Statistical Policy
Working Paper 16

A Comparative Study of Reporting Units in Selected Employer Data Systems

Prepared by
The Employer Reporting Unit Match Study (ERUMS) Work Group
Administrative Records Subcommittee
Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

May 1990

MEMBERS OF THE FEDERAL COMMITTEE ON

STATISTICAL METHODOLOGY

(April 1990)

Maria E. Gonzalez (Chair)

office of Management and Budget

Yvonne M. Bishop
Energy Information
Administration

Warren L. Buckler
Social Security Administration

Charles E. Caudill
National Agricultural
Statistical Service

John E. Cremeans
Office of Business Analysis

Daniel Kasprzyk
Bureau of the Census

Daniel Melnick
National Science Foundation

Robert P. Parker
Bureau of Economic Analysis

David A. Pierce
Federal Reserve Board

Thomas J. Plewes
Bureau of Labor Statistics

Zahava D. Doering

Smithsonian Institution

Joseph K. Garrett

Bureau of the Census

Robert M. Groves

Bureau of the Census

C. Terry Ireland

National Computer Security

Center

Charles D. Jones

Bureau of the Census

Wesley L. Schaible

Bureau of Labor Statistics

Fritz J. Scheuren

Internal Revenue Service

Monroe G. Sirken

National Center for Health

Statistics

Robert D. Tortora

Bureau of the Census

PREFACE

The Federal Committee on Statistical Methodology was organized by OMB in 1975 to investigate methodological issues in Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in their personal capacity rather than as agency representatives. The committee conducts its work through subcommittees and work groups that are organized to study particular issues and that are open to any Federal employee who wishes to participate in the studies. Working papers are prepared by the subcommittee/work group members and reflect only their individual and collective ideas.

The Employer Reporting Unit Match Study (ERUMS) Work Group of the Administrative Records Subcommittee was formed to conduct a study that compared employer and reporting unit data from the record systems of the Bureau of Labor Statistics (BLS), and the Social Security Administration (SSA), supplemented with employer level information from the Internal Revenue Service (IRS). To carry out

the match study, interagency agreements were developed between BLS and SSA and between BLS and IRS. These agreements were the bases for sharing the microdata. The purpose of the match was to obtain more precise information on the differences and similarities in the coverage and content of the data in these systems.

Although the study was limited in scope, the results serve to point in the direction of future work which needs to be done in understanding various establishment microrecord systems. Also in the context of possible future sharing of microrecords, further studies need to be carried out.

The Employer Reporting Unit Match Study Work Group was chaired by Warren L. Buckler of the Social Security Administration, Department of Health and Human Services.

Members of the ERUMS Workgroup

Administrative Records Subcommittee

(November 1989)

Warren Buckler*, Chair

Social Security Administration

Lois Alexander	Ken LeVasseur
Social Security Administration	Bureau of Labor Statistics
Marlene Einstein	Bruce Levine
Bureau of Labor Statistics	Bureau of Economic Analysis
Jerry Gates (observer)	Tom Petska
Bureau of the Census	Internal Revenue Service
Maria Gonzalez* (ex officio)	John Pinkos
Office of Management and Budget	Bureau of Labor Statistics

Tom Grzesiak — Bureau of Labor Statistics
Vern Renshaw — Bureau of Economic Analysis

Tom Jabine — Committee on National Statistics
Alan Zempel — Internal Revenue Service

* Member, Federal Committee on Statistical Methodology

ACKNOWLEDGEMENTS

This report represents the culmination of the collective efforts of many individuals who,have been involved with the ERUMS project throughout the course of its development and

implementation. A designated individual Workgroup member had the primary responsibility for each section of the report. In several cases, significant contributions were made by others, as shown below:

Section	Responsible author and other contributors
Exec. Sum.	Tom Jabine (CNSTAT)
Ch I	Tom Jabine (CNSTAT)
Ch II,A,1	Marlene Einstein (BLS), Ken LeVasseur (BLS), Karen Mainzer (BLS)
Ch II,A,2	Warren Buckler (SSA), Cheryl Williams (SSA)
Ch II,A,3	Alan Zempel (IRS), Charles Day (IRS)
Ch II,B & C	Tom Jabine (CNSTAT)
Ch II,D,1	Lois Alexander (SSA)
Ch II,D,2	Warren Buckler (SSA)
Ch III,A	Vern Renshaw (BEA)
Ch III,B	Tom Jabine (CNSTAT)
Ch IV,A	Vern Renshaw (BEA)
Ch IV,B	Tom Jabine (CNSTAT)

The data processing and tabulation preparation operations were

performed at BLS by Marlene Einstein, assisted by Suzie Yen, and by Joel Packman at SSA. Tom Jabine, CNSTAT, developed the outline for the format of the report and served as contents editor. All of the current members of the Workgroup reviewed successive drafts, offered comments and suggestions, and approved this final report. In addition, a number of improvements to the preliminary draft that was submitted to the Federal Committee on Statistical Methodology (FCSM) resulted from comments and suggestions made by the principal reviewers for that committee, Tom Plewes and Bob Parker, and by Fritz Scheuren and Dan Kasprzyk. .

The Workgroup would like to express its deep and sincere appreciation to all of the dedicated individuals who have been a part of this project. In addition to the current Workgroup members and other contributors to various sections of the report who have been previously cited, several former members of the Workgroup, as well as other staff of the participating,agencies, are to be recognized for their contributions. This group includes: Brian MacDonald, Linda Hardy, Michael Searson, John Pinkos, E.J. Filardi and Alan Tupek of the Bureau of Labor Statistics; Jackie Veach,

Linda Dill, Cres Smith, Barry Bye, and Shirley Piazza of the Social Security Administration; Fritz Scheuren of the Internal Revenue Service and Alfred Nucci of the Census Bureau.

The Workgroup would also like to express its appreciation to Maria Gonzalez for her patience, sound advice, and the guidance she provided throughout the project and to Tom Plewes for his unwavering support and constant encouragement for the work we were doing.

TABLE OF CONTENTS

page

LIST OF EXHIBITS

Page

Exhibit

Text

Appendix

(note: Attachments to the above agreements (B-1, B-2) are

not included with this report, but are available

upon request.)

LIST OF TABLES

Appendix

Introduction (Chapter I)

The Employer Reporting Unit Match Study (ERUMS) was a pilot record linkage study carried out under the auspices of the Federal Committee on Statistical Methodology (FCSM), Office of Management and Budget. The study linked records of employers and their reporting units from three agencies: the Bureau of Labor Statistics (BLS), the Social Security Administration (SSA) and the Internal Revenue Service (IRS). The primary linkages involved samples of the agencies, records for employers in the State of Texas, covering their activities in 1982.

The ERUMS project was planned and carried out by an interagency workgroup under the general guidance of the Federal Committee on Statistical Methodology. Planning began in 1983 and the project operations were completed in 1989. The motivation for ERUMS came from earlier work of the FCSM Subcommittee on Statistical Uses of Administrative Records, which had determined that effective and efficient statistical uses of administrative records were being hampered by the existence of noncompatible

systems for reporting employer information at the establishment level.

The goal of ERUMS was to demonstrate the feasibility of matching employer and reporting unit data from different agency record systems as a means of obtaining more precise information about differences in the coverage and content of the data in those systems. The study focussed on the BLS and SSA record systems, with employer-level data from IRS being used primarily to reconcile and explain BLS-SSA differences. It was expected that ERUMS, as a demonstration study, would provide valuable experience with the technical aspects of data linkage and the administrative requirements for gaining access to the data and carrying out the matching operations.

The record systems that were linked (Chapter II, Section A)

The primary source of data for ERUMS from BLS was the first quarter 1982 Unemployment Insurance (UI) Address File. For each

State, the UI Address File contains data for individual employers and their reporting units, which are often but not always equivalent to establishments. The data for this file are submitted annually (more recently quarterly) to BLS by the State employment security agencies that operate the Federal-State UI Program. The BLS uses the data submitted by the States as a basis for periodic statistical reports on employment and wages and uses the UI Address File as a national sampling frame for its establishment surveys.

The principal SSA files used for ERUMS were files developed for statistical uses within SSA. They included an edited file of

Form W-3 annual wage reports for 1982 and the Single Unit and Multi Unit Code Files. The Form W-3 file provided wage data for

individual employers and, in some cases, for each of their reporting units, which are frequently but not always equivalent to establishments. The Single Unit Code File, which is updated annually, contains a record for every entity that has filed an application for an Employer Identification Number (EIN), excluding non-employing entities and household employers. The Multi Unit Code File contains a record for each reporting unit of multi unit employers who are participating in the Establishment Reporting Plan, a voluntary program under which employers report their annual wage information on Form W-3 separately for each of their reporting units.

The main source of IRS data used for ERUMS was a Census-edited file based on Forms 941 and 943 for Tax Years 1981-83. These forms are used by employers to report each quarter (annually for Form 943) to IRS on income taxes withheld from wages and other payments to employees and on taxes under the Federal Insurance Contributions Act (Social Security taxes). Extracts of data from these forms are provided annually by IRS to the Census Bureau for use in the latter's County Business Patterns Program and other statistical

purposes. The Census Bureau edits the files to use the best available industry code for each employer and impute certain missing data. A copy of the edited file has been made available to the IRS Statistics of Income Division for use in its statistical programs. Data from this Census-edited file were obtained for most of the employers in the Phase II ERUMS sample (see below). In addition, copies of Form 940, Federal Unemployment Tax Return, for 1982 or 1983 were obtained for a substantial proportion of the Phase II sample cases.

The study design (Chapter II, Sections B and C)

Because of the ERUMS Workgroup's limited resources, the study was restricted to a single State, Texas, and a small sample of employers and their reporting units from that State. The sampling unit was the employer, identified by a unique EIN. A probability sample of all EINs active in the State of Texas in 1982 was selected from the BLS and SSA files described above. Employers were considered to be active in the BLS system if they had one or more records in the 1982 UI Address File and in the SSA system if

they had filed a W-2/W-3 wage report for 1982.

The sample was selected in two phases. The sampling fraction for Phase I was 6 in 100, and the selection was based on the 7th and 8th digits of the EIN. The BLS sample, which was selected first, contained 16,336 distinct EINS. The BLS sample was compared to the SSA files and an additional sample was selected (using the same pairs of digits) of 3,628 EINs which had at least one Texas reporting unit, had wage reports for 1982 and did not appear in the 1982 UI Address File. The Phase I sample EINs were stratified by match status (match, SSA only, BLS only)

and single/multi unit status. A Phase II sample of 401 EINs was selected from the Phase I sample, using disproportionate stratified

sampling, with equal probability systematic selection within each stratum. Nonmatch and multi unit EINs were oversampled in Phase II because of their greater interest for the purposes of ERUMS.

The Phase II sample provided the basis for the detailed analyses presented in this report. For matched cases, BLS and SSA geographic and industry codes were compared. The industry codes from both sources were compared with those in the IRS/Census-edited Form 941 file. The status of unmatched EINs was clarified by reviewing additional data sources in the agency for which the EIN did not show up in the initial match. Several of the EINs not located initially in the SSA edited 1982 W-3 file were found among groups of delinquent reporters or cases for which the W-2/W-3 wage report and IRS Form 941 data were being reconciled. In addition, several of the Phase II sample employers originally classified as SSA multi unit were reclassified as single unit because it could not be established that they reported 1982 wages for two or more reporting units in Texas. As a result of these reviews and changes, the final distribution of the sample EINs by match status and single/multi unit classification differed substantially from

the preliminary distribution of the Phase II sample.

Administrative arrangements (Chapter II, Section D)

For the ERUMS Workgroup to,gain access to the data sets needed for the study, it was necessary to develop working arrangements that complied with the provisions of confidentiality statutes, regulations and policies of the Federal and State agencies that controlled these data sets. After protracted negotiations, this was accomplished primarily through the development of two bilateral agreements (shown in Appendix B).

In one of these agreements, the IRS contracted with BLS for the performance of those parts of the ERUMS project that required access to tax data, including the wage report information that was to be provided by SSA. Under this agreement, SSA staff could be designated as special agents of BLS to carry out their part of the linkage and analysis operations. By law, the purposes of IRS participation in the project and its service contract with BLS had

to be related to IRS administration of the tax laws.

The second agreement was a conditions of use agreement between SSA and BLS which allowed SSA to release relevant data from its employer files to BLS and authorized BLS to link data from these files with data from the UI Address File and certain data to be furnished by IRS, and prohibited any other linkage. Both agreements incorporated several safeguards, with emphasis on limiting access at each stage of the project to those persons who needed to use identifiable data, keeping the number of such

persons to a minimum and having them sign non-disclosure affidavits.

To meet the statutory confidentiality requirements of the

State of Texas, BLS obtained the permission of the Texas State Employment Commission to use the 1982 Texas UI Address File microdata for the ERUMS study.

Results (Chapter III,A)

All results based on the ERUMS sample are estimates weighted to account for the disproportionate sampling used in the selection of the Phase II sample, unless otherwise noted. The main quantitative results are shown in Tables IIIA-1 through 8 at the end of Section III,A)

Of the Texas EINs that were active in 1982 in the BLS or SSA systems, 67.1 percent were active in both systems, 27.6 percent were active only in the SSA system and 5.3 percent were active only in the BLS system (Table IIIA-1). Only about 1.0 percent of all active EINs were classified as multi unit in one or both systems, and most of these were classified as multi unit only in the BLS system (Table IIIA-4).

For the matched single unit EINS, i.e., those that were active in both systems, an estimated 81.6 percent had the same State and county codes in both systems. The remaining cases were about equally distributed in three categories: same State, different county; same State with no county code in the SSA file; and different State (Table IIIA-5). An estimated 70.2 percent of the matched single unit cases had the same two-digit industry codes. About half of the remaining cases were not classified by industry in the SSA system (Table IIIA-5). When matched against the IRS/Census-edited Form 941/943 file, about three-fourths of the matched single units from both the BLS and SSA files had two-digit industry codes that agreed with those in the IRS/Census file. However, when the SSA unclassified cases were excluded from this comparison, the proportion of SSA cases that agreed with the IRS/Census two-digit code was somewhat greater than the corresponding proportion for the BLS matched single unit cases (Table IIIA-8).

Only a few EINs (nine sample cases) were classified as multi

unit in both the BLS and SSA systems. Matching individual reporting units for these cases proved to be difficult. Overall, the nine sample employers had 105 Texas reporting units in the BLS system and 60 in the SSA system for 1982.

Of the active SSA EINs not found in BLS's first quarter 1982 UI Address File, it was estimated that 69.2 percent had reported no first quarter employment to IRS on Form 941 and therefore would not normally be expected to appear in the BLS system (Table IIIA-6). For another 10 percent of these employers, the analysis suggested that they may not have met requirements for UI coverage

in Texas either because they had no operations in Texas, because of nonprofit status or because their payrolls were too small. For the

remaining 20 percent, the reasons for their absence are not always clear, but it may have resulted in part from lags in incorporating new employers in the UI State agency and BLS files.

Most of the employers who were included in the 1982 UI Address File but did not file 1982 W-2/W-3 wage reports (22 sample cases) appeared to have ceased hiring employees, gone out of business, or gone through other changes that altered their reporting to IRS and SSA. Half of the employers in this group reported no employment in the 1982 UI Address File. Many of the remainder had filed their final Form 941 with IRS (at least for the period 1981-1983) for a quarter in 1981.

An analysis of the sample EINs that appeared in SSA's Multi Unit Code File provided some indication of the extent to which multi unit employers were participating in SSA's Establishment Reporting Plan (ERP) in 1982 (Table IIIA-7). An estimated 35.9 percent of these EINs had been incorrectly added to the Multi Unit Code File as the result of a processing error that has since been corrected. Most of the remaining employers had initially agreed to participate in the ERP, but more than half of this group did not

provide separate data for each reporting unit in their W-3 wage reports for 1982.

Limitations of the study (Chapter III,B)

Several factors limit the broad applicability of the ERUMS findings. The results reflect the reporting requirements and operating procedures associated with the agency record systems in 1982. There have been significant changes since then. In particular, BLS has taken several steps to improve the timeliness and the completeness and accuracy of data in its UI Address File.

The study was based on data for a single State, Texas, and on a small sample of employers and reporting units. The UI system gives the States some latitude in their record-keeping practices, so indications of the coverage of employers in the record systems of the Texas State Employment Agency in 1982 should not be assumed to apply fully to the UI systems of other States at that time. The small sample size means that estimates based on the Phase II sample

are subject to relatively large sampling errors. Because of limited resources and the complexity of the Phase II sample design, we were able to compute sampling errors only for a few key estimates (see Table IIIA-4).

The analysis of the results was complicated by differences in concepts and coverage in the record systems used in the study. These differences occurred in the basic filing requirements for the UI and SSA/IRS systems, the time reference of the basic BLS and SSA files used for matching, the definition of reporting units in the BLS and the SSA/ERP systems, and the structures of the BLS and SSA industry classification systems. In addition,

certain file deficiencies and operational problems made the

analyses more difficult. About 1.3 percent of the records in the 1982 UI Address File for Texas did not have EINs and therefore were not included in the Phase I sample of EINs from that file. In the SSA files, a significant proportion of employers lacked county and industry codes. The most serious problem was that a high proportion of multi unit employers were not reporting separately in 1982 for each reporting unit, so that we were unable to do a thorough comparison of reporting units for multi unit employers active in both the BLS and SSA systems.

Although these differences and file deficiencies made the analyses more difficult, the fact that we succeeded in identifying and documenting them is an indication that the ERUMS project succeeded in its main goal, which was to demonstrate the feasibility of doing matching studies as a means of evaluating the suitability of administrative record systems for statistical uses.

The data on amounts of employment and payroll available from SSA, BLS and IRS files were used in reviewing the unmatched sample cases and trying to understand why they were not present in both

SSA and BLS files. However, the employment and payroll data were not added to the data file for the 401 sample EINs that were used to develop the estimates presented in this report. Therefore, all of the results shown are estimates of numbers of employers or reporting units, classified by attributes such as match status, and geographic and industry codes in the different systems included in the study. We did not attempt to estimate what proportions of aggregate employment or payroll were accounted for by employers who were unmatched or had different geographic or industry codes.

Findings (Chapter IV,A)

The detailed analyses of the ERUMS data did not suggest that large numbers of employers who report wages in one of the payroll tax systems were failing to report in the other system when they should have been. They do, however, suggest that late reports and different procedures for processing the reports in the two systems created potential problems for using both of the systems' data files for statistical purposes.

Perhaps the clearest finding was that it is not possible to maintain a usable establishment reporting unit plan for multi unit employers in the absence of systematic procedures,for monitoring employer reporting and updating files for changes in the number, location and industry of each employer's reporting units. SSA's Establishment Reporting Plan clearly lacked the necessary resources to do this in 1982 and there is no reason to think that the situation has improved since then.

There was a moderately high but by no means perfect correspondence between county and two-digit industry codes for

single unit employers included in both the BLS and SSA systems. A substantial proportion of the differences arose from the absence of

county or industry codes in the SSA system. Comparisons of industry codes at the three and four-digit level were not attempted because of the differences in the industry classification systems used by the two agencies.

With some qualifications, we were successful in matching the records of employers, as defined by their EINS, in different systems. However, we were not successful in matching BLS and SSA records for reporting units, the main reason being the incompleteness of SSA's data for reporting units provided under the voluntary ERP. Other reasons were the lack of a common identifier, analogous to the EIN at the employer level, for reporting units and the slight differences in the reporting unit definitions used by BLS and SSA.

We learned what we believe are some important lessons for others who may wish to match business records from different agency sources, whether for research or operational purposes. First, the plans and the necessary interagency agreements should be developed well ahead of the earliest date at which the files to be linked are

expected to be available. In particular, the development of interagency agreements for the exchange of identifiable records is a painstaking process and considerable time may be needed for their completion and approval.

Second, successful matching requires in-depth knowledge of all of the record systems involved and of the specific files that exist within those systems. An interagency team approach, with full exchange of information, is essential because there is unlikely to be a single individual who has all of the necessary information, even for the files of a single agency.

Finally, whenever possible, it is essential to pretest matching procedures before embarking on large-scale operational applications.

Recommendations (Chapter IV,B)

ERUMS was designed primarily as a demonstration project and

was therefore limited in its coverage and scope. Nevertheless, the Workgroup believes that the study results, along with other information acquired in the course of the study, justified the inclusion in its report of five formal recommendations addressed specifically to the BLS and SSA record systems for employers and reporting units. These recommendations were:

1 - SSA should undertake a full review of the current status and uses of the Establishment Reporting Plan and decide either to continue it with adequate resources for maintenance and improvement of quality or to discontinue it entirely.

2 - BLS should review the State Employment Security Agencies' procedures for identifying employer births (including those resulting from mergers and changes of organization) and seek ways

of reducing the apparent lag between filing of applications for EINs and inclusion of new employers on State Agency and BLS lists used as frames for statistical surveys and reports.

3 - Data in the UI Address File on employment and wages paid should be labelled to distinguish imputed data from data reported by employers.

4 - The EIN should be identified as a key item in the UI Address File and efforts should be made to achieve 100 percent reporting initially and current reporting of changes in EINS.

5 - BLS and SSA (if it continues the Establishment Reporting Plan) should strive to obtain data from employers for their establishments as defined in the 1987 Standard Industrial Classification (SIC) Manual. Both agencies should code industry for all establishments, without exception, at the 4-digit SIC level of detail. Whether or not the Establishment Reporting Plan is continued, SSA should code all employers identified on Forms SS-4 at the 4-digit level of detail.

In a broader context, the ERUMS Workgroup concluded that current efforts to collect economic data at the establishment level are dispersed among Federal and State,agencies, are poorly coordinated, and place unnecessary burden on employers. The Workgroup believes that further, more intensive and extensive interagency matching studies have an important role to play in resolving these problems and in determining the possible effects on statistical programs of prospective major changes in administrative reporting systems for employers. We therefore recommend that:

6 - Further matching studies should be directed at acquiring information that will support the eventual development of a mandatory reporting system to meet the needs of all Federal and State statistical programs for establishment lists, including SIC codes. An interim goal should be that all agencies requiring or requesting employers to provide data at the establishment or reporting unit level adopt common definitions of units and data items to be submitted for these units.

Three agencies -- the BLS, the Census Bureau and the National Agricultural Statistics Service -- play a dominant role in the direct collection of establishment-level economic data. Recent initiatives of these agencies, under the general guidance of OMB's Statistical Policy Office, have been directed at greater coordination of their respective list-building and maintenance activities. Further integration of business lists will require fuller understanding of the similarities and differences of the three systems, based on matching of individual establishments and reporting units in the different. systems.

CHAPTER I - INTRODUCTION

This working paper is a report on the Employer Reporting Unit

Match Study (ERUMS), a pilot record linkage study carried out by Federal agencies under the auspices of the Federal Committee on Statistical Methodology, Office of Management and Budget (OMB). The report describes the design, procedures and findings of the study and presents recommendations based on the findings.

