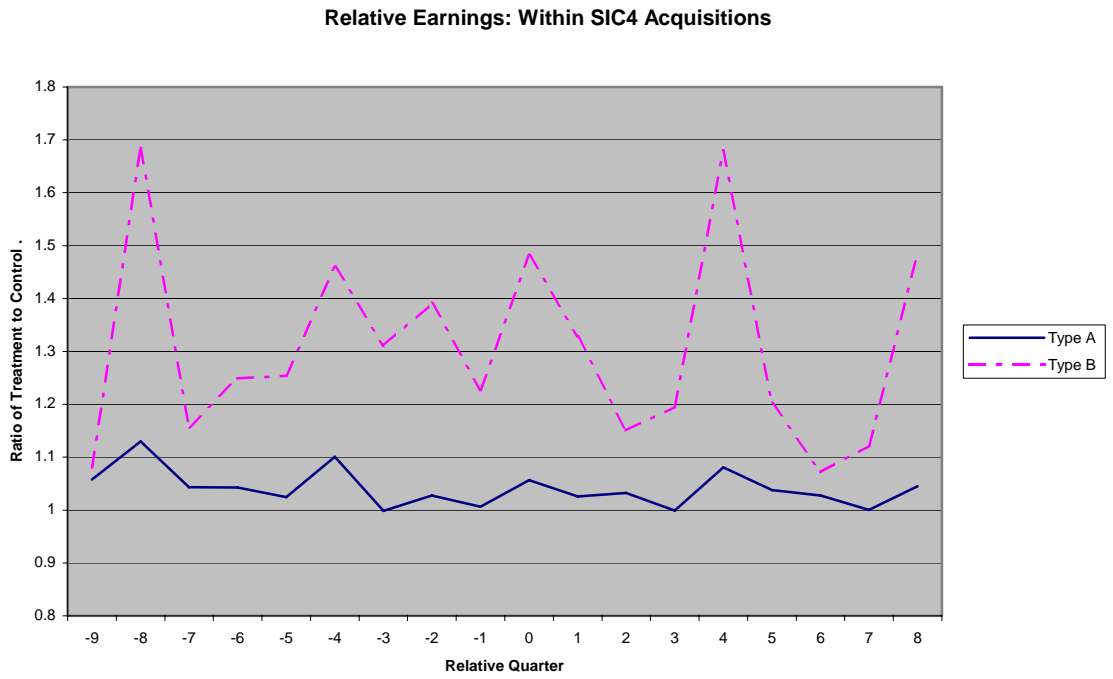


Figure 2.1b

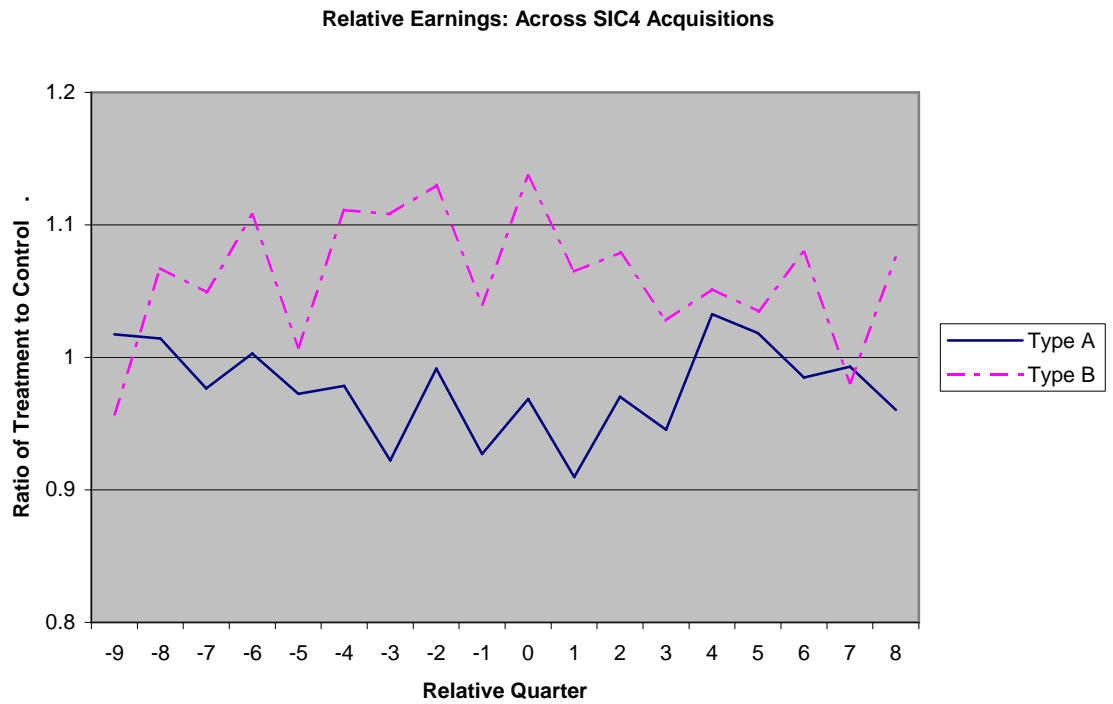


Figure 2.2a

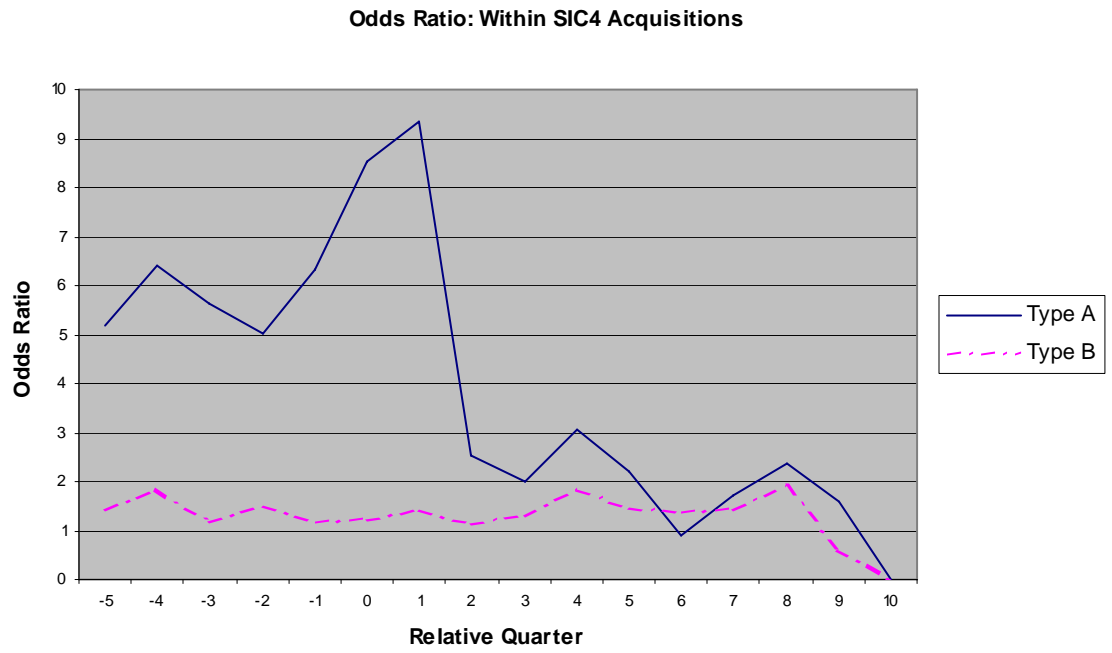


Figure 2.2b

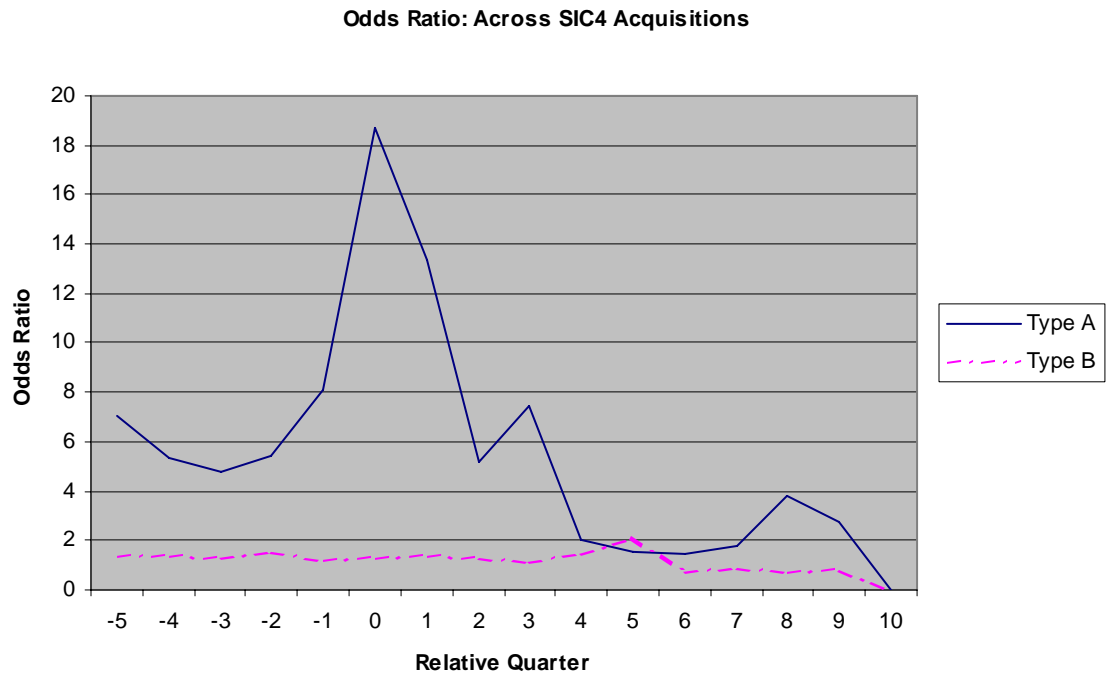


Figure 2.3

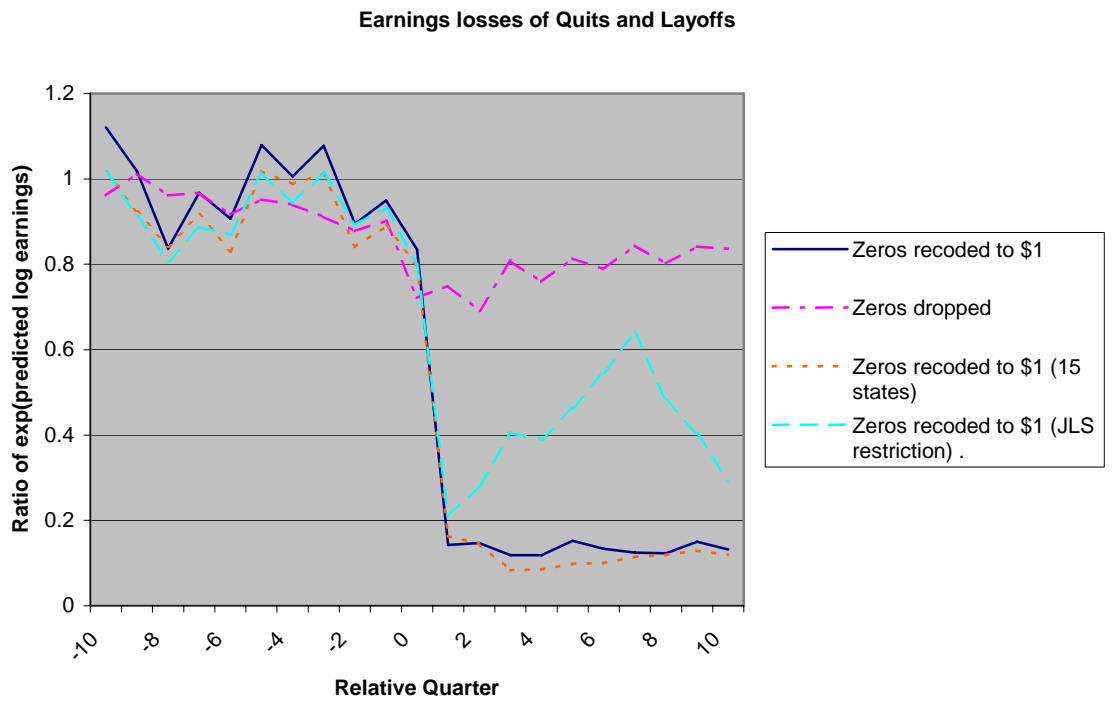


Figure 2.4a

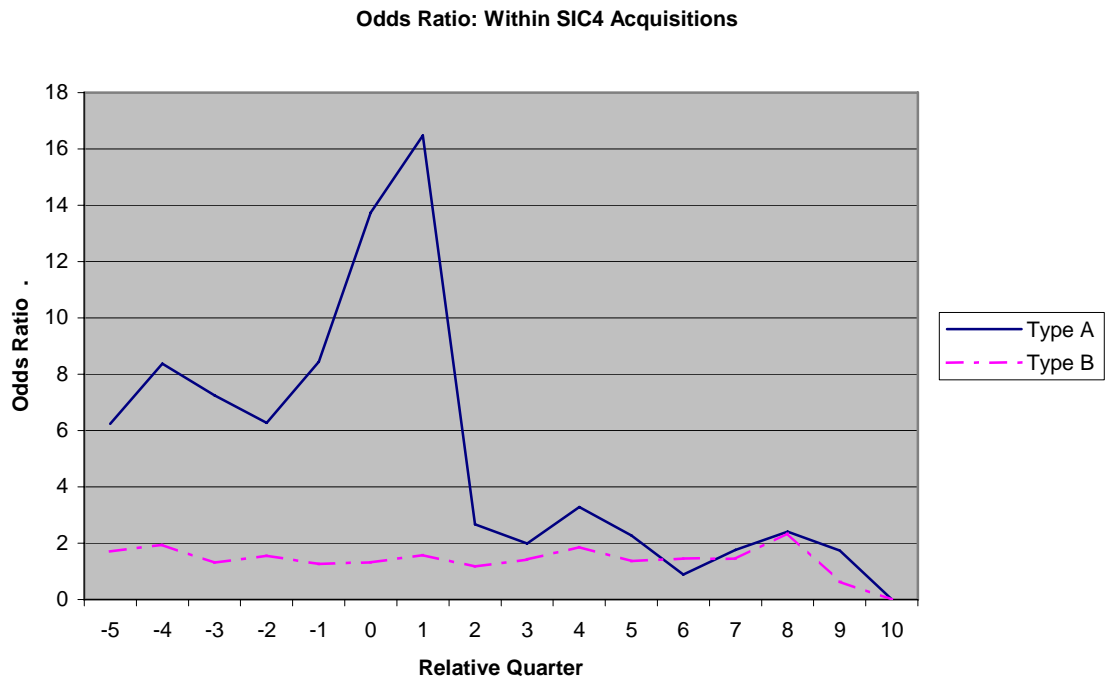
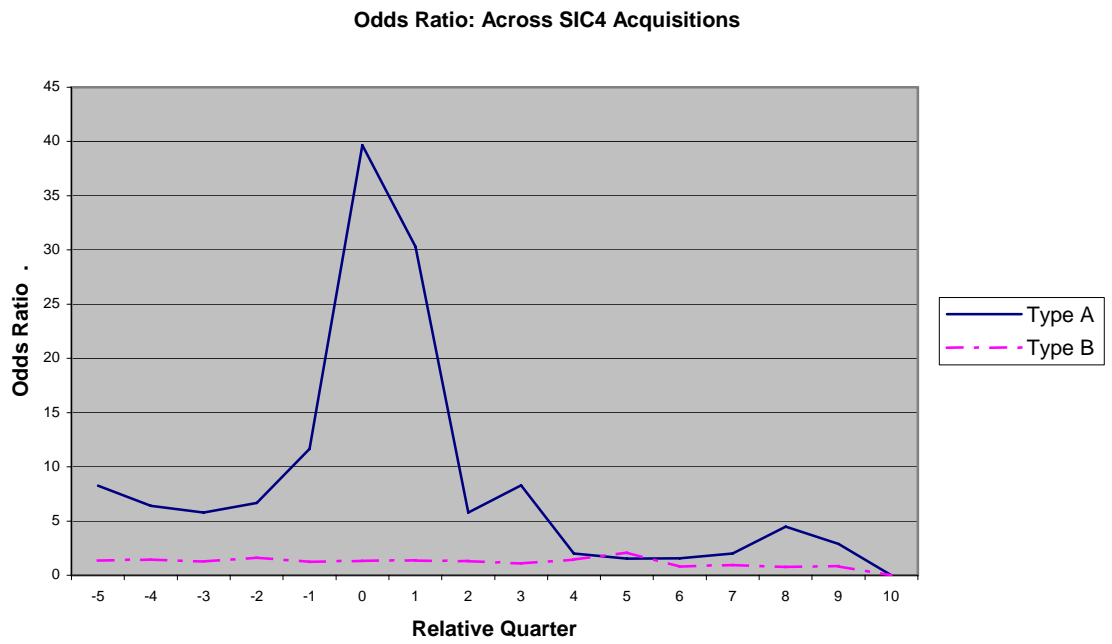


Figure 2.4b



Technical Description of the SIPP/SSA/IRS Public Use File Project^{*†}

3.1 Executive Summary

3.1.1 Purpose and brief history

The creation of public use data that combine variables from the Census Bureau's Survey of Income and Program Participation (SIPP), the Internal Revenue Service's (IRS) individual lifetime earnings data, and the Social Security Administration's (SSA) individual benefit data began as part of ongoing collaborative research at the Census Bureau and SSA. The current project had its genesis with the formation of a joint committee containing representatives from the Census Bureau, SSA, IRS, and the Congressional Budget Office (CBO) that designed a prospective public use file. Aimed at a user community that was primarily interested in national retirement

^{*}This chapter is a slightly abbreviated version of "Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project," John Abowd, Gary Benedetto, and Martha Stinson. LEHD Technical Paper.

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

and disability programs, the selection of variables for the proposed SIPP/SSA/IRS-PUF focused on the critical demographic data to be supplied from the SIPP, earnings histories from the IRS data maintained at SSA, and benefit data from SSA’s master beneficiary records.

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

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

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

The SIPP/SSA/IRS-PUF contains data from the records of individuals who responded to the SIPP panels conducted in 1990-1993 and 1996. A standardized extract of approximately 125 variables from all waves of each of these panels was created. We included the following demographic variables: gender, marital status, race (black), five categories of education, Hispanic ethnicity, birth date, death date, disability

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

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

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

FICA taxable maximum;¹

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

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

¹These data, as well as the Detailed Earnings Record data cited in the next bullet, are also confidential under the protocol defined in Title 26 of the U.S. Code. Prior permission from the IRS disclosure officer is required before they can be used in a project in combination with Title 13 confidential data. Permission to conduct the present research is monitored by the Census Bureau under Administrative Records Tracking System project 458, which contains a copy of the IRS approval.

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

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

There is a third source of missing data in the Gold Standard file. Some data items are structurally missing because it is not logically possible for the item to have a value; for example, no data are available concerning the second marriage of individuals who

never married or married only once. Structurally missing data remain in the Gold Standard file and in the synthetic data implicates that constitute the SIPP/SSA/IRSPUF.

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

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

The top level of this binary tree contains all variables that exhibit no logical dependencies on any other variables in the file, for example birth date. The tree has nine levels. At each level below the top, variables depend upon their parents, and are only processed when appropriate. In the intermediate levels of the tree, a variable can be both a parent and a child, for example, whether or not there is a second marriage is a child of the same variable for the first marriage and a parent of the variable for the third marriage. The terminal level and all leaves of the binary tree contain only child variables.

For each iteration of the missing data imputation phase and again during the synthesis phase, we estimate a joint posterior predictive distribution for all of the required variables according to the following protocol. At each node of the parent/child tree, a statistical model is estimated for each of the variables at the same level. The

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

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

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

3.1.4 Development of the weights

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

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

Our method for creating an ex-post weight for the merged SIPP panels involved seven steps. First, we reproduced the major component of the 1996 sampling frame (the unit frame) in the Census 2000 micro-data, updating the SIPP reference pop-

ulation to April 1, 2000. Next, we divided the Decennial individual records (long and short form) into strata according to the 1996 SIPP sampling plan. Then we standardized all five SIPP panels with respect to geographic definitions and strata used in the 1996 sampling plan. Each individual observation in the merged panels was then placed in a stratum according to the 1996 SIPP sampling plan. The fifth step was to link each SIPP person to a person in the Census 2000 reference population. This match was accomplished using probabilistic record linking. Most observations could be linked on the basis of the PIK² that had been assigned to the SIPP or Decennial individual. For those SIPP individuals who did not link by PIK, a cruder probabilistic record linkage based on characteristics used to define the sampling strata was used. Having accomplished this linkage for all in-scope individuals in the 1990-1993 and 1996 SIPPs, we created a preliminary weight as the ratio of in-scope individuals in Census 2000 to in-scope individuals in the merged SIPPs within each final (stage-2) SIPP sampling cluster. The final weight was created by raking the preliminary weights to agree with official U.S. population control totals for the sex/age/race/ethnicity demographic breakdown of the reference population, as supplied by the Census Bureau's Population Estimates Division. This final raking was controlled to exactly the same population categories as the official 1996 SIPP weights.

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

Because copying the final weight to each implicate of the synthetic data would

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

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

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

3.1.5 Analytical validity testing

Although synthetic data are designed to solve a confidentiality protection problem, the success of this solution is measured by both the degree of protection provided and the user's ability to estimate scientifically interesting quantities reliably. The latter property of the synthetic data is known as analytical (or statistical) validity. Analytical validity exists according to Rubin Rubin (1987) when, at a minimum,

estimands can be estimated without bias and their confidence intervals (or the nominal level of significance for hypothesis tests) can be stated accurately. The estimands can be summaries of the univariate distributions of the variables, bivariate measures of association, or multivariate relationships among all variables.