The study linked records of employers and their reporting units from three agencies: the Bureau of Labor Statistics (BLS), the Social Security Administration (SSA) and the Internal Revenue Service (IRS). The primary linkages involved samples of the agencies' records for employers in the State of Texas, covering their activities in 1982.

The study was designed and most of the work undertaken by members of the ERUMS Workgroup, whose members represented the three agencies whose records were linked, plus the OMB, the Bureau of Economic Analysis and the Committee on National Statistics, which has had a continuing interest in encouraging more effective statistical uses of administrative records. Bureau of the Census representatives attended many of the workgroup meetings as observers. The ERUMS Workgroup reported periodically to and

received guidance from the Federal Committee on Statistical Methodology (FCSM). The chair of the FCSM attended most of the Workgroup meetings.

A. Background

Establishment-based economic and business statistics in the United States are derived in large part from reporting systems developed to administer the Federal Income Tax and Social Security systems and the Federal-State Unemployment Insurance system. BLS statistical series on employment and total wages are a by-product of administrative reporting systems established at the State level to support the Unemployment Insurance (UI) system. SSA uses information derived from records of employer taxes on earnings to classify persons included in its Continuous Work History Sample by industry and place of work. IRS uses samples of income tax and information returns for corporations, partnerships and sole proprietors to produce annual data for these units in its Statistics of Income program. The Census Bureau uses data from

business tax returns for small units in lieu of direct data collection from these units in the quinquennial economic censuses and as a source of current employment and payroll data for its County Business Patterns Program.

In addition to their direct uses for statistical purposes, these administrative reporting systems provide lists of business units (sometimes called frames) that are used by statistical

agencies primarily the BLS and the Bureau of the Census, to determine which units to cover in periodic censuses and current surveys of economic establishments.

The extensive use of data from these administrative reporting

systems for statistical purposes is cost-effective and reduces the reporting burden on business. However, use of administrative records also has its problems. A primary difficulty is that reports by businesses for administrative purposes are generally needed only at aggregate levels. Reports of earnings to IRS and SSA for the Social Security system are for employers, i.e., all activities covered by a unit with a single Employer Identification Number (EIN). Employer reports of earnings to a State employment security agency for the Unemployment Insurance system frequently cover all activities by the employer (EIN unit) in that State. Likewise, reports submitted by employers to IRS on Form 940 under the Federal Unemployment Tax Act provide aggregate data, by State, on covered wages.

Data at this level of aggregation have limited value for statistical analyses. Many corporations and employers have activities in several different locations and in several different categories of industry. Detailed statistical analysis of economic activity calls for information on inputs and outputs at the establishment level, i.e, separate data for each kind of economic

activity at each physical location. The establishment, as formally defined by OMB, is the basic reporting unit for the Census Bureau's economic censuses and surveys.

To meet the need for establishment-type data, both BLS and SSA have developed voluntary statistical reporting systems to supplement their administrative reporting systems. BLS has a statistical reporting program, mandatory in 20 States and voluntary in the rest, under which employers submit quarterly reports to State employment security agencies with quarterly wage and monthly employment information by reporting unit. This information is used with data on single establishment firms to update BLS' Universe File, which is its frame for establishment surveys.

SSA has its voluntary Establishment Reporting Plan, under which participating employers filing their annual reports of earnings covered by Social Security provide separate information for each reporting unit. In 1982, the reference year for this study, the SSA reporting unit definition was similar to but not exactly the same as the one used by BLS. Both differed significantly from the OMB establishment definition used by the

Census Bureau in its statistical programs. There are also some differences in how each of the agencies has adapted OMB's Standard Industrial Classification for use in its own statistical programs (OMB, 1984; Jabine, 1984).

To meet its own requirements, the Census Bureau conducts an

annual survey, the Company Organization Survey, to collect current information about the location and activities of the establishments associated with multi unit employers. This information is used to update Census' Standard Statistical Establishment List (SSEL), which serves as the frame for all of its economic censuses and surveys.

There have been several studies comparing aggregate data on employment and earnings published by BLS, IRS, SSA and the Census Bureau (e.g., Bureau of the Budget, 1961; Bureau of Economic Analysis, 1972; Office of Federal Statistical Policy and Standards, 1980). As might be expected because of the differences in coverage and definition of the various administrative and statistical reporting systems, significant differences in data by industry and location have been observed in these studies. There have been few micro-level interagency comparisons of establishment-type data, especially in recent years. Those that have been undertaken (e.g., Bureau of the Census, 1965) have shown many differences in establishment reporting in the systems that were compared.

In summary, the effective and efficient use of administrative records for statistical purposes has been impeded by the existence of no-compatible systems for reporting of employer information at the establishment level. Serious problems exist because of differences in coverage, reporting unit definitions, and industry classification systems. These differences lead to lack of comparability in the economic statistics produced by different

agencies in our decentralized statistical system.

B. Prior Activities of the FCSM

The FCSM has been concerned with statistical uses of administrative records since 1977: several subcommittees and working groups have examined different aspects of this topic. The Subcommittee on Statistical Uses of Administrative Records (Office of Federal Statistical Policy and Standards', 1980) made a broad review of the quality of administrative data and their suitability for statistical applications. The Subcommittee recommended further efforts to: promote the use of standard identifiers, concepts and definitions in administrative reporting programs; identify and resolve problems of access to data in these systems for statistical applications; and establish government-wide coordination and support of relevant collection programs and research activities. A continuing Administrative Records Subcommittee was formed to pursue these goals.

Under the Administrative Records Subcommittee, an Establishment Reporting Work Group was formed early in 1981 to make a more detailed study of three major record systems: the Unemployment Insurance record systems maintained by the States under rules and procedures established by the Department of Labor; the annual W-2 and W-3 wage reports submitted by employers

to SSA and used by both SSA and IRS for administrative purposes; and the Census Bureau's Standard Statistical Establishment List (SSEL), which serves as the frame for,that agency's economic censuses and surveys. The Work Group succeeded in documenting the structural differences among these three systems but was unable, for various reasons, to undertake a planned record matching study to shed additional light on the factors contributing to statistical inconsistencies among the three systems. However, the final

recommendation of the Work Group to do further work in this area was heeded and the ERUMS Workgroup was formed early in 1983 (Cartwright, Levine and Buckler, 1983).

C. Goals of the ERUMS Project

Members of the ERUMS Workgroup felt that little more could be done to develop detailed recommendations for improved establishment reporting without first obtaining more precise information, at the micro-level, about inconsistencies among the major administrative reporting systems. Therefore, the Workgroup determined that its main goal would be to conduct a pilot study based on matching of data from employer wage reporting and establishment reporting systems of BLS, IRS and SSA. The study would focus on differences between the BLS and SSA systems, with employer-level data from IRS being used primarily to reconcile and explain BLS-SSA differences. For full coverage of the major establishment-based statistical programs, it would have been desirable to include the Census Bureau's SSEL in the matched data set, but the predecessor

workgroup had not been able to arrange to do this, and it was decided not to pursue this effort as part of the ERUMS project.

It was expected that ERUMS, as a pilot study, would provide valuable experience with both the technical aspects of matching data from the three systems and the administrative requirements for gaining access to the data and carrying out the matching operations. In short, ERUMS was planned as a learning experience, and that is exactly how it turned out. Members of the Workgroup, in addition to getting hands-on experience in interagency matching of employer and establishment records, gained new insights into the strengths and weaknesses of their own agencies, record systems.

D. Organization of this Report

Chapter II of our report describes the study design and execution. Section A provides a detailed description, for each of the three agencies, of the systems and files used in the ERUMS project. Because resources were limited, matching could only be

done for a sample of units in one State. Section B describes the sample design. The study design involved a relatively complex sequence of sample selection and matching operations; these are described in detail in Section C. Section D describes the administrative arrangements that were developed to gain access to identifiable records needed for ERUMS, to comply with the

agencies' requirements for maintaining confidentiality of the records, and to carry out the various phases of the study.

Chapter III presents the statistical results of ERUMS and an evaluation of the design that was used and its execution. Findings and recommendations are presented in Chapter IV. Section A presents the Workgroup's interpretation of statistical and other

results from the study, and Section B presents recommendations based on these findings. A list of references follows the text of the report. Detailed tables are included in Appendix A.

CHAPTER II - STUDY DESIGN AND EXECUTION

This chapter provides a detailed account of the design of ERUMS and how the study was carried out. The chapter has four sections. Section A describes the sources of the data for employers and reporting units that were matched. The data came from three agencies: the Bureau of Labor Statistics (BLS), the Social Security Administration (SSA) and the Internal Revenue

Service (IRS). A subsection for each of these agencies provides a broad description of the programs requiring the administrative record systems used in the study, followed by a description of the specific data files that were used for the ERUMS project. The subsection on SSA records also discusses the relationship between the SSA and IRS records used in the administration of the Old-Age, Survivors and Disability Insurance programs.

Because of the limited resources available for ERUMS, the matching had to be done for a sample of employers, identified by their Employer Identification Numbers (EINs). Section B describes the design of the sample. Section C provides a detailed account of the sample selection and matching procedures. Section D explains the administrative arrangements for the ERUMS project. Subsection 1 describes the formal interagency agreements that were developed to permit the necessary exchanges of identifiable records between agencies, subject to their confidentiality requirements. Subsection 2 describes the working arrangements for the project: meetings of the ERUMS workgroup and the development and maintenance of a project timetable.

For a good understanding of the results presented in Chapter III, it is recommended that all readers look at Sections A and B of this chapter. Those not interested in the detailed procedures and working arrangements may then wish to proceed directly to Chapter III.

A. Description of systems and files

1. Bureau of Labor Statistics (The Unemployment Insurance System and Address File)

The Unemployment Insurance (UI) program was created by the Social Security Act of 1935 to provide temporary income assistance to workers who become involuntarily unemployed. The UI system is a social insurance program that covers employees of commercial and industrial employers, most State and local government employees, and employees of specified nonprofit organizations. Employees of the Federal Government are covered by the Unemployment Compensation for Federal Employees (UCFE) program. The UI and UCFE programs

currently cover 97 percent of all wage and salary workers in the U.S.

The UI system covers, with certain exceptions, those employers with one employee on 1 day in each of 20 different

weeks in a calendar year, or who paid $1,500 or more in wages in one quarter in the current or previous calendar quarter. Those workers not covered by UI fall into a number of different categories. Agricultural workers are covered only if the employer has employed at least 10 workers in 20 weeks of the past or present calendar year, or has paid cash remuneration of $20,000 or more in any calendar quarter in the past or present year. Domestic workers employed in private homes, college clubs, or fraternities are

covered only if their employer pays more than $1,000 in cash in any quarter for such services. Patients, student nurses, and interns employed by a hospital are excluded from coverage. Also excluded are self-employed persons; insurance agents working on commission; and students and spouses of students working for the school, college, or university where the student is enrolled. An officer of a corporation is considered an employee of the corporation and, therefore, is eligible for unemployment benefits unless the officer is unemployed due to the sale of the corporation and the officer was directly involved in the sale. The same holds true for members of partnerships and proprietors: they are covered unless they are unemployed due to the sale of their business and they were directly involved in the sale. A small number of State and local government employees are not covered, including elected officials, legislators, members of the judiciary, persons in policymaking and advisory positions, temporary emergency employees, and members of the State National Guard and Air National Guard. The extent of coverage discussed in this paragraph pertains to Texas in 1982, and most States have similar, although not identical, provisions for coverage.

The UI program is authorized by both Federal and State laws. The U.S. Department of Labor (DOL) oversees the State UI programs and carries out the Federal obligation of financing the administration of the programs. While DOL insures that each State's program complies with the minimum standards set by Federal law, each State is entitled to develop a program suited to its own conditions. Each of the 50 States, as well as the District of Columbia, Puerto Rico, and the Virgin Islands, has enacted laws to determine its own tax structure, eligibility requirements, benefit levels, and coverage provisions. The administration of the UI program is the responsibility of the State Employment Security Agency (SESA) in each State.

The UI system is financed primarily through taxes assessed by both Federal and State governments on employers for wages paid to their employees. The provisions for the financing were established by the Federal Unemployment Tax Act (FUTA), Chapter 23 of the Internal Revenue Code. Currently, the gross FUTA tax is 6.2 percent of the first $7,000 per year paid to each employee ($434 maximum). (In 1982, the Federal taxable wage base was $6,000; it

was increased to $7,000 in 1983.) States levy employer UI taxes at rates determined by State law. If the State tax rate is at least 6.2 percent, employers receive a 5.4 percentage point credit against the FUTA tax, resulting in a net

Federal tax of 0.8 percent.

The Unemployment Insurance Address File is one of the statistical files produced under the Bureau of Labor Statistics (BLS) Federal/State ES-202 Program by the SESAs. The ES-202 Report (Quarterly Report on Employment, Wages, and Contributions) measures the extent of coverage under the various State Unemployment Insurance Programs. Its original use was to determine whether a State's program was in compliance with Federal law. The ES-202

Report represents the largest and most complete universe of monthly employment and quarterly wage information by industry, county, and State regularly available in this country. BLS funds and administers the ES-202 Program and provides conceptual, technical, and procedural guidance for all program activities.

The Unemployment Insurance Address File is a micro-level employer file prepared annually by each SESA. It contains first quarter information for each reporting unit subject to Unemployment Insurance reporting requirements in the State. A reporting unit is the most detailed economic unit for which data are submitted by the employer to the SESA. An establishment is an economic unit, generally at a single location, which is engaged primarily in one activity. In the case of a single establishment employer, the reporting unit and the establishment are identical. For many of the multi-unit employers, two or more establishments may comprise a single reporting unit. This cat occur when the establishments are engaged in similar activities (i.e., are in the same industry) and are located in the same county, or when the employment in the secondary industries and/or counties is not significant (i.e., less

than 50).

For any given quarter, typically about 10 percent of the reporting units show zero employment for all 3 months. Some of these zero employment figures are estimated (as discussed later in this report), although the great majority come from actual employer reports. (Some employers maintain an account even if no business is conducted during the quarter.) Data from some new businesses which came into existence during the first quarter may not be included in the UI Address file. This can occur if there is a substantial time lag between when the business started and when the employer submitted the completed status determination form (required from all newly established businesses) to the SESA (Grzesiak and Lent, 1988; Montana Department of Labor and Industry, 1987).

The 1982 Texas UI File examined by ERUMS contained a total of 270,612 unique accounts and a total of 303,582 individual records (reporting units). Of the 303,582 records, 4,020 had a blank or zero-filled Federal Employer Identification Number (EIN) and were ignored for the purposes of this study. The accounts examined included 267,487 single unit accounts (equal to 267,487 records)

and 3,125 multi-unit accounts comprised of 32,075 records.

The standardized UI Address File includes the following information for each reporting unit: name and address, State UI Account number, EIN, Standard Industrial Classification (SIC) code, Federal Information Processing Standards (FIPS) county code (township code for the New England States), ownership code, monthly employment levels for the payroll period including the 12th day of the month, and total quarterly wages.

Employer identifying information that enters the UI tax system, and eventually the UI Address File, is originally obtained from the initial status determination form. This form is used to

collect information concerning the business name, location, ownership, anticipated number of employees, and primary product or activity. On the basis of this information, the employer is assigned an account number and the various codes by the SESA.

Each reporting unit in the UI File is assigned a four-digit industry code from the SIC Manual on the basis of its primary activity. The primary activity is determined by the primary good produced or distributed or the primary service provided. SIC code 9999 is assigned as a temporary holding code when there is insufficient information on the State's initial status determination form for assigning a specific-industry code. Those reporting units assigned SIC code 9999 are requested to complete and return an SIC Refiling Form, with more detailed information, on a flow basis but no later than the next Annual Refiling Survey. There are a few exceptions to the 4-digit SIC coding requirement. Currently, States have the option to code employers in seven different 3-digit industry groups (representing 25 industries) to only the 3-digit level. These exceptions were created because adequate employer records may not be available to code to the 4-

digit level of detail or because reporting units in these industry groups frequently switch back and forth between 4-digit industries. These exceptions are as follows: SIC 074 (Veterinary services), SIC 078 (Landscape and horticulture services), SIC 152 (Residential building construction), SIC 154 (Nonresidential building construction), SIC 581 (Eating and drinking places), SIC 651 (Real estate operators and lessors), and SIC 721 (Laundry, cleaning, and garment services). SIC 421 (Trucking, local and long distance) and SIC 513 (Apparel, piece goods, and notions), comprised of a total of eight industries, were also coding exceptions in 1982.

In addition to an SIC code, the reporting unit is also assigned an ownership code according to legal proprietorship denoting Federal, State, Local, or International government,, or the private sector. A FIPS county code is assigned based upon the reporting unit's location or place of business. Besides the valid FIPS codes, there are additional codes which may be used: 996, 997, 998, and 999. County code 996 indicates a reporting unit located outside the U.S., Virgin Islands, and Puerto Rico but which reports to a SESA. County code 997 is assigned to reporting units with

locations in more than one

county but not Statewide. Reporting units located in a State other than the State to which they report are assigned county code 998. Finally, those reporting units with Statewide locations or unidentified locations are assigned county code 999.

To maintain accuracy of data on an ongoing basis, reporting units are asked to complete an SIC Refiling Form every 3 years to verify or update much of the identifying information (e.g., SIC, county, ownership) first collected on the initial status determination form or updated in the last Annual Refiling Survey. One-third of the universe of employers is surveyed in each of the 3 years of the Annual Refiling Survey.

Employers subject to State Unemployment Insurance laws are required to complete quarterly contribution reports and submit them to the appropriate SESA. The information from the quarterly contribution report submitted by the employer for the first quarter is used in the preparation of the UI Address File. The contribution report provides current information on the name, address, and UI account number of an employer; monthly employment levels; total wages paid; taxable wages; and contributions (taxes). Multi-establishment employers are also asked (required in 20 States, but not in Texas) to complete a statistical supplement questionnaire for each quarter furnishing similar information for each of their reporting units. The SESA uses the data supplied on the contribution reports and statistical supplements to create the UI Address File.

The SESAs are responsible for editing and estimating data items missing from employer accounts. These data are missing because the employer either fa iled to complete all of the entries on the contribution report or statistical supplement or failed to submit a contribution report or statistical supplement altogether.

Data missing from incomplete contribution reports and data for accounts delinquent 12 weeks after the end of the quarter are estimated. Estimates are generated for all delinquent accounts (including multi-establishments), unless the account is delinquent for two or more consecutive quarters. These delinquent accounts are contacted to determine if they are still active. Only if they are confirmed to be active are estimates prepared. Estimates are replaced on the State file when the actual data have been received and edited, but once estimated data items have been transmitted to BLS, they are not replaced with actual data.

Thus, the SESAs are responsible for editing and extracting data from their UI Tax file, collecting supplemental data, and maintaining the accuracy of the SIC and other codes for the UI Address file and ES-202 Report. After BLS reviews and edits the UI Address file transmitted by the State, that edited file is used to update the BLS Universe File. The Universe File is then used as a national sampling frame for BLS establishment surveys, including the Industry and Area Wage Surveys, Occupational Safety and Health Statistics, and Producer Price Index programs.

BLS is currently in a transitional period with respect to the UI Address File. For data through 1988, the SESAs were required to provide the UI Address File to BLS for only the first quarter of the year. Beginning with data for the first quarter of 1989, however, all States will be required to submit the file on a quarterly basis (6 months and 5 days following the end of the reference quarter). In addition, the UI Address File format will be expanded to contain supplementary information, including predecessor and successor UI Account and Reporting Unit numbers, expanded ZIP codes, address type indicators (e.g., physical location or corporate headquarters), multi-unit indicators, and telephone numbers.

Coinciding with the above improvement is the initiation of the new BLS Business Establishment List (BEL) Improvement Project (MacDonald, 1989). The fundamental goal of the BEL project is the collection of establishment level data, including physical location addresses or both single and multi-unit employers. These more detailed data will also be included in the UI Address File.

2. Social Security Administration

The Social Security Act of 1935 established a requirement that the Social Security Administration (SSA) perform the recordkeeping necessary to reflect accurately the earnings of workers in employment covered by the Act. As amended in 1939, the Act required detailed information on the continuity of employment by calendar quarter and covered wage amounts. The accumulation of quarters of coverage and quarterly wage amounts are used as the basis for determining eligibility for and amounts of program benefits. The law originally required all workers in industry and commerce, except railroad workers, to be covered. This coverage has been broadened over the years and self-employment has been

added. Now the only large segments of uncovered jobs are Federal civilian employees who have chosen to remain covered under the U.S. Civil Service Retirement system, and employees of State and local governments who are not covered by a Federal-State agreement. The program currently covers over 95 percent of wage and salary jobs and the self-employed.

The Old-Age, Survivors and Disability Insurance (OASDI) programs administered by SSA provide monthly benefits to retired and disabled workers and their dependents and to survivors of insured workers. Benefit payments are financed principally through taxes collected from employers, employees and the self-employed. Taxes ate paid based on earnings up to an indexed statutory taxable maximum which began at $3,000 in 1937 and is $51,300 in 1990. The method chosen for collection of the taxes is through employer reporting which was required quarterly in the beginning of the program and annually beginning in 1978. In 1978, employer reporting of Social Security covered wages was combined with the existing W-2 (Wage and Tax Statement) income tax reporting that employers are required to complete for the

Internal Revenue Service (IRS). Details of the reporting process are discussed below.

In 1937, SSA began a process to enumerate workers and employers to facilitate its record-keeping process. Workers received a Social Security Number (SSN) and employers received a nine-digit identification number to be used in the reporting process. The worker identification information and subsequent wage reports became part of SSA's Summary Earnings Record. The employer information collected at the time of issuance of the identification number was made part of the Employer Registration File. In 1958, the IRS was given responsibility for issuing Employer Identification Numbers (EINS) and constructed a file called the

Business Master File (BMF) that is currently used by SSA to identify employers. The employer information collected from the beginning of this enumeration process included geographic location and industrial activity. These particular items of information were not a direct part of SSA earnings processing, but were collected to help study the new emerging Social Security program. The additional information on employers evolved into a set of files used by SSA's Office of Research and Statistics (ORS) for special studies. These are the Single Unit and Multiunit Code Files that are discussed below along with the employer wage-reporting system that provided the source of employer information used in the ERUMS project.

Prior to January 1978, employers filed their tax and wage reports with the IRS on a quarterly basis, using Forms 941 (regular) and 942 (household work), and annually using Form 943 (agricultural work). Attached to these forms were Schedules A showing the detailed amounts of wages for each employee by SSN. These Schedules A were used by SSA to post wages each quarter to the workers' earnings records. Public Law 94-202 (Combined Old Age

Survivors and Disability Insurance Income Tax Reporting Amendments of 1975) enacted January 2, 1976, provided for annual, rather than quarterly, wage reporting. These amendments were effective for tax years beginning

- 1978 for United States domestic employers (other than State and local governments),

- 1979 for employers in Guam, American Samoa, Virgin Islands, and Puerto Rico (other than State and local governments), and

- 1981 for State and local government employers .

Under the Combined Annual Wage Reporting process, employers continue to file Forms 941 and 942 quarterly and Form 943 annually with IRS, but no Schedule A is required. Instead, Forms W-2 are filed by the employer as the annual wage report for the employees. These reports, in the form of Copy A of the Form W-2, along with a copy of the employer transmittal and Form W-3 are filed with SSA annually on or before the last day of February in

the year following the wage reporting year. Employers filing via magnetic media submit W-2 and W-3 data on electronic records plus transmittal Form 6559. In processing the Forms W-2/W-3, SSA performs the following functions: data entry, balancing the sum of the money fields on Forms W-2 to totals on the Form W-3, microfilming, posting the Forms W-2 data to the master earnings records of individuals and transmitting the Social Security and income tax data to the IRS. In addition, SSA creates a W-3 tape file for purposes of reconciling differences between wage information reported to IRS and SSA and locating annual wage reports on the microfilm.

To insure that SSA has received and accurately recorded all FICA wages (wages as defined by the Federal Insurance Contributions

Act), SSA's W-3 file is compared with IRS's 941 records annually, in a process known as reconciliation. This is an electronic comparison of SSA-processed employer FICA wage totals with the amount of FICA wages on which employers have paid taxes to the IRS. From this comparison, cases are identified in which IRS has a record of receiving taxes, but SSA has no record of having processed an annual wage report (W-2/W-3s) or SSA's processed wage totals for the employer are less than IRS's. Some other reasons for cases to be in reconciliation are: 1) the employer sent IRS wage information using one EIN and the Forms W-2/W-3 that were sent to SSA were processed using different EINS; 2) the employer transposed or used an incorrect digit in the EIN; and 3) IRS and/or SSA miskeyed the EIN. SSA corresponds with the employers of these reconciliation cases in an attempt to resolve the discrepancies.

As a byproduct of the employer reporting system, SSA maintains files that are used in ORS statistical programs. The Single Unit Code File and Multi Unit Code File contain coded information on the employer's geographic location and industrial activity. These coding files are updated each year with data from a special version

of the Form W-3 file, which has been edited to exclude certain records which are not required in ORS statistical operations (e.g., non-FICA, household 'employers, delinquent reports). The primary purpose of the Single and Multi Unit Code Files is to provide geographic and industry data for records of workers in statistical files, e.g., the Continuous Work History Sample (CWHS) which is the source of data for a variety of statistical studies and analyses, making revenue estimates and in tables in publications of SSA program data and research reports.

The Single Unit Code File (SUCF) contains one record for each entity that has filed a Form SS-4, Application for an Employer Identification Number (EIN), with the exception of nonemploying entities (e.g., trust funds, fiduciaries and estates) and household employers. EINs are assigned by the IRS and the forms are forwarded to SSA where they are coded for geography, industry, class (i.e., individual, corporation, partnership, etc.), employer size and reason for application. The geographic

classification of the entity is based on the physical location of the business as provided by the employer on the Form SS-4, otherwise, the mailing address is used. When a location is not available, the entity is given a State code based on the Internal Revenue District (IRD), the first two digits of the EIN,, in which the number was issued and a statewide county code. The SSA has its own industry classification system based on the Standard Industrial Classification (SIC). In 1982, full four-digit SIC codes were used for most industries. There were exceptions for major groups 01 (agricultural production--crops) and 02 (agricultural production--livestock) and division J, public administration. For each of these three categories, SSA used only a single code. In addition, for 63 four-digit industries in other categories, "foldback codes" for groups of four-digit industries were used when there was insufficient information to assign a specific four-digit code.

The SUCF is an historical file that includes both active (employers reporting annual wage reports in the current tax year) and inactive units (those employers no longer reporting annual wages, e.g., out of business). The file for the year ending December 1987 contained 21,325,091 EINS. It is updated annually with data from the coded Forms SS-4.

The Multi Unit Code File contains one record for each reporting unit of multi unit employers who are participating in a voluntary program, the Establishment Reporting Plan (ERP), conducted by the SSA. Excluded from the file are seasonal agricultural employers and Federal, State and local government employers. Employers are identified for participation in the ERP when the Form SS-4 indicates that the employer has more than one place of business and 100 or more employees or an annual wage report is received for 100 or more employees. Eligible employers are requested to participate in the ERP by providing SSA with a Form SSA-5019 (List of Establishments or Reporting Units) on which the employer lists his establishments and assigns a four digit unit number to each one. In addition, the employer must group his

employees under these same unit numbers on his annual wage report. Forms SSA-5019 are coded for industry, geographic location, auxiliary units, non-profit coverage and employer size. Each unit is geographically classified based on either the physical location of a reporting unit or the countywide, Statewide or nationwide location of a payroll grouping. The industry classification used for the ERP coding of multiunit employers is also based on the Standard Industrial Classification. The Multi Unit Code File is an historical file which contained 33,957 EINs and 116,613 reporting units for the year ending December 1987. This file is updated on an annual basis with information from the coded Form SSA-5019.

For the ERUMS project, SSA provided records from the Single Unit and Multi Unit Code Files and the 1982 Form W-3 file. A detailed description of how these files were used in the project is included in Section C of this chapter.

3. Internal Revenue Service

Requirements to file the Form 940 for 1982

The Federal Unemployment Tax Act (FUTA) established a Federal-State unemployment compensation system financed by separate Federal and State payroll taxes on Employers. Administrative funds are derived from the Federal payroll tax and benefits are paid mainly from State payroll taxes.

The Form 940 is the Employer's Annual Federal Unemployment Tax (FUTA) Return. A copy of the 1982 Form is shown as Exhibit IIA-2. This is the form on which the employer reports the State, or States, where contributions are required to be made and the wage information necessary to compute the FUTA tax and the credit reduction for payments made to a State or States. In general, the form must be filed by every employer who either paid wages of

$1,500 in any calendar quarter, or who had one or more employees for some part of a day in 20 different weeks.

Agricultural employers must file if they paid cash wages of $20,000 or more to farm workers during any calendar quarter, or employed 10 or more farmworkers during some part of the day for at least one day during any 20 different weeks.

Households which paid wages of $1,000 or more in any calendar quarter for household work in a private home were also required to file. For this purpose, household work in local college clubs and in the local chapters of college fraternities or sororities is included.

For purposes of counting its employees, a partnership does not count its partners.

Employers are authorized to claim a credit for contributions to a certified State unemployment fund by the due date for filing the Form 940. For this purpose, State was defined to include Puerto Rico and the Virgin Islands. "Contributions" are payments

that State law requires an employer to make to an unemployment fund. The credit can be claimed for these "contributions" only to the extent that they are not deducted or deductible from the employees' pay.

The forms are filed with the IRS at a service center determined by the location of the employer's principal business office or agency. Penalties are assessed for late filing or late deposit unless reasonable cause for the delay can be shown. There are also penalties for failure to file, failure to pay the tax or filing fraudulent returns.

For FUTA purposes "wages" and "employment" do not include every payment and every kind of service an employee may perform. In general, payments excluded from wages and payments for services excepted from employment are not subject to tax.

Examples include benefit payments for sickness or injury under a worker's compensation law, insurance plan and certain employer plans, certain family employment, certain fishing activities and noncash payments for farm work or work in a private home and meals and lodging.

For 1982, only the first $6,000 in wages paid to an employee was used for the FUTA calculation. The Federal FUTA tax rate on this part of wages was 3.4 percent. Amounts in excess of the wage base were exempt from the FUTA calculation, but not necessarily from the State unemployment tax calculation. If a State's unemployment compensation program met the requirements of Federal law, employers in the State received a 2.7 percent credit against the 3.4 percent Federal FUTA tax for 1982. (For information on the current wage base and tax rates, see Section II,A,l.)

The Employer's Quarterly Federal Tax Return (Form 941) File

In order to facilitate the collection of social security and federal income taxes, employers are required to withhold some portion of each employee's wages, and to deposit that portion in a timely fashion to the credit of the Treasury. At the end of each calendar quarter, nonagricultural employers (excepting those who have only household employees) are required to file an Employer's Quarterly Federal Tax Return, Form 941 (Form 941E for employers who report only withheld income tax, such as certain State and local governments) with the IRS. The information on this form includes a record of their federal tax liability throughout the quarter, along with a summary of their employees, wages, tips, and other compensation which was subject to withholding, the amount of taxes withheld, a summary of wages subject to Federal Insurance Contributions Act (Social Security) taxes, and the Social Security tax paid. Once a year, each employer is required to report the number of persons he employed during the week of March 12. A copy of the Form 941 for the first quarter of 1982 is shown as Exhibit IIA-3.

The Tax Years 1981-83 Form 941 File

Each year the IRS prepares an extract of its Forms 941 and 943 data for the Census Bureau. This extract contains Employer Identification Number, payroll, employment, industry, and legal form of organization information. The Census Bureau edits the payroll, employment, and industry data and makes any needed amputations. For Tax Years 1981-83, the IRS and Census agreed that Census would return the edited extracts to the Statistics of Income Division (SOI) of IRS. These edited extract files were the ones used for ERUMS. Definitions of the items in the files were as follows:

Employment- For purposes of income tax withholding, a common-law employee is defined as follows:

Under common-law rules, every individual who performs services that are subject to the will and control of an employer, as to both what must be done and how it must be done, is an employee. It does not matter that the employer allows the employee considerable discretion and freedom of action, so long as the employer has the legal right to control both the method and,the result of the services.

Two of the usual characteristics of an employer-employee relationship are that the employer has the right to discharge the employee and the employer supplies the employee with tools and a place to work.

If an individual's relationship with an employer fits this description, then the employer is required to withhold federal income tax and Social Security tax from the employee's pay, and to report such withholding on Form 941. Employees who fall into the

following categories are defined as statutory employees:

1) A driver who distributes meat, vegetable, fruit, or bakery products or beverages (other than milk) or picks up and delivers laundry or dry cleaning, if the driver is the employer's agent or is paid on commission.

2) A full-time life insurance sales agent whose principal business activity is selling life insurance or annuity contracts, or both, primarily for one life insurance company.

3) An individual who works at home on materials or goods which an employer supplies and which must be returned to the employer or a person the employer names, if the employer also furnishes specifications for the work to be done.

4) A full-time traveling or city salesperson who works on the employer's behalf and turns in orders to the employer from wholesalers, retailers, contractors, or operators of hotels, restaurants, or other similar establishments. The goods sold must

be merchandise for resale or supplies for use in the buyer's business operation. The work performed for the employer must be the salesperson's principal business activity if: a) The service contract states or implies that almost all of the services are to be performed personally by the contractor; b) The investment in the facilities (other than in facilities for transportation) used to perform the services is not substantially the individuals; and c) The services are performed on a continuing basis.

Employers are required to withhold Social Security tax, but not federal income tax, from the wages of statutory employees. Individuals who are either common-law or statutory employees are to be reported as employees.

There is anecdotal evidence from exact match studies and from IRS audits that some firms, particularly in the oil and gas extraction industry, were not complying with these reporting

rules in Tax Year 1982. These firms attempted (illegally) to treat all of their employees as independent contractors for tax purposes; therefore, no taxes were withheld, and no Forms 941 filed by these firms. No estimate of the number of such nonfilers is available, but the problem is believed to be of little significance in other industries.

Payroll- The payroll field on the extract comes from line 2 of Form 941. The instructions for this line read as follows:

> Enter the total of: all wages paid, tips reported, taxable fringe benefits provided, and other compensation paid to your employees, even if you do not have to withhold income or Social Security taxes on it. Do not include pensions, annuities, third-party sick pay, supplemental unemployment compensation benefits, or gambling winnings, even if you withheld income tax on them.

Legal Form of Organization- The IRS maintains, as part of its computerized Master File system, a record for each business which files a Form 941. This same record also contains information on the other tax returns which the business files, if the returns are posted to the Business Master File (BMF). (Note that sole proprietors report their income on Schedule C attached to their Form 1040, which posts to the Individual Master File. Thus, while a sole proprietorship with employees is represented in the BMF as a Form 941 filer, it was not possible to positively identify it from the BMF as a sole proprietorship in Tax Year 1982.) A portion of this record contains entity information,'for example, the name of the business, its address, its industry, and a set of codes indicating the type(s) of forms it is required to file. These filing requirement codes are a part of the Form 941 extract, and allow the identification of the legal form of organization of a business. A nonzero filing requirement code indicates that a business must file a form in the indicated series. Filing requirement codes,exist on the extract for Form 1120 (Corporation), Form 1065 (Partnership), and Form 990 (Nonprofit organization). As

explained earlier, Sole Proprietorships are not directly identifiable from these codes, but few other types of entities may operate a business.

Industry- Each extract record sent to the Census Bureau contains an industry code assigned during IRS revenue processing. The IRS industry codes, while based on the Standard Industrial Classification (SIC), are considerably less detailed than those used by BLS and SSA. Four-digit codes are used; however, most of them represent groupings of several SIC four-digit industries. The particular groupings used differ by type of organization: corporation, partnership and sole proprietorship. In 1982, roughly 200 categories were coded separately for each of the three types of organization. As a part of its data editing process, Census assigns industry codes from the following sources in order of preference: 1) the most recent economic census, 2) the Census Bureau's Current Business Surveys, Annual Survey of

Manufacturers, Company Organization Survey and County Business Patterns Program, 3) the Social Security Administration birth code based on the EIN application, Form SS-4, or 4) the original IRS industry code. Sources 1) and 2) are used only for single-establishment EINS. If only the original IRS code is available, Census uses a conversion program to convert it to a standard SIC format. In some such cases, SIC codes can only be assigned at the 2- or 3-digit level of detail. The codes used for ERUMS were the codes assigned by the Census Bureau. These codes were provided to IRS under the authority of the 1953 Opinion by Attorney General James P. McGranery, 41 Op. A.G.120. Under this Opinion, the Census Bureau can check industry classifications assigned by another agency against its own and either certify or correct the other agency's classifications.

Improvements Subsequent to 1982

The greatest improvement in the Form 941 information is coming from changes in the data collection method. Census Bureau representatives report that the number of changes made during edit and imputation have fallen dramatically as the IRS has implemented scanning of paper documents and filing on magnetic media as an alternative to keying data from paper documents. Also, problems with firms attempting to treat employees as independent contractors (which caused employee data to be underestimated) have been greatly reduced through effective enforcement efforts.

B. Sample design

1. Design considerations

The criteria that governed the choice of a sample design for ERUMS were:

- The study should be limited to one State.
- Within the selected state, probability sampling procedures should be used.
- The sample size should take into account the resources available to the ERUMS Work Group for computer and manual matching and other processing activities.
- All units in the selected state that were active during the study reference period in either the BLS or SSA reporting systems should have a chance of selection.
- Cases of greater interest, for example, those found in only one of the two systems (unmatched cases) and those involving more than one reporting unit (multi units) should be oversampled.

ERUMS was a pilot study, designed to develop and test procedures for linking and comparing employer and reporting unit data from different administrative record systems. The agencies participating in the study could provide only limited staff time and other resources. These considerations dictated the Workgroup's

decision to limit the study to one State and to a fairly small sample in that State.

Within the selected State, Texas, the use of probability sampling at all stages of selection provided two benefits. it ensured that sample results could be used to produce unbiased estimates for the study population and it made possible estimation of sampling errors from the sample. Although we recognized that sampling errors would be relatively large for most estimates, we felt it would be useful, for both analytical and methodological purposes, to produce weighted estimates.

One possible approach to the study design would have been to select a baseline sample from a single agency system, say the BLS UI system, and search for the sample units from that system in the SSA and IRS systems. However, that approach would have failed to provide any information about units that were in the SSA and IRS systems, but not in the BLS system. it proved to be feasible to use a design that sampled both the BLS and SSA systems, so that units existing in either one of these systems but not in the other would be represented. The Workgroup decided that it was not feasible to

sample the IRS system independently, given the complexity of the system and the administrative

difficulties in gaining access to it for such a purpose. Therefore, the final sample does not represent any units that may have been included in the IRS system but not in the BLS and SSA systems. Units in the final combined BLS/SSA sample were matched against the IRS files described in Section A of this chapter, so that we do have IRS data for the BLS and SSA sample cases that were found in the IRS files.

The requirement that all in-scope units in the BLS and SSA systems should have a chance of selection was not completely fulfilled. Because the Employer Identification Number (EIN) was to be the primary basis for matching records in all three systems, the group of reporting units covered by a single EIN was chosen as the sampling unit for both the BLS and SSA systems. However, in the

1982 Texas UI Name and Address File, 4,020 reporting unit records (1.3 percent) out of a total of 303,582 did not have EINS. These units were not included in the initial sample selection from the BLS UI file.

Oversampling of unmatched and multi unit cases was dictated by the exploratory nature of ERUMS. If proportional sampling had been used, about 70 percent of the sample cases (as it turned out) would have been matched single units, for which the processing was expected to be straightforward. The unmatched and multi unit cases were expected to present more difficulties and the Work Group wanted to have enough of these cases to learn what the situations were and to test methods of dealing with them.

2. The sample design adopted

The sample design and the matching procedures were closely interrelated. A summary of the sample design is presented here; details of the sample selection and matching procedures are given in Section C below.

The main steps in sample selection and matching were:

(1) Select samples of EINs from the BLS and SSA frames.

(2) Match each EIN in both agency's samples against the other agency's frame to determine whether it was included in that frame, i.e., whether it was a matched sample unit.

(3) From the combined samples after steps (1) and (2), select a subsample of EINS, with subsampling rates that varied, depending on initial match status and classification as a single unit or multi unit.

(4) Match the subsample units against selected IRS files and, for those located in the IRS files, add relevant IRS data to the data base for the subsample.

A key feature of the sample design was the use of a digital

sampling procedure, based on EINS, in step (1). The EIN is a unique nine-digit number assigned to each employer. Sampling based on the final (9th) digit is not recommended because the nature of the issuance process has resulted in an excess of EINs ending in 0 and 5 (Harte, 1986). For this reason, we selected, from both the BLS and SSA frames, all EINs that had one of six randomly selected pairs of digits in the 7th and 8th position. Using the same sets of digits for both the BLS and SSA samples made it possible to complete step (2) by matching the two samples against each other, rather than by matching each sample against the other agency's complete frame.

The Workgroup decided that the final sample size should be about 400 matched and unmatched EINs and that about one-half of these should be EiNs classified as multi unit in one or both systems. EIN counts obtained for the Texas UI File prior to the initial sample selection were:

Single unit	267,487
Multi unit	3,125

Total EINs 270,612

A sampling rate of 6 in 100 would produce an expected sample of about 188 multi unit EINs from the BLS frame: this was the rationale for using 6 out of 100 possible pairs of ending digits.

The initial sample selected by this method from the BLS and SSA frames contained a total of 19,964 EINS, of which 16,336 were selected initially from the BLS Texas UI file for 1982 and the remaining 3,628 were EINs from SSA's Single or Multi Unit Code Files that had all of the following characteristics:

- Wages reported for 1982.

- One or more reporting units in Texas shown on SSA's Single unit or multi Unit Code File.

- Not included in the BLS Texas UI file for 1982. (However, the employer could have been in the UI file without an EIN.)

All cases in the initial sample were then classified by match status and whether they were identified as single or multi unit EINs in the BLS and SSA files. On the basis of these classifications, 9 major strata were formed. Two of the strata that involved BLS multi unit EINs were subdivided, putting employers with 20 or more reporting units in a separate stratum in each case. Using varying sampling fractions, subsamples were selected from each of the 11 strata to produce a final sample of 200 EINs involving only single units and 201 EINs initially classified as multi unit by BLS, SSA or both.

The initial match and the BLS and SSA single/multi unit

classifications were used to form the strata from which the subsamples were selected. These classifications were later modified for analytical purposes, as will be explained in Section

C. However, the weights applied to the sample cases to produce estimates depend on which of the strata they were selected from. Weighting by the reciprocal of the subsampling fractions produces estimates at the level of the first-stage sample. These estimates can be used to calculate percent distributions, because EINs in the first-stage sample were selected with equal probability. To produce estimates of totals for the universe, the first-stage estimates have to be further weighted by the reciprocal of 0.06, the sampling fraction used to select the first-stage sample.

After the selecti on of the second-stage sample, it was discovered that an additional 2,608 EINs should have been included in the first stage sample from the SSA frame, but were inadvertently omitted. This problem w as dealt with by reweighting the second stage sample cases for the strata that were affected. Further details are given in Section C of this chapter.

Sampling errors were calculated for a few key estimates and are shown in Tables IIIA-4 and A1. For the latter table, in which the estimates were based on the full first-stage sample, the actual

sample of 22,572 EINs was treated as a fixed size simple random sample, selected without replacement, and the sampling errors were estimated under that assumption. The estimates in Table IIIA-4 were based on the second-stage sample. The calculation of sampling errors for these estimates treated the first-stage sample of 22,572 cases as the universe and the second-stage sample as though it had been a stratified random sample selected without replacement from that universe. These assumptions result in a slight understatement of the sampling errors, since they do not take into account the contribution of the first stage of sampling to the overall sampling errors.

Exhibit IIB-1 summarizes the main features of the ERUMS sample design. A more detailed description of the sample selection and matching procedures is given in Section C. Section D describes the administrative and working arrangements for carrying out the study. Readers who are mainly interested in the results may wish to proceed directly to Chapter III.

Exhibit IIB-1

Summary of the ERUMS sample design

FRAMES

BLS: EINs in Texas UI file for first quarter 1982

SSA: EINs in Single Unit and Multi Unit Code Files that:

(1) Had wage reports for 1982 and

(2) Had at least one Texas reporting unit and

(3) Did not appear in the BLS frame.

FIRST-STAGE SAMPLE

Selection method: Equal probability, based on 7th and 8th digits of EIN

Sampling fraction: 6 in 100

Sample size:	BLS frame	16,336
	SSA frame	3,628*
	Total	19,964

SECOND-STAGE SAMPLE

Selection method:	Stratified systematic, equal probability within stratum	
Sampling fractions:	Varied by stratum from take all to 1 in 173.78	
Sample size:	Multi unit in BLS or SSA	201
	All other	200
	Total	401

* Plus 2,608 cases that were inadvertently omitted. See discussion in Sections B and C of this chapter.

C. Sample selection and matching procedures

There are two reasons for providing a detailed account of the ERUMS sample selection and matching procedures. The obvious reason is that the results, like those of any research study, are dependent on the procedures used and anyone interested in the results is entitled to a full description of how the study was carried out. The other reason, equally or perhaps more important, is that ERUMS was a venture into uncharted territory and we believe that future projects of this kind will benefit from the availability of a detailed road map of the procedures that were developed to match and compare employer and reporting unit records from BLS, SSA and IRS for statistical purposes.