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

Section 3.6 gives a complete summary of the inference framework and computational formulae for assessing analytical validity. From a theoretical framework, the synthetic data will be analytically valid for the precise set of relations embodied in the posterior predictive distributions used for the synthesis. This theoretical result is reassuring to the extent that substantial computational power and flexible methods for estimating complex multivariate distributions can produce reliable posterior predictive distributions. Given the limits of current technology, however, the analytical validity of a particular synthetic data product must be directly assessed. Our method of assessment proceeds as follows. Parallel analyses of a large number of estimands are conducted on the completed Gold Standard data and on the synthetic

data. The estimand is averaged over all implicates: four in the case of the completed confidential data and 16 in the case of the synthetic data. Next, the within and between implicate components of the estimand’s variance are combined, according to rules that depend upon the precise multiple imputation method used, to generate an estimate of the total variance. The square roots of the diagonal elements of the total variance matrix and the appropriate degrees of freedom are used to form 95% confidence intervals for the completed and synthetic data estimates of the estimand. Ideally, the confidence intervals computed from the synthetic data should cover the confidence intervals computed from the completed data. At a minimum, there should be substantial overlap in the confidence intervals. The point estimates should also be “close,” but this result is a by product of confidence interval coverage. In general, the confidence interval in the synthetic data will be wider than the interval computed from the completed confidential data for a specified nominal level, and this loss of precision is part of the cost of confidentiality protection. However, the width of the synthetic confidence interval can be reduced by increasing the number of synthetic implicates. A summary of our analytical validity results follows.

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

and over on April 1, 2000 from the synthetic data as from the completed data.

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

Summary statistics for all workers and for OASDI beneficiaries Although our synthesizer automatically develops models for subgroups when there are adequate sample sizes, the order in which the subgroups will be formed and tested for sample size adequacy is specified in advance. Consequently, one cannot say *a priori* that results will be analytically valid for all subsamples. We compared the results for all workers and for all OASDI beneficiaries using subsamples constructed on demographic variables and benefit type. This testing focused on important earnings and benefit measures. Work histories, average annual earnings, average indexed monthly earnings (AIME) or average monthly wages (AMW), primary insurance amount (PIA), lifetime earnings, and personal savings account accumulated balances are very similar between the synthetic and completed data files for all major demographic subgroups and all types of benefits. In general, the univariate confidence intervals in the synthetic data cover those in the completed data and are not excessively wide. These results hold whether the reference group is all persons age 18 or older or only OASDI benefit recipients. Overall, the version 4.0 synthetic data have almost complete analytic validity for these tests. This is a notable improvement over all previous versions and, in particular, over version 3.1. See section 3.6.4 for the detailed summary of the results for all workers and section 3.6.4 for the detailed summary of the OASDI recipient results.

Summary statistics by education We also studied summary statistics for several important variables by three-way interactions of race, gender, and education category. This analysis focused on earnings and benefits in 2000. Again, most point estimates were very close and synthetic confidence intervals covered the completed

data intervals. In earlier versions, there were problems with certain educational categories. Some problems remain with the groups having no high school diploma or graduate degrees. These problems are usually that the confidence interval in the synthetic data is excessively wide—indicating that the synthesizer had trouble simulating these relationships and reflected a great deal of model uncertainty in the posterior predictive distribution. This is not surprising since these education categories, when cross-classified by race and gender, contain relatively few individuals in the Gold Standard file. See section 3.6.4 for a detailed summary of these results.

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

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

See section 3.6.4 for a detailed discussion of the regression results.

3.1.6 Disclosure avoidance assessment

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

In order to test the effectiveness of the synthetic data in controlling disclosure risk, we conducted two distinct matching exercises between the synthetic data and the Gold Standard. Since the Gold Standard contains actual values of the data items as released in the original SIPP public use files, the Gold Standard variables are the equivalent of the best available information for an intruder attempting to re-identify a record in the synthetic data. Successful matches between the Gold Standard and the synthetic data represent potential disclosure risks. However, for an actual re-identification of any of records that were successfully matched to an existing SIPP public use file, an additional non-trivial step is required—the intruder must make another successful link to exogenous data files that contain direct identifiers such as names, addresses, telephone numbers, *etc.* Hence, the results from our experiments are very conservative estimates of re-identification risk. Nevertheless, we find that the re-identifiable records represent only a very small proportion of the candidate

records, less than three percent using the most aggressive technology, and that these correct re-identifications are swamped by a sea of false re-identifications, which a real intruder would not be able to distinguish from the true re-identifications.

The Census Bureau Disclosure Review Board has adopted two standards for disclosure avoidance in partially synthetic data. First, using the best available matching technology, the percentage of true matches relative to the size of the files should not be excessively large. In our case, the true match rate never exceeds three percent of the relevant candidate records. Second, the ratio of true matches to the total number of matches (true and false) should be close to one-half. We have performed two types of matching exercises, probabilistic and distance-based. The first criterion ensures that very few candidate re-identifications occur. The second criterion ensures that those candidate re-identifications are surrounded by substantial uncertainty as regards their correctness.

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

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

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

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

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

disclosure review officers is also required before the file can be officially released.

3.1.7 Using the SIPP/SSA/IRS-PUF

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

3.1.8 Next steps

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

When we began this project, there was a great deal of uncertainty over whether

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

3.2 Project Background

3.2.1 Purpose and brief history

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

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

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

3.2.2 Overview project description

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

The first major design decision was that the file would contain records for individuals surveyed in one of five SIPP panels, 1990, 1991, 1992, 1993, and 1996, but the panel of origin for each individual would not be revealed. The decision to suppress the panel of origin for the individual was part of the overall confidentiality protection plan for the new PUF. The second major design decision was that the number of variables on the new public use file that came from the SIPP would be limited and would be chosen to facilitate national studies by retirement and disability researchers. The third major design decision was that the primary disclosure avoidance method would be to produce partially synthetic micro-data that could not be re-identified

in the existing SIPP public use files. Thus, instead of containing the actual values of SIPP-reported variables, administrative earnings and benefits, the file would contain values that were draws from the joint posterior predictive distribution of the underlying variables conditional on the existing confidential data. The process of synthesizing data is described in detail in section 3.4.

The committee began its work by selecting the variables to include on the file. The selection process involved detailed discussions between all four agencies and consultations with outside researchers. As part of the process, the Census Bureau created a standardized extract of variables from each SIPP panel and merged these extracts with individual administrative earnings and benefits records. These extracts were combined and named the “Gold Standard” file. (See section 3.3 for a detailed description of this file.) The Gold Standard file has been revised many times during the past five years as new variables have been added, old ones dropped, and formatting for some variables changed. This file serves as the basis for the creation of the SIPP/SSA/IRS-PUF. It establishes the metadata for each variable, determines the sample of people to be included, and serves as the source data for the modeling required to create the synthetic data. The Gold Standard file contains data that are Title 13, Title 26, and Title 42 confidential because it commingles Census Bureau , IRS, and SSA data.

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

After each preliminary public use file was produced, committee members from

each agency were responsible for reviewing the file to assess analytical validity and disclosure risk. Analytical validity tests have consisted of comparing univariate distributions, cross-tabulations, moments, and regression coefficients calculated from the synthetic and the completed Gold Standard data. (See section 3.6 for a detailed discussion of the analytical validity assessment.) Disclosure risk analysis has included probabilistic and distance-based record linking between the synthetic and the Gold Standard files. (See section 3.7 for a detailed discussion).

3.3 Creation of the Gold Standard File

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

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

3.3.1 SIPP data

We chose the following demographic variables to be included on the Gold Standard file: gender, marital status, race (black/African-American), five categories of education, Hispanic ethnicity, birth date, death date, disability status, number of children, marital history, foreign born indicator, decade arrived in the United States if foreign born, and a spouse identifier that links to the marriage partner if the respondent is married and the spouse was also surveyed. We took the values for these variables at a

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

The marital history variables are some of the most complicated historical vari-

ables on the Gold Standard file. Most of the information in this history came from the marital history topical module collected in wave 2 of each panel. We supplemented this information by creating a short marital status history that covered the period of the panel from wave 2 forward by using the marital status reported in each month. In the topical module, individuals could report 0, 1, 2, 3, or more than 3 marriages. Dates for the beginning and end of first, second, and most recent marriages were then collected, as well as the reason for a marital termination (death or divorce/separation). If an individual had more than 3 marriages, no dates for those marriages between the second and most recent were collected. We used our short history from the panel period post-wave 2 to check for additional marital events: beginning of a new marriage or ending of an existing marriage. We took account of at most one additional marital history event for an individual. We summarized all this information in a set of 16 variables. They include: *mh_category*, a categorical variable that classifies individuals according to their number of marriages and the type and order of the endings of those marriages; *mh1 – mh7*, a set of flags that provides the same information as *mh_category* but broken down by event; *flag_mar4t*, an indicator for whether the individual was missing a marriage because the SIPP only collected information on three marriages; age at the time of every reported marital history event.

It is important to understand that the marital history variables may differ from the marital status variable described earlier. In particular if a person reports being married in wave 2, month 8 but is not married at the end of his or her history, this is because a divorce or death occurred over the course of the SIPP panel. Similarly, if a person reports not being married in wave 2, month 8 but is married at the end of his or her history, this is because a marriage occurred during the course of the SIPP panel. Although a person may report only 3 marriages during the topical module, it is possible to have a fourth marriage as part of the marital history because the last

marriage occurred during the course of the SIPP panel. However no more than 4 marriages can be recorded in the SIPP from all available sources. There are also some things that must be consistent between the marital history and the wave 2 marital status. If the person reports being divorced or widowed in wave 2, month 8 then at least one divorce or widowhood must occur during that individual's marital history. Likewise if the person reported being married, then the history must contain at least one marriage.

Birth date and death date are unique variables in both their source and treatment. We originally extracted birth date from the first self-reported value in the SIPP survey. However, after several discussion between the Census Bureau and SSA about the measurement error likely to be contained in this variable, we switched to using an administrative birth date. Thus, in the final version of the Gold Standard, we create a variable called *birthdate_pcf* from the Census Personal Characteristics File (PCF), an administrative database that has as one of its inputs the Social Security Numident file. Any individual that has applied for a Social Security Number (SSN) has a record in the Numident file that contains, among other things, SSN and birth date. The administrative record birth date variable (*birthdate_pcf*) serves as the basis for the synthesis process and is comparable to the variable *birthdate* in the synthetic public use files. We chose the administrative source for this important variable in order to insure as much consistency as possible between the administrative earnings and benefits variables and age. Using administrative birth date helps to avoid cases where it appears that people receive retirement benefits prior to age 62, a legal impossibility caused by self-reported birth dates that are several years later than the actual dates.³ We also included a variable called *birthdate_sipp* on the Gold Standard file in order

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

to facilitate re-identification tests that attempt to link the synthetic data back to the Gold Standard. Since the administrative birth date is not an existing SIPP public use variable, anyone attempting to link the new synthetic SIPP/SSA/IRS-PUF to the existing SIPP public use data would have to use the SIPP reported birth date for this purpose.

Death date was extracted from the same PCF file as administrative birth date. In this case, however, the original sources were the Numident file and SSA's Death Master File (DMF), another supplementary file that the Census Bureau receives from SSA that reports deaths every month. The link between SIPP respondents and the PCF was performed using a validated SSN, a process described in more detail in section 3.3.2.

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

Some SIPP variables were purposely omitted from the public use file in order to minimize disclosure risk. Specifically, no data are provided on geography. We include

a state of residence variable Gold Standard file but will not release this variable on the public use file. The exact linkage of spouses is the only family relationship data on the file. No other family relationship data are provided on either the Gold Standard or public use files. No panel dating information is provided on the public use file although we retained the panel source variable in the Gold Standard data to facilitate evaluation and testing.

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

3.3.2 IRS/SSA earnings data

Administrative earnings data were extracted from the Master Earnings File (MEF), a historical compilation of earnings reports filed with the IRS by employers (most commonly using the W-2 form). This administrative database is maintained by SSA for the purpose of calculating benefits when workers retire or become disabled. We receive earnings in two forms: Summary Earnings Records (SER) and Detailed Earnings Records (DER). The SER data contain total personal earnings capped at the FICA taxable maximum for each year from 1951-2003. The DER data contain

uncapped earnings broken out by employer from 1978-2003. In the Gold Standard file we include the entire annual SER history plus total earnings from 1937 to 1951. From the DER, we create total annual earnings from FICA covered jobs by summing earnings from each employer that was required to withhold social security tax. We also create total annual deferred earnings from FICA covered jobs by summing deferred wages (*i.e.*, contributions to 401(k) plans) from these same employers. We create an analogous set of earnings and deferred earnings variables that pertain to jobs not covered by FICA. Thus, in each year from 1978-2003, the SER earnings variable indicates the amount of FICA covered earnings in a year, up to the taxable maximum, and the set of four DER earnings variables indicates total earnings, uncapped, split between deferred and paid, and FICA and non-FICA covered jobs. The sum of the two DER earnings variables that represent paid wages gives total wages and salary that an individual would report on IRS Form 1040 and which would be taxable under federal income tax laws.

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

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

proximately 12% of individuals in the gold standard did not have valid SSNs and were, consequently, missing MBR, SER and DER data.

3.3.3 SSA data

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

3.3.4 Weight creation and use

One concern that arose early in the process of creating the public use file was the provision of proper weights for a file that pooled SIPP respondents from five separate samples. The 1990-1993 panels contain some overlapping time periods. The official SIPP public use file documentation explains how to pool the published

weights for those panels in order to construct a weight that has a well-defined reference population and reference date. However, there is no design guidance for pooling the individuals from all five SIPP panels in order to produce estimates representative of a well-defined target population at a known reference date. In addition, the different SIPP panels over-sample low income individuals and other groups at differential rates. Hence, these survey data can only be used to construct estimates representative of the U.S. civilian non-institutionalized population as of a particular date if an appropriate weight is provided. Thus, another major data activity for this project has been to create an *ex post* weight for the individuals in the Gold Standard file such that each person's weight indicated how many persons in the reference population that SIPP person represented as of a known date. The designated reference population is all individuals age 18 or older in the civilian non-institutionalized U.S. population as of April 1, 2000, the reference date for Census 2000. A full report on the details of this process is provided in section 3.5.

3.4 Data Completion and Synthesis

3.4.1 General methodology

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

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

the rows of Y that correspond to the rows for which $I_i = 0$. So Y_{exc} contains the data for all the unsampled entities. Similarly, let R_{obs} ($n \times p$) be the submatrix of R corresponding to the item missingness for the sampled entities; *i.e.*, those rows for which $I_i = 1$. Finally, define the submatrices Y_{obs} and Y_{mis} as follows

$$Y_{obs,ij} = \left\{ \begin{array}{l} Y_{ij}, \text{ if } I_i = 1 \text{ and } R_{ij} = 1 \\ \text{undefined, otherwise} \end{array} \right\}$$

and

$$Y_{mis,ij} = \left\{ \begin{array}{l} Y_{ij}, \text{ if } I_i = 1 \text{ and } R_{ij} = 0 \\ \text{undefined, otherwise} \end{array} \right\}$$

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

General Definitions

N = number of individuals in the population

X ($N \times f$) = population characteristics of f variables for N individuals,
 f variables are known for all N individuals in the population

p = number of variables for which survey/admin. systems will
collect data (13)

Y ($N \times p$) = data on p variables for N individuals; only sampled values
available (14)

I ($N \times 1$) = identifies which individuals from the population were
sampled, *i.e.* tells which rows of Y are non-missing

R ($N \times p$) = identifies non-missing elements, *i.e.* tells which
variables are non-missing for which individuals (15)

Y_{obs} = observed data, submatrix of Y ($N \times p$) that contains
only elements where individual was sampled and
provided data on specific variable (16)

Y_{mis} = missing data, submatrix of Y ($N \times p$) that contains
elements where individual was sampled but did not
provide data on specific variable (17)

D = $\{X, Y_{obs}, I, R\}$ or all known data about individuals
in survey sample (18)

In the context of our public use file, the above notation applies as follows: Y ($N \times p$) is a matrix with one row for every member of the U.S. population age 15 and

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

Specific Definitions for SIPP/SSA/IRS-PUF

N = 287 million, i.e. population of U.S.

X ($N \times f$) = race, gender, and birth date, known for all individuals on
Census short form

p = 173 SIPP variables, 443 SER/DER variables, 5 SSA
benefit variables

Y ($N \times p$) = data on all the above variables for entire U.S. population

I ($N \times 1$) = identifies which individuals from the population were
sampled by the SIPP and included in gold standard

R ($N \times p$) = identifies which SIPP and administrative variables are
non-missing for which individuals

Y_{obs} = observed data, submatrix of Y ($N \times p$) that contains
only elements where individual was sampled by SIPP
and data is non-missing

Y_{mis} = missing data, submatrix of Y ($N \times p$) that contains
elements where individual was sampled by SIPP but
did not provide data on specific variable

D = $\{X, Y_{obs}, I, R\}$ or all known data about individuals in
the SIPP samples

In the classic Rubin (1987) missing data application, Y_{mis} is imputed m times by sampling from $p(Y_{mis} | D)$, the posterior predictive distribution of Y_{mis} given D . The completed data consist of m sets $D^{(\ell)} = \{D, Y_{mis}^{(\ell)}\}$, where $Y_{mis}^{(\ell)}$ is the ℓ^{th} draw

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

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

The major focus of the synthesizing process is to obtain a good estimate of the posterior predictive distribution (PPD) for all the variables to be completed and synthesized. We now discuss the computational formulae for estimation and sampling from the PPD. More general methods exist, such as Markov Chain Monte Carlo, but

the methods summarized herein are the ones used by this particular project.

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

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

⁵An exception is the process used to create and synthesize the *ex post* weight, which is described in section 3.5. In that process the full matrix X was used.

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

3.4.2 Bayesian Bootstrap

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

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

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

In other words, beginning with a data matrix, X , that contains values for the k variables of interest, this process assigns a probability of choosing a given observation from X to provide data to a corresponding observation in Y for the k variables. The set of probabilities constitutes a non-parametric representation of the posterior distribution from which the sampling is done. In a conventional bootstrap, because of the assumption that the sample CDF is equivalent to the population CDF, each

observation in X would be assigned probability $\frac{1}{n}$ of being chosen. There would be no uncertainty in what probability would be assigned to a given observation. However, the Bayesian bootstrap accounts for the fact that the sample CDF is not the population CDF and hence does not assign equal probability to each observation. To better understand this concept, consider the example of $k = 1$ where the variable of interest, x_1 , is an indicator variable. Suppose that for 75% of the sample of individuals, $x_1 = 1$ and that $x_1 = 0$ for the remaining 25%. In a conventional bootstrap, with each individual assigned a probability of $\frac{1}{n}$ of being chosen, the CDF used for sampling would always give $x_1 = 1$ a 0.75 probability and $x_1 = 0$ a 0.25 probability. The resulting target matrix Y would not necessarily have a realized 75%/25% frequency distribution for the two values for x_1 but all the bootstrap samples would have been drawn from such a distribution. In a Bayesian bootstrap, when each source record is assigned a unique probability whose expected value is $\frac{1}{n}$, the CDF used for BB sampling might have 73% versus 27% probability of drawing $x_1 = 1$ or 0. The next BB might have 76% versus 24%. The variation in the BB probabilities reflects the fact that the sample proportion of 75% in X is an estimate of the probability that $x_1 = 1$.

Bayesian bootstrap application Choose grouping variables such that the rows of (X, Y_{obs}) can be assumed to come from the same joint distribution within each group defined by the unique combinations of values of the grouping variables. Some collapsing of categories may be required and is described later under implementation details. What is required is the creation of G groups based on the values of the variables in the grouping variable list. It is essential to the success of the Bayesian bootstrap in accurately replicating statistical properties of the data that the observations in a given source (donor) group and a given target (donee) group be as homogenous as possible. Thus, ideally, a large list of grouping variables should be chosen initially.

One of the main advantages of the Bayesian bootstrap is that the group sizes do not have to be as large as groups where parametric modeling is done. Another advantage that is described below is that groups of dependent variables can be done together. This method also helps to preserve the statistical properties of the data by keeping intact relationships among variables.

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

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

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

$h = 1$.

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

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

3.4.3 Sequential Regression Multivariate Imputation

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

complications for the case of longitudinally linked employer-employee data. The algorithms are summarized below ignoring the complications associated with dependent sampling.

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

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

For each dependent variable, the analyst selects grouping variables, conditioning variables and ignored variables. The grouping variables stratify the estimation into G mutually exclusive and exhaustive groups as illustrated in section (3.4.2). The conditioning variables may include all columns of $(X, Y_{\sim j})$, including columns that are created to allow for nonlinearities in the conditional relations. The ignored variables are all columns of $(X, Y_{\sim j})$ not included among the conditioning variables. We wish to generate m implicates $Y_{mis}^{(\ell)}$. SRMI is an iterative procedure. Denote the interim values of implicate ℓ as $Y_{mis}^{(\ell, s)}$. Initialize $\ell = 1$ and $s = 1$. Initialize $Y_{mis}^{(1, 0)}$ using Bayesian bootstrap methods.

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

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

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

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

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

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

2. If converged then

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

(b) Increment ℓ .

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

(d) Reinitialize $s = 1$

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

The test for convergence is not formal. In practice s is often limited to 10 or less. The algorithm estimates the joint distribution $p(Y_{mis}|D)$ by iterating over each conditional distribution $p(Y_{mis,j}|D)$ and filling the “data matrix” with imputed values based on the previous iteration’s estimate of $p(Y_{mis}|D)$. Once the estimation has converged, the implicates are all drawn from the same estimate of $p(Y_{mis}|D)$. However, the completion of D for each implicate results in different conditioning data for the draws. In the implementation of the algorithm, one cycles over the grouping variables $g = 1, \dots, G$ performing the entire algorithm for each homogeneous group. In steps 1.a and 1.b only the conditioning variables appropriate for $Y_{mis,j}$ in conditioning group g are actually included in the conditioning set. The initial selection of these variables

is dependent on the analyst. However after the variables are tentatively included in the model the Bayes Information Criterion (BIC) is used to reduce the variable list by eliminating variables that have a posterior odds ratio below a pre-specified level. The posterior odds ratio cutoff for the BIC value in this variable selection mechanism can be controlled by the analyst. See Abowd and Woodcock (2001) for details.

3.4.4 Summary of synthetic data production

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

Regular missing data, which we multiply impute, result from item non-response or an out-of-scope survey year. Structurally missing data occur when an item is missing due to the logical structure of a set of variables in the survey or administrative record. Structurally missing data still exist in our completed and synthetic data—every individual will not necessarily have a value for every variable. For example, an individual who was born in the United States will have structurally missing data for the variable that indicates which decade the person immigrated to the United States. For survey data, structurally missing values occur when the skip logic of the survey dictates that

a question is not appropriate because of the response given to a prior question. Administrative record data have a similar, albeit implicit, structure. Statisticians usually call such values “structural zeroes.” Structurally missing data are never completed (*i.e.*, imputed) because they do not represent missing information. In contrast, regular missing data, which we complete by multiple imputation, do represent a failure on the part of the survey or administrative records to capture certain information. In this report we use the term “missing” to mean “missing-to-be-completed” and will explicitly describe any other data that are missing as “structurally missing.”