Exhibit IIC-1 gives an overview of the ERUMS sample selection and matching operations that will be discussed in this section. The subsection numbers used in this section correspond to the operation numbers on the chart (1.0 to 10.0). Most of the 10

operations are relatively simple and therefore easy to describe; however operation 3.0, covering Phase I sample selection operations at SSA, was complex and required a separate chart (Exhibit IIC-2) for clarification.

An important consideration in developing the procedures was the large size of the administrative record files from which the samples were selected and relevant data for the sample cases extracted. This dictated a strategy of minimizing the number of runs of these large files and extracting only the sample units and data needed for the study so that working files would be of manageable size and could be processed on a microcomputer accessible only to BLS personnel cleared to work on the ERUMS project. In operation 3.0, for example, single runs of SSA's Single and Multi Unit Code Files were made to extract all of the data needed for the Phase I sample selection at one time.

Certain of the procedures used were necessary to comply with policies of the participating agencies concerning access to identifiable records from their systems. In particular', it can be seen in Exhibit IIC-1 that in operation 2.0, BLS transmitted only

the stems (digits 1-6,9) of the sample EINs rather than the full 9-digit EINs to SSA. This was done because it was not considered appropriate to identify specific UI filers in an administrative record system operational environment. Later, when only SSA personnel cleared to participate in ERUMS had access to the working files for the study, full 9-digit EINs were included.

Once the study specifications had been agreed on and the interagency agreements approved, the project operations depicted in Exhibit IIC-1 occupied a period of about three years. The initial sample selection operations at BLS and SSA (steps 1.0 to 3.0) were completed during a relatively short period in mid-1986. The elimination of nonsample EINs and the electronic merge of SSA

and BLS data for the Phase I sample (steps 4.0 and 5.0) were completed at BLS in January 1987. The selection of the Phase II sample (step 6.0) was completed at BLS in October 1987. For the most part, the acquisition of additional BLS, SSA and IRS data for

the Phase II sample cases (steps 7.0 to 9.0) was-completed by April 1988. Final review and analysis continued until the end of 198'9.

1. Selection of BLS Phase I sample

The first step, once the overall design for the study had been agreed on by the Workgroup, was to select the Phase I sample from the BLS UI Address File for the State of Texas for the first quarter of 1982. This file, which had been transmitted from the State to BLS in October 1982, contained records for all covered Texas employers who had filed their ES-202 statistical reports for the first quarter of 1982, plus records for some employers who had not filed but for whom employment had been imputed based on reports for prior quarters. The file included a few employers who had filed reports but reported zero employment for the first quarter of 1982.

The sample selection, as reported in the previous section, was based on the EIN as the sampling unit. Therefore, the 1.3 percent

of records with no EINs reported were excluded from the sample selection.

All records having any one of six randomly selected pairs of 7th and 8th digits in their EINs were included in the sample. (To minimize disclosure risks, the specific pairs are not identified in this report.) If an EIN had only one reporting unit (RU) associated with it, it was classified as a BLS single unit EIN; if it had more than one associated RU, it was classified as a BLS multi unit EIN.

The Phase I BLS sample contained 16,336 EINS. The expected take was 0.06 x 270,612 = 16,237. For this sample.of 16,336 EINS, data items needed for ERUMS were extracted from the source file.

2. Listing of EIN stems for BLS Phase I sample

The "EIN stem" is defined as digits 1 to 6 and 9 of the full 9-digit EIN. BLS created and transmitted to SSA a file containing only the EIN stems of the 16,336 sample EINS. Some stems appeared

more than once in this file. A listing of unique stems subsequently created by SSA contained 11,655 records.

As explained earlier, the reason for the use of EIN stems at this stage was to avoid identification to SSA operating staff, not cleared to participate in the study, of employers reporting to the UI system.

3. SSA Phase I sample selection operations

Exhibit IIC-2 shows the details of operation 3.0, the steps carried out at SSA to extract SSA data for EINs in the BLS Phase I sample and for other EINs meeting the criteria for sample selection but not included in the BLS sample. In the exhibit, operations are represented by rectangles; input and output files are represented by parallelograms.

More specifically, the goal of operation 3.0 was to produce

two files and transmit them to BLS for further processing and Phase II sample selection. One output file consisted of full 9-digit EINs and data for stem matches, i.e., single and multi unit records from SSA's Single and Multi Unit Code Files which:

- Had the same stem (EIN digits 1-6,9) as at least one of the BLS Phase I sample EINs and;

- Were associated with employers who had filed W-3 Wage Reports for 1982 (active SSA employers).

This stem match file contained three types of records:

- Records for EINs corresponding to full 9-digit EINs in the BLS Phase I sample, i.e., matched cases.

- Records for EINs not corresponding to full 9-digit EINs in the BLS sample, but eligible for the study by reason of having one of the six designated pairs in digits 7 and 8, and having a Texas code. These records are referred

to as sample nonmatches.

- All other records, i.e., nonsample nonmatches. These were of no further interest for the study.

The second output file contained 9-digit EINs and data for sample nonmatches, i.e., records from the Single Unit and Multi Unit Code Files that did not match any of the BLS stems and:

- Had one of the six designated sample pairs of digits in positions 7 and 8 of the EIN;

- Had a Texas code; and

- Were associated with employers who filed W-2/W-3 Wage Reports for 1982.

All of the records in this file were designated as sample nonmatches. Note that sample nonmatches could occur in either of the two output files. However as explained under step 4.0, the

sample nonmatches in the stem match file were not included in the Phase I sample.

To understand the SSA operations described in this subsection, it is necessary to make a distinction between employers and reporting units. Each record in SSA's Single Unit Code File has a unique EIN, representing a single employer. All employers who completed Form SS-4 and were issued EINs should be included once in this file, regardless of the number of reporting units they have.

The records in the Multi Unit Code File represent reporting units, so that the same EIN can be associated with more than one record -in that file. Employers with one or more records in the Multi Unit Code File have been identified at some stage as having more than one reporting unit, but they do not all currently participate in SSA's voluntary Establishment Reporting Plan program and report their wages separately by reporting unit. Therefore, it is possible to have EINs with only one record in the Multi Unit

Code File. All EINs appearing in the Multi Unit Code File should also appear in the Single Unit Code File, although there may be a few exceptions.

The steps in operation 3.0 were as follows:

Step 3.1 - The list of unduplicated BLS stems and the list of the six randomly selected sample pairs of digits were compared with each of the EINs in the Single Unit Code File to produce two extract files. The stem match extract file contained records for all EINs having one of the BLS sample stems. The sample nonmatch extract file contained records for all EINs with nonmatching stems that had a Texas state code and one of the sample pairs of digits in positions 7 and 8. The number of records in each of these extract files is shown in Exhibit IIC-2.

Step 3.2 - Essentially the same procedure was followed for the Multi Unit Code File. The stem match extract file contained all reporting unit records for every EIN having one of the BLS sample stems. The sample nonmatch extract file contained records that had

Texas state codes and were associated with EINs that had nonmatching stems and one of the sample pairs of digits in positions 7 and 8. Thus, for sample nonmatch EINs for employers with reporting units in more than one State, only their Texas reporting units were included in the extract file. The number of records in each of these files is shown in Exhibit IIC-2.

Step 3.3 - The stem match extract files from the Single and Multi Unit Code Files were compared on the basis of EIN. Records in the single unit extract file having EINs that also appeared in the multi unit extract file were eliminated.

Step 3.4 - The records remaining from step 3.3 were compared with an edited W-3 Wage Report File for 1982, on the basis of EIN. Records for with EINs having no 1982 wage reports in this file were eliminated. The output of this step was a file of 182,536 records that were potential matches to the BLS sample EINS.

Step 3.5 - The sample nonmatch extract files from the Single and Multi Unit Code Files were compared on the basis of EIN. Records in the single unit extract file having EINs that also appeared on one or more records in the multi unit extract file were eliminated. The number of records eliminated at this point was quite small, probably because many of the EINs appearing in the multi unit extract file had records in the Single Unit Code File with non-Texas state codes, hence these EINs had not been included in the sample nonmatch file that was extracted from the Single Unit Code File in step 3.1.

Step 3.6 - The records remaining from step 3.5 were compared with the edited W-3 Wage Report File for 1982, on the basis of EIN. Records associated with EINs having no 1982 wage reports in this file were eliminated. The output file of sample nonmatches contained a total of 3,658 records.

Following completion of these steps, the final output 'files of stem (potential) matches and sample nonmatches were transmitted to BLS. In addition to full 9-digit EINS, these files included SSA

geographic codes (State and county) and the first two digits of the SIC codes.

4. Elimination of non-sample EINs from SSA output files

All EINs in SSA's sample nonmatch output file were included in the final Phase I sample. However, as explained in subsection 3, some of the EINs in the stem match file did not meet the criteria for inclusion in the Phase I sample. BLS matched the full 9-digit EINs from its initial sample against the 9-digit EINs in the stem match file and retained in the Phase I sample only those EINs that matched. At that time, no one recognized that the stem match file could also include sample nonmatch cases. As a result, nonmatch cases that had stems appearing in BLS's initial sample were not included in the ERUMS Phase I and Phase II samples. When this oversight came to light, it was found that an additional 2,608 SSA nonmatch cases, of which 2,576 were single unit and 32 were multi unit, should have been included in the Phase I sample. As explained in subsection C,11, below, the weights for the affected

strata were revised to compensate for their being undersampled.

5. Merge of BLS and SSA data for Phase I sample EINs

The output file from operation 4.0 was merged with the data file for the BLS Phase I sample from operation 1.0. Data elements for each EIN appearing in both files were combined on a single record for that EIN. The EINs in the merged file, whether or not appearing in both the BLS and SSA samples, constituted the final Phase I sample.

6. Phase II sample selection

The Phase I sample EINs were divided into 11 strata, as shown in Table IIC-1.

Table IIC-1 - PHASE I SAMPLE COUNTS BY STRATUM

Stratum	BLS status	SSA status	Other classifiers	No. of EINs
1	single	single	Match on county and 2-digit SIC	8,689
2	single	single	Different county or 2-digit SIC	4,392
3	single	NWR		2,698
4	NWR	single		3,559
5	multi	Single	<20 RUs in BLS	88
6	multi	single	20+ RUs in BLS	6
7	single	multi		356
8	multi	NWR		41
9	NWR	multi		69
10	multi	multi	<20 RUs in BLS	60
11	multi	multi	20+ RUs in BLS	6
			TOTAL	19,964

The definitions used in classifying EINs by strata were as follows (NWR stands for "no wage report"):

BLS status

Single	One reporting unit with EIN in Texas UI file for 1982
Multi	2+ reporting units with EINs in Texas UI file for 1982
NWR	No reporting unit with EIN in Texas UI file for 1982

SSA status

Single	W-3 Wage Report for 1982, not included i n SSA Multi Unit Code File
Multi	W-3 Wage Report for 1982, included in SSA Multi Unit Code File
NWR	No W-3 Wage Report for 1982.

The sample counts shown in Table IIC-1 were reviewed by the ERUMS work group, which decided to allocate the Phase II sample as follows: take all EINs in strata 6,8 and 11; select 50 EINs from each of strata 1 to 4; select 34 EINs from stratum 5 (giving a total of 40 from strata 5 and 6 combined); select 40 EINs from stratum 7; and select 34 EINs from stratum 10 (giving a total of 40 from strata 10 and 11 combined). The specified number of EINs was then selected from each stratum systematically, using a random starting point and the sampling interval needed to achieve the desired sample size. The sampling intervals used and the sample sizes by stratum are shown in Table IIC-2.

Table IIC-2 - PHASE II SAMPLING INTERVALS AND SAMPLE SIZES

Stratum	Sampling interval	EINs selected

1	173.78	50
2	87.84	50
3	53.96	50
4	71.18	50
5	2.59	34
6	1.00	6
7	8.90	40
8	1.00	41
9	1.73	40
10	1.76	34
11	1.00	6
TOTAL		401

7. Listing of EINs for Phase II sample

For the relatively small Phase II sample, it was now possible to assemble information from several sources for use in the final analysis, which had several goals: to assign each sample EIN to a definitive final match status; to compare the characteristics, such

as industry classification and geographic location, for matched units; to explain, to the extent possible, why no matches were found for some EINS; and, for EINs with more than one reporting unit in either or both systems, to examine the relationships between individual reporting units. To acquire such information, BLS prepared lists of the 401 Phase II sample EINs and transmitted them to SSA and IRS for extraction of additional data for the sample employers.

8. Acquisition of additional SSA records for Phase II sample

The principal SSA sources of information about the sample EINs were the Single Unit Code File (SUCF), the Multi Unit Code File (MUCF) and a unedited file based on 1982 W-3 wage reports. From each of these files, listings were prepared of information

for each of the 401 sample EINs that appeared in that file. All

but 2 of the sample EINs appeared in at least one of the three files.

From the SUCF two listings were prepared: a listing of employer names and addresses, in EIN order, and a listing containing geographic and industry codes, plus some codes not used in the ERUMS analysis, also in EIN order.

From the MUCF a listing of reporting units was prepared for each of the 125 Phase II sample EINs that appeared on that file. For each reporting unit, the listing included an establishment number, geographic and industry codes, size codes, and date and source codes. The MUCF is a permanent file, so the existence of a reporting unit in that file did not necessarily mean that wages had been reported for that unit in 1982.

The unedited file based on 1982 W-3 wage reports included information for some employers not present in the edited W-3 file used for the electronic match prior to selection of the Phase I sample for ERUMS. In particular, information was available for

delinquent reporters, employers whose workers were not subject to Social Security taxes and household employers. Data for employers whose wage reports were being reconciled with their Forms 941 were obtained from yet another source.

The listing prepared from the unedited W-3 file had one or more lines for each of the 399 sample EINs that was found in SSA records. If no 1982 wage reports had been received, this was stated on a single line. For each EIN with 1982 wage reports, the listing included one or more lines, each showing establishment number, wages reported and number of employees. For some EINs with two or more lines, there were no establishment numbers, and for some there were establishment numbers which did not fully correspond with the establishment numbers shown for that EIN in the listing from the MUCF.

9. Acquisition of IRS records for Phase II sample'

The goal of the IRS record acquisition process was to obtain,

for each of the 401 Phase II sample EINS, data from Employer's Quarterly Tax Returns (Form 941), Farm Employer's Annual Tax Returns (Form 943) and Federal Unemployment Tax Returns (Form 940) for 1982 (see Section A of this chapter for a description of the purposes and filing requirements for these IRS forms). It was expected that the IRS records would be useful in the analysis of nonmatched BLS and SSA sample cases and also -in exploring the relative coverage of employers and reporting units in all three systems.

As explained in Section A of this chapter, Forms 941 and 943 data for the Phase II sample cases were obtained from a file containing extracts of Form 941, 941E and 943 data for tax years 1981 to 1983 that had been edited by the Bureau of the Census and

returned in edited form to the Statistics of Income Division of IRS. Data were obtained from the edited extract file for 385 of the 401 Phase II sample cases. The listings for these 385 cases included geographic and industry codes, quarterly information on

payroll, and annual information on first-quarter employment for each of the three years for which the employer had filed returns. For some EINS, the industry codes based on information reported to IRS had been replaced by Census Bureau codes.

Computer-generated listings containing the desired information from the Forms 940 were not available, as the main information of interest for the ERUMS project, the allocation of taxable wages by State, had not been keyed from the forms. Therefore, hard copies of the forms were requested from the 10 IRS service centers. It turned out that the retention period for the 1982 Forms 940 ended in January 1988, consequently, most of the returns received were for tax year 1983. A total of 227 Forms 940 were received, 16 for 1982 and 211 for 1983. The service centers also provided a total of 306 Forms 941 for the 401 sample EINS. Of these, 26 were for 1982 and 280 for 1983.

10. Final review and analysis

The results of the final review and analysis of the 401 Phase II sample cases are presented in Section A of Chapter 3. The general approach and methods used will be described briefly here. This step required working with individually identifiable records from the three agencies. Consequently, the work was performed by members of the ERUMS Workgroup who had been cleared for access to such data under the relevant initial and supplementary interagency agreements. This subgroup, known as the Microdata Access Group, consisted of the BLS and SSA members of the ERUMS Workgroup, plus one each from the Bureau of Economic Analysis and the Committee on National Statistics.

A major element of the final review was the determination of final match status codes with respect to status of each EIN in the BLS and SSA systems. There were eight possible classifications:

Group no.	Status in: BLS	SSA
1	Single	Single

2	Single	Inactive
3	Inactive	Single
4	Multi	Single
5	Single	Multi
6	Multi	Inactive
7	Inactive	Multi
8	Multi	Multi

EINs that were inactive in both systems had no chance of entering the ERUMS sample.

The classifications for BLS EINs were straightforward. An EIN was considered active if it appeared in the 1982 UI Address File for Texas. If an active EIN had only one reporting unit, it was classified as single unit; if it had two or more reporting units with EINs in Texas, it was classified as multi unit. These classifications did not change at any time during the sample

selection and analysis phases.

There were numerous cases, however, in which the classifications initially assigned on the SSA side were changed as a result of the final review. The definition of active for SSA was that the employer filed a W-2/W-3 wage report for 1982. The edited SSA file originally used to make this determination was incomplete: certain types of filers, such as those whose reports were delinquent or were still in the process of reconciliation with Forms 941 submitted to IRS, were not included. Working with a more nearly complete file in the final review, 44 of 91 EINs originally classified as SSA inactive were reclassified to active.

In the initial single unit/multi unit classification for active SSA EINS, a broad definition of multi unit was adopted: any EIN that appeared in the SSA Multi Unit Code File was classified as multi unit. In the course of the analysis, however, it became evident that this definition was far from comparable with the definition used for multi unit on the BLS side. consequently, the SSA multi unit category was redefined to include only those EINs

for which it could be clearly established that two or more reporting units in Texas had been identified in the W-2/W-3 wage reports for 1982. Use of this much narrower definition reduced the number of SSA multi unit EINs in the Phase II sample drastically, from 120 to 10. All of the other 110 EINs were reclassified as single units.

This decision did not mean that none of the Ill EINs reclassified as single units had more than one reporting unit, as defined for SSA's Establishment Reporting Unit Plan (ERP). As discussed further in Chapter III,A and shown in Table IIIA-7, some of these employers, although eligible, were not currently participating in the ERP. Some appeared to have filed as multi unit employers for 1982, but because their W-3s did not include establishment numbers appearing in the Multi Unit Code File, the location of their reporting units could not be determined. Some reported under the-ERP but simply did not report for two or more units in Texas.

A second part of the review and analysis was the comparison of

industry (SIC) classification and geographic location (State and county) for EINs classified as active single units in both systems. BLS and SSA use slightly different adaptations of the OMB's Standard Industrial Classification (SIC) system, so failure

to match at the 4-digit SIC level does not always mean that the two agencies have placed a unit in substantively different categories. Comparisons of geographic location had to deal with the fact that SSA county codes were not available for about 6.5 percent (weighted) of the matched single unit EINS.

A third important aspect of the review was to look at the cases classified as inactive in one of the two systems to try to determine the reasons for non-coverage. For the BLS inactive EINS, SSA industry codes and employment data were reviewed to identify employers who might be exempt from the UI filing requirements. UI Address Files for later years were examined to see if the employer had reported to the system after the first quarter of 1982, and the listing of Form 941 data by quarter was examined to identify

employers who may have entered the system in 1982 after the first quarter.

For the SSA inactive EINS, BLS examined the employers" records from the 1982 and 1983 UI Address Files and found possible explanations for some of the nonmatches, including EINs that appeared in the 1982 file but had zero employment reported for each of the three months covered and cases where the same employer reported under a different EIN in the 1983 file. As in the case of the BLS inactive EINS, data on the Form 941 listings for 1981 to 1983 could be examined to see during which years of this period they reported employment to IRS.

For EINs classified as multi unit in both systems, it had been planned to compare industry and geographic location for the individual reporting units, as was done for the matching single units. This turned out to be impractical for a variety of reasons. There were only 9 EINs classified as multi unit in both systems, and even for these it turned out in most cases to be difficult to establish correspondence between individual reporting units in the

BLS and SSA records.

Finally, the BLS and SSA industry codes for the matched single unit EINs were compared with the industry codes for these units in the IRS/Census file. The IRS data are all provided at the employer level (except for the allocation of wages by State on Form 940), so comparisons of industry and geographic classification were meaningful only for single unit employers. For reasons explained earlier, comparisons of industry codes were limited to the SIC 2-digit level. Some results of these comparisons are presented in Table IIIA-8 and discussed in Chapter III, Section A,3.

11. Reweighting to account for missed SSA nonmatch cases

As explained earlier, after most of the analyses had been completed it was discovered that SSA nonmatch cases included in the stem match file produced by SSA had not been extracted from that,file for inclusion in the Phase I sample. Adding cases to the Phase I and II samples at that point would have further

delayed completion of the study, so we, decided to reweight the Phase II sample cases in the affected strata and rerun the results tables that would be affected by the changes in the weights. This procedure was potentially biased, since it meant that a certain subset of SSA nonmatch cases had no chance of selection. However, we could think of no reason why the subset of 'SSA nonmatch cases that matched the 7-digit EIN stem of a case in the BLS initial sample should differ in any significant way from the SSA nonmatch cases that were included in the Phase I and II samples.

The strata affected were those for which there were no reporting units in the Texas UI file for 1982, namely strata 4 and 9 (see Tables IIC-1 and 2). The new weights were calculated as follows:

Stratum 4 w = 71.18 + 2,516/50 = 122.70

Stratum 9 w = 1.73 + 32/40 = 2.53

For each stratum, the first term is the old weight and the second term is the number of additional sample cases divided by the number of cases (which was not changed) in the Phase II sample.

All tables in this report that were affected by these changes have been rerun, using the new weights.

D. Administrative arrangements

1. Confidentiality protection and interagency arrangements

Confidentiality Protection: the Challenge.

From the beginning, the ERUMS project was faced with obstacles imposed by confidentiality statutes and rules. Those legal restrictions were enacted to protect records about private

organizations that the government collects and maintains, but they often fail to address the realities of interagency coordination required to perform the government's necessary statistical operations.

To meet the challenge of confidentiality laws, it was necessary to devise procedures and to develop interagency agreements for data exchange that would satisfy both state and federal requirements. throughout the project, even in the final phases after the data linkage had been completed, those arrangements had to be reexamined and in some instances renegotiated to comply with complex legal restrictions before the analysis plan could be carried out.

Data Sharing: the Issues.

Interagency exchange of identifiable microdata was the essence of ERUMS. Such data sharing is greatly restricted by Federal confidentiality laws, which generally permit agencies to disclose

statistical information only in summary or other unidentifiable. form. Since the ERUMS study was designed to link and compare information about individual employers collected separately by the different agencies, the Workgroup had to develop and carry out lawful methods of transferring data about identifiable business units among the participants. A related task was to minimize the disclosure of identifiers in making those transfers and linkages.

In studying individual employers, the Workgroup was particularly interested in differences in the way a given employer may report establishment or multi-unit enterprise data to various State and Federal agencies, with resulting discrepancies among the agencies in State and county levels of wage and employment detail for that employer. To examine and evaluate these differences, the Workgroup needed to compare employers' reports to the BLS through State UI programs, the SSA in FICA reporting, and the IRS on employment tax returns. Members of the Workgroup included employees of those three agencies, plus employees of the Bureau of Economic Analysis, OMB, the Bureau of the Census, and a contractor to the Committee on National Statistics, National Academy of Sciences.

In the ERUMS study, the Employer Identification Number (EIN) was the identifier that was common to all the reporting systems.

It was used to define the sample drawn by BLS. In addition it provided the basis for retrieving, linking and comparing records containing information from the SSA and IRS files. By law the EIN is a tax identification number. Even when standing alone, the EIN is protected by Internal Revenue Code confidentiality restrictions if its source was a tax record.

The Workgroup planned to analyze the similarities and differences in the information that corresponded to each EIN as it was reported to each of the agencies. The analysis and findings would be entirely statistical, with no reference to individual identifiable cases. Nevertheless each step in defining, selecting, matching, verifying, editing and developing analysis plans required

access by some persons to identifiable data from protected sources.

Confidentiality Considerations.

State and Federal confidentiality restrictions were an impediment to the interagency exchange of microdata that was an essential element in the purpose and methodology of the project. Much of the detailed work of matching and reconciling BLS, SSA and IRS employer reports had to be performed manually, and it was anticipated that most of the group's members would need at least limited access to microdata at some time during the project.

Since the Workgroup was composed of employees from various agencies and organizations, confidentiality laws did not apply to them uniformly. In varying degrees, certain laws, regulations and policies affected each agency's access to identifiable records from particular sources and provided differential access to various individuals in the Workgroup. A recurring theme was the necessity at each stage in the process to identify the particular persons who needed to use identifiable data and to ensure that others did not

have access at that time.

Confidentiality of Federal Tax Records.

The study called for access to data from W-3 records which by law are Federal tax records that are processed and maintained at SSA in connection with the computation of Social Security retirement benefits. In addition, the EINS, which had a central function in the matching process, are Federal tax identification numbers, thus requiring compliance with the confidentiality restrictions in the 1976 Tax Reform Act amendments to the Internal Revenue Code (26 U.S.C. 6103). The status of these records as tax records made it necessary to satisfy IRS that the selection by SSA of sample cases, SSA's disclosure of W-3 data to BLS, and the use of employer data by other members of the Workgroup met the requirements of the Internal Revenue Code.

Confidentiality of State UI Records.

BLS selected Texas as the State whose records it would sample and it obtained written permission from the State Employment Security Agency to use Texas UI records in the project. The Texas Unemployment Compensation Act (section 11(g)) requires Texas employers to maintain records and file reports to the Texas Employment Commission with detailed information about the business operations and the number and amount of compensation of employees. The law prohibits disclosure except for administering the Act, and it makes improper disclosure of that information punishable with fines or imprisonment.

As explained in Section II,A,l, the State agency periodically submits to BLS a UI address file that compiles identification data for all reporting units to the most detailed level that is available from employers reporting to it. BLS collects these reports under a pledge of confidentiality that allows the data to be used only by authorized persons for statistical purposes or for other purposes made known in advance to the respondent. Further, since the EINs that defined the employer sample are tax

information, the state records and identifying EINs require special treatment to comply with the Internal Revenue Code requirements.

At the outset, the Workgroup had established "need to know" as the basic rule to control access to identifiable state microdata. In addition, SSA employees who needed identifiable UI microdata would be required to sign a non-disclosure affidavit before BLS would provide them with access to state UI data. They also would acquire special status for access to tax data that IRS made available to BLS.

Technical Safeguards.

Besides affidavits and other written procedures to protect the confidentiality of records, certain technical methods of minimizing disclosure risk were adopted. The first of these methods was to avoid identifying actual sample cases by EIN to persons who performed program or operational services for the participating agencies but were not directly associated with the Workgroup. This

method was adopted to conform to the Internal Revenue Code requirements for tax information under the agreement BLS had with the State. At BLS this led to a decision not to select the sample or compile data on the mainframe computer system that is operated by a private organization under,contract to the Department Of Labor. Instead BLS stored and used the data on a mini-computer accessible only to regular BLS employees who were in the Workgroup.

Once BLS selected the Texas sample, it had to create a finder list so that SSA could extract corresponding records from its W-3 and related files for employers in the sample. To avoid identifying actual cases in the sample, BLS furnished SSA with a

listing of digits 1 to 6 and 9 of all sample EINS. (As an extension of this safeguard, the specific pairs of 7th and 8th digits used for sampling have not been reported outside the Workgroup.) SSA operational staff then extracted records from the W-3 and related files for all records in which those 7 digits appeared, with no way of knowing which particular employers were actually in the BLS

sample. This procedure effectively masked the identities of sample cases derived from state UI files, and significantly limited the number of SSA employees who were required to sign BLS non-disclosure affidavits.

Agreements for Interagency Data Sharing.

To accomplish the necessary interagency data transfers, the Workgroup originally planned a tripartite arrangement through interagency agreements of SSA and BLS with IRS. However, IRS counsel raised object ions that quickly made it evident that a multi-party agreement would be unduly cumbersome, and approval would probably not be forthcoming. As an alternative, IRS proposed to contract exclusively with BLS for the performance by BLS of services that required access to tax data. SSA staff would be designated as special agents of BLS to process the data. Work was then begun to draft bilateral BLS/IRS and BLS/SSA agreements.

The drafting of these agreements proved to be a more delicate

task than had been anticipated. By law, the purposes of IRS participation in the project and its service contract with BLS had to be related to IRS administration of the tax laws. Section 6103 of the Internal Revenue Code is the provision that allows IRS to use contractors, but only to the extent necessary . in connection with activities performed for purposes of tax administration. One of those purposes is the conduct of statistical studies based on return information, which Section 6108 of the Internal Revenue Code authorizes IRS to perform.

The first revision, proposed by IRS, to the statement of purpose for ERUMS, drafted to meet the advice of IRS counsel as to the requirements of the law, did not satisfy Workgroup participants from other agencies. They felt that the IRS redraft did not fairly describe the purposes of the ERUMS project or SSA's role in it, and consequently they asked for further revision. In the following draft,' care was taken to define contractual purposes in language that covered the statistical purposes of the several participating agencies, and that provided for the exchange of records to create a common pool of data for a variety of analytical purposes, including those related to tax administration.

At the same time SSA drafted a companion document, a Conditions of Use Agreement, that was acceptable to IRS and would enable BLS to use SSA files for the ERUMS project. Under this agreement, SSA would furnish BLS with SSA's Single Unit Code

File, Multi Unit Code File and Employer Report (W-3) Record. The agreement authorized BLS to link data from these statistical files with data in the BLS Unemployment Insurance Address File and with certain data to be furnished by IRS, and prohibited any other linkage.

Finally the terms of a contract between IRS and BLS were agreed upon. The contract enabled BLS to receive tapes containing tax information extracted by SSA from W-3 and Employer Identification records and records extracted from the IRS Business Master File, and to combine them with records in the UI Address File maintained by BLS. It imposed strict safeguard procedures and

required BLS to provide IRS with a list of all persons permitted to see confidential tax return data. This list included SSA employees who were required to sign affidavits as agents of BLS. As soon as the contract and the Conditions of Use Agreement were signed by officials of the participating agencies, the way was cleared for the data transfers to begin. (Copies of the two agreements are shown in Appendix B.)

In retrospect, the signing of interagency agreements between BLS and IRS and between BLS and SSA had the appearance of breaking a log-jam that had threatened to block the study. It would be a mistake, however, to regard those documents as magic incantations that moved the project. Rather, they documented a process of negotiation by which the study plan was adapted to the requirements of the various confidentiality laws that impinged on it, and by which a combination of technical and procedural safeguards were fitted to those requirements.

The Microdata Access Group (MAG).

In the planning and matching stages of the project, the persons who needed to have access to microdata were those members of the Workgroup who were performing the manual and electronic matching and verification. At Workgroup meetings, members generally reviewed data in the form of frequencies. and other summaries to track the progress of the matching operations and to plan future steps. Occasionally discrepancies appeared, or questions arose concerning classification of a particular employer or possible mismatch of data. Those matters were usually referred to particular members to resolve, with access to microdata as needed on an ad hoc basis.

When the matching steps were completed and time came to plan the analysis, new arrangements were needed to enable a different group of persons to examine identifiable microdata. The Microdata Access Group was formed for this purpose. At this point, IRS agreed that its contractor, BLS, would be permitted to make Workgroup members its agents as needed for the analysis stage. This enabled employees of BEA and the contractor to the Committee

on National Statistics to become sworn agents who would be permitted to examine and analyze microdata when necessary. This group or subgroups of it met periodically to plan and

perform the analysis and to prepare findings. The Microdata Access Group then reported its activities and findings back to the full Workgroup.

2. Working arrangements and schedule of operations

Starting with eight individuals representing five agencies (BLS, SSA, IRS, BEA and OMB), the first meeting of the ERUMS Workgroup, which was held in February 1983, was devoted to setting out the ground rules for how the group would conduct its business. There was agreement on a format of rotating future meeting sites among the agencies, which was followed throughout the course of the project. These regular Workgroup meetings were to be a forum for discussion of issues in pursuit of fulfilling the charge of the

Federal Committee on Statistical Methodology (FCSM), with assignments being made to the appropriate representatives to be worked on between meetings and reported on for discussion at a subsequent meeting. At this first meeting there was also a discussion of the group's organizational affiliation, objectives, data access, and data processing issues.

Meeting eight times over the 12 month period which ended March 1984, the Workgroup focussed its efforts on: 1) developing a formal statement regarding the purpose of the Workgroup, 2) outlining plans for conducting the study, 3) preparing a project description, 3) documenting potential data files, and 4) defining specific tasks that needed to be done. During this period some personnel changes took place among the BLS and IRS representatives. In addition, a representative from the Committee on National Statistics and an observer from the Bureau of the Census joined the Workgroup.

By the end of March 1985 this expanded group had met eight more times, resulting in the following accomplishments: 1) development of electronic matching criteria, 2) selection of a

state for the study, 3) obtaining universe counts UI of records for the state, 4) development of the sample design, and 5) preparation of the first drafts of the interagency agreements covering the conditions of the data exchanges and work to be done.

The first five of the eight meetings that were held between April 1985 and March 1986 were devoted almost exclusively to resolving serious concerns that had surfaced regarding the interagency agreements and how these concerns could be dealt with to the satisfaction of the parties to those agreements. A number of redrafts of these agreements were prepared, culminating in September 1985 with the final versions having the approval of all Workgroup participants. In the last three meetings in this period the group concentrated on refining the specifications for selecting records from BLS and SSA files and the electronic matching of the records selected. Some further changes among the individuals representing BLS and IRS on the Workgroup occurred during this time.

The frequency of regular Workgroup meetings declined somewhat over the next 24 months (April 1986 - March 1988), with a total of thirteen held, during which efforts were directed toward: 1) operations surrounding the selection of the cases for the final sample and preparing these records for the manual matching and classification operations, 2) performing the manual match, and 3) documenting and presenting the results. In conducting the manual match, BLS and SSA provided additional staff of individuals authorized to access the microdata records. This special group met several times for about a month to complete the manual matching operations and then presented the results to the Workgroup.

The time between March 1988 and March 1989 was spent refining the results, developing alternative approaches to presenting them, preparing descriptive and analytical tabulations, and planning for the preparation of the Workgroup's final report. There were eight regular Workgroup meetings held during this period along with several additional meetings of the special group having access to the microdata records.

For the six months of the project that ended in September 1989, the Workgroup concentrated on completing outstanding assignments that were needed for the final report as well as the actual drafting, reviewing and redrafting of sections of the report.

After the ERUMS project had been underway for awhile, the workgroup agreed on the need for a project timetable, with target completion dates for each task, in order to establish concrete goals and make it easier to evaluate the current status of the work and identify problem areas. The timetable, with initial and revised target dates and actual completion dates, is shown as Exhibit IID-1.

As can be seen in the exhibit, a draft workgroup report was not produced until about three and one-half years after the initially scheduled date. Several factors accounted for this delay. Approval of the interagency agreements took considerably longer than expected, and the exchanges of data and actual matching could not begin prior to their approval. Once the agreements were approved, the selection of the Phase I BLS and SSA samples (Tasks

16 and 17) proceeded expeditiously; however, the subsequent steps that led to the selection of the Phase II subsample (Task 20) were delayed further from the initial target dates.

What is not obvious from the timetable is that review of the initial set of statistical tabulations (produced in Task 26) by Workgroup members suggested the need for several additional analyses in order to explain some unexpected findings. In some instances, these additional analyses indicated a need to change the definition of match status and other classifiers used in the initial set of tabulations, thus making it necessary to redo the

initial tabulations. As explained earlier in this chapter, some of the tabulations had to be redone to compensate for the omission of a significant number of SSA nonmatch cases from the Phase I and II samples. Pursuing these additional lines of investigation, while it caused additional delays in issuing this report, was very fruitful in bringing to light additional information relevant to the goals

of the ERUMS project. We did not complete all of the analyses that we would like to have done, but we reached the point where we felt it would be more productive to concentrate our efforts on issuing this report containing our main findings and recommendations.

Exhibit IID-1

ERUMS - PROJECT TIMETABLE

Task	Completion Dates estimated date	revised date	actual date
1. Purpose statement		06/83	
2. Files documentation		03/84	

3. Project d ascription	04/84
4. Criteria for match operations	04/84
5. Draft Interagency Agreements	04/84
6. Select state	05/84
7. Preliminary sample design	05/84
8. Specifications for universe counts	06/84
9. Obtain universe counts	11/84
10. Finalize sample design	01/85
11. Finalize BLS/SSA Agreement	
02/85 1/	05/85
12. Finalize BLS/IRS Agreement	

02/85 1/ 09/85 2/ 09/85

13. Obtain agency approvals on agreements

03/85 1/ 10/85 2/

03/86 3/

04/86 4/ 04/86

14. Specifications for sample selection

02/85 1/ 11/85 2/

03/86 3/ 03/86

15. Specifications for electronic match

03/85 1/ 11/85 2/

03/86 3/ 03/86

16. Select sample cases from BLS files

04/85 1/ 01/86 2/

04/86 3/ 06/86

17. Select sample cases fran SSA files

05/85 1/ 02/86 2/

		05/86	3/	07/86

18. Electronic match and counts

06/85	1/	04/86	2/	
		06/86	3/	01/87

19. Determine subsampling criteria

06/85	1/	04/86	2/	
		06/86	3/	
		09/87	5/	09/87

1/ established 01/85

2/ revised 09/85

3/ revised 02/86

4/ revised 03/86

5/ revised 09/87

6/ revised 10/87

ERUMS - PROJECT TIMETABLE

Task	Completion Dates: estimated date	revised date	actual date
20. Select subsample	07/85 1/	05/86 2/	
		07/86 3/	
		10/87 6/	10/87
21. Listings of final sample cases for manual matching operations	07/85 1/	05/86 2/	
		07/86	
		12/87	12/87
22. Specifications for manual matching and classification operations	08/85 1/	06/86	

08/86

01/88 6/ 01/88

23. Obtain IRS data for final sample

10/85 1/ 08/86

09/86 3/

12/87 6 / 04/88

24. Manual matching and classification

10/85 1/ 09/86 2/

11/86

02/88 6/ 02/88

25. Analyze and document results of manual matching operations

12/85 1/ 12/86 2/

02/87 3/

03/88 6/ 03/88

26. Initial set of statistical tabulations

12/86 7/ 02/87 3/

04/88 6/ 03/88

27. Additional investigation, analysis and revised/additional tabulations a.s.a.p. 8/ 05/89

28. Draft Workgroup Report (preliminary)

02/86 1/ 02/87 2/

05/87 3/

06/87 4/

06/88 5/ 06/89

29. Re-draft Workgroup report (final?)

a.s.a.p. 9/ 09/89

1/ established 01/85

2/ revised 09/85

3/ revised 02/86

4/ revised 03/86

5/ revised 09/87

6/ revised 10/87

7/ established. 09/85

8/ established 04/88

9/ established 06/89

CHAPTER III--RESULTS

A. Substantive results

1. Introduction

This section of Chapter II - I presents and discusses the principal results of the ERUMS project. Data highlights are shown in text tables. More detailed tables appear in Appendix A.

Some distributions of employers by match category have been influenced by differences in definitions and coding conventions used by the BLS, SSA and IRS systems. Perhaps the most significant differences in definitions affecting basic match categories involve

the definitions of active employers in the BLS and SSA systems. Employers were considered active in the SSA system if they submitted an annual W-3 wage report for 1982. Employers were considered active in the BLS system if they were present on the first quarter UI Address File. In general, inclusion on the UI file required a first quarter payroll tax report, although as subsequent discussion of individual match categories will reveal, some employers appear on the file even though they show zero employment and some employers apparently remained in the file with estimated employment and payroll values for a few quarters after they ceased filing quarterly payroll reports. In general, employers operating any time during 1982 were supposed to file a W-3 wage report even if they were not operating in the first quarter and therefore were not likely to appear in the UI file. In practice, however, not all W-3 wage reports that were filed were easy to find for use in the ERUMS study. As noted in Chapter II, a number of employers were reclassified as active in the manual match stage when it was possible to locate W-3 wage reports that had been filed late or were still being reconciled with the quarterly Form 941 reports.

For the most part, the definitions of reporting units for multi unit employers are the same for the BLS and SSA systems. Both systems request reporting by county-industry combination. But as has also been noted in the discussion of reclassifications in Chapter II, the status of records in SSA's files made it difficult to derive a concept of multi unit employer that would permit meaningful comparisons with BLS multi unit employers. The SSA definition of multi unit employer that was finally settled on required that an employer had filed a 1982 W-3 wage report on which two or more reporting units could be clearly identified as Texas units. Since the SSA establishment reporting plan has not been well maintained, this definition turned out to yield a very small number of sample SSA multi unit employers in the final match stage of the project. All other active SSA employers were classified as single unit. Undoubtedly some of these employers actually had more than one active reporting unit in Texas, but did not file their 1982 W-3's in such a way that this could be determined.

Because the ERUMS Phase II sample was so small, it is important to keep in mind that sampling errors are relatively large for many of the estimates of the distribution of employers by match category. This is especially true for match categories involving SSA multi unit employers where the reclassification process left some categories empty or nearly empty. Nonetheless, most of the discussion in this section is focused on the distributions of weighted sample totals. The next part of Section A discusses distributions by broad match status. Subsequent parts discuss more detailed match categories divided between a) employers that were active in both systems and matched on single/multi unit status and b) employers that were not active in one of the two systems and/or did not match on single/multi unit status.

2. Distribution by final match status

Table IIIA-1 shows the distribution of employers in the ERUMS sample by match status with respect to activity in the BLS and SSA 1982 files. On a weighted basis, 67 percent of employers were

active in both systems, 28 percent were active in only the SSA system, and 5 percent were active only in the BLS system. Because of the lack of strict comparability in the definition of active status in the two files, however, these percentages overstate the extent to which employers tend to be active in only one of the two payroll tax systems. For example, as noted above, employers that filed W-3's with SSA for 1982 did not necessarily operate in the first quarter of 1982 and therefore may not have been expected to appear in the BLS file.

Tables IIIA-2 and IIIA-3 show the single/multi unit match status of the ERUMS employers. Table IIIA-2 shows the match status for all employers classified as active in the BLS system and Table IIIA-3 shows the match status for all employers classified as active in the SSA system. These tables show clearly the relatively small number of multi unit employers in the final classification scheme (particularly in the SSA system) and also how imperfect the matching on multi unit status appears to be (although small samples and the definition of multi unit that was used often make interpretation of multi unit match status problematic).

Overall, about 1 percent of all active SSA employers were classified as multi unit in the BLS system, and only about 0.1 percent of all active BLS employers were classified as multi unit in the SSA system. Only about 7 percent of BLS employers classified as multi unit also met the final SSA multi unit criteria. Because the number of SSA employers ultimately classified as multi unit was very small (yielding large relative sampling errors), Table IIIA-3 tells us little about the extent to which SSA multi unit employers also report as multi unit in the BLS system.

The remainder of this section discusses the eight specific final match categories that are identified in Table IIIA-4. The discussion is divided between the categories that match on single/multi unit status (groups 1 and 8) and categories that do not match on that status (groups 2 to 7). For the matched groups the focus is on determining the extent to which the information on

geography and industrial activities also matches. For the nonmatched groups the focus is on identifying, to the extent possible, the reasons for nonmatch. An examination of information from IRS Forms 940 and 941 was particularly helpful in this latter endeavor.

3. Characteristics of matched cases

Table IIIA-4 shows that employers classified as single unit in Texas in both the BLS and SSA systems constituted 66.2 percent of all employers on a weighted basis. Table IIIA-5 provides breakdowns of this group by match status with respect to geographic location and industrial activity.

The weighted counts in Table IIIA-5 show an estimated 82 percent match rate on county. The group 1 employers that did not match on county are divided into three categories: those that did not match because they had been given a "statewide" code by SSA, those that were coded into different counties within Texas, and

those that were coded into a State other than Texas in the SSA system. Each of these three categories accounted for about 6 percent of the group I employers.

On a weighted basis, about 77 percent of the single unit employers that matched on county locations also matched on 2-digit SIC industry. Nearly all of the employers that failed to match on county because of statewide coding, however, also failed to match on 2-digit industry. This phenomenon reflects the fact that SSA often codes employers both "statewide" in terms of location and unclassified in terms of industry when an employer is assigned an employer identification number without first filing an application form (SS-4) that requests information on location and industry. The remaining single unit employers that did not match county code occupied an intermediate position with respect to the proportions matching on 2-digit industry and receiving "unclassified" industry codes.

The most telling feature of the employers that ended up classified as multi unit in both the BLS and SSA systems (group 8) is how few of them there are. They constitute only 0.1 percent of

all employers on a weighted basis (see Table IIIA-4). Of the nine sample cases in group 8, five had more BLS than SSA reporting units in Texas, two had more SSA than BLS reporting units, and two had the same number of BLS and SSA units. In all, the nine employers had 105 Texas reporting units in the BLS system and 60 Texas reporting units in the SSA system for 1982. For 53 of the 60 SSA reporting units, there was a corresponding BLS reporting unit for the same employer in the same county, and

in most of these cases the SIC codes of the reporting units also matched at the 2-digit level.