Completing data involves choosing a model for each variable with missing data. We used the SRMI methodology to impute missing values for most of the SIPP variables. The few exceptions are described in 3.4.5 where we give details about the modeling for each variable. We used the BB technique to handle missing data due to missing SSNs. When an individual failed to provide an SSN that could be validated, we could not link that individual to the administrative databases (PCF, SER, DER, and SSA benefits) and, as a result, several hundred administrative variables were missing. One approach to this problem would have been to use the SRMI methodology to model each individual administrative variable and impute missing values. However the magnitude of this task and concern about the need to preserve internal consistency among all the administrative variables, led us to choose the BB completion method for the SSN variable. This method allowed us to choose an appropriate donor record with a non-missing SSN which provided the complete set of administrative variables: PCF (birth date, death date), SER and DER (earnings), and MBR (SSA benefits). Once the SSN had been completed, we treated all administrative data as completed. If a validated SSN did not have a record in a particular administrative database, we treated these data as structurally missing. In other words, no Master Beneficiary Record meant the person had not received benefits from SSA under a program that would generate an MBR record and no DER job records meant

the person had not earned federally taxable income since 1978. Once again, in the completed administrative data, an individual does not have a value for every variable. But individuals who were originally missing SSNs now have donated SSNs which link to administrative data that is either present or structurally missing.

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

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

more cases available for model estimation. The end product of the SRMI process is a data set that contains completed administrative and SIPP variables.

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

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

the 1996 panel. Our hope was that these improved starting data would lead to better models for our completion and synthesis of the wealth variables.

The SRMI method estimates the posterior distribution of the regression parameters (coefficients and variance of the error) and draws from this distribution to obtain parameters used to impute values. We impute multiple times, meaning we take multiple draws from the posterior distribution of the regression parameters. The data product that results is actually a set of files called the completed data implicates. Each implicate has an identical structure (same number of observations, variables, *etc.*) and contains identical data in cases where the information was originally non-missing. For example, if total net worth was non-missing for 75% of the individuals in the sample, then 75% of the observations in each implicate file would have identical values for total net worth. The remaining 25% of the observations would have different values of total net worth across implicate files because of the multiple imputation. The implicates are generated by 4 separate SRMI processes, which is necessary because of the inter-related nature of the variables. Once a variable has been completed, its updated value is used as a right-hand-side variable in the imputation process for other variables. For example, once total net worth has been completed, its updated value will be used to impute a missing value for total income in 1990. Thus, in order to maintain internal consistency within an implicate file, each implicate must be generated separately. For version 4.0, we created four missing data implicates.

Because of the many iterations necessary to complete the data, the majority of the computing time spent creating a synthetic data set is actually spent dealing with missing values. Once the data are completed and contain no missing data except for structurally missing items, the final step of actually synthesizing all the data is takes much less time (*i.e.*, several weeks versus several months). Synthetic implicates are just like completed data implicates except that every individual has his or her values

imputed, variable by variable, conditional on the completed data.⁸ For example, in the case of total net worth, in the data completion phase 25% of individuals received imputed values to replace originally missing data. In the synthesizing phase, 100% of individuals received an imputed value to replace their original data, whether it was missing or not. Synthesizing data is in essence like doing one more iteration of missing data completion except everyone's data has to be completed.

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

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

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

Otherwise, the value from the 9th iteration of completed data was used.

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

3.4.5 Modeling details

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

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

For continuous variables, the imputation model is a normal linear regression, which means that the posterior predictive distribution is based on the normal/inverted gamma posterior distribution for the parameters of a normal linear regression. Under

an appropriate uninformative or conjugate prior, the posterior predictive distribution for the variable under study is normal (given the conditioning variables and the standard error of the equation). If the univariate distribution of the variable we are trying to synthesize, y_k , differs greatly from conditional normality, the distribution of the synthetic values will differ from that of the confidential values. To handle this situation, we transform the confidential data so that they have an approximately normal distribution, estimate the posterior predictive model on the transformed data, and perform the inverse transformation on the imputed values.

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

¹⁰This explanation is simplified. We take proper account of the inverted-gamma distribution on the standard error of the regression. Our procedure samples from the posterior distribution of

the \tilde{y}'_k to have zero mean and unit variance, compute the inverse transformation $\tilde{y}_k = \hat{K}^{-1}(\Phi(\tilde{y}'_k))$. The imputed values \tilde{y}_k are distributed according to \hat{K} , preserving the univariate distribution of the underlying confidential data. Further details of this procedure can be found in Woodcock and Benedetto (2006).

For binary discrete variables, the PPD is based on the asymptotic posterior distribution of the parameters of a logistic regression model. As described in section 3.4.3, we first split our sample of SIPP respondents into homogenous sub-groups using a set of grouping variables (sometimes called by-variables because they specify the subsets of observations that will be used for a particular model). Next, we estimated logistic regression models for each sub-group. We encountered problems with this approach when some sub-groups did not have enough variation to make the computation of a unique maximum likelihood estimate feasible. In other words, for some sub-groups, there were some combinations of right-hand-side variables that perfectly determined some value of the dependent variable. This problem, which is well known in the logistic regression literature see Albert and Anderson (1984), created a continuum of maximizers and prevented convergence in the algorithm used to maximize the likelihood function.¹¹ Because of this problem, known as quasi-separation, the results of the logistic regressions were sometimes unreliable and the coefficients had very large standard errors. The problem of partial ordering of the dependent variable in a logistic regression, which causes the log likelihood function not to have an interior maximum even though it is globally concave, is usually handled by respecifying the logistic regression. Failure to do so causes numerous problems with our synthesizer—in particular, the BIC-based automatic variable selection drops too many variables, if not all of them, and the draws from the posterior predictive distribution are extremely

the standard error of the regression, conditional on the sample error sum of squares and degrees of freedom. The sampled value of the regression equation standard error is used in the conditional normal posterior distribution of the regression coefficients.

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

dispersed.

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

For categorical variables, the PPD is based on the asymptotic posterior distribution of the parameters of a binary tree of logistic regression models that are used to model each level of the categorical variable successively as branches in the binary tree. The categorical variable modeling program looks for an equal split of individuals across categories, thus lumping some of the original categories together, and then models the probability that a person falls in either the first group or the second group. Then, within these two groups, another split is done and the probability that

a person falls into one or the other of these subcategories is modeled. The binary tree continues until all the original categories have been modeled. Finally, the binary tree is synthesized and the synthetic values are used to recreate a synthetic value of the original categorical variable. For example, when the industry variable with four categories is modeled, the program might first split people into groups based on those with $ind_4cat = \{1, 2\}$ and those with $ind_4cat = \{3, 4\}$. It will then split the groups again in order to model $ind_4cat = 3$ versus 4 and $ind_4cat = 1$ versus 2. After the modeling is finished, a new synthesized ind_4cat variable is created that takes on values 1 to 4.

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

Our first tool is to specify parent-child relationships. We define parent variables as those that restrict which observations of another variable are present and which observations are structurally missing. These parent-child relations formalize the skip patterns in the SIPP survey instrument and the logical dependencies in the administrative records. A parent variable determines the universe of observations that are in scope to estimate the model for the associated child variable and will receive an imputed value following the estimation. If the parent variable indicates that the child variable is structurally missing (out of the universe) for an individual, then this observation will not be included in the estimation nor will it receive an imputed value. Instead, it will be set to SAS missing. An example of this type of relationship can be constructed from the variables *foreign_born* and *time_arrive_usa*. *Foreign_born* is the parent variable and takes a value of zero or one for everyone in the data set. It controls whether an individual is in scope to have a value for *time_arrive_usa*, the

child variable. If a person was born outside the US, then that person should have a value for decade of arrival in the U.S. This value may be originally missing or not, but when $foreign_born = 1$, the person is in scope to contribute data to the estimation of the model for $time_arrive_usa$ and will receive an imputed value for this variable that either replaces the missing data or synthesizes the original data. In this manner, we can prevent structurally missing data from skewing our modeling and we can also ensure that only the appropriate people receive a value for $time_arrive_usa$. In this example, the child variable is in-scope only when the parent variable takes a specific value ($foreign_born = 1$). However, the method generalizes so the parent can take on a range of values. For instance, a person is in-scope to have a value for weeks worked part-time if weeks worked with pay is greater than or equal to one and less than or equal to five. In other words, as long as weeks worked with pay is positive, the person is in-scope to have a value for weeks worked part-time. If a person works a full month but never part-time, that person will have weeks worked with pay equal to four or five (depending on the month) and weeks worked part-time equal to zero. If a person does not work at all in a month, that person will have weeks worked with pay equal to zero and weeks worked part-time will be SAS missing.

Our second tool for handling relationships among variables is to place restrictions or constraints on some variables. Constraints do not restrict which observations are used in estimation nor do they restrict which observations receive an imputed or synthetic value. Instead, constraints specify a minimum and maximum value that restricts the range of draws from the posterior predictive distribution. For example, we synthesize birth date for every individual regardless of the value of any other variables. Thus, there is no parent variable for birth date. However the synthesized value for birth date must be consistent with the age requirements for any SSA benefits received by an individual. For example, if the individual began receiving retirement benefits in 1980, he or she must have been born by 1918 at the

latest in order to be at least 62 years old by the time initial retirement benefit receipt. Thus, restrictions are imposed on birth date by the initial type of benefit and date of initial entitlement variables. Our programs impose these constraints by calculating what we term “utility variables” that contain these maximum and minimum values for every constrained variable. When we draw from the posterior predictive distribution for a constrained variable, the candidate sampled value is compared to the maximum and minimum for this individual and if the candidate draw falls outside the specified range, another draw is taken. This comparison and re-sampling is repeated until the candidate sampled value satisfies the constraints or 100 candidate draws have been performed—at which point the value is set equal to the closest boundary (*i.e.*, if the value is over the maximum on the 100th candidate draw, it is set equal to the maximum).

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

variables, *etc.* In most cases, any variable with either a parent or restrictions of some type will be either a level two variable or higher.

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

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

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

Grouping and conditioning variables Finally, as described in sections 3.4.2 and 3.4.3, the analyst chooses both grouping variables and conditioning variables. Group-

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

Problems develop when the grouping variables produce sub-groups that are too small to estimate a statistically reliable PPD. We use the rule that the number of observations in any sub-group must be at least 15 times the number of conditioning variables or 1,000, whichever is greater. To implement this rule, the programs begin with the complete set of grouping variables, form all possible sub-groups, and then check their sample sizes. Sub-groups that are too small are collapsed along specified dimensions and then split into sub-groups again, using a list of grouping variables that is shorter and produces fewer groups. Hence, the analyst actually specifies multiple lists of grouping variables and conditioning variables for each model. Each set of grouping variables is defined by progressively fewer variables as variables are dropped in order to create sub-groups of larger sizes. As variables are dropped from the grouping variables list, they are added to the list of conditioning variables. Hence, each list of conditioning variables becomes progressively longer. For example, the analyst might originally use *black*, *male*, and *age_cat_expand*, an 11 category age variable, as grouping variables. This would produce 44 groups (2 categories for *black*, 2 categories for *male*, and 11 categories for age). The program would form these 44 sub-groups and check the sample size of each group against the minimum of 15

times the number of conditioning variables. If the analyst included 7 conditioning variables, each sub-group would need at least $\max(1000, 105) = 1000$ observations. If the analyst included 100 conditioning variables, then each sub-group would require at least $\max(1000, 1500) = 1500$ observations. Any sub-group that was large enough would be sent directly to the modeling step using the specified conditioning variables. All groups that were too small would be combined and then split again using a the next set of grouping variables specified by the analyst. In this case the analyst might use only *black* and *male* as grouping variables and then include *age_cat_expand* in the list of conditioning variables that corresponds to this second list of grouping variables. This process continues until all the sub-groups meet the minimum observation requirements or until the list of grouping variables provided by the analyst is exhausted, at which point all groups that are still too small are combined and sent to the regression modeling step.

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

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

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

¹²See `varlists_description_version_4_0.xls` in the appendix to this report.

Thus, this original list of unsynthesized variables was chosen to minimize the number of cells in the Gold Standard file with fewer than 10 people when cross-classified by all the unsynthesized variables.

During this final year of the project, the Census Bureau and SSA conducted lengthy discussions about the possibility of including unsynthesized SSA benefit variables on the file. Although these variables were administrative and hence did not have direct equivalents in the original SIPP survey files, the Census Bureau was concerned that adding more unsynthesized variables to the file would create even more small cells that would allow a user to link across synthetic implicates. If the synthetic implicates were linked, they could be averaged and something resembling the original record could possibly be re-created. The Census Bureau felt that this possibility presented too much disclosure risk and preferred to keep the number of unsynthesized variables small enough to avoid large numbers of cells with fewer than 10 people.

Discussions between the two agencies produced the following compromise. Gender, marital status, and the spouse-link would remain unsynthesized. In addition, we would add two important SSA benefit variables to the unsynthesized list: type of benefit at time of initial benefit receipt and type of benefit in April 2000. These two categorical variables quantify fact of receipt as well as the reason and are hence the most fundamental of all the SSA benefit variables. Thus, the list of unsynthesized variables in the final version of the synthetic public use files is gender, marital status, initial type of benefit, type of benefit in 2000, spouse initial type of benefit and spouse type of benefit in 2000 (both created using the unsynthesized spouse link), and the spouse identifier variable.¹³ The resulting configuration of unsynthesized variables

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

creates no small cells using only the variables originating from Gold Standard SIPP variables. Furthermore, there are only approximately 130 cells with fewer than 10 individuals when stratifying using the full list, which includes the two SSA-provided type of benefit variables that are not present on any current SIPP public use file. See Table 3.1 for a full break down of small cells created by various configurations of unsynthesized variables.

The existence of unsynthesized variables required the imposition of some constraints on other variables. In particular, receipt of certain types of benefits imposed constraints on an individual’s age at a given point in time and marital status at the time of the survey imposed constraints on the marital history of an individual. We describe how we handled these restrictions in sections 3.4.5 and 3.4.5, respectively.

Birth date, death date, and dates of benefit receipt One of the most important variables in the file, from the perspective of both disclosure risk and usefulness in analyses, is *birthdate*.¹⁴ It was essential that this variable be adequately protected yet synthesized well enough to reproduce appropriate age distributions for many sub-groups. We used the administrative value of the date of birth (from SSA administrative records) whenever we could. The administrative *birthdate_pcf* was missing in cases where the individual did not have a validated SSN and was completed using the couple-level Bayesian bootstrap described in 3.4.4. We modeled the variable in the data synthesis phase as a continuous variable with restrictions. If a person received benefits in April 1, 2000 ($tob_2000 = \{1, 2, 3, 5, 100\}$), we forced the synthetic *birthdate* to be such that the individual would be appropriately old enough for the benefit received. Individuals with retirement benefits had to be at least 62 by April 1, 2000, individuals with aged spouse benefits ($tob_2000 = 3$) had to be at

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

least 62, individuals with aged widow benefits ($tob_2000 = 5$) had to be at least 60, and individuals receiving disability benefits ($tob_2000 = 2$) could not be 65 years old or older. In addition, we restricted draws for the synthetic *birthdate* such that it was no more than a year in either direction from the original administrative birth date (*birthdate_pcf*). So that

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

where we note that date variables are measured in days.

Because *tob_2000* and *tob_initial* are unsynthesized, further consistency restrictions were imposed on *birthdate*. If an individual's initial benefit types were retired or retired spouse ($tob_initial = \{1, 3\}$) and unsynthesized *date_initial_entitle* is before April 1, 2000, then *birthdate* must be consistent with age at April 1, 2000 greater than 62. The reason for this restriction is that when *date_initial_entitle* is synthesized, there will be support for a synthetic value that is consistent with these types of benefits starting before April 1, 2000. If an individual's initial benefit types were retired or retired spouse ($tob_initial = \{1, 3\}$) and unsynthesized *date_initial_entitle* is on or after April 1, 2000, then *birthdate* must be consistent with the individual turning 62 (and thus being eligible for these types of benefits) before December 31, 2002.¹⁵ The same process is repeated for aged widow benefits ($tob_initial = 5$) using an age cut off of 60. Finally, for disabled benefits ($tob_initial = 2$) we reverse this procedure to keep *birthdate* consistent with being less than 65 years old when this type of benefit is collected. If $tob_initial = 2$ and unsynthesized *date_initial_entitle* is on or after April 1, 2000, then the minimum synthetic *birthdate* is May 1, 1935 so that there is support for a synthetic *date_initial_entitle* on or after April 1, 2000 and the individual would be age-eligible for disability benefits at that time.

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

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

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

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

initial entitlement date and April 1, 2000, we ensure that the explanation for this loss of benefit is the date of death. If there is no benefit amount reported for the entire month of April 2000 ($tob_2000 = \text{SAS missing}$), the initial benefit type is present ($tob_initial > 0$), and the unsynthesized initial entitlement date is on or after May 1, 2000 ($date_initial_entitle \geq \text{May 1, 2000}$), then the minimum value for *deathdate* is May 1, 2000. We do this to create support for a draw of the synthetic date of initial entitlement that is consistent with receiving no benefits in the month of April 2000 and the synthetic date of death.

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

value was forced to be no more than 2 years in either direction from the original value.

Administrative earnings After completing missing SER and DER data using the couple-level Bayesian bootstrap, the administrative earnings variables were synthesized in two parts. We first modeled whether the SIPP individual had positive earnings in a given year and then only modeled actual earnings for those with a positive earnings indicator. Thus, the earnings indicator was the parent variable and the actual earnings variable was the child. We synthesized the earnings indicators using a Bayesian bootstrap, done one year at a time. We used leads and lags for previous and future years as grouping variables as well as demographic variables and summary earnings measures. We began with SER earnings (capped at the FICA maximum) in 1951. Using the bootstrap, we created a synthetic value for every individual for the variable *ser_posearn_1951*. For those with *ser_posearn_1951*=1, we then used a bootstrap to create a synthetic value for whether each individual had reached the FICA taxable maximum in 1951 (*ser_maxearn_1951*). For those with *ser_maxearn_1951* = 1, we automatically set *totearn_ser_1951* equal to the maximum. For those with *ser_maxearn_1951* = 0, we modeled earnings using our continuous variable techniques, including the two-sided KDE transform. After 1951 was completed, we moved to 1952 and repeated the process. When creating grouping variables for 1952, we used the new synthetic values for 1951 and the completed data for 1953 and after. We moved through the entire array in this manner until the year 1978.

The DER array of earnings begins in 1978. Beginning with this year, we synthesized total earnings. We used a similar process to the one used for the SER earnings except that we synthesized four separate time series: non-deferred total earnings at FICA covered jobs (*nondefer_der_fica_{year}*), deferred total earn-

ings at FICA covered jobs ($defer_der_fica_ \{year\}$), non-deferred total earnings ($nondefer_der_nonfica_ \{year\}$) at non-FICA covered jobs, and deferred total earnings at non-FICA covered jobs ($defer_der_nonfica_ \{year\}$). After each year of DER earnings was synthesized, we calculated SER earnings as the lesser of total non-deferred and deferred earnings at FICA covered jobs or the FICA taxable maximum:

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

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

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

Social Security benefits We synthesized two SSA benefit variables: monthly benefit amount for the month of initial entitlement and monthly benefit amount for April 2000. Each of these variables was the child of the corresponding type of benefit variables. Only individuals with a positive initial type of benefit received a synthesized value for the initial MBA ($mba_initial$) and likewise for mba_2000 . However since neither type of benefit variable was synthesized, the set of people with positive $mba_initial$ and mba_2000 values was the same in the completed and synthetic data. Both MBA variables were synthesized using continuous variable methods and were restricted such that synthetic values had to be no more than \$50 less than or greater than the original values:

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

and similarly for *mba_2000*.

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

For individuals who reached age 62 before 1979, we calculated the *AMW* and for those who reached age 62 after 1979, we calculated the *AIME*. To compute the *AMW*, we first calculated the number of years between age 21 (or 1951 if later) and age 62, subtracted five years, and multiplied by 12 to get the number of months at risk. We then summed earnings between age 21 and age 62, dropping the five lowest years. Total summed earnings were then divided by the number of months at risk to give the Average Monthly Wage. There was one exception. For men (but not women) born before 1911, the calculation was performed using the years between age 21 and age 65 because the retirement age for men was three years older prior to

1973. The *AIME* calculation was essentially the same as the *AMW* but earnings were indexed to the year in which the individual turned 60.

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

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

SIPP time series arrays The synthetic data includes 13 time series of SIPP variables: weeks with pay, weeks part-time, total annual hours, family poverty cut-off, family total income, personal total income, personal total earnings, family welfare participation, family welfare income, private health program participation, private health program income, general health insurance coverage, employer-provided health

insurance coverage. In addition, weeks with pay, weeks part-time, annual hours, family income, personal income, and personal earnings have corresponding arrays of indicator variables that serve as parent variables and tell whether the continuous variable takes on a value or not. We use a Bayesian bootstrap to complete and synthesize all the indicator arrays. We then use continuous variable methods to complete and synthesize the remaining variables with the indicators serving as parent variables.

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

Marital history variables The challenge in synthesizing the marital history variables was to ensure that the historical variables were consistent with the reported marital status and with each other. To accomplish this, we used a Bayesian bootstrap to both complete and synthesize marital history variables. We first bootstrapped a group of variables that summarized the history (*mh_category*, number of marriages, number of divorces, and married at end of history) using marital status as one of the grouping variables. This guaranteed that individuals would receive donated values of *mh_category* and the three other summary variables only from other individuals with the same marital status so no inconsistencies would arise. We then used an additional Bayesian bootstrap for *flag_mar4t* with *mh_category* as one of the grouping variables. Once these variables had been modeled, we created a set of

indicator flags that indicated whether the individual should have an age at time of first marriage, duration of first marriage, duration of end of first marriage, duration of second marriage, duration of end of second marriage, duration of third marriage, duration of end of third marriage, and duration of fourth marriage based on the events that occurred in his or her history. Individuals with at least one marriage in their history were modeled to have an age at time of first marriage. Individuals whose first marriage had ended were modeled to have a duration of first marriage and duration of first marriage end. Individuals with at least two marriages were modeled to have a duration of end of first marriage (*i.e.*, time between first and second marriages) and duration of second marriage and so on until the fourth marriage. The age and duration variables were modeled using our continuous variable techniques and were children of the indicator flags.

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

50% of the time, the interval between marriages accounted for 25% of the time, and the second marriage accounted for 25% of the time, then age at time of first marriage ending would be 35 (1993) and age at time of second marriage would be 40 (1998).

3.5 Weight Creation and Synthesis

3.5.1 Introduction and background

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

In order to determine how many people in the reference population each SIPP person represented, we used the 1996 SIPP sampling plan as our guide and divided the Decennial reference population into the same strata (*i.e.*, groups) from which

SIPP individuals were originally sampled. We then located each SIPP individual in the Decennial reference population. Once we knew how many SIPP people were in each stratum, the preliminary weight calculation was straight forward: each SIPP person's weight equals the number of Decennial persons in that particular stratum divided by the number of SIPP persons in the same stratum. For example, if the tenth stratum contained 100 Decennial persons and two SIPP sample individuals, then each SIPP person in the tenth stratum received a preliminary weight of $50=100/2$. The final weight was calculated by raking the preliminary weight to match official U.S. civilian non-institutional population estimates as of April 1, 2000 based on the same control total categories used for the 1996 SIPP weights in the current public use files. The validity of the final weight was tested by computing univariate statistics for key SIPP and SSA variables and comparing them to independently derived estimates from other sources. The results of this testing are reported in Table 3.2.

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

The strata from the SIPP sampling plan had several levels. The first stratification level (or grouping level) was Primary Sampling Units (PSUs), which were created by

grouping geographic counties together. The SIPP Survey Design Branch (SIPPSDB) in the Demographic Statistical Methods Division (DSMD) provided us with a file that assigned geographic counties to PSUs. Large counties were assigned a unique PSU while smaller counties were grouped together to form a single PSU. The second stratification level was by stage-1-clusters, which were simply created by grouping PSUs together. Some PSUs were self-representing, meaning that they were the only PSU in their stage-1-cluster and were sampled with certainty. Other PSUs were non-self-representing, meaning they were grouped with other PSUs and were sampled with probability less than one. The SIPP Survey Design Branch provided us with a file that assigned the 1,928 PSUs to 217 stage-1 clusters. These stage-1 clusters were then used to select PSUs from which individuals would be sampled. Once PSUs were selected, individuals in high poverty strata were over-sampled in each selected PSU. Therefore, our final stratification level was defined by whether an individual was in the high poverty stratum or the low poverty stratum according to the definitions of high and low poverty in the SIPP Sampling Plan. The final stratification which combined the location of an individual in a stage-1-cluster and a poverty stratum was called a stage-2-cluster. The number of SIPP and Decennial persons in each stage-2 cluster was used to calculate the preliminary weight according to the above formula. Raking was then applied directly to the preliminary weight to create the final weight. Finally, a synthetic version of the weight was created for each of the synthetic implicates.

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

3.5.2 Summary of the weight creation process

Our method for creating an ex-post weight for the merged SIPP panels involved seven steps. Parts A and B describe the method of creating the Census 2000 reference population and dividing it into strata according to the 1996 SIPP Sampling Plan. Parts C and D do the same for the SIPP, describing the method by which the SIPP was divided into strata according to the 1996 SIPP Sampling Plan. Part E describes the method by which each SIPP person was located in the Decennial reference population. Part F describes the creation of the preliminary weight according to the formula mentioned above. Part G describes the creation of the final weight by raking (*i.e.*, adjusting) the preliminary weights to agree with official U.S. population control totals for the sex/age/race/ethnicity demographic breakdown of the reference population, as supplied by the Census Bureau's Population Estimates Division. The next two subsections (3.5.11 and 3.5.12) describe some geography and birth date issues that arose during the weight creation process. The next subsection (3.5.13) discusses the overall evaluation of the Gold Standard weight, and the final two subsections (3.5.14 and 3.5.15) describe the creation of the synthetic weight and discuss the results of the analytical validity testing of this weight.

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

Part A describes the creation of a poverty stratification variable for Census 2000 records according to original 1996 SIPP stratification rules. Households were assigned to a poverty stratum based on either household income or household composition. For long form households (Sample Census Edited File, SCEF), an income variable was available and households/records were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty threshold. For long form respondents for whom income data was not available and for short form respondents (Hundred percent Census Edited File, HCEF), household composition was used to

proxy poverty status. A household was assigned to the high poverty stratum if it had any of six characteristics such as a black householder under age 18 or over age 64 (see 3.5.3 below for the full list of characteristics).

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

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

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

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

population) in the following manner. Each SIPP person was first assigned a set of Decennial candidate records (candidates for a match) that agreed exactly with that SIPP record's values for each variable in a set of blocking variables. Then, one of the Decennial candidates was chosen as a match for the SIPP record based on how closely that Decennial record's values agreed with that SIPP record's values for each variable in a set of matching variables. There were two blocking passes through the data. The first blocking pass used 6 blocking variables and 7 matching variables (see 3.5.8 below for the complete list). Any SIPP record that had 30 or fewer Decennial candidates was considered unmatched and sent through the second blocking pass, which used 3 blocking variables and 10 matching variables.

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

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

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

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

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

person to be the householder. This file was then merged to each person's record.

Data sources for long-form respondents Information about long-form individuals came from three SCEF data files: block-level, housing-unit level, and person-level files. From the person-level file we obtained the same demographic and geographic variables as from the short-form: indicators for householder, child of householder, spouse of householder, gender, black, Hispanic, age groups (<18, 18-64, >64; and <18, 18-62, >62), birth date, state, county, and approximate tabulation geography. In addition, we obtained information on education (college, some high school) and income (total annual personal income, 1999). From the block-level file we also obtained the same variables as from the short-form: county, state, population count, housing count, place code by *geocodefull* (unique tabulation block identifier) as well as housing counts and population counts. Finally, from the housing-unit file we obtained an indicator for family-type housing versus group quarters, count of persons living in family-type housing, number of children under age 18, and an indicator for monthly rent below \$300. We also used the person-level data to create some additional housing-unit information, in particular an indicator for family-type housing versus group quarters, count of persons living in family-type housing, number of children under age 18, householder information (female, black, Hispanic, age: <18, 18-64, >64), and an indicator variable for households with female householder and no spouse present. In cases where no person was assigned to be the householder (*e.g.*, group quarters have no householder in the Decennial), we assigned the oldest person to be the householder. Using the person-level income variable, we created a variable for total annual housing unit income in 1999. All household information was again attached to each person's record.

Data source for MSA variable The Population Division provided us with a file that included an indicator for "Living in a central city (MSA)". This indicator

was merged to the Census 2000 records by state, county, and Census place code. Accordingly, 82,249,968 persons lived in a central city and there were 636 unique central cities.

3.5.4 Poverty stratum assignment

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

- if one-person-housing-unit, age of householder ≤ 64 years, and no children under 18 years then poverty threshold 1999(hhpov1999)=8667;
- else if one-person-housing-unit, age of householder > 64 , and no children under 18 years then poverty threshold 1999 =7990;
- else if two-person-housing-unit, age of householder ≤ 64 years, and no children under 18 years then poverty threshold 1999 =11156;
- else if two-person-housing-unit, age of householder > 64 , and no children under 18 years then poverty threshold 1999 =10070;
- else if two-person-housing-unit, age of householder ≤ 64 years, and 1 child under 18 years then poverty threshold 1999 =11483;
- else if two-person-housing-unit, age of householder > 64 , and 1 child under 18 years then poverty threshold 1999 =11440;
- else if three-person-housing-unit and no children under 18 years then poverty threshold 1999 =13032;
- else if three-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =13410;
- else if three-person-housing-unit and 2 children under 18 years then poverty

threshold 1999 =13423;

- else if four-person-housing-unit and no children under 18 years then poverty threshold 1999 =17184;
- else if four-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =17465;
- else if four-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =16895;
- else if four-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =16954;
- else if five-person-housing-unit and no children under 18 years then poverty threshold 1999 =20723;
- else if five-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =21024;
- else if five-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =20380;
- else if five-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =19882;
- else if five-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =19578;
- else if six-person-housing-unit and no children under 18 years then poverty threshold 1999 =23835;
- else if six-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =23930;
- else if six-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =23436;
- else if six-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =22964;

- else if six-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =22261;
- else if six-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =21845;
- else if seven-person-housing-unit and no children under 18 years then poverty threshold 1999 =27425;
- else if seven-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =27596;
- else if seven-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =27006;
- else if seven-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =26595;
- else if seven-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =25828;
- else if seven-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =24934;
- else if seven-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =23953;
- else if eight-person-housing-unit and no children under 18 years then poverty threshold 1999 =30673;
- else if eight-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =30944;
- else if eight-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =30387;
- else if eight-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =29899;
- else if eight-person-housing-unit and 4 children under 18 years then poverty

threshold 1999 =29206;

- else if eight-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =28327;
- else if eight-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =27412;
- else if eight-person-housing-unit and 7 children under 18 years then poverty threshold 1999 =27180;
- else if >=9-person-housing-unit and no children under 18 years then poverty threshold 1999 =36897;
- else if >=9-person-housing-unit and 1 child under 18 years then poverty threshold 1999 =37076;
- else if >=9-person-housing-unit and 2 children under 18 years then poverty threshold 1999 =36583;
- else if >=9-person-housing-unit and 3 children under 18 years then poverty threshold 1999 =36169;
- else if >=9-person-housing-unit and 4 children under 18 years then poverty threshold 1999 =35489;
- else if >=9-person-housing-unit and 5 children under 18 years then poverty threshold 1999 =34554;
- else if >=9-person-housing-unit and 6 children under 18 years then poverty threshold 1999 =33708;
- else if >=9-person-housing-unit and 7 children under 18 years then poverty threshold 1999 =33499;
- else if >=9-person-housing-unit and >= 8 children under 18 years then poverty threshold 1999 =32208;

When income data were not available for long-form households, household composition was used to proxy poverty status. A household was assigned to the high

poverty stratum if it had any of the following characteristics:

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

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

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

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

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

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

cluster. We then combined the stage-1 cluster with the poverty stratum created in Part A and created stage-2 clusters.

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

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

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

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

- Montana: Yellowstone National Park (30-113) was annexed to Gallatin (30-031) and Park (30-067) counties in November 1997. Park County and Yellowstone National Park were assigned in 1990 to the same PSU and stage-1-cluster, Gallatin

was assigned to a different PSU and stage-1-cluster. Because most people were moving to Park County from Yellowstone National Park and only very few people were living in Yellowstone National Park, no changes were made to the sampling frame, except that the record for Yellowstone National Park was taken out.

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

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

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

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

Dropping Census 2000 records that were out-of-scope for SIPP samples Because of the differing nature of a census and a program survey, we recognized the need to exclude some Decennial records as out-of-scope to be sampled for the SIPP. The SIPP Quality Profile 1998, third edition states:

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

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

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

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

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

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

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

We then created the poverty stratification variable for each SIPP record. Individuals were assigned to strata based on either household income or household composition. For Gold-Standard respondents surveyed in the 1996 SIPP panel, 1999 household income was available (81,409 respondents) and they were assigned to the high poverty stratum if 1999 household income was below 150 percent of the poverty

threshold for their household type. Thresholds were defined according to criteria used for the Census 2000 records (see 3.5.3).

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

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

It was not possible to assign SIPP respondents to the high poverty stratum based on whether they lived in the central city of an MSA (as was done for the Decennial respondents) because this variable depended upon knowing state, county, and Census place code information for each household, and we did not have Census place code on the internal SIPP file. The final stage of adjusting the weight to correct population control totals within sex, race, ethnicity, and geographic location (see 3.5.10) handled this problem.

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

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

As with Census 2000 records, we used the information provided by the SIPP Survey Design Branch to assign Primary Sampling Units (PSUs) to geographic entities. To assign each SIPP individual to a PSU, we needed state and county information. Unfortunately, county level geography was very difficult to obtain for the 1990-1993 SIPP panels. Given the likelihood that an individual's county had changed between

the early 1990s and 2000, we did not invest in obtaining SIPP county information for the early panels. Instead we used the state variable recorded for respondents during the last year of their panel and then randomly assigned county and the corresponding PSU. For respondents from the 1996 panel, state and county geography was available and PSUs were assigned as they were for Census 2000.

Once SIPP respondents were placed in PSUs, the creation of stage-1 and stage-2 clusters proceeded as outlined in 3.5.5. At this point, we used the link between the Decennial and the SIPP to flag SIPP individuals who matched to a Decennial record that had previously been determined to be out-of-scope, as explained in 3.5.5. The Gold Standard version 4.0 file contained 263,793 people, 177,165 of which were matched by PIK (*i.e.*, replacement SSN) to a Decennial record. Of these 177,165 records, 2,229 were matched to a Census 2000 record that was out-of-scope for the SIPP, meaning that these SIPP records received a zero weight in the final weight calculation. The remaining 261,564 SIPP in-scope records were used in calculating the weight. The link between the Decennial and the SIPP essentially served to indicate when a person interviewed in the 1990s had experienced a life-change by 2000 that removed them from the reference population.

After removing the SIPP records that matched to out-of-scope Decennial records, we made the appropriate SIPP cell counts for the 434 stage-2-clusters. The 261,564 SIPP in-scope observations translated into a 602.68 mean cell count. The largest cell contained 4,210 observations and the smallest cell contained 2 observations. The strata with very small numbers of SIPP observations could have presented a confidentiality problem when the weight is used on the SIPP/SSA/IRS public use file. We addressed this issue by synthesizing the weight in the Preliminary PUF 4.0 (see 3.5.14).

3.5.8 Part E: Matching SIPP individuals to Census 2000 records

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

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

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

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

The 7 matching variables were

- a. birth month
- b. children under 18 (dummy variable)
- c. no spouse present (dummy variable)
- d. female householder (dummy variable)
- e. black householder (dummy variable)
- f. Hispanic householder (dummy variable)
- g. age of householder (<18,18-64,>64)

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

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

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

record within the same set of blocking variables had values for the matching variable that agreed exactly.

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

Once m and u probabilities were created for each matching variable and for each SIPP record, agreement and disagreement weights were created for each Decennial candidate as follows: agreement weight = $\ln(m/u)$ and disagreement weight

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

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

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

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

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

3.5.9 Part F: Creation of preliminary weight

After all SIPP records were matched to Decennial records, a preliminary weight was calculated. In order to calculate this weight, we used the Decennial stage-2 cluster counts from 3.5.5 and the SIPP stage-2 cluster counts from 3.5.7. The preliminary weight was calculated using the following formula:

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

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

3.5.10 Part G: Creation of the final weight

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

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

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

by the preliminary weight to create the final weight for each individual. The final weights ranged from 30.90 to 32,625.69, with a mean of 780.64.

3.5.11 Geography issues

Different geography concepts on HCEF and other Census 2000 files In order to establish the poverty indicator we used information from the HCEF and from other files that were merged onto the HCEF either through person IDs or geography (*e.g.*, central city indicators, PSUs). Merging by person IDs did not pose any problems, but merging by geography did. We were working with an internal HCEF file that had not yet been converted to the “tabulation” geography concept that the SCEF (as well as the other files) used. Our HCEF file had geography that was on collection geography level, which made it easier for the enumerators to perform the interviews. The SCEF file we used (as well as the files that we received from the pop-division and the SIPP Survey Design Branch) had tabulation geography (which was the geography concept that all the Census 2000 tabulations on the web used, for example). While state information was the same for “collection” and “tabulation” geography, county information could be different. Merging therefore was not straightforward. On our internal version of the HCEF was another geography variable (Current geography). This was not “Tabulation geography” either but matches it reasonably well. We used this variable to merge by geography.

Changing geography boundaries between the 1990s and 2000 Geography and, especially, county boundaries changed between 1990 and 2000. There were boundary changes as well as the deletion and creation of new counties during that time frame. This affected the use of SIPP-sampling units, because the SIPP sampling units for the 1991-1996 panels were created using the 1990 Census, and the SIPP sampling units for the 1990 SIPP sample used geography from before the 1990 Census. Also,

people moved across county boundaries and therefore were counted in different PSU units in 2000 compared to the time they started participating in the SIPP survey. The changes that were made to accommodate the additions and deletions of counties were written down in detail in 3.5.5. We have not made any changes for counties that purely changed boundaries.

3.5.12 Birth date issue

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

3.5.13 Overall evaluation of Gold Standard weight

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

Table 3.2, Columns B and C presents the results of testing of the Gold Standard weight. We chose several selected statistics from the 2001 SSA Annual Statistical Supplement and calculated these same statistics using our weighted Gold Standard data. Our weighted Gold Standard file reproduces all of these selected statistics fairly closely. In particular, the number of workers receiving retirement benefits in December of 2000 in the Gold Standard data is lower than the number reported by SSA by only one million. The number of widows and widowers receiving benefits in the Gold Standard is lower than the corresponding SSA statistic by only 300,000, and the number of disabled receiving benefits is higher by only 800,000. The average monthly benefit received by these various sets of workers falls within 3%, 7%, and 6% of the

SSA reported average monthly benefit for retired workers, widows and widowers, and disabled workers, respectively. The number of permanently insured individuals in December 2000 in the Gold Standard data falls within 1% of the corresponding number reported by SSA, and the number of wage and salary workers with taxable earnings for 2000 falls within 3% of the SSA reported number. The DER average earnings for 2000 in the Gold Standard is about \$3,000 higher than the DER average earnings reported by SSA, and the SER average earnings for 2000 in the Gold Standard is about \$1,400 higher than the SER average earnings reported by SSA. In general, we believe our Gold Standard weight does a particularly good job of reproducing these selected statistics from the 2001 SSA Annual Statistical Supplement.

3.5.14 Synthesizing the weight

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

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

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

where θ_j is the probability of taking on the j th outcome category; these probabilities

sum to one ($\sum_{j=1}^k \theta_j = 1$).¹⁶ The total number of observations is $\sum_{j=1}^k y_j = n$. The conjugate prior distribution for this multinomial distribution is known as the Dirichlet,

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

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

In our application, each person in the data can take on one of 55,552 possible values for the weight.¹⁷ The sum of the weights played the role of the “data sample size, and equaled 204,044,727. We used a noninformative prior distribution by spreading additional observations evenly across the 55,552 cells; this was the “prior sample size.”¹⁸ The sum of the “data sample size” and the “prior sample size,” is called the “posterior sample size” in the posterior Dirichlet distribution for the cell probabilities.

In practice, we replaced the likelihood counts, y_j , with their expected values. We used the SAS procedure PROC CATMOD to model the expected counts for each of the possible 55,552 cells created by the six strata variables (stage-1 cluster, poverty stratum, male, black, Hispanic, and age category). This procedure performs categorical data modeling of data that can be represented by a contingency table. We supplied the procedure with the weighted cell count data from the completed data, where each observation was a cell in the contingency table created by the complete

¹⁶It should be noted that the θ 's in this chapter are entirely separate from the θ 's in chapter 2.

¹⁷This number of different possible values for the weight comes from the fact that the weight differed only by the values of the following variables: stage-1 cluster, poverty stratum, male, black, hispanic, and age category. There were 217 stage-1 clusters, 2 values for poverty stratum, male, black, and hispanic, and 16 age categories, resulting in $217*2*2*2*2*16 = 55,552$ possible unique values for the weight.

¹⁸By agreement with the Census Bureau Disclosure Review Board, we do not disclose the prior sample size when a Dirichlet prior is used for confidentiality protection.

cross-classification of the six strata variables, and each cell count was the weighted sum of the number of persons in that cell. The procedure used maximum likelihood analysis to estimate a log-linear model and calculate the predicted cell frequencies. We computed the maximum likelihood estimates using an iterative proportional fitting algorithm rather than the usual Newton-Raphson algorithm because it allowed us to obtain the predicted cell frequencies without performing time-consuming parameter estimation. The log-linear model included all six main effects (one for each stratum variable), all two-way interaction effects, and a single three-way interaction effect between the poverty stratum, black, and Hispanic variables.

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

3.5.15 Evaluation of the synthesized weight

Table 3.2, Columns C and D present the results of comparing the weighted completed data to the weighted synthetic data for the same published SSA statistics as were chosen for the testing of the Gold Standard weight. The results from the synthetic data very closely match those from the completed data. Column E shows that the percentage difference between these statistics for the two types of data is very small, ranging from no difference in the number of disabled workers receiving benefits to 4.4% difference in the number of widows and widowers receiving benefits. More specifically, the estimated number of individuals in the reference population

receiving retirement benefits in December of 2000 in the synthetic data is lower than the estimated number in the completed data by 700,000. The estimated number of widows and widowers receiving benefits in the synthetic data is lower than the corresponding statistic in the completed data by only 200,000, and the number of disabled receiving benefits is exactly the same in the synthetic and completed data. The average monthly benefit received by workers in the synthetic data falls within 1% of the average monthly benefit in the completed data for all three types of workers. The number of permanently insured individuals in December 2000 in the synthetic data falls within 2% of the corresponding number in the completed data, and the number of wage and salary workers with taxable earnings for 2000 falls within 1% of the corresponding statistic in the completed data. The DER average earnings for 2000 in the synthetic data is about \$1,400 higher than the DER average earnings in the completed data, and the SER average earnings for 2000 in the synthetic data is about \$800 higher than the SER average earnings in the completed data. Overall, we have shown that our weighted synthetic data does a very good job of matching our weighted completed data on these selected statistics from the 2001 SSA Annual Statistical Supplement.

3.6 Analytical Validity

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

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

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

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

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

3.6.1 Complete data estimation

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

Estimates of Q are random because they are based on D , which involves sampling from the finite population and incomplete observation of Y in the sample. We can only calculate an estimate of the average 1990 income because of the sampling involved with the SIPP and because not all SIPP individuals provided 1990 income data. When all sampled individuals provide data on all p variables, there are no item missing data. However, an estimator of Q is still random because of the sample design embodied in I . Even if all SIPP individuals in our sample reported 1990

income, the sample design of the SIPP would still make the average 1990 income a random variable. We will call the complete data estimator $q(D)$ and its variance estimator $u(D)$. Notice that because of the definition of complete data, q and u depend only on (X, Y_{obs}, I) and not on R . The analyst is assumed to have an inference system for $q(D)$ and $u(D)$. In particular, complete data inference can be based on $(q(D) - Q) \sim N(0, u(D))$, which may be exact or an approximation but is assumed to be appropriate in what follows.