Since BLS and SSA assign industry and geographic codes independently, some discrepancies are to be expected in SIC and county codes for employers that match on single or multi unit status. The extent of nonmatch between the BLS and SSA systems is also affected by the fact that SSA has not had adequate resources to follow up with employers to clarify and update initial

information obtained concerning the geography and industry of reporting units. The high incidence of employers and reporting units that were not assigned specific industry and county codes in the SSA system reflects, in part, the lack of resources for follow-up with employers that initially supply inadequate information. in addition, since SSA does not have a program like BLS's 'or updating the geographic and industry codes of employers on a regular cycle, more of the characteristics indicated in SSA files would be expected to be obsolete than in BLS files. In the case of multi unit employers in group 8, for example, lack of a systematic SSA updating program probably contributed to the smaller number of identifiable Texas reporting units in the wage reports of the SSA system than in the wage reports of the BLS system.

4. Characteristics of nonmatched cases

On a weighted basis, the largest of the nonmatched categories was group 3, the SSA single units that had no first quarter wage report for 1982 in the BLS system. Because the SSA wage reports

(W-3's) covered all of 1982 and the BLS wage reports covered only the first quarter of 1982, many of the group 3 cases could have been employers that commenced operations some time in the final three quarters of 1982, or for other reasons did not have wages to report for the first quarter. To determine how many sample cases fit this category, a check of the IRS records was made to see how many of them had not reported first quarter 1982 wages on Form 941. As shown in Table IIIA-6, about 69 percent of group 3 cases showed no first quarter 1982 IRS wages (weighted estimate). About half of these cases either had no wages at all reported to IRS for the period 1981-83, or only had wages reported on the annual Form 943 for agricultural employers. (As explained in Chapter II, Section A,1, agricultural workers are only partially covered by the UI system.)

On a weighted basis, about 3 percent of group 3 sample employers had first quarter IRS wages, but had incomplete or ambiguous geographic information that precluded verification of active operations in Texas in 1982. Most of the cases in this category were employers on SSA's Multi Unit Code File who appeared

to have most of their operations outside of Texas and for which it was not possible to determine from the W-3 file whether or not they maintained operations in Texas in 1982. IRS Form 940 provides an alternative means of checking to see if an employer reports wages in a particular state. Unfortunately, not

all Form 940s were available for ERUMS sample cases. But a check of available Form 940s yielded three multi state employers in group 3 with reported wages in Texas. In one of these three cases, the reason for the initial nonmatch was found to be an incorrect EIN in the UI name and address file.

An additional (weighted) 8 percent of group 3 sample employers had first quarter 1982 IRS wages, but did not appear to meet UI payroll tax coverage requirements because of nonprofit status or a payroll that was too small. (See Chapter II, Section A,l for detailed information about coverage requirements of the UI system.) For the remaining 20 percent of group 3 sample employers (10 sample

cases) with first quarter 1982 IRS wages, the reasons for absence from the UI address file were not clear. One of the 10 cases was found in the UI system under a different EIN, suggesting an error in reporting or recording the EIN. About half of the cases were found in the 1983 UI Address File. A search of the Texas State agency files might provide additional information about the status of these 10 cases, but the interagency agreements for ERUMS (see Appendix B) do not provide for the disclosures that would have been necessary for that purpose.

Apart from employers with no BLS wage record, the largest number of employers in a nonmatch category in Table IIIA-4 is in group 2--i.e., employers that were single unit in the BLS system but lacked SSA wage reports for 1982. On a weighted basis, about 5 percent of employers fell into group 2.

Most of the employers that lacked SSA wage records seem to represent businesses that ceased hiring employees, went out of business, or went through other changes that altered their reporting to IRS and SSA. Half of the group 2 employers (11 of 22)

reported no employment in the 1982 BLS file. Most of these cases (9 of 11) had dropped out of the BLS system by 1983, and an additional 23 percent (5) of the group 2 cases had positive employment in the 1982 BLS file, but had no record in the 1983 file. most of the group 2 cases that had either no employment in the 1982 BLS file or no record in the 1983 BLS file (or both) had filed their last Form 941 with IRS for a quarter in 1981 (usually the third or fourth). This pattern is consistent with a BLS policy of continuing to estimate employment and payroll for employers who appear to be late in filing until the reason for nonfiling can be determined or until a specified number of quarters (now set at two) has passed without a filing.

For the remainder of group 2 (6 cases or 27 percent of the total), it is more difficult to explain the discrepancies between the BLS and IRS/SSA files. Three of 6 cases apparently filed no IRS tax forms (using the sample EIN'S) over the period 1981-1983. one case filed partnership returns with IRS, but did not report employment or payroll. one case reported employment to IRS for 1981 and 1983, but not for 1982, and one case reported employment to IRS for 1981, but not for 1982 or 1983. The case that had IRS

employment in 1981 and 1983 could have been carried with a positive employment imputation in the 1982 BLS file if it had not filed UI tax forms in the quarters when it didn't have payroll. The other cases may reflect EIN discrepancies between the BLS and IRS/SSA systems that arise because of clerical errors or because business reorganizations create new legal entities (with new EIN'S) that may not have been reflected in the records from the UI Address File used for this study.

Group 6, like group 2, contains sample cases with no SSA wage records for 1982. But group 6 involves multi rather than single unit BLS employers, and accounts for only 0.1 percent of all employers on a weighted basis. Four of the 10 group 6 sample cases had dropped out of the BLS file by 1983, including 2 cases that "dropped out" because they_appeared in the 1983 file under a new EIN, indicating some kind of business reorganization. Of the remaining 6 cases in group 6, 4 apparently filed no IRS tax returns

(under the sample EIN) over the period 1981-1983, and the other 2 cases had reported employment and payroll to IRS for 1981, but not for 1982 or 1983. Most of these 6 cases were large enough to make it unlikely that they would have simply failed to file IRS tax returns while continuing to file UI tax returns. Thus, business reorganizations or other factors leading to discrepancies between the EIN's in the BLS and the IRS/SSA systems may well have resulted in their presence in group 6.

On a weighted basis group 4 accounts for 1 percent of all employers in Table IIIA-4. Group 4 includes employers classified as multi unit in the BLS system and as single unit in the SSA system. On a weighted basis, 53 percent of the cases in group 4 did not appear on SSA's Multi Unit Code File at all. The 'rest of the cases did appear on the Multi Unit Code File, but did not have W-3 wage reports that clearly identified more than one Texas reporting unit. The relatively large number of BLS multi unit employers that did not appear at all on SSA's multi unit code file probably reflects largely the lack of SSA resources to monitor employer status over time in order to identify employers that expand from single unit to multi unit status as they grow. It may

also reflect inadequate monitoring of initial applications for EIN's to identify potential candidates for establishment reporting.

Except for group 7 (which was left with no sample employers after reclassification), the smallest among the eight match groups in Table IIIA-4 is group 5, which had only one sample employer after the reclassification process. Group 5 includes employers classified as multi unit in the SSA system and as single unit in the BLS system. Unfortunately, the small sample size for SSA multi units (that resulted partly from the reclassification after the final sample was drawn) precludes meaningful conclusions about the extent to which SSA multi units tend not to be classified as multi unit in the BLS system.

5. SSA's establishment reporting plan

As explained in Section A,2 of Chapter II, some multi unit

employers report their employees' annual wages to SSA separately for each of their individual reporting units (which are usually equivalent to establishments), under a voluntary arrangement called the Establishment Reporting Plan (ERP). This system makes it possible for SSA to code reporting units and the employees working in those units by county, State and industry.

Potential multi unit employers are identified at the time they file their applications for Employer Identification Numbers on Form SS-4. They are asked to complete a Form 5019 on which they list each of their reporting units, with information about geographic location and industrial activities. Subsequently, they are expected, if they agree to participate in the plan, to report annual wages separately, on Forms W-3, for each of their reporting units and to inform SSA, by submitting new Forms 5019, of any changes in the number and characteristics of their reporting units.

The results of the ERUMS project have provided, as a by-product, some quantitative information about participation of multi unit employers in the ERP as of 1982. This information is

summarized in Table IIIA-7. The table shows information for all sample EINs that were in SSA's Multi Unit Code File and were active in 1982 according to SSA or BLS or both.

About three-fifths (weighted estimate) of these employers had filed a Form 5019 at some time in the past, indicating their willingness to participate in the ERP. However, among those who reported wages for 1982, about three-fourths reported as single units, i.e., they filed a single Form W-3 covering all of their employees. It is possible that a few employers in this group no longer had more than one reporting unit, but the likely explanation for most of them is that they no longer chose to file separate W-3s for each reporting unit.

An additional 6 percent of the employers who had filed Forms 5019 in the past filed multiple W-3s for 1982, but for at least some of their W-3s the reporting unit numbers used did not match the numbers they had provided on Form 5019, so that location and industry codes could not be assigned to those reporting units. As a result, only an estimated 18 percent of the multi unit employers

who had initially agreed to participate in the ERP could b e regarded as full participants in 1982.

About two-fifths (weighted estimate) of the sample employers appearing in SSA's Multi Unit Code File had not filed Forms 5019. Under certain circumstances, employers who appear to be filing Forms W-3 for more than one reporting unit are added to the Multi Unit Code File, even though they have never filed Form 5019. in such cases, the employers may be contacted to solicit their participation in the ERP. However, after some investigation of

these cases, it was concluded that most of the 37 sample employers in this group had been incorrectly added to the Multi Unit Code File. The procedural error which led to these erroneous additions was subsequently corrected.

6. Results of matching BLS and SSA industry codes to IRS industry codes

For the BLS/SSA matched single unit employers, the industry codes from both systems were compared with the industry codes in IRS's Form 941 file for tax years 1981-83. The comparisons were made only at the two-digit SIC level, primarily because the IRS industry coding structure does not provide full detail at the three and four-digit levels. The comparisons were made only for single unit employers (as defined for ERUMS) because the IRS system does not provide separate codes for establishments or reporting units of multi unit employers.

The results of the comparisons are shown in Table IIIA-8. They cover 164 of the 167 matched BLS/SSA single units: the remaining 3 sample EINs were not included in the IRS Form 941 file. There were some cases for which the SSA or IRS systems, or both, did not have an industry code. These are shown separately in the table. There were no sample EINs in this group for which BLS did not have an industry code.

About three-fourths of both the SSA and BLS codes matched the IRS codes at the two-digit level. However, looking only at the

cases with no missing codes in one or both systems, the proportion of matches was somewhat higher for the SSA industry codes: 89 percent versus 79 percent for the BLS codes.

Given the small sample size and other limitations of the ERUMS design, it would be improper to suggest any definitive explanations for the results shown in Table IIIA-8. Keeping in mind that some of the codes in the IRS Form 941 file came from Census sources (see "The Tax Years 1981-83 Form 941 File" in Chapter II, Section A,3) and that some of the Census codes may have come initially from SSA, one possibility is that there is a greater degree of independence between the BLS and IRS sources of industry codes that there is between SSA and IRS code sources.

The non-matches at the two-digit level may have resulted in part from the fact that the single unit definitions used for ERUMS were based only on reporting units in Texas. Thus some of the "single unit" employers included in the comparison may have had additional reporting units in other States, and the IRS industry codes may have reflected a predominant activity that differed from

the one carried on by their Texas reporting units.

Table IIIA-1 - Distribution of EINs by final match status

Active in:		No. of sample EINs	Weighted percent
UI	SSA		
Yes	Yes	279	67.1
Yes	No	32	5.3
No	Yes	90	27.6
Total		401	100.0

Definitions of active:

UI - Included in UI address file for 1st quarter 1982.

SSA - Submitted Forms W-2/W-3 for 1982.

Table IIIA-2 - Distribution of active 1/ BLS EINS by final match status

BLS category	SSA category	No. of sample EINs	Weighted percent of BLS category
Single unit	Single unit	167	92.6
	Multi unit	1	0.1
	No 1982 W-3 2/	22	7.3
	Total		100.0
Multi unit	Single unit	102	88.3
	Multi unit	9	6.7
	No 1982 W-3	10	5.0
	Total	121	100.0
All active	Single unit	269	92.6

BLS EINs	Multi unit	10	0.1
	No 1982 W-3	32	7.3
	Total	311	100.0

Definitions:

BLS multi unit - two or more reporting units in Texas.

SSA multi unit - wage reports for two or more Texas reporting units clearly identified from 1982 Forms W-3.

BLS and SSA single unit - any EIN not meeting multi unit definition.

Footnotes:

1 - For definition of active, see Table IIIA-1.

2 - Includes two EINs not found in any SSA files.

Table IIIA-3 - Distribution of active SSA EINs by final match status

SSA category	BLS category	No. of sample EINs	Weighted percent of SSA category
Single unit	Single unit	167	70.0
	Multi unit	102	0.8
	Not in UI file	90	29.2
	Total	359	100.0
Multi unit	Single unit	1	39.6
	Multi unit	9	60.4
	Not in UI file	-	--
	Total	10	100.0

All active	Single unit	168	69.9
	Multi unit	111	0.9
	Not in UI file	90	29.2
	Total	369	100.0

For definitions, see:

Active - Table IIIA-1

Single and multi unit - Table IIIA-2

Table IIIA-6 - Distribution of EINs not in 1982 UI File by 1982 IRS/SSA status

Status in IRS/SSA system	Number of EINs in sample	Weighted share of total (%)

No IRS employment reported for first quarter 1982	43	69.2
IRS employment reported for first quarter 1982	47	30.8
Geographic location unclear in IRS/SSA system*	33	3.3
UI coverage unlikely, based on IRS/SSA data**	4	7.8
All others***	10	19.7
TOTAL	90	100.0

* Mainly multiunit employers that did not supply enough information on their 1982 W-3 reports to determine if

they had active units in Texas.

** Because of nonprofit status or small payroll.

*** Includes employers incorporated into the UI system with a lag, or not at all.

Table IIIA-7 - Status of SSA employers 1/ included in the Multi Unit Code File (MUCF)

Category	No. of sample EINs	Weighted percent of: Total	With Form 5019 and filed W-3 for 1982

With Form 5019	88	61.4	
W-3 for 1982	84	60.6	100.0
Filed as multi, all codable 2/	15	10.8	17.9
Filed as multi, other	7	3.8	6.3
Filed as single	62	46.0	75.8
No W-3 for 1982	4	0.8	
No Form 5019	37	38.6	
Probable multi unit	3	2.6	
Incorrectly added to MUCF	34	35.9	
TOTAL	125	100.0	

Notes:

1 The sample for this table represents Texas employers who were included in SSA's Multi Unit Code File and were active in BLS and/or SSA systems in 1982.

2 "Codable" employers are those for whom each unit reported on Form W-3 had an establishment number that corresponded to one appearing in the Multi Unit Code File, so that industry and county codes could be assigned.

Table IIIA-8 - Distribution of matched BLS and SSA single units by result of match of their SIC codes against IRS's at the two-digit level*

Source of SIC code matched to IRS	Result of match to IRS code: Match	Non-match	One or both codes missing	Total

SSA

No. of cases	118	20	26	164
Weighted percent				
All cases	76.7	9.6	13.7	100.0
Cases with no missing codes	88.9	11.1	n.a.	100.0

BLS

No. of cases	113	47	4**	164
Weighted percent				
All cases	77.2	20.9	1.9	100.0
Cases with no missing codes	78.7	21.3	n.a.	100.0

* As explained in the text, some of the SIC codes in the IRS records came from Census Bureau sources. See "The Tax Years 1981-83 Form 941 File" in Chapter II, Section A,3.

** The BLS records for the sample cases had no missing SIC codes. IRS did not have SIC codes for these 4 cases.

B. Limitations of the ERUMS design and execution

We feel that the ERUMS project has provided valuable information about the BLS, IRS and SSA record systems that were linked, especially the characteristics of the systems that have a bearing on their uses for economic statistics programs. The difficulties encountered and the ways in which they were at least partially overcome have, we believe, important implications for future initiatives in the area of interagency data sharing for statistical purposes.

It is important that readers be aware of the limitations of the study results and of the significant problems that inhibited some of the analyses that we had hoped to do. Therefore in this section we will describe: features of the study design that impose restrictions on how far the findings can properly be generalized; interagency differences in record system coverage and content that complicated the analysis of the matched units; and operational problems encountered in the course of the study. We hope that this information will lead to a better understanding of what the results mean and will be helpful to anyone designing similar record linkage projects.

1. Limitations on the generality of the study findings

Factors that limit the broad applicability of the ERUMS findings are the time reference, the limited geographic coverage and the relatively small sample size.

The study was based primarily on administrative and statistical files from the three agencies for calendar year 1982. The results reflect the reporting requirements and the operating and quality assurance procedures associated with the agency record systems at that time. As explained in Chapter II, Section A, BLS is presently shifting from an annual to a quarterly update procedure for the UI Address File, and additional identifiers are being included in the file. The BLS's future plans call for a shift from a reporting unit to an establishment record system, with physical location addresses for both single and multi unit employers.

With respect to the IRS records based on Form 941, there is evidence of improvement since 1982 in the completeness and accuracy of the records because of a shift to scanning of paper documents and filing on magnetic media as alternatives to keying data from paper documents. It is likely that these trends are also leading to improvements in the SSA files based on Forms W-2 and W-3. On the other hand, in view of the limited resources available in recent years for updating and maintenance of SSA's Establishment

Reporting Plan (ERP) records, it is quite possible that the quality of SSA reporting unit data is even less satisfactory today than it appears to have been in 1982.

Because of the Federal/State character of the UI program, the BLS records used in the study had to be limited to those available from a single State, the State of Texas. This restriction limits the generality of the findings with respect to BLS and State records maintained for the UI system. The Department of Labor imposes certain guidelines that must be followed by all States, but the States are also allowed some latitude in their record-keeping practices. Conclusions, based on the study results, concerning the coverage of employers and reporting units in the record systems of the Texas State Employment Agency in 1982 should not be assumed to apply fully to UI record systems of other States at that time.

The use of BLS records for a single State also meant that for the ERUMS project it was not possible to identify, in the UI system, employers who reported only one reporting unit for the

State of Texas, but reported one or more additional units in other States. For this reason, we eventually reached the conclusion that our analysis of multi unit employers in the UI system had to be based on a very restricted definition of multi unit, namely, an employer with two or more reporting units included i n the Texas UI Address File for 1982. As explained more fully in Chapter II, Section C, this meant that many SSA employers originally classified as multi units because there were records for them in SSA's Multi Unit Code File were reclassified as single units prior to the final analysis, in order to make the SSA definition of multi unit comparable to the one used for BLS. Thus, the data for multi unit employers shown in the results tables apply only to a very restricted subset of those employers who would be classified as multi unit in a national context.

Finally, a relatively small sample of EINs was used for the ERUMS study because of the limited resources available for manual review of matched and unmatched records from the three systems. The use of a disproportionate stratified sampling scheme, based on preliminary classification of sample EINs by single/multi unit and

match status, made the effective sample size for overall estimates even smaller. As a result, all of the estimates shown in the results tables that are based on the Phase II sample are subject to fairly large sampling errors. Sampling errors are shown in Table IIIA-4 for the estimated proportions of EINs in each of the final match categories. Because of the complexity of the sample design and the limited resources available, sampling errors were not computed for the other estimates based on the Phase II sample.

2. Interagency differences in concepts and coverage

As explained in Chapter II, Section A, there are some differences in the basic filing requirements for employers under the Unemployment Insurance and Social Security programs. Indeed, these differences explain some of the instances in which we found employers in one of the two systems who were not covered by the other. The discussion here will be limited to those features of

the UI and SSA record systems that have been established primarily for statistical purposes.

First, the time reference of the basic BLS and SSA files used for matching was different. The SSA W-2/W-3 files included employers who had filed wage reports covering all or any part of the entire year 1982. Although there were some exceptions, the 1982 UI Address File was supposed to include only employers who had filed UI wage reports for the first quarter of 1982. As discussed in Section A of this chapter, this difference in time reference accounted for some of the cases in which employers reporting to SSA were not found in the UI Address File. The quarterly-Form 941 data obtained from IRS for the Phase II sample EINs were helpful in determining the reasons for failures to match.

Second, the reporting unit definitions used by BLS and SSA in their respective systems, although similar, are not identical. Basically, the reporting unit in each case is a single establishment or a group of two or more establishments under the same employer in the same county and industry. However, there are

subtle differences in the two agencies' definitions and in the manner in which they are applied (for further details, see Office of Management and Budget, 1984 and Jabine, 1984).

Third, there are minor differences in the structures of the BLS and SSA industry classification systems. Both are based on the Standard Industrial Classification (SIC), but both agencies have grouped certain of the approximately 1,000 four-digit industry categories in the SIC. The amount of grouping is. somewhat less in the BLS system. As a result of these differences in code structures, even in the absence of reporting or coding error the two agencies do not always assign the same code to an employer or reporting unit. (For further detail see Chapter II, Section A and the two references given above.) Because of these differences, we have limited the analysis of industry codes for matched EINs to comparisons at the two-digit SIC level.

3. File deficiencies and operational problems

The only significant problem found in using the Texas UI Address File was that 1.3 percent of the records did not have EINs and therefore could not readily be included in the initial selection of a sample of BLS records. In the final review of unmatched SSA sample cases, a search was made for records in the UI file, including those with no EINS, that had matching addresses, but no additional UI matching records were found. The absence of EINs could have affected the classification of BLS sample EINs as single or multi unit. This could have happened if an employer had two or more reporting units in Texas but the EIN was shown in the UI Address File for only one of these units. it is not known whether or not this actually occurred.

In the SSA files, several employers lacked industry and county codes. For the latter a "State-wide" code was substituted for the county code. These missing codes are reflected in Tables IIIA-5, A-3 and A-4, which show the results of comparing BLS and SSA industry codes for matched single units. A probable explanation

for many of the missing industry codes is that some employers report wages to IRS on Forms 941 or 943 without ever having applied for an EIN on Form SS-4. In such cases an EIN is assigned by IRS and no attempt is made to obtain the industry and geographic information that is normally reported on Form SS-4.

Another difficulty was the incomplete coverage of the SSA file used in the initial electronic match (in mid-1986) to determine which sample EINs were active in 1982. It did not include some employers who were delinquent in filing their wage reports for 1982 or whose wage reports were still being reconciled with the amounts of wages reported on their IRS Forms 941. A more complete W-3 file was available for the review of the Phase II sample, but the delay in access to records for this subset of active SSA employers may have some implications for statistical uses of SSA wage data when timeliness is an important consideration.

A more serious problem was that, for many active SSA employers included on the Multi Unit Code File, it was not possible to determine from the W-3 reports how many reporting units, if any, they had in the State of Texas. In some cases there were no

establishment numbers associated with W-3 wage report listings for these employers. In other cases some or all of the establishment numbers shown on the W-3 listings did not appear on the Multi Unit Code File, so that there was no way to determine the States in which these reporting units were located. In the final analysis, all such employers were classified as single unit, even though many of them may have actually had two or more reporting units in Texas.

The retrieval of information needed for the ERUMS project from IRS Form 940 was difficult for two reasons. First the information of interest, the breakdown of taxable payroll by State in Part V of Form 940, is not keyed by IRS, so we had to request copies of the forms, which are filed in the IRS service centers. Second, by the time we requested the forms for the Phase II sample cases early in 1988, many of the Forms 940 for 1982 had been destroyed, so we were given copies of 1983 forms instead.

The electronic Patching and subsequent sampling operations were made more complex by the confidentiality policies which led BLS to release to SSA only the EIN stems (digits 1 to 6 and 9)

instead of the full 9-digit numbers for the employers in their Phase I sample. Probably because of this complexity, some SSA sample nonmatch cases were unintentionally excluded from the Phase I and II samples, as explained in Chapter II, sections C,4

and C,11. A potentially biased reweighting procedure was used to compensate for the exclusion of these cases.

The BLS and IRS (Form 941) employment data were not added to the data base for the Phase II sample, so we have not been able to compare BLS and IRS reports of employment for matched cases. As -a consequence, all of the estimates presented in this report are counts of employers or reporting units, with no indication of their relative sizes. Thus, in comparing industry codes for matched single unit employers, we were unable to estimate what proportion of their total employment or payroll belonged to units for which the industry codes did not agree at the two-digit level.

CHAPTER IV--FINDINGS AND RECOMMENDATIONS

A. Findings

1. Relative coverage

The ERUMS sample match rates suggest the possibility of significant coverage problems in both the IRS/SSA and UI payroll tax systems. After reclassifications, 28 percent of the employers with, evidence of 1982 activity in either or both of the two systems had no evidence of activity in the 1982 UI file used in the match, while 5 percent of the employers had no evidence of activity in the 1982 SSA files used. The good news is that more detailed analysis does not suggest that large numbers of employers who report wages in one of the payroll tax systems are failing to report wages in the other system when they should be. The not-so-good news is that late employer reports and different procedures for processing the reports in the two systems create potential

problems in using both of the systems' data files for statistical purposes.

At the initial matching stage, it appeared that over twice as large a proportion of a I employers failed to file SSA W-3 wage reports as was found after reclassifications occurring at the final stage. The reclassifications were made because the SSA file used at the first stage of matching did not include some employers who were later found to have filed delinquent reports or reports that had been pulled from the normal processing cycle because of difficulties in reconciling the W-3s with IRS Forms 941. The employers with no W-3s for 1982 after reclassification appeared to be mainly employers who were going out of business or were going through some kind of reorganization that might have been accompanied by an EIN change in the IRS/SSA system. Many of these cases also dropped out of the UI reporting system shortly after dropping out of the IRS/SSA system. The tendency for employers to be dropped more slowly from,the UI system is probably a result of a policy of estimating employment and payroll for employers who appear to be late in filing their quarterly reports--pending verification of the reason for failure to file on time. Employers

who may have had an EIN change because of some type of reorganization (e.g., incorporation) were difficult to identify with certainty in either the IRS/SSA or UI systems. In the early 1980s, the IRS/SSA system provided no systematic basis for tracing such EIN changes for small employers, and because the UI number rather than the EIN is the primary identifier in the UI system, EIN changes for employers in the UI Address File will not necessarily be introduced into the UI records in a timely manner.

A comparison of IRS Form 941 quarterly,employer records with the sample cases that were not on the 1982 UI Address File (but had filed annual W-3s for 1982) revealed that about 70 percent either started business after the first quarter (on which the UI

Address File was based) or otherwise did not file first quarter wage reports. For another group of the sample cases (about 10 percent), IRS/SSA data suggested the possibility that the employers may not have met requirements for UI coverage in Texas either

because they had no operations in Texas, because of nonprofit status, or because their payrolls were too small. (See Chapter II, Section A,l for detailed information about coverage requirements of the UI system.) For the remaining cases (about 20 percent), there was some indication of presence in the UI system for time periods other than the first quarter of 1982. Some of these latter cases would appear to represent employers that were incorporated into the UI system with a lag, but further research would be needed to clearly separate any problem of lagged introductions from other reporting and processing problems in the UI and IRS/SSA systems.

Although a few new employers obtain UI account numbers before they hire their first employees, there is usually some lag' between the time the first employees are hired and the issuance of a UI account number to the employer. A study for the State of New York, covering new employers in the file as of June 1987, showed that the median lag was about I month and that about 90 percent of employers had received UI accounts within 5 months of the time the first employees were hired (Grzesiak and Lent, 1988). A similar study by the Montana Department of Labor and Industry (1987) for the period

June 1984 to June 1986 showed that the lag was 90 days or less for about two-thirds of the new employers during this period.

There is, of course an additional lag between the time an employer receives a UI account from the State agency and the time the employer's name and identification data are submitted to BLS for inclusion in its UI Address File. Now that BLS is requiring the States to submit new inputs for the Address File each quarter, rather than once a year, the average length of this lag time will be much shorter than it was during the time period covered by the ERUMS project.

2. Multi unit employers: acquisition and updating of reporting unit information

The clearest finding of the ERUMS study is that it is not possible to maintain a usable establishment reporting plan for multi unit employers without systematic procedures for monitoring employer reporting and updating files for changes in the number,

location, and industry of employers, reporting units. Since both SSA and BLS request that multiestablishment employers break their, employment and payrolls down by reporting units on a similar county/industry basis, it might seem logical that employers would find it convenient to organize their establishment reporting so that they use the same reporting units for both systems. But there was little evidence that employers tried to do this, and to the extent that they did, the lack of systematic procedures in the SSA system for monitoring changes in the number, location,

etc., of employers, reporting units made it very difficult to interpret many of the employer reports filed under SSA's ERP. The extent of the problems is reflected in the fact that in the SSA system only 0.1 percent of the EIN's (weighted estimate) could be clearly identified as reporting for multiple reporting units in Texas, compared to 0.9 percent in the UI system. Even among the small number of employers who did have multiple Texas units in the SSA system and could be matched to employers in the UI system, the SSA system listed only about half as many identifiable Texas

reporting units as the UI system.

Because of the deficiencies in SSA's ERP, it was not possible to use the small ERUMS sample to identify,problems in the UI establishment reporting system. But the problems in the SSA system are not good news for other establishment reporting plans. Good establishment reporting requires considerable effort. while the UI system employs more resources in monitoring and updating its system than does SSA, the UI system is unlikely to be perfect in an environment in which the reporting requirements of different agencies are administered in quite different ways and may, taken together, appear to be confusing and burdensome to many of the multi unit employers filing payroll tax reports.

3. Content differences for matched units

It was not possible to make meaningful comparisons between employment and payroll from the first quarter UI Address File with employment and payroll measures from the annual W-3 wage reports

submitted to SSA. As noted earlier, employment data from first quarter IRS Form 941 reports were used in refining . estimates of coverage differences between the UI and IRS/SSA systems. However, we did not try to do a thorough comparison of UI and Form 941 employment measures because Form 941 employment is not available for the separate units of multi unit employers.

For single unit employers who matched on EIN and were active in both the UI and SSA systems, the content analysis focused on comparisons of county and industry codes. There was a moderately high, but far from perfect, rate of correspondence between codes in the two systems (about 80 percent agreed on county and about 70 percent agreed on 2-digit SIC). A significant share of cases that did not agree on county and industry codes were cases that had "statewide county codes and "unclassified" industry codes in the SSA system. In most of these cases, employers had apparently been assigned EINs without filing a Form SS-4 to apply for an EIN, and supply the information needed for coding county and industry. SSA has not had the resources to follow up with these employers, or with other employers who supply incomplete information on location

and industry. In addition, SSA does not have a program for updating its county and industry codes on a regular basis as is done for the UI system. Thus, the content differences between the UI and SSA systems with,respect to county and industry codes for matched single unit employers would seem,to reflect largely

the lack of resources f or thorough coding and updating of codes in the SSA system.

The SSA and BLS two-digit industry codes for the single unit matched cases were also compared with two-digit industry codes in a file that had been initially created from the IRS Business Master File and Forms 941 and 943 data and for which the industry codes had been edited by the Census Bureau, using the source of information that Census considered to be the most reliable in each case. Both the BLS and SSA industry codes matched the IRS/Census codes at the two-digit level for about three-fourths of the sample EINS, but when the unclassified SSA cases were eliminated, the rate

of agreement was somewhat higher for SSA than for BLS industry codes.

In summary, there were some fairly substantial discrepancies among the BLS, SSA, IRS and Census systems with respect to geographic and industry classification for matched single unit employers. These findings do not provide a basis for evaluating the relative levels of accuracy of such information in the four data systems. Such an evaluation would have required that we reconcile the differences and determine the correct information in each case, and we did not have the resources to do that.

4. The role of IRS records in the matching process

IRS is not involved directly in either the SSA or UI establishment reporting plans for payroll. But both the social security and unemployment insurance payroll taxes are closely linked to IRS payroll taxes from an administrative perspective. Specifically, IRS wage-withholding taxes for the individual' income tax are reported using the same (quarterly) Forms 941 and (annual)

W-3/W-2 wage report forms used to report Social Security payroll taxes, and the IRS Form 940 is used to collect information on payrolls and UI taxes paid by State as a part of the process of determining federal unemployment insurance tax obligations.

IRS records were vital in evaluating the SSA and UI reporting systems because the SSA establishment data (from W-3 forms) were available only for the calendar year 1982, while the UI establishment data (from the name and address file) were available only for the first quarter of 1982. As noted above, quarterly IRS Form 941 data were I used directly to determine which of the employers who filed annual W-3 forms, but were not 'included in the first quarter UI file, had reported first quarter wages in the IRS/SSA system. It was hoped that the annual IRS Form 940 wage reports could be used to determine how many of the employers not included in the first quarter UI file had reported UI taxes for the calendar year. Unfortunately, however, the request for the Form 940s was filed near the end of the scheduled retention period for 1982 returns and it was not possible to obtain 1982 Form 940s for enough employers to reliably resolve

discrepancies between the annual SSA data and the quarterly state UI data.

The administrative procedures for assigning EIN's and for processing payroll tax records in the IRS/SSA systems were found to have some deficiencies from a statistical perspective. For example, a significant number of EINs appear to have been assigned directly (usually by IRS) with no request for the employers in question to submit Forms SS-4 or to otherwise submit information suitable for coding county and industry and determining if the employers were just beginning operations or had merely gone through a reorganization that required assignment of a new EIN. In addition, as noted earlier, a relatively large number of employers appeared to be missing W-3 wage reports in the initial match phase of the project, but were subsequently reclassified when a more thorough search of IRS and SSA record systems located,their W-3s. The fact that a record of 'all wage reports submitted could not be found in a single convenient data file suggests that greater

coordination between statistical and administrative users of the records may be required if reasonably complete data files are to be created for statistical uses. Finally, although it was not possible to use the IRS Forms 940 in a systematic manner in the study, the fact that much of the Form 940 information needed for comparisons with the data from the other systems was no longer available in any form suggests scope for improving coordination between the Form 940 system and both the state UI systems and the national W-3/Form 941 systems.

5. Feasibility of interagency matching of employer and establishment records

The feasibility of interagency matching of business records depends on the purposes and scope of the proposed linkages. A small-scale matching study for evaluation is quite different from a large-scale operational system for using records from various sources to develop and maintain frames for economic surveys. Much depends on the characteristics of the particular agency record

systems and files that provide the records to be matched. Our main findings about feasibility are specific to the purposes and scope of the ERUMS project and the agency record systems that were used. However, we believe that we also learned some lessons that could be useful for other kinds of matching activities.

Our most important finding concerns the types of u 'nits to be matched. With some qualifications, we were successful in linking data for employers, as defined by their EINS. We were not successful in linking BLS and SSA data for reporting units.

The main reason for our inability to match records for reporting units was the incompleteness of SSA's current data for reporting units provided on the W-3 wage reports. Other significant problems were:

- BLS and SSA do not have a common numerical identifier, analogous to the EIN at the employer level, for reporting units. Consequently, matching requires the use of other

identifiers, such as name and address, and auxiliary information, such as county code, industry classification, employment and payroll.

- BLS and SSA use slightly different definitions of reporting units.

Thus, even if the SSA reporting unit information had been complete and current, matching with BLS reporting units would have been difficult and costly, at best.

Other findings specific to the ERUMS experience were:

- Matching national files in one system against State files in another system leads to problems of coverage and interpretation of findings. In particular, the fact that we had no information, for the sample employers, about their BLS reporting units in States other than Texas forced us to use a restrictive, nonstandard definition of multi unit employer for the study. Also, we were not in

a position to determine whether any of the observed differences in coverage could have been explained by multi unit employers reporting on units physically located in Texas to State Employment Security Agencies in other States. We note, however, that with complete reporting of,EINs in the UI system, the development of a national employer file based on inputs from the UI State agencies would be possible.

- The IRS Form 941 file, which provided quarterly data on employment and payroll for 1981, 1982 and 1983, proved to be very useful as an aid in reconciling differences in coverage between the BLS and SSA files for 1982.

- In planning such a study, one needs to consider the period of availability for each file or set of records to be used: when does it first become available and after what date is it no longer available? As noted in Chapter II, Section C, the W-2/W-3 wage report file first used for matching was incomplete: several EINs initially

classified as not active in SSA were reclassified to active when a more complete file became available. Also, because our request to IRS for copies of Forms 940 for the Phase II sample employers was not made until late in the project, we found that most of the Forms 940 for 1982 had already been destroyed.

- As explained in Chapter II, Section C, the initial matching operations at SSA were carried out using only 7 of the 9 digits of the EINs in the BLS Phase I sample. This restriction was deemed necessary in order to meet BLS's confidentiality requirements. The result was a substantial increase in the complexity of the sample selection and matching operations and, as it turned out, the inadvertent omission of a significant group of in scope EINs from the Phase I sample. The experience suggests that the use of such special procedures be limited to those which are clearly essential to meet

agency confidentiality requirements and that all specifications and procedures be fully documented and carefully reviewed.

There have been and are many other examples of matching,of business records for statistical purposes, especially for the construction of frames for statistical purposes, a notable example being the use of IRS and SSA records by the Census Bureau in the development of its frames for economic censuses and surveys. Some are large-scale ongoing activities and it would be presumptuous to suggest that the limited ERUMS experience can offer important new insights on how to carry out such activities. Nevertheless, there are a few points that may be worth emphasizing:

- Matching activities are better carried out prospectively, i.e., the plans and the necessary interagency agreements should be developed well ahead of the earliest date at which the files to be linked are expected to be available. This is important for both research-oriented and operational matching activities, but especially the latter.

- The development of interagency agreements for exchange of identifiable records is a painstaking process. It requires: identification of all laws and regulations that may affect the proposed exchange; identification of all persons who will examine or process data from another agency; and development of a step-by-step description of each and every transfer or exposure of information called for by the proposed matching activity. Adequate time must be allowed for the completion and approval of such agreements.

- Successful matching requires an in-depth knowledge of all record systems involved and of the specific files that are generated from those systems. Usually no one person has all of this information and an interagency team approach, with full exchange of information, is essential. Whenever possible, procedures should be pretested or pilot tested before. embarking on largescale operational applications.

B. Recommendations

1. Introduction

This report contains the basic findings of the ERUMS project and it is the desire of the workgroup and the Federal Committee on Statistical Methodology that this information and the accompanying recommendations be put before the statistical community now. With its limited resources, the Workgroup has not been able to exploit the ERUMS data base as fully as it would have liked to. However, the main goals of the project were achieved and we believe that whatever resources can be made available for future matching studies should be devoted to prospective studies using currently available business lists.

The findings of the ERUMS project have confirmed the importance of earlier recommendations by the Subcommittee on

Statistical Uses of Administrative Records (1980) in Statistical Policy Working Paper 6 and the Establishment Reporting Work Group (Cartwright, Levine and Buckler, 1983). As stated in Chapter I, ERUMS represented an effort to build on and extend the work of those two interagency groups. Specifically, the ERUMS project was responsive to Recommendation 2 in Working Paper 6:

> The quality of administrative records to be used for statistical purposes should be evaluated systematically to determine the appropriateness of the records for the proposed use.

ERUMS was, of course, limited in its scope and objectives. It was a demonstration project designed to show how matching of administrative records from different agencies could provide a basis for evaluation of their suitability for statistical uses. Nevertheless, it is the ERUMS Workgroup's view that the study findings, in combination with related information from other sources, provide adequate justification for the recommendations presented in this section.

Most of our recommendations, presented in Subsection 2, are directed specifically to BLS and SSA and concern the administrative and statistical business lists maintained-by those agencies. A single recommendation concerning future matching studies is presented and discussed in Subsection 3.

2. Recommendations to SSA and BLS

The recommendations in this section refer to the SSA and BLS systems for the collection of economic data at the establishment or reporting unit level. We are conscious of the limitations of the ERUMS study with respect to coverage and sample size and, especially, the fact that the findings refer to the status of those systems eight years ago, in 1982. BLS, as described in Chapter II, Section A, has made and is making various changes

designed to upgrade the quality of statistical data based on the UI

system. SSA, on the other hand, appears to have-done little, since 1982, to improve the quality of reporting under its Establishment Reporting Plan (ERP) or even to maintain quality at the 1982 level, which was clearly unsatisfactory. The following recommendations attempt to take account of the current status of these systems, insofar as we are aware of it.

Recommendation 1-- SSA should undertake a full review of the current status and uses of the Establishment Reporting Plan and decide either to continue it with adequate resources for maintenance and improvement of quality or to discontinue it entirely.

The level of compliance with the ERP is so low that it is clearly of little value for its intended uses. If continued at this level, it would represent an unjustifiable burden on those employers who continue to participate.

Discontinuance of the ERP would affect the level of detail available for code individuals by industry and geography in SSA's

Continuous Work History Sample (CWHS). Industry could continue to be coded,but in a single unit context. County codes based on ERP reporting unit locations could be replaced by county codes based either on W-2 addresses or on taxpayer addresses in the IRS individual master file, provided the necessary arrangements could be worked out with the IRS.

The ERUMS Workgroup has been informed that a full evaluation of the ERP is now underway. We strongly support the undertaking. We suggest that the review include interviews with a small sample of multi unit employers, including some who have not been reporting usable establishment-type data. The interviews should explore employers' reasons for noncompliance or incomplete or incorrect reporting under the ERP, as well as their interest in the development of greater uniformity in establishment reporting standards of SSA, the UI system, the Census Bureau and other agencies that collect disaggregated data from employers.

We noted in Section A of this chapter that a substantial proportion of SSA single unit employers in Texas lacked industry

codes. For, some of these cases no Form SS-4 (application for an EIN) was ever obtained by SSA and for some no industry code could be assigned on the basis of the information on the SS-4: SSA as made some attempts to obtain industry information by mail from larger, active employers in this group, but with limited success. If SSA decides that it wishes to continue maintaining industry information for all employers, greater efforts will be needed to reduce the proportion of employers whose industry is unknown.

With respect to the UI Address File, the main problem we found, based both on the ERUMS comparisons with SSA and IRS records and on the more recent New York and Montana studies cited in Section A of this chapter, was the delays in adding births to

and deleting deaths from the system. To the extent that the UI Address File is being used as a frame for sample surveys at the national or state level, the delays in adding births are more likely to have the more serious consequences. The lag question

will assume added importance if, as has been proposed by OMB, the BLS is designated as the single Federal agency responsible for the collection of business identification information for the nonagricultural sector of the economy.

Recommendation 2 - BLS should review the State Employment Security Agencies, procedures for identifying employer Births (including those resulting from mergers and changes of organization) and seek ways of reducing the apparent lag between filing of applications for EINs and inclusion of new employers on State Agency and BLS lists used as frames for statistical surveys and reports.

We note that the new requirement that states submit UI Address Files to BLS for each quarter is one step in this direction.

As discussed in Chapter III, Section A,4, delays in deleting deaths from the UI Address File were apparently due in part to the States' practice of imputing employment and payroll for employers who appear to be late filing their quarterly reports.

Recommendation 3 - Data in the UI Address File on employment and wages paid should be labelled to distinguish imputed data from data reported by employers.

We have been informed that as of the first quarter of 1989, 40 states had adopted this practice. A related issue which needs to be considered is whether the actual data for these employers, when available to the States, should be submitted to BLS to replace the imputed data in its files.

We also noted that slightly more than one percent of the records in the 1982 UI Address File for Texas did not have EINS. The absence of EINs could cause problems for linkages of data for the same employer between states within the UI system or for any linkages with other systems that might be undertaken.

Recommendation 4 - The EIN should be identified as a key item in the UI Address File and efforts should be made to achieve 100 percent reporting initially and current reporting of changes in EINS.

We have been informed that BLS has put increased emphasis on complete reporting of current EINS.

As noted in Chapter I, the reporting unit definitions used by BLS and SSA are similar but not identical. Under its new Business Establishment List project, the BLS will be moving toward the collection of establishment-level data, using the OMB definition of establishment. We have also noted that BLS and SSA

use somewhat different adaptations of OMB's Standard Industrial Classification for their own classification of employers and reporting units by industry.

Recommendation 5 - BLS and SSA (if it continues the Establishment Reporting Plan) should strive to obtain data from employers for their establishments as defined in the 1987 Standard Industrial Classification (SIC) Manual. Both

agencies should code industry for all establishments, without exception, at the 4-digit SIC level of detail. Whether or not the Establishment Reporting Plan is continued, SSA should code all employers identified on Forms SS-4 at the 4-digit level of detail.

Implementation of this recommendation would be consistent with the broad recommendation in Working Paper 6 for agencies to follow consistent procedures in coding reporting unit characteristics (Subcommittee on Statistical Uses of Administrative Records, Office of Federal Statistical Policy and Standards, 1980, Recommendation 3).

The goals of BLS's Business Establishment List Improvement Project, which is being implemented, include obtaining reports at the establishment level from all employers and elimination of the present limited number of 3-digit coding exceptions (Chapter II, Section A,I).

3. Future matching studies

The collection of economic data at the establishment level is an important function of the Federal statistical system and of state statistical units. Current efforts to collect such data are dispersed and poorly coordinated and place unnecessary burden on employers. In particular, the inability of Federal and state agencies to share business lists for statistical purposes is a well recognized problem of long standing (American Statistical Association, 1980). Many of the establishment-level data collection programs, including those associated with the Unemployment Insurance system (in some states) and W-2/W-3 wage reporting, are voluntary.

It is also important that the overall reporting burden on employers, for both administrative and statistical purposes, be held to a minimum. The SSA's strategic plan for the year 2000 calls for exploration of:

> ... the possibility of replacing the existing employers, wage reporting requirements with agreements by which the states

would share with SSA, through electronic media, the wage data reported by employers for unemployment insurance and program purposes. (Social Security Administration, 1988)

In exploring this possibility, and any other proposed changes in administrative reporting systems, it is essential not to lose

sight of the statistical requirements of SSA, BLS and the State Employment Security Agencies, as well as any other statistical programs that may be linked to or in any way depend on the unemployment insurance and employer wage reporting systems.

The ERUMS workgroup believes that further, more intensive and extensive interagency matching studies have an important role to play in resolving the difficulties cited above and in determining the possible effects on statistical programs of prospective major changes in, administrative reporting systems for employers. The design of such studies will be helped by agreement on and adherence

to a set of basic goals.