3.6.2 Inference frameworks using multiple imputation

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

statistic calculated on each implicate file:

$$q^{(\ell)} = q(D^{(\ell)}).$$

In our example the function q is the average of 1990 income across all individuals in the sample. This average is calculated separately for each implicate and then

averaged across implicates as the next formula indicates:

average of the statistic across implicates:

$$\bar{q}_m = \sum_{\ell=1}^m \frac{q^{(\ell)}}{m}.$$

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

variance of the statistic across implicates:

$$b_m = \sum_{\ell=1}^m \frac{(q^{(\ell)} - \bar{q}_m) (q^{(\ell)} - \bar{q}_m)'}{m - 1}$$

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

variance of the statistic on each implicate file:

$$u^{(\ell)} = u(D^{(\ell)})$$

and

average variance of the statistic across implicates:

$$\bar{u}_m = \sum_{\ell=1}^m \frac{u^{(\ell)}}{m}.$$

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

total variance of the average statistic across implicates:

$$T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

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

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

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

Missing and partially synthetic data In order to analyze synthetic data that were created from data that originally contained some missing values, the missing data imputation and the synthetic data sampling must be done sequentially. First, complete m versions of D by sampling from $p(Y_{mis}|D)$. Denote the m completed data sets as $D^{(\ell)} = \{X, Y_{obs}, Y_{mis}^{(\ell)}, I, R\}$, $\ell = 1, \dots, m$. Let the vector Z ($n \times 1$) denote entities i for which any values of Y_{obs} have been synthesized. So, $Z_i = 1$ if any of the values of $Y_{obs,i}$ have been synthesized. Partition Y_{obs} into Y_{nrep} containing the rows

where $Z_i = 0$ and Y_{rep} containing the rows where $Z_i = 1$. Then, for each completed data set, partially synthesize r implicates by sampling from $p(Y_{rep}|D^{(\ell)}, Z)$. Denote the r completed partially synthetic data sets as $D^{(\ell,k)} = \{X, Y_{nrep}^{(\ell)}, Y_{rep}^{(\ell,k)}, I, R, Z\}$, $k = 1, \dots, r$ and where $Y_{nrep}^{(\ell)}$ corresponds to the rows of $(Y_{obs}, Y_{mis}^{(\ell)})$ for which $Z_i = 0$ and $Y_{rep}^{(\ell,k)}$ corresponds to the rows of $(Y_{obs}, Y_{mis}^{(\ell)})$ for which $Z_i = 1$. Note that $Y_{nrep}^{(\ell)}$ contains no synthetic data but may contain missing data imputations whereas $Y_{rep}^{(\ell,k)}$ may contain both missing data implicates (an element of $Y_{rep}^{(\ell,k)}$, say ij , for which item j is missing for entity i but not synthesized; entity i is in this set because $Z_i = 1$ whenever any element of Y_{inc} is synthesized) and synthetic data (an element of $Y_{rep}^{(\ell,k)}$, say ij , for which item j is missing for entity i and is synthesized; entity i is in this set because $Z_i = 1$ and element j element of $Y_{inc,i}$ is synthesized).

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

statistic calculated on each implicate file:

$$q^{(\ell,k)} = q(D^{(\ell,k)}).$$

Then, we average across the 2 synthetic implicates that correspond to a given missing

data implicate creating $\bar{q}^{(1)}, \bar{q}^{(2)}, \bar{q}^{(3)}, \bar{q}^{(4)}$ according to the formula:

average of the statistic across the synthetic implicates:

$$\bar{q}^{(\ell)} = \sum_{k=1}^r \frac{q^{(\ell,k)}}{r}$$

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

average of the statistic across synthetic and missing data implicates:

$$\bar{q}_M = \sum_{\ell=1}^m \sum_{k=1}^r \frac{q^{(\ell,k)}}{mr} = \sum_{\ell=1}^m \frac{\bar{q}^{(\ell)}}{m}.$$

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

variance of the statistic due to variation in synthetic implicates:

$$b^{(\ell)} = \sum_{k=1}^r \frac{(q^{(\ell,k)} - \bar{q}^{(\ell)}) (q^{(\ell,k)} - \bar{q}^{(\ell)})'}{r - 1}.$$

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

We then average this variance over the missing data implicates:

average of $b^{(\ell)}$ over missing data implicates:

$$b_M = \sum_{\ell=1}^m \sum_{k=1}^r \frac{(q^{(\ell,k)} - \bar{q}^{(\ell)}) (q^{(\ell,k)} - \bar{q}^{(\ell)})'}{m(r - 1)} = \sum_{\ell=1}^m \frac{b^{(\ell)}}{m}.$$

The next source of variation comes from the multiple implicates due to missing data

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

variance of the statistic due to variation in missing data implicates:

$$B_M = \sum_{\ell=1}^m \frac{(\bar{q}^{(\ell)} - \bar{q}_M) (\bar{q}^{(\ell)} - \bar{q}_M)'}{m - 1}.$$

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

variance of the statistic on each implicate file:

$$u^{(\ell,k)} = u(D^{(\ell,k)}),$$

average variance of the statistic across synthetic implicates:

$$\bar{u}^{(\ell)} = \sum_{k=1}^r \frac{u^{(\ell,k)}}{r}$$

and

average variance of the statistic across synthetic and missing data implicates:

$$\bar{u}_M = \sum_{\ell=1}^m \sum_{k=1}^r \frac{u^{(\ell,k)}}{mr} = \sum_{\ell=1}^m \frac{\bar{u}^{(\ell)}}{m}$$

The total variance is, once again, a weighted sum of the difference sources of variation—

between synthetic implicate, between missing data implicate, and within implicate:

total variance of the average statistic across implicates: .

$$T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{b_M}{r} + \bar{u}_M$$

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

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

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

3.6.3 Application to the SIPP/SSA/IRS-PUF

Version 4.0 of the public use file consists of 16 implicates. We created four implicates in the missing data completion phase and then created four synthetic implicates per missing data implicate, thus $m = 4$ and $r = 4$. We chose to focus on two types of statistics—regression coefficients and univariate statistics (means, variances and percentiles) because these are most likely to be of interest to the potential users of our public use file. When showing regression results, we report \bar{q}_m and \bar{q}_M as vectors of regression coefficients. To calculate \bar{q}_m we run the same regression on each of the four missing data implicates and then average the coefficients across implicates. To calculate \bar{q}_M we run the same regression on each of the 16 synthetic implicates and then average the coefficients across these implicates. We also report the vari-

ance associated with each average coefficient in the form of vectors that contain the diagonal elements of the covariance matrices T_m and T_M . In the same format we report the standard error (square root of diagonal elements of T_m and T_M), t -ratio (each coefficient divided by the standard error), degrees of freedom (calculated using formulae above), and upper and lower bounds of the 95 percent confidence interval. To show the effect of the two types of implicates on the total variance calculation, we also report the component pieces of the overall variance: diagonal elements of B_M, b_M, \bar{u}_M for the synthetic data, and b_m and \bar{u}_m for the missing data. Univariate statistics are reported in the same manner except the results are scalars instead of vectors.

3.6.4 Results

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

intervals, the synthetic variable will need to be carefully examined to determine what might have caused the discrepancy.

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

It is important to note one more detail about the univariate and regressions results we present here. We have used the weight that we created by matching individuals in

our sample to the Census 2000 micro-data. Hence, in both the completed data and the synthetic data, all the statistics we report are weighted and should be interpreted as representative of individuals from the civilian non-institutional U.S. population age 18 or older as of April 1, 2000.

Summary statistics for OASDI beneficiaries Tables 3.3-3.12 give results comparing means of important earnings variables by demographic group and type of benefit for individuals who became OASDI beneficiaries during the time period covered by these data (*i.e.*, had date of initial entitlement between 1951 and 2002). Tables 3.3-3.10 show results for SER work indicators (positive FICA covered earnings in a year) and SER earnings (total FICA covered earnings up to the maximum). As in version 3.1, the percentage of individuals who worked in a given year is very close, on average, for all the groups and across all the years and the confidence intervals overlap. In addition, average earnings are now much closer for all the groups. For example in 1995 average earnings for white males who retire at some point were \$10,347 in the synthetic data and \$11,012 in the completed data. For white females who retire the correspondence is even closer: \$5,495 versus \$5,566. Particularly strong improvement was made for black males. In 1995, black males who retire at some point earned \$8,856 on average in the synthetic data and \$8,564 in the completed data and there is almost complete overlap in the confidence interval. Synthetic earnings data for this group was particularly problematic in earlier versions so this result represents a significant step forward in our modeling. Figures 3.1 and 3.2 show the time trend for labor force participation for the four main demographic groups for individuals who retire at some point. Figures 3.3 and 3.4 show the same time trend for earnings for the same groups. Labor force participation and earnings trends are the closest for white women, followed by black women and black men. White men have a slightly higher discrepancy between synthetic and completed earnings in 1985. Still the trend

is the same and other years have closer correspondence.

As shown in Tables 3.11-3.12, Total SER earnings summed over all years 1951-2003 and total number of years with positive earnings are also very close for most groups. White females who retire at some point earned on average \$192,468 over this time period according to the synthetic data compared to \$198,303 in the completed data and they worked a total of 26.17 versus 26.69 years. None of the individuals who retire or receive disability benefits have total years off by even a full year when comparing the synthetic and completed data. Total earnings differ by between \$1,000 (black males) and \$25,000 (white males).

Summary statistics for all workers Figures 3.5-3.8 show comparisons of trends in employment and earnings between the synthetic and completed variables described above but for all workers instead of just OASDI beneficiaries. Specifically, Figures 3.5 and 3.6 show proportions of individuals who worked and average earnings for the years 1965, 1975, 1985, and 1995. These comparisons show very close correspondence between the synthetic and completed data. Average earnings for white males in 1995 is \$17,047 using synthetic data and \$17,241 using completed data. Of the white males in our sample, 67.1% had positive FICA covered earnings in 1995 according to the synthetic data versus 67.5% according to the completed data. These results are consistent across years and demographic groups. For whites, the synthetic and completed time trends lie on top of each other. For blacks, there are a few more differences, in particular earnings for black males seem to diverge a bit in 1995, but generally the time trends are close and show the same pattern.

Figures 3.7 and 3.8 compare total earnings and years worked from 1951-2003. The group with the closest correspondence on average between the synthetic and complete data is white females (total earnings of \$211,817 versus \$212,751 and total years 17.76 versus 17.99). The group with the largest difference is black males (\$257,525 versus

\$240,933 and 18.99 versus 18.41 years). None of the groups differ by more than half a year in the total number of years worked and both black females and white males differ by less than \$10,000 in total earnings.

Summary statistics by education categories We next consider means of several important variables stratified by race, gender, and education category. In our analyses of version 3.1, we found that the relationship between education and other variables had not always been well preserved in the synthetic data. In this version of synthetic data, we find some improvements in this respect. Tables 3.13-3.15 show proportions for foreign-born, Hispanic, and disabled individuals by race, gender, and education. The percentages of individuals who are foreign-born and Hispanic are very close for the demographic and education sub-groups. Again the three middle education categories show particularly close correspondence between the synthetic and completed data. For example 9.3% of white males with some college are foreign-born according to the synthetic data compared to 9.4% in the completed data. The synthetic and completed data both give 9.0% of individuals as being Hispanic for the same group. In both cases there is complete overlap in the confidence intervals. In past versions, Hispanic was a particularly difficult variable to synthesize but these results seem to indicate that we have made significant progress modeling this variable. Percentages of individuals who report being disabled in the SIPP are also relatively consistent between the synthetic and completed data. White males are the closest across all education categories ($\% \text{disabled synthetic} - \% \text{disabled completed} < 1\%$ for all groups except graduate degrees) and black males are the most different (but still $\% \text{disabled synthetic} - \% \text{disabled completed} < 2\%$ for all groups except graduate degrees), but in all cases there is significant overlap in the confidence intervals.

Summary statistics for marital histories Table 3.16 shows means and confidence intervals for six marital history variables: number of marriages, percent ever

divorced, percent ever widowed, duration of 1st marriage, duration of 2nd marriage, and age at first marriage. The first three variables are nearly indistinguishable on average between the synthetic and completed data, clearly the result of a successful Bayesian bootstrap of *mh_category*. The durations are shorter in the synthetic data than in the completed data for both the point estimates and the confidence intervals by 2-3 years. Age at first marriage is approximately 23 years in both data types. The consistent synthesis of these marital history variables is another major step forward given that past versions of the synthetic data contained synthetic values that did not even meet minimum consistency standards with the unsynthesized and other synthesized variables.

Age at time of retirement Of particular interest when considering the synthesis of birth date and year of initial entitlement is whether these two variables are consistent enough with each other to produce an expected distribution of retirement ages. Figures 3.9 and 3.10 chart both weighted and unweighted counts of individuals who retired (*i.e.*, had *tob_initial* = 1) at different ages. The first important thing to note is that the completed data have some discrepancies between recorded retirement age and legal retirement age. There are almost 5,000 individuals in our sample whose original administrative birth date and year of initial entitlement imply that they retired between age 61 and age 62. It also appears that in the completed data there are large numbers of individuals retiring at age 62 and at age 64. We had expected the spike at age 62 but thought the later spike would be at age 65. In our synthetic data, we attempt to impose the restriction that retirees must be at least 62 and are successful in all but a few cases. Hence the group retiring between ages 61 and 62 vanishes in our synthetic data. The synthetic data also have a high point at age 62 but then taper off more uniformly across ages 63, 64, and 65. Ideally the counts of individuals retiring at age 63 in the synthetic data might have dropped off more quickly.

However the modeling is difficult here because the completed data are not entirely as expected and we are forcing some data consistency that does not exist in the original data. Given our careful modeling of date of initial entitlement and its close correspondence on average between the synthetic and completed data, more research is needed to determine the exact cause of the differences in these distributions.

Selected regression results We begin our discussion of regression results with Tables 3.17-3.20 where the dependent variable is the log of total DER earnings (sum of deferred and non-deferred at FICA and non-FICA jobs) in the year 2000. We ran four separate regressions for each of the major demographic groups: white males, black males, white females, and black females. The closest correspondence between the synthetic and completed regression coefficients is in the education variables which always have the same sign and generally have significant overlap in the confidence intervals. The exceptions for overlapping confidence intervals are usually the graduate degree indicator, not surprising given the results in the means presented earlier. The demographic group with the closest synthetic and completed education coefficients is white males. The coefficient on high school degree only in the synthetic data regression is .214 compared to .230 using the completed data, and for some college, the coefficients are .400 and .431 respectively. In both these cases the confidence intervals in the synthetic data contain the confidence intervals in the completed data. In comparison the high school degree only coefficients for black females are .263 and .347 for synthetic and completed data respectively and for some college the coefficients are .494 and .587. The confidence intervals overlap to a great extent but not completely.

The other SIPP demographic variables, Hispanic, disabled, and foreign-born, are not as consistently similar between the synthetic and completed data but they have improved significantly compared to prior versions of the synthetic data. Foreign-born and disabled always have the same sign and Hispanic has the same sign in

the regressions for white males and black females. For white males and females the confidence intervals for foreign-born and disabled overlap, and for black males and females the confidence intervals for all three variables overlap. The magnitudes of the coefficients differ but the confidence intervals give reason to be hopeful that the synthetic data are not producing estimates that are entirely different from the completed data.

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

Table 3.21 shows results for a regression of the log of the *AIME/AMW* variable on various demographic characteristics. The results for this summary measure of earnings generally show point estimates that are quite close between the completed and synthetic data. The race/gender interaction terms have overlapping confidence intervals except for black females and even in this case the point estimates and the intervals are not very different (-.928 versus -.995). The education coefficients all have overlapping confidence intervals with the exception of graduate degree. The Hispanic and marital status indicators are all very close both in terms of confidence intervals and point estimates. Only disabled shows significant bias. The age coefficients are slightly different between the synthetic and completed data but the confidence intervals do overlap.

Univariate distributions of continuous variables Table 3.22 examines univariate distributions for most of the continuous variables in our sample (we only show a handful of years from the various arrays to make the table a reasonable size). The continuous variable synthesis techniques used in this project generally did a very good job of modeling the overall univariate distributions of a variety of variables. The percentiles of the synthesized variables match closely with the percentiles of the corresponding completed variables, capturing the general shape of the distribution; although, very sudden spikes and cliffs in the distributions do get smoothed out a bit. Some of the variables had their synthetic draws restricted to rather narrow windows making the close match not too surprising, but even the variables whose synthesis was unrestricted resulted in very similar univariate distributions.

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

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

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

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

The DER earnings arrays present some of the same challenges as the wealth variables only to a lesser degree. They also have some very large outliers and are heavily skewed. As a result, the synthetic values display some of the same problems as the synthetic wealth variables, but again, to a lesser degree. The lower ends of the distri-

butions tend to be slightly underestimated and the upper ends slightly overestimated. The deferred earnings arrays (not shown in the table) have extremely small sample sizes and struggle a lot more than the non-deferred earnings arrays, but once again, the means and general shape of the distributions are preserved very well for all the years.

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

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

Counts and percentages of categorical variables Finally, Table 3.23 shows weighted and unweighted counts and percentages of some of the basic demographic and benefit variables in the synthetic and completed data. Included variables are: *male*, *black*, *Hispanic*, marital status, *tob_initial*, *tob_2000*, home ownership, foreign-born, education category, age category in 1990, age category at time of initial entitlement, and age category at time of retirement. We include these as a help to those seeking to do basic comparisons between the synthetic and completed data.

3.7 Assessing Disclosure Risk

3.7.1 Overview

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

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

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

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

3.7.2 Matching based on probabilistic record linking

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

Probabilistic matching requires caching a set of blocking and matching variables that are common to both files. We implemented one blocking strategy using the unsynthesized variables for blocking. For married individuals we use the unsynthesized variable *male* for each member of the couples. For unmarried individuals we use the two unsynthesized variables, *male* and *maritalstat*. The latter can be either

widowed, divorced/separated, or never married ($maritalstat = \{2, 3, 4\}$). In other words, for two records to be a match, they must necessarily have identical values for marital status and gender since these two variables were not synthesized. After this has been determined to be the case, other variables can be compared to determine the probability that two records represent the same person.

The probabilistic record linking was performed using the Census Bureau's internal record linking software, which is maintained by the Statistical Research Division. The discussion in this section describes the technical settings used for that software. We set the blank filter flag equal to 0 so that if the variable is missing, the record will automatically be considered to agree on that field. Matching for the two groups, married and unmarried, was done separately. Blocking variables help to reduce the number of records used for comparison; however, in any given run all records in the same blocking group of the synthetic implicate and the Gold Standard files are compared. Thus, record linking computation is quadratic with run times dominated by the size of the largest block. In this latest version of the SIPP/SSA/IRS-PUF, the block sizes are very large. For this reason, the matching is done within corresponding segments of the Gold Standard and PUF files. Internally we know when segments of the Gold Standard and PUF files (single implicate) correspond to the same individuals, because we make use of the common artificial person identifier (*personid*) that is on both files. Without the information contained in *personid* (which is not on the actual PUF), an intruder would have to compare many more record pairs to find true matches and would not find any more true matches (the true match is guaranteed to be in the blocks being compared) and would almost certainly find more false matches. For this reason our approach leads to a conservative measure of the disclosure risk.

When the SIPP/SSA/IRS-PUF is finally publicly released there will be no link between the Gold Standard data and the synthetic implicate files. However for testing purposes, we have maintained this link by keeping the common person identifier on

the Gold Standard file and the PUF implicate files. Thus, by naming this person identifier in the sequence field of the record linking software, we can check which matched record pairs with a given score are correct matches and which are false matches by comparing this person identifier. When the person identifier is the same, the matching algorithm was successful in finding the person in the Gold Standard file to whom the synthetic data record belonged. When the person identifier is different, the matching algorithm was unsuccessful. This technology is also used for the distance matching discussed in section 3.7.3.

Automatic searches for matches occur only within those records sharing the same values on the blocking variables. Matches agree exactly on values for the blocking variables and, additionally, they agree on values for the matching variables. An input file to the matching software specifies the agreement criterion for each of the matching variables. Two numbers have to be specified for each of the matching variables. The first number represents the conditional probability that the two records agree on the matching field value given that the two records represent a match, called the m probability. The second number represents the conditional probability that the two records agree on the matching field value given that the two records do not represent a match, called the u probability. This technology was also used in creating the weight; see 3.5.8.

From the agreement criterion, the software computes a score. The agreement score for a match on a particular variable from two comparison records is based upon $\ln(m/u)$. A larger ratio implies a stronger distinguishing power for that matching field. Presumably, the ratio $m/u > 1$. When using Census Bureau matching software for the un-duplication of a file, one is trying to identify specific duplicate pairs, so more precise probability estimates may be helpful. However, when using this software for extracting subsets of plausible matches from a large file, the conditional agreement probabilities can be rough general estimates. To lean towards a more conservative

assessment of disclosure risk, we obtained the best possible m and u estimates by using the *personid* variable that is common between the files. Given that the public will not have access to this variable, an intruder trying to match the two files cannot possibly obtain better results using matching software that is at least as efficient as the Census Bureau software.

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

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

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

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

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

The agreement and disagreement conditional probabilities for those variables used for matching individuals with spouses are shown in Table 3.24. All matching fields were assigned the exact matching comparison type. This caused the program to assign full agreement/disagreement scores according to whether the fields agree or disagree. The corresponding agreement probabilities for single individuals are just

slightly different and are shown in Table 3.25.

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

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

variable *flag_mar4t*, which is used to identify individuals who reported more than three marriages. When two records agree on this variable, and they are a match, the cumulative matching score increases by 5.317686217. If the records are not a match, but agree on this variable, then the cumulative score decreases by -4.609063992 .

The output cutoff flag for the cumulative matching score provides the comparison points for the matching score. In our testing we declare any pair of records with a cumulative score between -20 and 20 to be a potential match. That is, we consider matching two records whenever their agreement score exceeds -20 even though most applications of probabilistic record linking use a positive cut-off for the automatic selection of potential matches. Thus, we declare records to be candidate matches based on an aggressive matching strategy. From either Table 3.24 or 3.25 we can see that the total matching scores cannot be outside of this range. Essentially, we allow every record in the synthetic file to have candidate matches in the Gold Standard. The output files are sorted by decreasing cumulative agreement score; then, the best two matches are kept. Finally, the proportions of true matches and the ratio of true to false matches are obtained.

The number and proportions of false and true matches, for each of the segments of the file, are given in Tables 3.26 and 3.27. The number of true and false matches in each segment are reported in column 3 and sum to equal the total number of records in each segment. The ratio of true to total matches and false to total matches gives the percentage of true matches and false matches in each segment and is reported in column 4. In Table 3.26, there are no data segments that have a true match rate over 1% and the ratio of true to false matches is extremely low. In Table 3.27, the percentages of true matches are slightly higher but the highest value is still just over 1% (1.18% for segment 2).

3.7.3 Distance matching

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

Our tests consider the case of an intruder who uses distance-based re-identification to match the source records from the Gold Standard to synthetic SIPP/SSA/IRS-PUF observations. Such re-identification methods calculate the distance between a given record in the Gold Standard and every record in the synthetic implicate. The j closest records are then declared potential candidates for a match to the source record. In our analysis we consider $j = 3$.

Our distance-based re-identification proceeds in two stages. First we split both the Gold Standard and the first synthetic implicate ($m = 1$ and $r = 1$) into groups based on the unsynthesized variables. In this case, marital status and *male* are the only two unsynthesized variables. We next split each blocking group into smaller segments of approximately 10,000 observations in order to decrease the processing time, which is quadratic in the size of the largest files compared. We performed the segment split on both the Gold Standard and synthetic files so that the correct match in the Gold Standard was always in the same block and segment of the synthetic data used for comparison. In other words, we forced the segmentation of the files to guarantee that the correct match could always be found in the block/segments being

compared. This is the same assumption as we used in section 3.7.2 to segment the comparison files in that analysis. The segmentation of the blocks uses our prior knowledge of which records are actual matches and hence our matching results are conservative—overestimates as compared to a distance record link that could not segment the comparison files because the intruder did not have access to the true *personid*. After splitting the data into blocking groups and segments, we then calculate the distance between a given Gold Standard record and every record in the synthetic file in its corresponding blocking group and segment using a set of 163 matching variables. The three closest records are then declared possible matches.

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

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

We consider four specific cases of the general distance. In the first case we assume that the intruder can properly calculate the $Cov(A, B)$. We denote this distance $MAHA1$, and note that it is a true Mahalanobis distance; hence we expect that this distance measure will give us the highest match rates since it uses all of the available information, including the correct covariance structure of the errors in synthesizing all 163 variables. In the second case, we assume that the $Cov(A, B) = 0$. This is equivalent to assuming that we do not know how to link the observations across the A and B files and cannot compute $Cov(A, B)$. A real intruder would not have

access to $Cov(A, B)$. We denote the second distance *MAHA2*, and note that it is a “feasible” Mahalanobis distance. In the third case, we assume $[Var(A) + Var(B) - 2Cov(A, B)] = I$, where I is the identity matrix. We denote the third measure as *EUCL1*, which is a Euclidian distance with unstandardized inputs. For the fourth measure, we transform all of the matching variables in the A and B files to $N(0, 1)$ variables. Call the transformed files \tilde{A} and \tilde{B} . We then calculate the distance using $[Var(\tilde{A}) + Var(\tilde{B}) - 2Cov(\tilde{A}, \tilde{B})] = I$. We denote this fourth metric *EUCL2*, and note that it is a standardized Euclidian distance.

Tables 3.28-3.29 shows the results of the re-identification exercises for each of the four metrics. Table 3.28 shows the results using the Mahalanobis distance measures and Table 3.29 shows the results for the Euclidian distance measures. For each metric there are six columns. Match rate 1 (closest two records in A and B), match rate 2 (second closest two records in A and B), ratio of 2/1, match rate 3 (third closest two records in A and B), ratio of 3/2, and ratio (3+2)/1. Match rate j is calculated as the number of successful matches within a blocking group based on the j th closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages). For example, match rate 2 is calculated as the number of successful matches within a blocking group and segment based on the second closest observation divided by the total number of observations in that group (multiplied by 100 to convert to percentages).

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

The three ratio columns give us a sense of how much better the closest match does than the second and third best matches. Ideally, we want to ensure that if an intruder looked at the top three matches, he or she would face sufficient uncertainty about which one was the correct match. If the second closest record is exactly as likely to be the correct match as the closest record, then the ratio of match rate 2 to match rate 1 would be unity. If this ratio is less than one, then the closest record is more likely to be the correct match. If this ratio is greater than one, then the second closest record is more likely to be the correct match. The other ratio columns have the same interpretation. For the *MAHA1* metric, the column Ratio (3+2)/1 ranges from 0.79 to 1.12. This suggests that the 2nd or 3rd closest matches are almost as likely to be correct as the closest match. The totals in the last row are essentially weighted averages of each column where the weights are the percentage of records in each group.

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

3.8 Using Synthetic Data

Many potential users may be concerned about how to begin using synthetic data and multiple implicate files. In this section we give some suggestions and advice for using these data sets to perform analyses and apply the formulae described in 3.6.

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

Each synthetic implicate has two variables that control the relationship between implicate files. The variable *m_implicate* tells which completed data implicate served as the starting basis for this particular synthetic data implicate. The variable *r_implicate* gives the synthetic implicate number for the file. There are four completed data implicates, so the variable *m_implicate* ranges from 1 to 4. There are four synthetic implicates per completed data implicate so the variable *r_implicate*

ranges from 1 to 4 also. The first synthetic implicate will have $m_implicate = 1$ and $r_implicate = 1$, the second synthetic implicate will have $m_implicate = 1$ and $r_implicate = 2$, and so on, until the fifth synthetic implicate, which will have $m_implicate = 2$ and $r_implicate = 1$. In this manner a user can tell which synthetic implicates stemmed from the same completed data implicate. This information is necessary in order to apply the combining formulae.

Any statistic of interest to a researcher can be calculated from the synthetic data by calculating it once per synthetic implicate and then averaging across the 16 implicates. If the researcher wants to know average earnings in a given year, he or she should calculate the average in each of the 16 implicates using standard methods and then calculate the simple average these 16 separate means to get one grand mean. If the researcher wants to know the variance of earnings in a given year, he or she should follow the same procedure: calculate the variance in each implicate and then calculate the simple average these 16 statistics to get one grand variance. Note, and this is very important, the grand mean of the variances is just one component of the estimated total variance required to compute a confidence interval for average earnings. The complete formula is contained in section 3.6. Point estimates for any statistic of interest from regression results to moments or percentiles of a distribution can be obtained in this manner. In the standard combining formulae, every implicate is equally weighted, so simple averaging is all that is required.

The calculation of the estimated total variance of a statistic of interest, from which one might compute a confidence interval or test statistic, is more complicated but still can be performed with standard software. In addition to the statistic of interest, the user should save the estimated sampling variance of this statistic for each of the 16 implicates. For example, if calculating one mean per implicate, the user should calculate the sampling variance of the mean once per implicate.¹⁹ The

¹⁹The reader is cautioned to be certain to perform all calculations on variances and not standard deviations. To compute a standard deviation or standard error, the square root operation should be

within-implicate sampling variances are then averaged to estimate the average within-implicate variance, one component of the total variance. The user must then make use of the *m_implicate* and *r_implicate* variables to calculate the between-completed-data-implicate variance and the between-synthetic-data-implicate variance according to the formula in 3.6.2. The user first calculates the variance of the statistic across the four *r* implicates associated with a particular *m* implicate. There will be four of these variances: one per completed data implicate. These four variances are then averaged to give the overall between-synthetic-data-implicate variance. The user then calculates the mean of the statistic of interest for all the synthetic implicates associated with a particular completed data implicate. Again, there will be four of these means. The between *m* implicate variance is then calculated as the average of the squared deviations of these four means from the overall grand mean. If the statistics of interest are saved in a data set, these calculations can be easily performed. The variance pieces are then combined to create the total variance and calculate degrees of freedom. In the case that the total variance becomes negative, we recommend not subtracting the between-synthetic-data-implicate variance when calculating the total variance. The confidence interval can be calculated using the asymptotic assumption of normality instead of the finite sample *t*-distribution.

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

performed on the total variance that has been computed by combining all of the component variances appropriately.

3.9 Conclusion

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

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

product, as funding resources permit.

Table 3.1: Small Cell Consequences of Selected Combinations of Non-synthesized Variables

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

Table 3.2: Analytic Validity of SIPP-PUF Weights -- Weighted Counts of Benefit Recipients

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

Table 3.3: SER work indicator for year 1965

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

Table 3.4: SER earnings in year 1965

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

Table 3.5: SER work indicator for year 1975

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

Table 3.6: SER Earnings for year 1975

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

Table 3.7: SER work indicator for year 1985

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

Table 3.8: SER Earnings for year 1985

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

Table 3.9: SER work indicator for year 1995

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

Table 3.10: SER Earnings for year 1995

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

Table 3.11: Total SER earnings 1951-2003

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

Table 3.12: Total years worked in SER (i.e. positive FICA earnings)

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

Table 3.13: Foreign Born Indicator

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

Table 3.14: Hispanic Indicator

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

Table 3.15: SIPP Disability Indicator

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

Table 3.16: Marital History Variables

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

Table 3.17: Log of Total DER Earnings in year 2000 for white males

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

Table 3.18: Log of Total DER Earnings in year 2000 for black males

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

Table 3.19: Log of Total DER Earnings in year 2000 for white females

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

Table 3.20: Log of Total DER Earnings in year 2000 for black females

Explanatory Variables	Coefficient		Confidence Interval		Standard Error		Synthetic DF Not Exist
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	
Intercept	8.236	7.194	8.101	7.059	0.080	0.082	1
highschool_only	0.263	0.347	0.197	0.282	0.032	0.039	0
somecollege	0.494	0.587	0.376	0.518	0.046	0.041	0
college_only	0.842	1.118	0.770	1.030	0.042	0.054	1
graduate	0.953	1.289	0.748	1.179	0.077	0.067	0
disab	-0.300	-0.453	-0.408	-0.543	0.052	0.054	0
foreign_born	0.114	0.183	-0.016	0.245	0.066	0.073	0
hispanic	-0.011	-0.007	-0.124	0.101	0.064	0.064	0
ser_totyrs_2000	0.086	0.297	-0.010	0.183	0.037	0.023	0
ser_totyrs_2000_2	-0.019	-0.175	-0.092	-0.207	0.031	0.019	0
ser_totyrs_2000_3	0.003	0.048	-0.019	0.024	0.010	0.006	0
ser_totyrs_2000_4	0.000	-0.005	-0.003	-0.006	0.001	0.001	0

Table 3.21: Log of Average Indexed Monthly Earnings (AIME) or Average Monthly Wage (AMW) for all individuals

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

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

Table 3.22: Percentiles of Selected Synthetic and Completed Variables

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
Date variables											
birthdate	completed	1/22/1955	1/12/1913	4/28/1922	9/6/1928	4/21/1943	6/13/1957	4/1/1969	2/1/1977	9/10/1979	4/20/1981
birthdate	synthesized	2/17/1955	4/24/1913	8/22/1922	3/23/1929	10/1/1943	7/2/1957	1/27/1969	8/25/1976	6/10/1979	3/7/1981
date_initial_entitle	completed	3/9/1988	12/9/1963	1/31/1970	10/9/1973	12/24/1980	10/24/1989	5/24/1996	6/1/2000	9/9/2001	9/1/2002
date_initial_entitle	synthesized	5/17/1988	3/3/1964	4/5/1970	11/21/1973	3/7/1981	12/21/1989	7/30/1996	6/20/2000	8/31/2001	9/29/2002
deathdate	completed	7/5/2001	4/12/2000	5/17/2000	7/16/2000	12/2/2000	7/3/2001	2/17/2002	6/24/2002	8/5/2002	9/14/2002
deathdate	synthesized	10/19/2000	2/4/1993	4/22/1996	8/6/1998	7/13/2000	3/18/2001	11/26/2001	6/2/2002	8/28/2002	12/7/2002
MBA Variables											
mba_2000	completed	643	28	91	156	353	611	919	1136	1260	1549
mba_2000	synthesized	642	37	94	155	350	609	921	1137	1259	1537
mba_initial_real	completed	609	34	105	165	337	549	873	1116	1236	1433
mba_initial_real	synthesized	612	45	110	170	339	551	880	1120	1237	1431
Marital History Variables											
age_mar1	completed	23.4	15.8	17.2	18.1	19.8	22.3	25.6	30	33.4	42.4
age_mar1	synthesized	23.1	15.7	17	17.9	19.6	22.1	25.2	29.2	32.4	40.9
duration_mar1	completed	14.5	0.3	1.2	2.2	4.7	9.6	20.1	36	44.7	55.5
duration_mar1	synthesized	13.4	0.3	1.3	2.2	4.4	8.8	18.2	33.1	42.3	53.6
duration_end1	completed	928	0	1	1	3	13	1961	1973	1977	1981
duration_end1	synthesized	963	0	1	2	4	242	1948	1975	2005	2070
duration_mar2	completed	1169	0	2	4	9	1955	1970	1975	1978	1981
duration_mar2	synthesized	1200	1	2	5	17	1941	1969	1976	1983	2065
duration_end2	completed	1074	0	1	2	4	1933	1960	1968	1972	1978
duration_end2	synthesized	1059	0	2	4	21	1860	1951	1988	2015	2062
duration_mar3	completed	1298	0	2	3	10	1953	1964	1970	1973	1979
duration_mar3	synthesized	1218	2	11	27	141	1803	1965	1987	2026	2113
duration_end3	completed	1814	1	6	1928	1942	1953	1961	1967	1969	1974
duration_end3	synthesized	1842	1	723	1929	1944	1954	1961	1966	1969	1973
duration_mar4	completed	1955	1925	1936	1941	1949	1956	1962	1967	1970	1972
duration_mar4	synthesized	1956	1924	1934	1942	1950	1957	1963	1968	1970	1973
Wealth Variables											
homeequity	completed	72314	-9000	4000	8000	22000	50000	100000	163625	215750	320000
homeequity	synthesized	74491	-26272	1327	5284	18539	48942	101062	178872	249326	380942
nonhouswealth	completed	74925	-7000	1000	2000	6000	17000	61000	181500	317250	765000
nonhouswealth	synthesized	72921	-75235	-513	1056	4695	15181	56042	177601	324071	877694
totnetworth	completed	119632	-33000	-6000	1000	9000	51500	141000	294500	449500	925000
totnetworth	synthesized	113145	-42844	-7631	-995	7525	49761	137418	287070	436222	879640

Table 3.22: Percentiles of Selected Synthetic and Completed Variables

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
nondefer_der_fica_1980	completed	11930	58	349	825	2872	8272	16014	25068	30336	53290
nondefer_der_fica_1980	synthesized	12041	171	729	1338	3369	8306	15972	24643	30185	56505
nondefer_der_fica_1985	completed	15815	70	479	1141	4044	11514	21920	34376	41800	74477
nondefer_der_fica_1985	synthesized	15998	137	658	1431	4190	11504	21660	33907	42649	78072
nondefer_der_fica_1990	completed	19588	90	596	1429	5122	14180	26505	41143	51300	96543
nondefer_der_fica_1990	synthesized	19555	156	814	1784	5139	13968	26231	40710	52098	98197
nondefer_der_fica_1995	completed	23562	90	600	1488	5771	16082	30322	48258	63701	131903
nondefer_der_fica_1995	synthesized	23918	276	1220	2420	6580	16215	30160	48015	64170	135003
nondefer_der_fica_2000	completed	32320	151	1249	2970	9484	21767	38395	61340	83931	182894
nondefer_der_fica_2000	synthesized	33828	390	1930	3957	10297	22575	39376	63049	87879	189390
				SER Earnings Arrays							
totteam_ser_1955	completed	2060	14	66	152	550	1878	3664	4200	4200	4200
totteam_ser_1955	synthesized	1933	25	105	195	509	1645	3358	4200	4200	4200
totteam_ser_1960	completed	2559	18	103	238	857	2473	4629	4800	4800	4800
totteam_ser_1960	synthesized	2438	35	142	261	728	2253	4346	4800	4800	4800
totteam_ser_1965	completed	2833	23	128	291	1048	3026	4800	4800	4800	4800
totteam_ser_1965	synthesized	2691	38	159	296	868	2739	4800	4800	4800	4800
totteam_ser_1970	completed	4213	28	177	401	1436	4106	7660	7800	7800	7800
totteam_ser_1970	synthesized	4014	53	222	422	1243	3765	7098	7800	7800	7800
totteam_ser_1975	completed	6458	43	243	551	1965	5683	10626	14100	14100	14100
totteam_ser_1975	synthesized	6194	75	315	601	1748	5333	10190	14100	14100	14100
				SIPP Arrays							
famwelamt1990	completed	2177	25	113	210	531	1194	2847	5722	7763	11454
famwelamt1990	synthesized	2312	42	172	298	624	1287	2979	5931	8020	12183
famwelamt1995	completed	1920	21	98	193	487	1082	2276	5090	7183	10815
famwelamt1995	synthesized	2102	28	119	214	478	1047	2406	5881	8166	13120
fpov1990	completed	129678	73957	79746	84031	99608	119033	157243	187777	208340	272232
fpov1990	synthesized	131862	74677	80581	86047	101730	121391	158244	188948	208597	267694
fpov1995	completed	149438	79258	88736	95207	111751	135717	183384	218590	241636	314551
fpov1995	synthesized	152346	80973	90013	96712	113828	139191	185755	221023	244702	315943
ftotinc1990	completed	36965	1044	5735	9236	17692	31328	49528	71261	88542	128397
ftotinc1990	synthesized	35271	658	5548	8978	17156	30137	47206	67433	83269	121216
ftotinc1995	completed	43020	1263	6615	10793	20432	37055	58670	83405	100230	142381
ftotinc1995	synthesized	40829	-997	4067	7914	17084	33432	56576	83688	103719	147841

Table 3.22: Percentiles of Selected Synthetic and Completed Variables

Variable Name	Type	Mean	P01	P05	P10	P25	Median	P75	P90	P95	P99
helamt1990	completed	2213	38	154	285	592	1148	2399	5048	7983	17238
helamt1990	synthesized	2051	32	144	256	538	1030	2094	4526	7480	17343
helamt1995	completed	3965	116	427	706	1346	2518	4694	8655	13187	21299
helamt1995	synthesized	3594	84	299	514	1066	2085	4035	7768	12943	23200
tofeam1990	completed	17643	107	697	1587	5355	13886	24881	37967	48411	77399
tofeam1990	synthesized	16465	85	562	1140	3881	12559	23673	36576	46880	75989
tofeam1995	completed	19857	285	1525	2943	7014	15132	27000	42197	54303	91142
tofeam1995	synthesized	19385	346	1568	2786	6364	14103	26438	42250	55008	92325
tohoursannual1990	completed	1671	44	181	346	972	1945	2197	2595	2915	3537
tohoursannual1990	synthesized	1521	19	103	220	729	1801	2141	2435	2691	3257
tohoursannual1995	completed	1620	111	320	506	972	1836	2135	2452	2726	3243
tohoursannual1995	synthesized	1550	147	366	545	962	1677	2089	2318	2542	3044
totinc1990	completed	16615	-829	152	1024	4628	12676	23770	36793	47160	74068
totinc1990	synthesized	16282	-764	99	827	3872	11736	23480	37345	48158	76832
totinc1995	completed	18995	-758	1081	2627	6400	14060	26131	41229	52884	87577
totinc1995	synthesized	18996	-636	1076	2364	5699	13036	25962	42851	56489	94769
wkspt1990	completed	15.8	0.3	1.1	2.2	5.2	12.3	22.5	36.7	44.5	52
wkspt1990	synthesized	15.7	0.2	1.2	2.2	5.1	12	22.4	36.7	44.2	51.1
wkspt1995	completed	16.5	0.6	2.2	3.8	8.2	13.9	23.5	32.4	40	49.4
wkspt1995	synthesized	18.1	0.7	2.3	3.9	8.5	15.7	25.1	36.6	42.2	49.5
wkswp1990	completed	41	3.2	10	17.3	35.3	47.5	51.2	51.9	52	52
wkswp1990	synthesized	40.7	2	7.5	13.8	34.7	48.5	51.2	51.9	52	52
wkswp1995	completed	41.2	6.8	15.1	20.2	35.4	46.5	50.5	51.6	51.8	52
wkswp1995	synthesized	41.3	4.5	12.4	19.4	35.4	47.5	50.7	51.6	51.8	52
Cardinal Categorical Variables											
time_arrive_usa	completed	5.56	1	1	2	4	6	8	8	8	8
time_arrive_usa	synthesized	5.44	1	1	1.19	4	6	8	8	8	8
tofam_kids	completed	0.9	0	0	0	0	0	2	3	3	5
tofam_kids	synthesized	0.9	0	0	0	0	0	2	3	3	5

Table 3.23: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)												
Variable	WEIGHTED						UNWEIGHTED					
	Count		Percentage		Count		Percentage		Count		Percentage	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed
Male												
0	106,949,968	106,470,823	52.41	52.18	138,357	138,357	52.45	52.45	138,357	138,357	52.45	52.45
1	97,094,759	97,573,904	47.59	47.82	125,436	125,436	47.55	47.55	125,436	125,436	47.55	47.55
Black												
0	181,725,962	180,358,529	89.06	88.39	232,401	233,326	88.10	88.45	232,401	233,326	88.10	88.45
1	22,318,765	23,686,198	10.94	11.61	31,392	30,467	11.90	11.55	31,392	30,467	11.90	11.55
Hispanic												
0	184,181,614	181,560,851	90.27	88.98	238,277	238,558	90.33	90.43	238,277	238,558	90.33	90.43
1	19,863,113	22,483,876	9.73	11.02	25,516	25,235	9.67	9.57	25,516	25,235	9.67	9.57
Maritalstat												
1	103,680,995	102,213,429	50.81	50.09	141,292	141,292	53.56	53.56	141,292	141,292	53.56	53.56
2	9,673,233	9,958,884	4.74	4.88	17,483	17,483	6.63	6.63	17,483	17,483	6.63	6.63
3	23,885,992	24,400,916	11.71	11.96	31,103	31,103	11.79	11.79	31,103	31,103	11.79	11.79
4	66,804,507	67,471,498	32.74	33.07	73,915	73,915	28.02	28.02	73,915	73,915	28.02	28.02
Tob_initial												
1	144,158,303	142,951,882	70.65	70.06	173,251	173,251	65.68	65.68	173,251	173,251	65.68	65.68
2	28,120,067	28,896,858	13.78	14.16	45,543	45,543	17.26	17.26	45,543	45,543	17.26	17.26
3	9,771,612	9,934,114	4.79	4.87	15,401	15,401	5.84	5.84	15,401	15,401	5.84	5.84
5	2,562,145	2,603,705	1.26	1.28	4,028	4,028	1.53	1.53	4,028	4,028	1.53	1.53
100	3,881,472	4,000,409	1.90	1.96	6,920	6,920	2.62	2.62	6,920	6,920	2.62	2.62
	15,551,128	15,657,759	7.62	7.67	18,652	18,652	7.07	7.07	18,652	18,652	7.07	7.07
Tob_2000												
1	150,570,926	149,502,998	73.79	73.27	196,028	196,028	74.31	74.31	196,028	196,028	74.31	74.31
2	26,411,823	27,162,166	12.94	13.31	34,306	34,306	13.00	13.00	34,306	34,306	13.00	13.00
3	5,873,327	5,926,142	2.88	2.90	7,571	7,571	2.87	2.87	7,571	7,571	2.87	2.87
5	1,860,289	1,880,083	0.91	0.92	2,433	2,433	0.92	0.92	2,433	2,433	0.92	0.92
100	4,311,070	4,460,705	2.11	2.19	5,781	5,781	2.19	2.19	5,781	5,781	2.19	2.19
	15,017,292	15,112,633	7.36	7.41	17,675	17,675	6.70	6.70	17,675	17,675	6.70	6.70