Recommendation 6 - Further matching studies should be directed at acquiring information that will support the eventual development of a mandatory reporting system to meet the needs of all federal and state statistical programs for establishment lists, including SIC codes. An interim goal should be that all agencies requiring or requesting employers to provide data at the establishment or reporting unit level adopt common definitions of units and data items to be submitted for these units.

To the extent possible, such a reporting system should derive most of its information from the major administrative reporting systems. All supplemental information required for statistical purposes should be collected as part of a fully integrated program, using concepts and definitions agreed on by all users.

Three agencies -- the BLS, the Census Bureau and the National Agricultural Statistics Service -- play a dominant role in the

direct collection of establishment-level economic data. Recent initiatives of these agencies, under the general guidance of OMB's Statistical Policy Office, have been directed at greater coordination of their respective list-building and maintenance activities. Further integration of business lists will require fuller understanding of the similarities and differences of the three systems, based on matching of individual establishments and reporting units in the different systems.

To be successful, future matching studies will require the full-time efforts of staff members from each of the agencies involved and provision of adequate support facilities and funding. It will be essential to have the 'cooperation of, the major suppliers of administrative lists: IRS, SSA and the State Employment Security Agencies.

Based on the ERUMS experience, present statutes, regulations and policies of the agencies involved are likely to present obstacles to the timely conduct of future matching studies. The ERUMS project has demonstrated that carefully constructed

interagency agreements can make it possible to conduct limited matching studies, and it is probable that some additional studies

could be conducted under similar arrangements; however, the Workgroup feels that certain studies may require changes in the relevant statutes and regulations.

The employer identification number (EIN) plays an important role in economic statistics programs. It is a key identifier for matching records from different systems. Application for an EIN is often the first indication of the existence of a new employer, and the application form (SS-4) provides initial information about the characteristics of the new employer. Existing employers frequently apply for new EINs as the result of changes in type of organization or corporate. reorganizations.

The EIN issuance procedures in effect during the reference period for ERUMS did not provide any reliable method for statistical agencies to track such changes. The current version of

Form SS-4 (adopted in August 1988) asks whether the applicant has previously applied for an EIN for the current or any other business and, if the answer is yes, to provide that EIN. This new information is potentially valuable for use in updating business lists and should be exploited for that purpose.

REFERENCES

American Statistical Association

1980 "Business Directories: Findings and Recommendations of the ASA Committee on Privacy and Confidentiality". The American Statistician, 34:8 10.

Buckler, W.L.

1985 "Employer Reporting Unit Match Study (ERUMS): A Progress Report". Proceedings of the Survey Research Methods

Section, American Statistical Association: 434-437.

1988 "Employer Reporting Unit Match Study (ERUMS) -- What have we learned?" Presented at the annual meeting of the American Statistical Association, New Orleans, LA.

Bureau of the Budget

1961 "Brief History of the Movement in the Federal Government for a Central Directory and of Related Efforts Aimed at Improving Quality and Comparability of Economic Statistics." Unpublished report, Office of Statistical Standards. Washington, DC: Bureau of the Budget.

Bureau of the Census

1965 "Final Results of BES Census Retail Payroll Reconciliation for the State of Delaware". Memorandum from Peter Ohs and Ralph Woodruff to Harvey Kailin and William Hurwitz, July 22. Washington, DC: U.S.

Department of Commerce.

Bureau of Economic Analysis

1972 An Evaluation of the Usefulness of the Social Security Administration's Continuous Work History Sample. Report prepared for the Manpower Administration, U.S. Department of Commerce. Washington, DC: Department of Commerce.

Cartwright, D., Levine, B. and Buckler, W.

1983 "An Update on Establishment Reporting Issues: Practical Considerations". Proceedings of the Survey Research Methods Section, American Statistical Association: 481-486.

Grzesiak, T. and Lent, J.

1988 "Estimating Business Birth Employment in the Current Statistics Program". Paper presented at the Annual Meeting of the American Statistical Association, New Orleans, August 21-25.

Harte, J.

1986 "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS". Proceedings of the Survey Research Methods Section, American Statistical Association: 603-608.

Jabine, T.

1984 The Comparability and Accuracy of Industry Codes in Different Data Systems. Committee on National Statistics, National Research Council. Washington, DC: National Academy Press.

MacDonald, B.

1989 "Progress Report, U.S. Bureau of Labor Statistics".
Paper prepared for the Fourth International Roundtable on
Business Survey Frames, Newport, Gwent, United Kingdom.

Montana Department of Labor and Industry

1987 Montana Business Birth-Death Study: 1984 to 1986.
Research and Analysis Bureau, Employment Policy Division.

Office of Federal Statistical Policy and Standards

1980 Report on Statistical Uses of Administrative Records:
Statistical Policy Working Paper 6. Washington, DC:
Department of Commerce.

Office of Management and Budget

1983 Establishment Reporting in Major Administrative Record Systems. Establishment Reporting Working Group, Administrative Records Subcommittee, Federal Committee on Statistical Methodology. Unpublished report, October 17. Washington, DC: Office of Statistical Policy.

1984 A Review of Industry Coding Systems: Statistical Policy Working Paper 11. Washington, DC: Office of Management and Budget.

Social Security Administration

1988 2000: A Strategic Plan. Washington, DC: Department of Health and Human Services.

Appendix B

Exhibit B-1

Agreement Between

Statistics of Income Division, Internal Revenue Service

and

Bureau of Labor Statistics

Department of Labor

A. INTRODUCTION AND PURPOSE

The purpose of this agreement is to provide the Bureau of Labor Statistics (BLS) with limited access to taxpayer data for the purpose of the Employer Reporting Unit Match Study (ERUMS). ERUMS is designed to study the types of problems, and potential benefits, resulting from matching employer administrative and statistical records from different agencies for statistical purposes. To carry out the study a small sample of records will be selected from the files listed below and extra&.s produced which will be subsequently matched:

1. Extract from the Employer Identification file and an extract from the Form W-3, Transmittal of Income and Tax Statements, file. The W-3 is an IRS document. An

extra&,. of these tape files, which are maintained by the Social Security Administration (SSA), will be used.

2. Extracts from several parts of the IRS Business Master File (BMF) System including limited data (e.g., industry codes) from income tax returns, plus data from the Form 940, Employer's Annual Federal Unemployment Tax Return and Form 941, Employer's Quarterly Federal Tax Return.

3. Extract of the Unemployment Insurance Address file for a specific state. This file is maintained by the Bureau of Labor Statistics.

Both computer and manual matching procedures will be employed. Once the match is completed by BLS, summary tables will be produced and an overall report will be written making recommendations about the development of a system using common identifiers in order to make such matches easier and to develop consistent procedures to be used In data collection and analysis.

In addition to a report on the general results of the match study, which will include recommendations regarding establishment reporting, the following specific products will result from this study:

1. Evaluation of SOI Industrial Classification System

In addition its extensive use of the Department of Treasury's Office of Tax Analysis, a major application of SOI data is in the development of the Department of Commerce's National Income and Product Accounts. The value of the SOI for this purpose is compromised, to some extent, because the industrial Classification system used in the SOI is not strictly comparable to the industrial classification system used in the other major source of income data for the National Income and Product Accounts. This other major source, wage data from the ES-202 reporting system, is administered by State Employment Security Agencies and coordinated by BLS. The ES-202 reporting system supplies wage and salary data based on reports made

Appendix B

Exhibit B-1

in conjunction with unemployment insurance (UI) payroll taxation. Not only are the SOI and ui industrial classification systems administered independently, but they also involve different reporting unit concepts. Whereas SOI data are based on business' income- taxpaying units, the UI reporting system is designed around 'reporting units" which provide greater geographic and industrial detail than is generally provided by taxpaying units. In particular, multi-establishment businesses are required, in the UI system, to report or, the basis of units that separate the employment and payroll of activities carried out in different counties and/or different Standard Industrial Classification (SIC) categories.

This study permits a direct comparison of reporting units in the UI system with IRS taxpaying units to provide knowledge Of the extent and nature of comparabilities and noncomparabilities between

the two data systems. In addition, the study will provide an evaluation of the joint IRS-SSA payroll reporting system. This system, which uses the IRS Form W-3 is of particular interest because it is conceptually designed on the basis of reporting units comparable to those in the UI system. Further=re, it could be linked to the SOI system on a regular basis bringing about considerable improvement in the quality of SOI industry coding and saving substantial resources currently used to manually correct defective industry codes. The W-3 reporting system, however, requires evaluation before it can be used in conjunction with SOI data because additional geographic and industrial detail requested in the reports of multi establishment taxpaying units is obtained on a voluntary basis and because relatively limited resources have been devoted to maintaining and improving the quality of the data supplied through the W-3 system. In the UI reporting system, by contrast, special efforts are made to obtain geographic (county) and industrial (4 digit SIC) reports, and a systematic program of data quality control has been implemented. A major question to be addressed in this study, therefore, involves whether or not the quality of W-3 geographic and industrial data is sufficiently high

to merit consideration of their use in conjunction with the development and application of SOI data.

Table I will provide basic comparisons among the UI, W-3, and IRS reporting systems. Except where clear problem can be demonstrated in the UI system or where UI data are unavailable, the UI reporting unit will be taken as the standard from which deviations in the W-3 and IRS systems will be compared. The table will be divided into seven parts to highlight the various potential causes of discrepancies among the reporting systems. The first part compares, for all UI reporting units, the extent of agreement or disagreement among industrial codes by major industry group. The second part of the table examines this issue for single unit businesses only. In this latter case, discrepancies can be assured to be due to differences in the coding process. In the third part of the table, reporting units are compared for multi-establishment businesses that contain a majority of their operations (as measured by payroll) in the state for which UI data have been sampled (Texas). This part includes businesses which may operate in more than one industry, but excludes businesses for which the sampled

data are unlikely to be representative of the their overall operations. The fourth part includes the UI reporting units for multi-establishment businesses with the majority of their

APPENDIX B

Exhibit B-1

operations outside of the sample state. The final three parts repeat the comparisons for all units and for the two categories of multi-establishment businesses but with reclassification of the UI and W-3 reporting units so that all industry codes for each business are the same (based on the code of the largest unit). These last parts, in conjunction with the earlier parts, will help determine the extent to which discrepancies between IRS data and the UI and W-3 data result from differences in coding rather than from lack of reporting unit detail in the IRS data.

2. Feasibility of Developing State Data from IRS Records

Because IRS taxpaying units often operate in more than one state, the ability to present SOI estimates on a state basis is problematic. But just as the W-3 and UI record systems can be used to evaluate the industrial classification of IRS records, they can also be used to determine the potential for developing usable geographic data from IRS records. Not only can the extent of multi-state operations by IRS reporting units be determined 9 but comparisons among the record systems can also be used to determine the potential for using geographic data from the W-3 reporting system in conjunction with SOI records to develop data by state. The UI data in this study are only for,the state of Texas, and there- Fare provide only a limited basis for assessing the quality of state-level geographic data in the W-3 system. The UI data, however, can be supplemented in the evaluation of the W-3 data . by the state data reported to IRS on Form 940 in connection with the Federal unemployment insurance tax. Form 940 requests data by state on taxable wages for multi-state firm. No breakdown of wages is available on Form 940 for substate areas or for the separate industries of multi-industry businesses. Moreover, the use of Form

940 data necessitates the estimation of total wages an the basis of taxable wages. But, in contrast to the voluntary geographic reporting in the W-3 reporting system, the state reporting on Form 940 is a legal requirement. Indeed, if total, wages by state can be estimated reliably from the taxable wages reported on Form 940, then it might prove feasible to use Form 940 data in conjunction with SOI data to develop a limited range of state, data within the SOI statistical framework.

Table II will compare estimated wages in Texas for the UI, W-3, and Farm 940 reporting systems for various categories of reporting employers. The table will also compare estimated non-Texas wages for the W-3 and Form 940 reporting systems. Three major categories of employers will be distinguished: 1) employers that operate only in Texas, 2) employers that operate outside of Texas but pay the majority of their wages in Texas (according to both the W-3 and Form 940 reporting systems), and 3) employers that operate outside Texas and pay the majority of their wages outside of the state (according to either the W-3 or the Form 940 reporting systems). Within each of these major categories, finer breakdowns

will be based on the extent, of agreement of reported employer characteristics among the three reporting systems. Categories of particular importance in evaluating the state-level reliability of reporting in the W-3 system, for example, will be categories indicating cases in which the W-3 reports cover fewer states than the Form 940 reports.

A proposed schedule of major tasks is included in Attachment I.

B. TERMS AND CONDITIONS

1. IRS will provide BLS with a computer file containing data from the Form W-3. BLS will use this file to locate W-3 information for reporting units randomly selected from the BLS file.

2. One-of the objectives of this project is,to determine the amount of overlap in reporting in the W-3 and BLS files. Since it is anticipated that in some cases more than one BLS State Employment Security Administration record may match with one W-3 record, multiple: matches will have to be resolved manually. The computer output required to do this match and analysis will consist of formatted printouts of the individual records. Security for this file will be. guaranteed by the contractor's agreeing to the provisions Of Section C, specifically paragraphs 1. a,, b. and c.

3. In an effort to add to the information in the these files, the Internal Revenue Service will extract from the BMF copies of Forms 940 and 941 records for -the units selected for this study. BLS will be provided only hardcopy output from these records. No computer copies will be made of these records. At the completion of the study, BLS will return the Forms 940 and 941 records to the IRS custodian.

4. No results of this study will be released until IRS certifies that the results are disclosure free. Disclosure free in this regard will be defined to mean that it will not be possible to identify data, either directly or indirectly, for an individual entity. As a minimum, prior to the release of any information, all data which can be identified as being based on fewer than three sampled items will be suppressed. Output which has been certified by the contractors to meet these criteria must be reviewed and approved for release by IRS. If IRS withholds it's approval for the release of the material, it will specify the areas in which the submitted material is found not to be free of disclosure.

5. Individuals designated by BLS as custodians of the files (see Attachment II) will be responsible for observance of all conditions of use and for the establishment and maintenance of security arrangements to prevent unauthorized access. If the custodianship is to be

transferred within the organization, written IRS concurrence will be required.

C. SAFEGUARDS AND RESTRICTIONS ON USE OF IRS DATA

Information will be furnished to BLS by IRS for the purpose of Section and as authorized by Section 6103(n) of the Internal Revenue Code and implementing Treasury Regulation Section 301.6103(n)-1(1). The conditions of receipt, use,, disclosure,, storage,, transmission, access and disposition of the return information is governed by the principles,contained in IRM 1(14)2(13).(11) as shown below.

1. Safeguards

In performance of this contract, the contractor (BLS) agrees to comply with and assume responsibility for compliance by its employees with the following requirements:

APPENDIX B

Exhibit B-1

a. The contractor certifies that the data obtained from IRS for purposes of this study shall be completely purged from all data storage components. If immediate purging of all storage components is not possible to accomplish, the contractor and subcontractor certify that such IRS data remaining on any storage component will be safeguarded to prevent unauthorized disclosure.

b. All return information will be accounted for upon receipt and properly stored before, during, and after processing. In addition, all related output shall be given the same level of protection as required for the source material.

c. All work will be performed under the supervision of the contractor and the contractor's responsible employees.

d. Any return information used, in any format, shall be used only for the purposes of carrying out the provisions of this contract, an information contained in such material shall be treated as confidential and shall not be divulged or made known in any manner to any person except as may be necessary in the performance of the contract. Disclosure to anyone other than an officer or employee of the contractor, except as expressly provided by this contract, shall require prior written approval of the Internal Revenue Service. Requests to make such disclosure should be addressed to the IRS Project Coordinator.

e. Any spoilage or any intermediate hardcopy printout which may result during BLS's processing of tax return data used in this project shall be given to the IRS representative. When this is not feasible, the contractor will be responsible for the destruction (shredding) of the spoilage or. any intermediate hardcopy

and printout and shall provide the IRS coordinator with a statement containing the, date of destruction, description of material destroy)ed, and the method used.

f. No work involving information furnished under this contract will be subcontracted to organizations other than BLS without the specific approval of the IRS Project Coordinator.'

g. The contractor shall provide the Internal Revenue Service with a list of people employed who are permitted to see confidential tax return information.

h. Failure to meet the above safeguards will result in termination of this agreement.

2. Criminal/Civil Sanctions

a. Each officer or employee of any person to whom returns or

return information is or may be disclosed shall be notified in writing that such returns or return information disclosed to such officer or employee can be used only for a purpose and to the extent authorized herein, and that further disclosure of,any such returns or return information for a purpose or to an extent unauthorized herein constitutes a felony, punishable upon conviction by a fine of as much as 15,000 or imprisonment for as long as 5 years, or both, together with the costs of prosecution. Such person shall also so notify each such office and employee that any such unauthorized further disclosure of returns or return information may also result in an

award of civil damages against the officer or employee in

an amount not less than $1,000 with respect to each instance of unauthorized disclosure. These penalties are prescribed by IRC 7213 and 7431 and set forth at 26 CFR 301.6103(n)-l.

b. Additionally, it is incumbent upon the 'Contractor to inform its officers and employees of the penalties for improper disclosure imposed by the Privacy Act of 1974, 5 USC 552a. Specifically, 5 Usc 552a(i)(1), which is node applicable to contractors by 5 USC 552a(ml,, provides that any officer or employee of a contractor who, by virtue of his/her employment or official position, has possession of, or access to, agency records which contain individually identifiable information, the disclosure of which is prohibited by the Privacy Act or regulations established thereunder, and who, knowing that disclosure of the specific material is so prohibited, willfully discloses the material in any manner to any person or agency not entitled to receive it, shall be guilty of a misdemeanor and fined not more than *5,000.

D. INSPECTION

The Internal Revenue Service shall have the right to send its officers and employees into the processing facilities of BLS for inspection of the facilities and operations provided for the performance of any work under this contract. On the basis of such inspection, the Internal Revenue Service shall have the right to stipulate specific measures needed to implement the safeguards contained in paragraphs I.(a) through I.(h) above, as determined essential by the Internal Revenue Service. See Attachment III, Publication 1075, Tax Information Security Guidelines.

E. PROJECT COORDINATORS

Mr. Thomas Petska, Statistics of Income Division 376-0761, is designated as the IRS Project Coordinator under this I contract. Ms. Linda Hardy, Division of Occupational and Administrative

Statistics, BLS, 523-1636, is designated as the BLS Project Coordinator under this contract. The IRS Project Coordinator will 'receive for the IRS all of the services called for in this contract and will represent the IRS in the technical phases of the work. The BLS Project Coordinator will receive for BLS all of the services called for in this contract land will represent BLS in the technical phases of the work.

F. AUTHORITY

Authority for the agreement is found in Sections 6103(n) and 6108 of the Internal Revenue Code and implementing Treasury Regulations. thereunder.

_____	_____
Fritz Scheuren (Date)	Janet L. Norwood (Date)
Director	Commissioner
Statistics of Income Division	Bureau of Labor Statistics
Internal Revenue Service	Department of Labor

Appendix B

Exhibit B-2

AGREEMENT BETWEEN SSA AND BLS

FOR EXCHANGE Of STATISTICAL INFORMATION IN

EMPLOYER REPORTING UNIT MATCH STUDY (ERUMS) PILOT PROJECT

Terms and Conditions:

1. The Office of Research, Statistics and International Policy (ORSIP) in the Social Security Administration (SSA) will furnish the Bureau of Labor Statistics (BLS) with tapes containing statistical data copied or derived from SSA's employer files to be used exclusively for the statistical purposes of this agreement.

2. The statistical purpose for BLS use of SSA data authorized by this agreement is to conduct a pilot study "Employer Reporting Unit Match Study" (ERUMS) designed to match information from employer wage reporting and establishment reporting systems at BLS and SSA, supplemented with reporting unit information from the Internal Revenue Service (IRS).

3. SSA will furnish BLS with the following files (described in Appendix 1) containing information for cases in Texas selected by BLS and SSA:

   a. Single Unit Code File

   b. Multi-unit Code File

   c. Employer Report Record

4. Brian McDonald will be custodian of the files for BLS to assure that the data are used only by persons authorized in

writing by ORSIP and BLS to carry out this agreements BLS will notify ORSIP in writing of any change of custodian. Copies or extracts of SSA data will be treated as if they were original data files obtained from SSA.

5. In accordance with the specifications set forth in Appendix 2, BLS is authorized to perform individual comparisons and linkages of these records with records selected from the BLS Unemployment Insurance (UI) employer name and address file for the purpose of categorizing records and preparing counts and listings for subsequent analysis and to perform individual comparisons and linkages with information supplied by IRS for the purpose of preparing statistical Durations. Persons authorized by ORSIP will have access to the linked records for the statistical purposes of this agreement.

6. Except as authorized by paragraph 5, no effort whatsoever may be made by any person to compare or link individual records

with names or identifying numbers or identifiable information from any source about, particular entities.

Appendix B

Exhibit B-2

7. No listings of data from individual records may be published or otherwise released by BLS.

8. Release of statistical data to anyone other than, persons authorized by this agreement will be only in summary form which is not potentially identifiable as to individual employers. Any distribution in a table should be based on the most stringent criteria for disclosure of statistics as applied by SSA, BLS, or IRS.

9. Adequate physical security procedures must be used to prevent access by unauthorized individuals and BLS will provide assurance satisfactory to SSA that such procedures are carried out, and will permit ORSIP to conduct site visits at reasonable times for this purpose.

10. Approximately 6 months will be scheduled to perform the matching operations and analyses of the results of the matches; approximately 3 additional months to produce statistical data and to perform disclosure analysis and suppression; and approximately 1 year to prepare a report on the results. When these operations have been completed, all tapes, copies, extracts, derivatives and printouts of microdata or other data restricted by this agreement will be returned to SSA or destroyed under SSA supervision

11. SSA will consult IRS before releasing statistical files based on tax return information to BLS under this agreement. SSA

and BLS may enter into other agreements consistent with the terms of this agreement as IRS or the Department of the Treasury may require with respect to such statistical information.

_____ _____

Date Jane L. Ross, Director

Office of Research, Statistics and

International Policy

_____ _____

Date Janet L. Norwood Commissioner

Bureau of Labor Statistics

Appendix 1: Description of Contents of Single Unit Code File,

Multi-unit Code File, and Employer Report Record.

Appendix 2: Specifications for Sample, Selection, Electronic Match and Related Operations, and "ERUMS" Project Timetable.

Reports Available in the

Statistical Policy

Working Paper Series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)

2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)

3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)

4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)

5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)

6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)

7. An Interagency Review of Time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)

8. Statistical Interagency Agreements (NTIS Document Sales, PB86-230570/AS)

9. Contracting for Surveys (NTIS Document,Sales, PB83-233148)

10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)

11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)

12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)

13. Federal Longitudinal Surveys (NTIS Document Sales, PB86-139730)

14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)

15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)

16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)

17. Survey Coverage (NTIS Document Sales, PB90-205246)

18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)

19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650