Table 3.23: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)											
Variable	WEIGHTED				UNWEIGHTED						
	Count Synthetic	Count Completed	Percentage Synthetic	Percentage Completed	Count Synthetic	Count Completed	Percentage Synthetic	Percentage Completed			
Own_home											
0	69,152,585	69,785,113	33.89	34.20	87,230	86,839	33.07	32.92			
1	134,892,142	134,259,614	66.11	65.80	176,563	176,955	66.93	67.08			
Foreign_born											
0	180,969,389	179,239,395	88.69	87.84	234,496	235,262	88.89	89.18			
1	23,075,338	24,805,332	11.31	12.16	29,297	28,531	11.11	10.82			
Educ_5cat											
1	43,271,926	42,898,854	21.21	21.02	56,851	55,903	21.55	21.19			
2	69,217,821	68,520,001	33.92	33.58	89,529	90,001	33.94	34.12			
3	51,100,854	52,842,654	25.04	25.90	64,585	66,115	24.48	25.06			
4	22,930,251	23,458,739	11.24	11.50	29,804	29,230	11.30	11.08			
5	17,523,875	16,324,479	8.59	8.00	23,024	22,544	8.73	8.55			
Age_cat12											
<=21	10,993,657	12,355,657	5.39	6.06	4,452	4,480	1.69	1.70			
22-24	11,083,762	11,637,091	5.43	5.70	8,296	8,484	3.14	3.22			
25-29	19,330,649	18,791,213	9.47	9.21	24,922	25,344	9.45	9.61			
30-34	20,827,765	19,839,766	10.21	9.72	28,311	27,657	10.73	10.48			
35-39	22,623,857	22,137,354	11.09	10.85	28,824	28,715	10.93	10.89			
40-44	22,702,993	22,384,522	11.13	10.97	28,493	28,579	10.80	10.83			
45-49	20,541,360	20,091,637	10.07	9.85	26,195	26,150	9.93	9.91			
50-54	18,114,622	17,779,895	8.88	8.71	23,761	23,797	9.01	9.02			
55-59	13,593,649	13,497,244	6.66	6.61	17,822	17,895	6.76	6.78			
60-64	10,666,518	10,893,885	5.23	5.34	14,636	14,469	5.55	5.48			
65-69	9,336,754	9,547,527	4.58	4.68	12,829	12,849	4.86	4.87			
>=70	24,229,138	25,088,937	11.87	12.30	45,255	45,375	17.16	17.20			
Agecat_initial_entitle (Tob_initial=1,2,3,5)											
<62	11,699,746	11,688,294	26.39	25.73	18,555	18,184	25.81	25.29			
>=62 and <63	7,273,818	18,508,709	16.41	40.74	11,673	29,067	16.24	40.43			

Table 3.23: Selected Variables – Weighted and unweighted counts and percentages (averaged across completed and synthetic implicates)												
Variable	WEIGHTED						UNWEIGHTED					
	Count		Percentage		Count		Percentage		Count		Percentage	
	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed	Synthetic	Completed
>=63 and <64	11,912,607	3,419,745	26.87	7.53	19,452	5,391	27.06	7.50				
>=64 and <65	7,337,083	2,757,925	16.55	6.07	11,964	4,961	16.64	6.90				
>=65 and <66	3,351,021	7,440,877	7.56	16.38	5,536	11,672	7.70	16.24				
>=66 and <67	1,514,780	482,726	3.42	1.06	2,587	798	3.60	1.11				
>=67	1,208,960	1,077,981	2.73	2.37	2,051	1,726	2.85	2.40				
>=80	37,283	58,829	0.08	0.13	73	90	0.10	0.13				
Agecat_retire (Tob_initial=1 only)												
<62	56,373	45,537	0.20	0.16	83	73	0.18	0.16				
>=62 and <63	5,843,165	15,661,703	20.78	54.20	9,187	24,252	20.17	53.25				
>=63 and <64	10,449,792	2,770,806	37.16	9.59	16,945	4,374	37.21	9.60				
>=64 and <65	6,457,351	2,351,612	22.96	8.14	10,486	4,237	23.02	9.30				
>=65 and <66	2,941,391	6,703,497	10.46	23.20	4,850	10,471	10.65	22.99				
>=66 and <67	1,323,219	406,445	4.71	1.41	2,259	663	4.96	1.46				
>=67	1,020,418	910,199	3.63	3.15	1,677	1,403	3.68	3.08				
>=80	28,357	47,059	0.10	0.16	56	71	0.12	0.16				

Table 3.24: Agreement Probabilities for Individuals with Spouses

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

Table 3.25: Agreement Probabilities for Single Individuals

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

Table 3.26: Match Rates for Married Individuals, Split into Data Blocks

Segment	Match Status	COUNT	PERCENT
1	FALSE	29939	99.30675335
	TRUE	209	0.69324665
2	FALSE	19660	99.5745543
	TRUE	84	0.425445705
3	FALSE	19517	99.62227554
	TRUE	74	0.377724465
4	FALSE	20202	99.71372162
	TRUE	58	0.286278381
5	FALSE	20017	99.71108344
	TRUE	58	0.288916563
6	FALSE	19811	99.6178408
	TRUE	76	0.3821592
7	FALSE	19658	99.65022558
	TRUE	69	0.349774421
8	FALSE	19564	99.70441341
	TRUE	58	0.295586587
9	FALSE	18305	99.62989169
	TRUE	68	0.370108311
10	FALSE	19724	99.72696936
	TRUE	54	0.27303064

Table 3.27: Match Rates for Single Individuals, Split into Data Blocks

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

Table 3.28: Mahalanobis Distance Matching Results

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

Table 3.29: Euclidean Distance Matching Results

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

Figure 3.1:
Comparison of Synthetic and Completed Annual Work Indicators
Retired White Males and Females



Figure 3.2:
Comparison of Synthetic and Completed Annual Work Indicators
Retired Black Males and Females

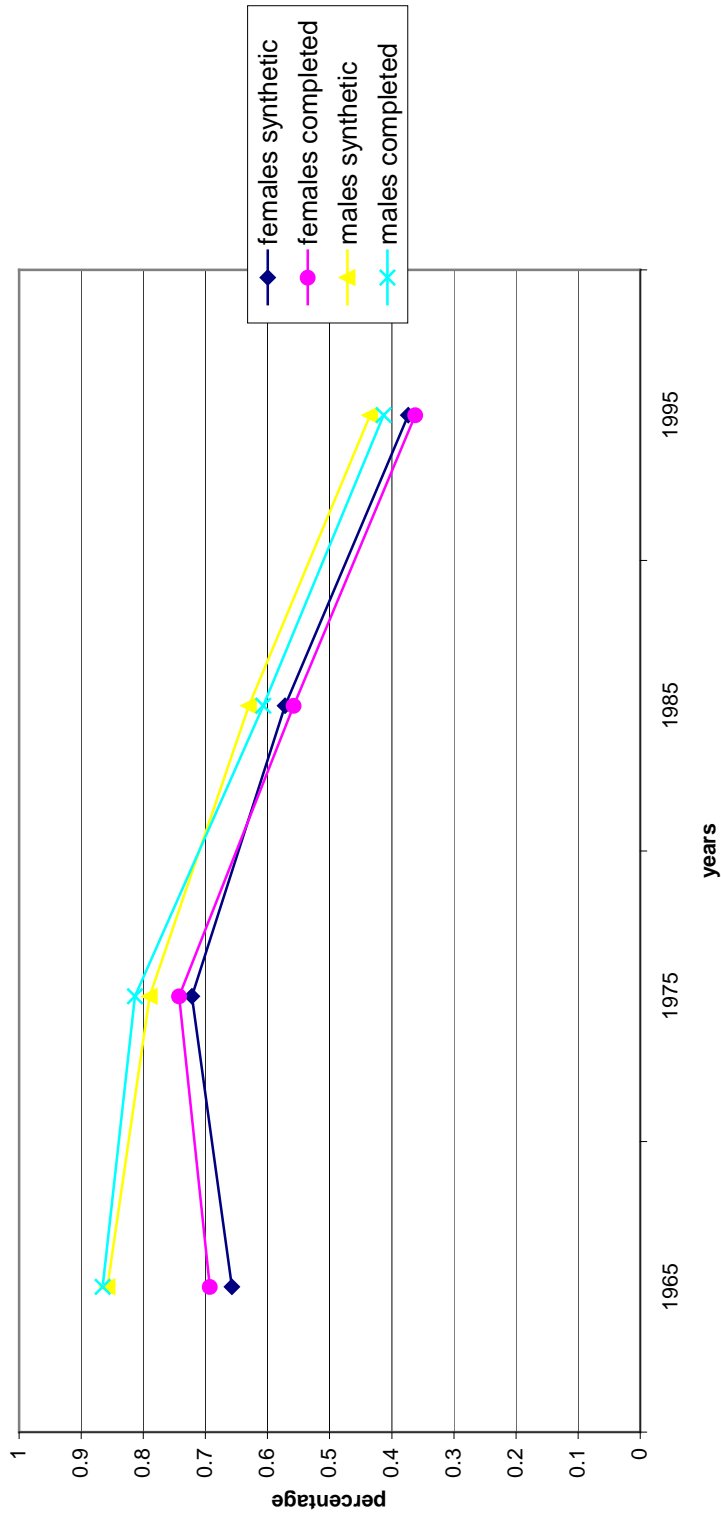


Figure 3.3:
Comparison of Synthetic and Completed Earnings
Retired White Males and Females

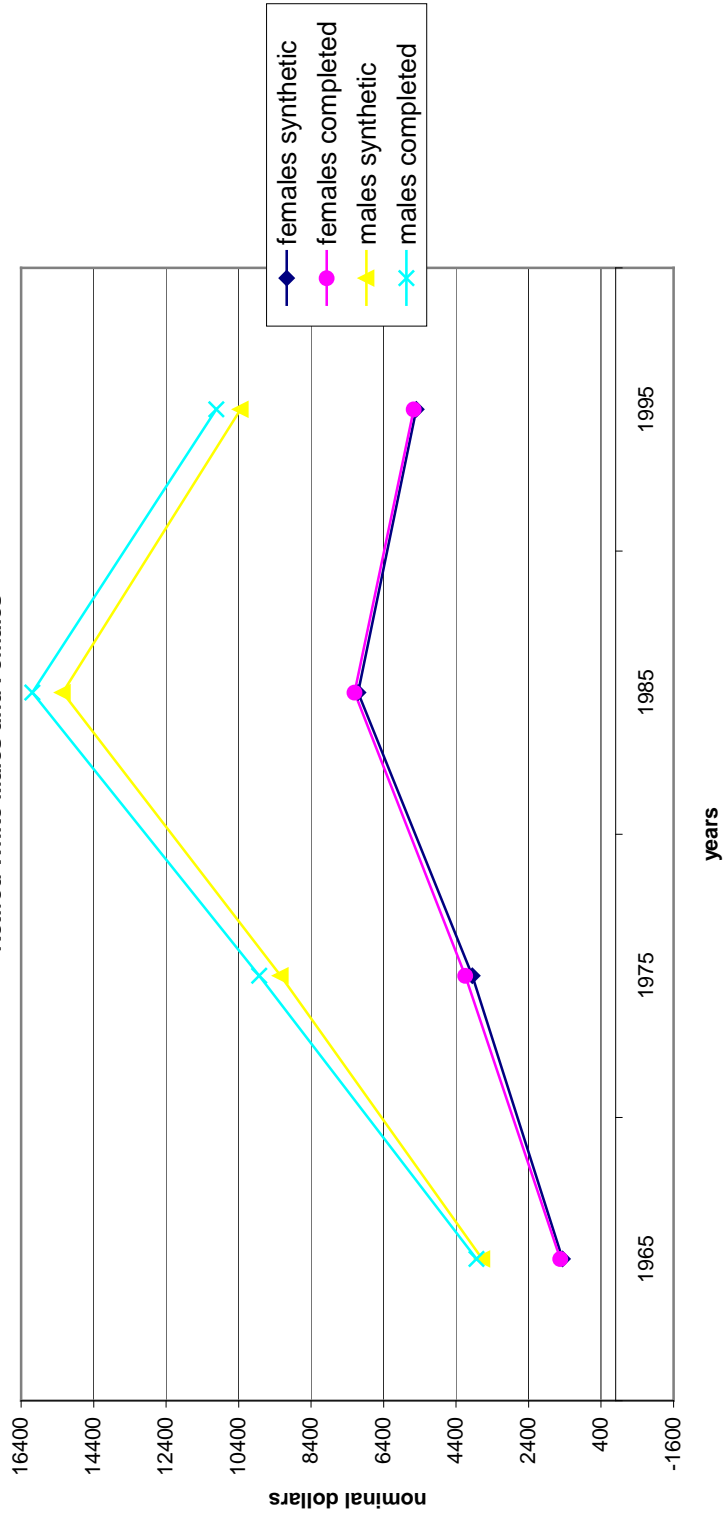


Figure 3.4:
 Comparison of Synthetic and Completed Earnings
 Retired Black Males and Females

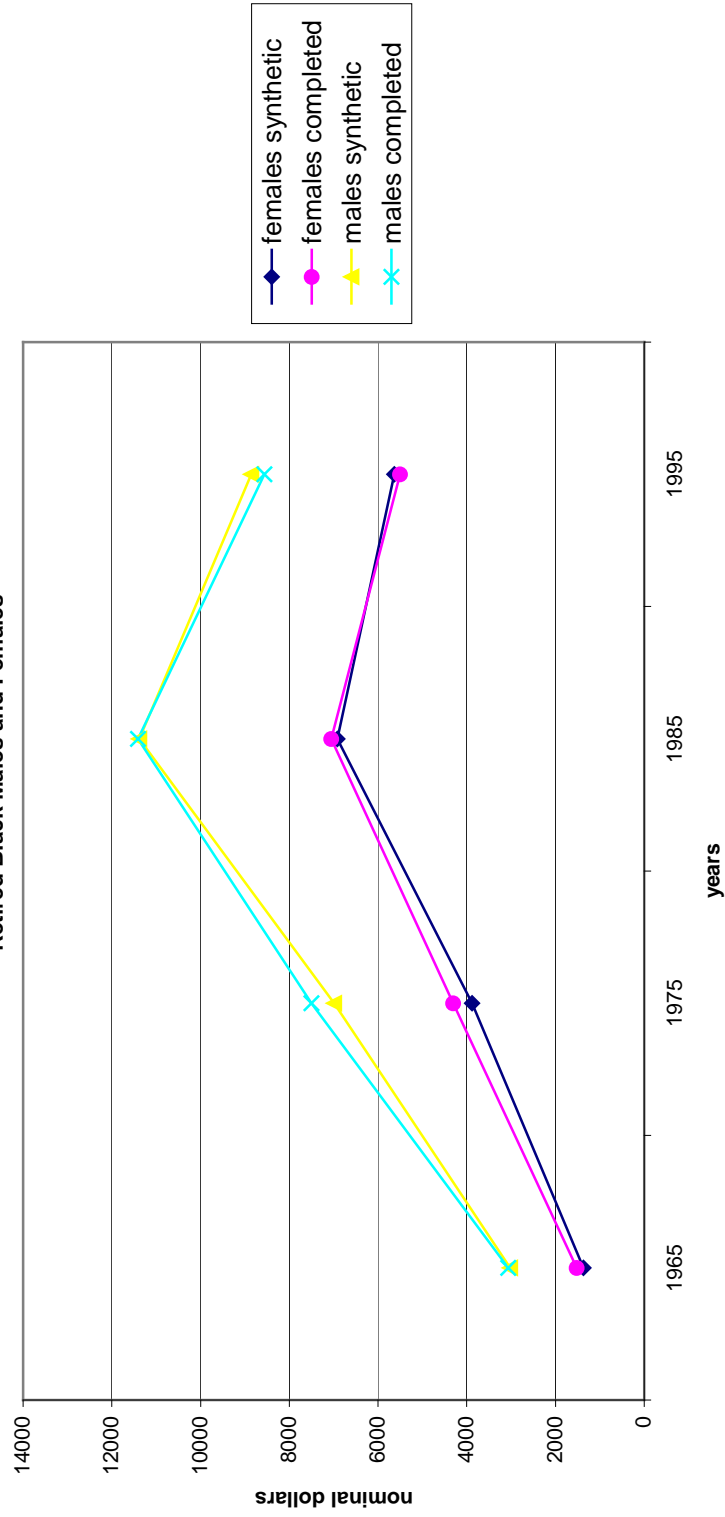


Figure 3.5:
Comparison of Synthetic and Completed Annual Work Indicators
White Males and Females

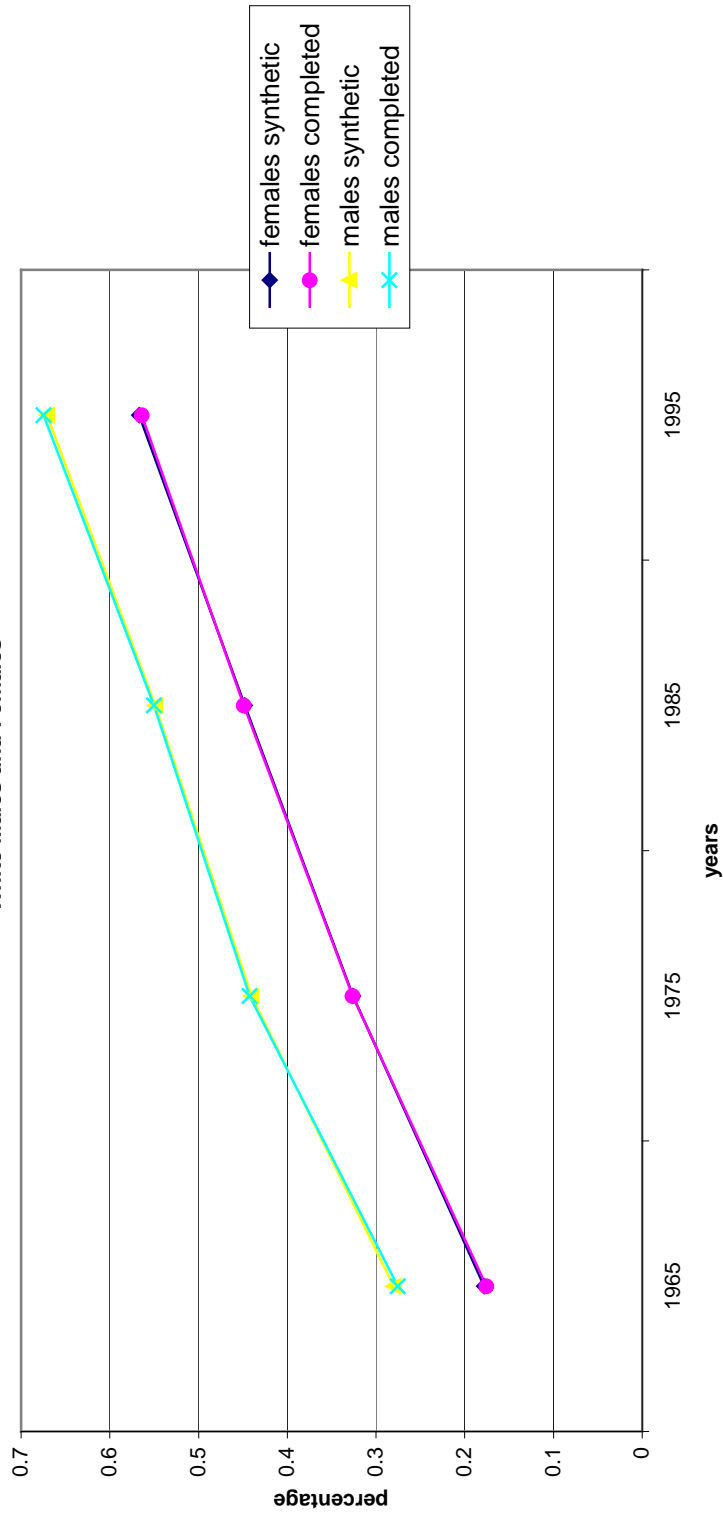


Figure 3.6:
Comparison of Synthetic and Completed Annual Work Indicators
Black Males and Females

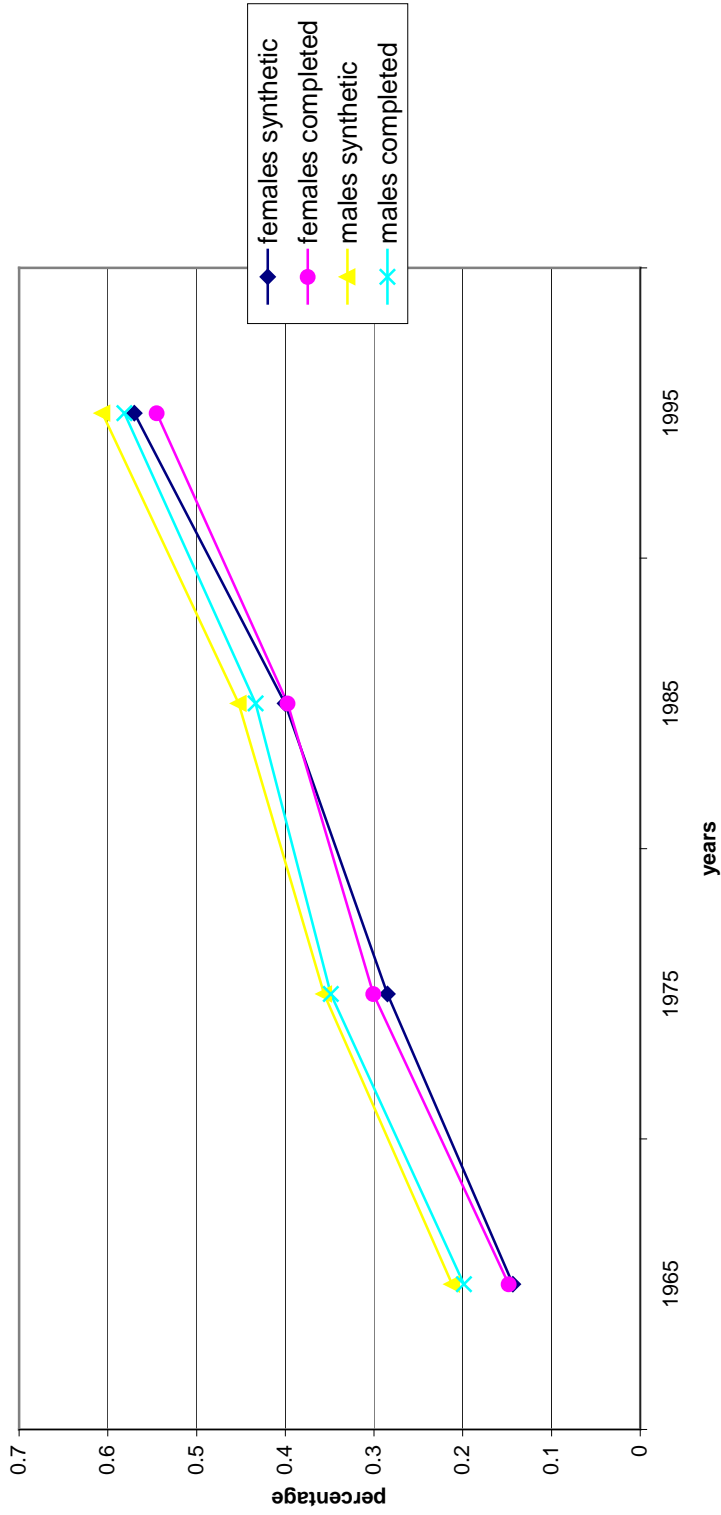


Figure 3.7:
Comparison of Synthetic and Completed Earnings
White Males and Females

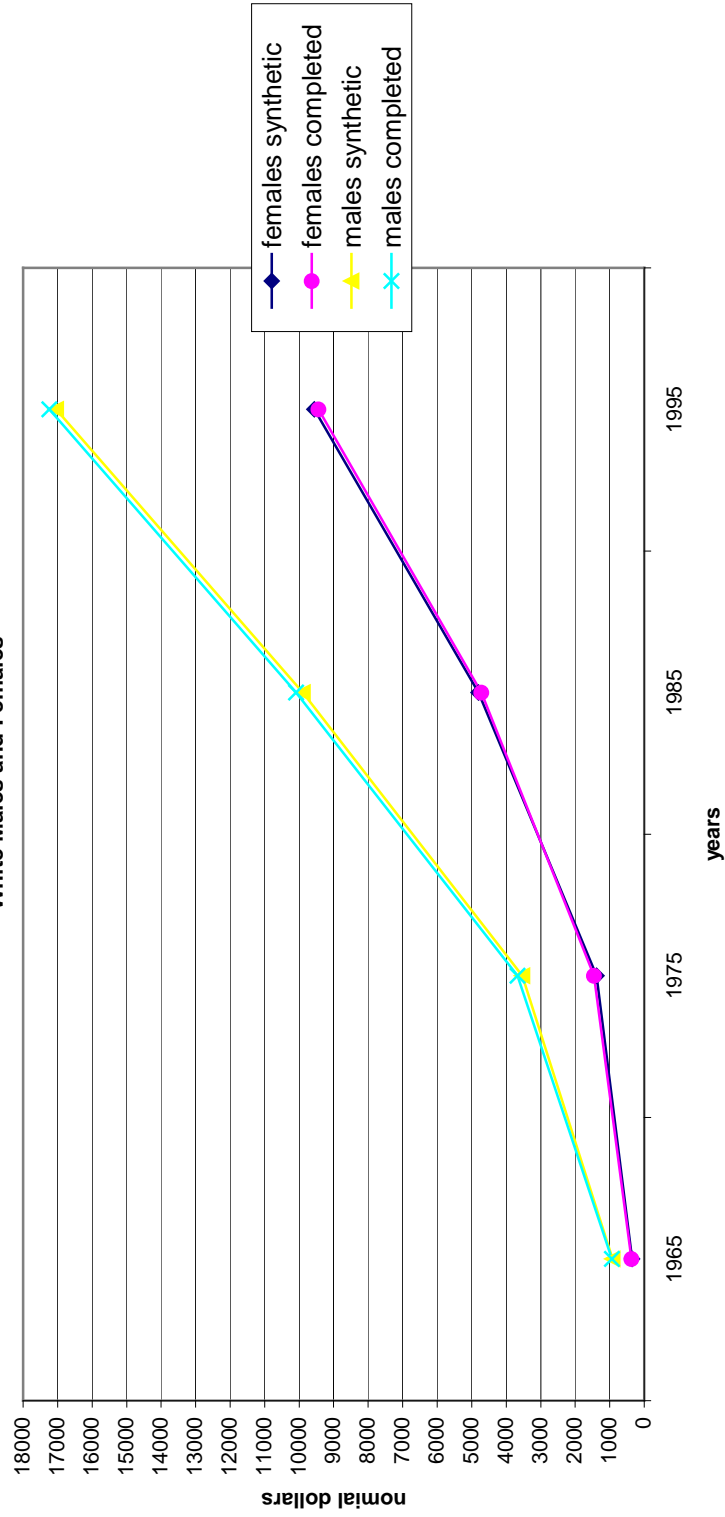


Figure 3.8:
Comparison of Synthetic and Completed Earnings
Black Males and Females



Figure 3.9: Age at Retirement, Weighted

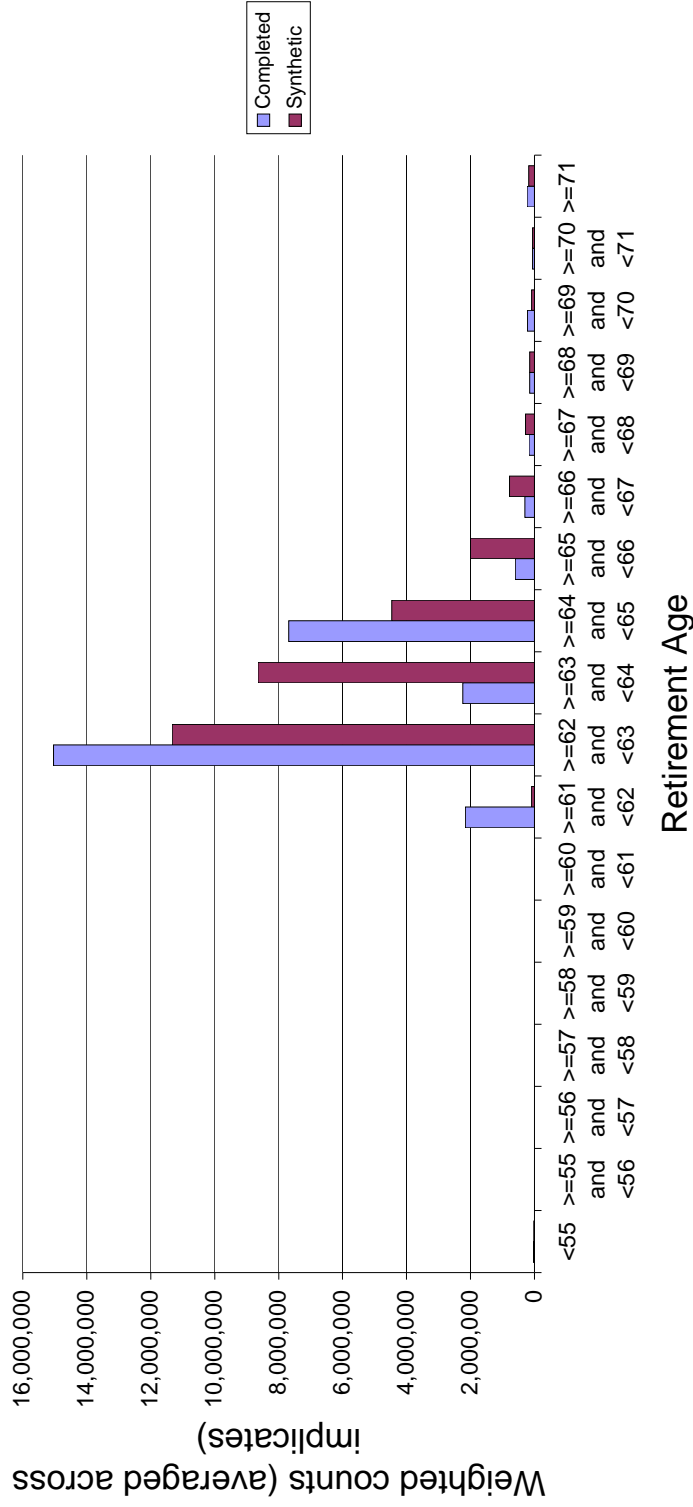
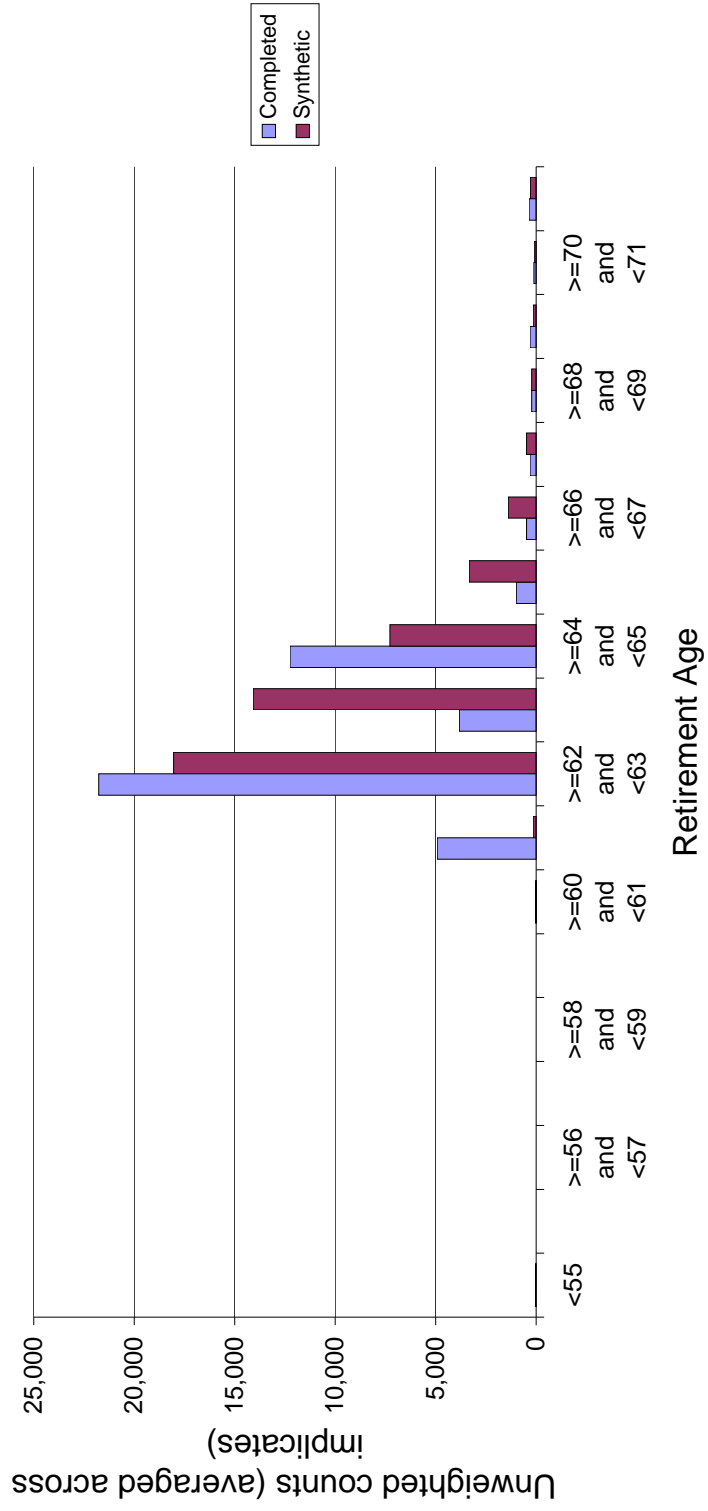


Figure 3.10: Age at Retirement, Unweighted



References

Abowd, J., Lengermann, P. and McKinney, K. (2002). Changing the boundaries of the firm: Changes in the clustering of human capital, *LEHD technical paper tp-2002-02*, U.S. Census Bureau.

URL: <http://lehd.did.census.gov/led/library/techpapers/tp-2002-02.pdf>

Abowd, J. M., Haltiwanger, J., Lane, J. and Sandusky, K. (2001). Within and between firm changes in human capital, technology, and productivity, *LEHD technical paper tp-2001-03*, U.S. Census Bureau.

URL: <http://lehd.did.census.gov/led/library/techpapers/tp-2001-03.pdf>

Abowd, J. M., Kramarz, F. and Margolis, D. (1999). High Wage Workers and High Wage Firms, *Econometrica* **67**(2): 251–333.

Abowd, J. M., Lane, J. and Prevost, R. (2000). Design and conceptual issues in realizing analytical enhancements through data linkages of employer and employee data, *Proceedings of the Federal Committee on Statistical Methodology*.

Abowd, J. M. and Stinson, M. H. (2006). Estimating measurement error in sipp annual job earnings: A comparison of census survey and ssa administrative data, *LEHD technical paper tp-2006-05*, U.S. Census Bureau.

URL: <http://lehd.did.census.gov/led/library/techpapers/tp-2006-05.pdf>

Abowd, J. M. and Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers, *Journal of Business Economic Statistics* **23**(2): 133–152.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure Limitation in Longitudinal Linked Data, in P. Doyle, J. Lane, J. Theeuwes and L. Zayatz (eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam: North Holland, pp. 215–277.

- Abraham, K. and Taylor, S. (1996). Firms' use of outside contractors: Theory and evidence, *Journal of Labor Economics* **14**(3): 394–424.
- Acs, Z. and Armington, C. (1998). Firms' use of outside contractors: Theory and evidence, *CES working paper ces-wp-98-9*, U.S. Census Bureau.
- Albert, A. and Anderson, J. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika* **71**: 1–10.
- Andrade, G., Mitchell, M. and Stafford, E. (2001). New evidence and perspectives on mergers, *Journal of Economic Perspectives* **15**(2): 103–120.
- Autor, D. (2003). Outsourcing at will: Unjust dismissal doctrine and the growth of temporary-help employment, *Journal of Labor Economics* **21**(1).
- Benedetto, G., Haltiwanger, J., Lane, J. and McKinney, K. (2007). Using worker flows to measure firm dynamics, *Journal of Business Economic Statistics* . (forthcoming).
- Brown, C., Haltiwanger, J. and Lane, J. (eds) (2006). *Economic Turbulence: Is Volatility Good for America?*, Chicago: University of Chicago Press.
- Brown, C. and Medoff, J. L. (1987). The impact of firm acquisitions on labor, *Nber working paper 2273*, NBER.
- Carroll, N., Hyslop, D., Maré, D., Timmins, J. and Wood, J. (2002). An analysis of new zealand's business demography database, *New Zealand Economic Papers* **36**(1): 59–62.
- Clayton, R., Akbar, S., Spletzer, J. and Talan, D. (2003). Business demographics: Measuring job creation and job destruction dynamics underlying net employment change. Presented at Joint Unece/Eurostat Seminar On Business Registers (Luxembourg, June 25-26).

- Couch, K. A. and Lillard, D. R. (1998). Sample selection and the intergenerational correlation in earnings, *Labour Economics* **5**(3): 313–329.
- Davis, S., Haltiwanger, J., Jarmin, R., Krizan, C., Miranda, J., Nucci, A. and Sandusky, K. (2005). Measuring the dynamics of young and small businesses: Integrating the employer and nonemployer universes, *CES working paper ces-wp-06-04*, U.S. Census Bureau.
- Davis, S. J., Faberman, R. J. and Haltiwanger, J. (2006). The flow approach to labor markets: New data sources and micro-macro links, *The Journal of Economic Perspectives* **20**(3): 3–26.
- Domingo-Ferrer, J., Abowd, J. M. and Torra, V. (2006). Using mahalanobis distance-based record linkage for disclosure risk assessment, *in* J. Domingo-Ferrer and L. Franconi (eds), *Privacy in Statistical Databases*, Springer-Verlag, p. forthcoming.
- Domingo-Ferrer, J., Torra, V., Mateo-Sanz, J. and Sebe, F. (2006). Empirical disclosure risk assessment of the ipso synthetic data generators, *Monographs in Official Statistics-Work Session on Statistical Data Confidentiality*, Eurostat.
- Doms, M. and Bartelsman, E. (2000). Understanding productivity: Lessons from longitudinal microdata, *Journal of Economic Literature* **38**(3): 569–594.
- Estevao, M. and Lach, S. (1999). The evolution of the demand for temporary-help supply, *Nber working paper 7427*, NBER.
- Faberman, J. (2001). Job creation and destruction within washington and baltimore, *Monthly Labor Review* **123**(9): 24–31.
- Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage, *Journal of the American Statistical Association* **64**: 1183–1210.

- Foster, L., Haltiwanger, J. and Krizan, C. (2001). Aggregate productivity growth: Lessons from microeconomic evidence, *in* E. Dean, M. Harper and C. Hulten (eds), *New Directions in Productivity Analysis*, Chicago: University of Chicago Press.
- Gaynor, M. and Haas-Wilson, D. (1998). The blessing and the curse of managed care: Vertical relations in health care markets, *in* M. Morrissey (ed.), *Managed Care and Changing Health Care Markets*, Washington, D.C.: American Enterprise Institute Press.
- Gelman, A. B., Carlin, J. S., Stern, H. S. and Rubin, D. B. (2000). *Bayesian Data Analysis*, Chapman and Hall.
- Gokhale, J., Groshen, E. L. and Neumark, D. (1995). Do hostile takeovers reduce extramarginal wage payments?, *The Review of Economics and Statistics* **77**(3): 470–485.
- Hamermesh, D. (1999). Leaping into the future of labor economics, *Labour Economics* **6**(1): 25–41.
- Houseman, S. (1997). Temporary, part-time, and contract employment in the united states: A report on the w.e. upjohn institute’s employer survey on flexible staffing policies, *Technical report*, U.S. Department of Labor.
- Hunter, L., Bernhardt, A., Hughes, K. and Skuratowicz, E. (2001). It’s not just the atms: Firm strategies, work restructuring, and workers’ earnings in retail banking, *Industrial and Labor Relations Review* **54**(2A): 402–424.
- Jacobson, L. S., LaLonde, R. J. and Sullivan, D. G. (1993). Earnings losses of displaced workers, *The American Economic Review* **83**(4): 685–709.
- Jarmin, R. and Miranda, J. (2002). The longitudinal business database, *CES working paper ces-wp-02-17*, U.S. Census Bureau.

- Jensen, M. C. (1988). Takeovers: Their causes and consequences, *The Journal of Economic Perspectives* **2**(1): 21–48.
- Jovanovic, B. and Rousseau, P. (2002). Mergers as reallocation, *Nber working paper 9279*, NBER.
- Lane, J., Burgess, S. and Theeuwes, J. (1998). The uses of longitudinal matched employer/employee data in labor market analysis, *Proceedings of the American Statistical Association*.
- Lengermann, P. and Vilhuber, L. (2001). Abandoning the sinking ship: The composition of worker flows prior to displacement, *LEHD technical paper tp-2002-11*, U.S. Census Bureau.
URL: <http://lehd.did.census.gov/led/library/techpapers/tp-2002-11.pdf>
- Little, R. J. (1993). Statistical analysis of masked data, *Journal of Official Statistics* **9**(2): 407–426.
- Minka, T. P. (2003). Bayesian inference, entropy, and the multinomial distribution, *Technical report*, Microsoft, Inc. accessed October 30, 2006.
URL: <http://research.microsoft.com/~minka/papers/minka-multinomial.pdf>
- Persson, H. (1999). Essays on labour demand and career mobility. PhD Dissertation, University of Stockholm.
- Pivetz, T. and Chang, H. (1998). Linking unemployment insurance wage records to es-202 establishment microdata to improve the accuracy of the bls' longitudinal business establishment database. Presented at the 1998 International Symposium on Linked Employer-Employee Data, Washington, D.C.
- Pivetz, T., Searson, M. and Spletzer, J. (2001). Measuring job and establishment flows with bls longitudinal microdata, *Monthly Labor Review* **124**(4): 13–20.

- Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V. and Solenberger, P. (1998). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Research Center, University of Michigan.
- Raghunathan, T., Reiter, J. and Rubin, D. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics* **19**(1): 1–16.
- Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets, *Survey Methodology* **29**: 181–188.
- Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation, *Survey Methodology* **30**: 235–242.
- Rubin, D. B. (1981). The bayesian bootstrap, *The Annals of Statistics* **9**: 130–134.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.
- Rubin, D. B. (1993). Discussion of statistical disclosure limitation, *Journal of Official Statistics* **9**(2): 461–468.
- Rubin, D. B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association* **91**(434): 473–489.
- Spletzer, J. (2000). The contribution of establishment births and deaths to employment growth, *Journal of Business and Economic Statistics* **18**(1): 113–126.
- Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, third edn, Springer-Verlag.
- United States Department of Labor, E. (2002). Unemployment insurance program letter no.34-02 (revised). accessed August 16, 2006.
- URL:** http://workforcesecurity.doleta.gov/dmstree/uipl/uipl2k2/uipl_3402.htm

Woodcock, S. D. and Benedetto, G. (2006). Distribution-preserving statistical disclosure limitation, *LEHD technical paper tp-2006-04*, U.S. Census Bureau. accessed October 31, 2006.

URL: <http://lehd.did.census.gov/led/library/techpapers/tp-2006-04.pdf>