EMPLOYER REPORTING UNIT MATCH STUDY (ERUMS) -- What have we learned?

Warren L. Buckler, Social Security Administration

INTRODUCTION

Work on a pilot study which was designed to match microdata from employer reporting systems of three federal agencies is nearing completion. The intention of this effort, sponsored by the Federal Committee on Statistical Methodology (FCSM), was to examine some of the issues surrounding the inadequate and inconsistent reporting of employer information at the establishment level that have impeded more effective and efficient uses of administrative records for statistical purposes. In addition, the group hoped to gain experience in attempting to accomplish interagency data/files exchange within the framework of current regulatory constraints.

This paper reports on the activities of the workgroup that has been conducting the study and what has been learned through the effort to date.

BACKGROUND

Late in 1977, the Federal Committee on Statistical Methodology (FCSM), which was then in the Office of Federal Statistical Policy and Standards (OFSPS), Department of Commerce and is now a standing committee in the Office of Management and Budget (OMB), formed a subcommittee to inquire into the statistical uses of administrative records along the general lines of:

1) evaluation of the quality of administrative data and their suitability for statistical applications; and

2) assessment of the problems of access to administrative records for statistical purposes and of needs for improved coordination between statistical and administrative components.

The Subcommittee on Statistical Uses of Administrative Records (SUAR) was comprised of representatives from various federal statistical agencies and statistical components of program agencies. The subcommittee drew upon the expertise and experience of its members to concentrate its investigation on administrative programs that collect important social and economic information from individuals and businesses. The subcommittee, in its final report in 1980, listed 11 recommendations for dealing with the issues it studied. These fell into three main areas:

1) Identifying and formulating solutions for common problems related to statistical standards for administrative information programs.

2) Identifying and meeting various problems related to access to administrative record systems.

3) Identifying collection programs and research activities requiring government-wide coordination and support.

Though the Subcommittee on SUAR issued its final report, work in this area was by no means finished. Early in 1981, another subcommittee was formed with the charge of attempting to implement some of the recommendations made by the original group.

This new group, now known as the Administrative Records Subcommittee, established several work groups to deal with specific recommendations. One of these, the Establishment Reporting Work Group, was formed to study possibilities and problems involved in implementing two of the recommendations which dealt with the manner in which employers file administrative reports for their establishments:

- Recommendation #1.--Common identifiers should be used whenever possible in collecting information pertaining to the same individuals or organizations, and

- Recommendation #3.--Consistent procedures should be used in administrative and statistical data collection efforts for defining reporting units, identifying and coding reporting unit characteristics and developing standards for data tabulation.

The principal reason for selecting these recommendations was that there was a direct tie in to the underlying desirability of increasing the use of administrative records for statistical purposes. The employer reporting systems were seen as areas where relatively low cost modifications to existing systems could be made that would yield overall benefits to both administrative and statistical programs by reducing respondent burdens, data collection costs, and data processing costs. The statistical applications to a coordinated employer reporting unit system has the potential for providing for the creation of powerful data bases that can be used to measure economic activity and demographic changes for subnational areas.

To address the issues before the work group, three major administrative record systems were selected for study:

1) Unemployment Insurance (UI) records collectted by each State under rules and procedures established and coordinated by the Department of Labor;

2) W-2/W-3 records submitted to and processed by the Social Security Administration (SSA) for both SSA and Internal Revenue Service (IRS) administrative purposes; and,

3) Census Bureau's Standard Statistical Establisbment List (SSEL) records.

Note: IRS tax return records for business were not selected because this information is collected on a company basis rather than an establishment and thus did not provide the breakdown of information needed for this study.

The work group identified three major tasks:

1) document the structural differences among the three systems;

2) study the factors contributing to statistical inconsistencies among the three systems; and

3) study the possibilities and problems involved in implementing the recommendations.

Task 1 was completed with little difficulty. As part of task 2, the work group planned a micro study to shed more light on the extent and nature of the establishment reporting problem. Though much developmental work was done, the micro study was not conducted, primarily due to confidentiality restrictions on access to data, limited resources and priority conflicts among the participating agencies. An implementation study to investigate the feasibility of converting SSA and BLS systems to the establishment units contained in Census's SSEL was planned for task 3. Here again, the problems that confronted the group were such that no formal proposals were developed.

The work of the Establishment Reporting Work Group is well documented in its final report which was completed in December, 1982. The fact that it stopped short of its original goals is evidence of the difficulties facing the statistical community in obtaining the needed improvements in administrative records. However, significant progress was made in identifying the issues that require attention and possibilities for further exploration. The report concluded with a recommendation for continued study of establishment reporting and completion of some of the unfinished business at hand. So, early in 1983 a new workgroup was formed within the Administrative Records Subcommittee to continue examining some of the issues identified by the previous group in a somewhat different direction and with a new focus and, at the same time, partially address a topic identified by the FCSM for study -- the need for interagency sharing of statistical data/files.

The purpose of this paper is to describe the project undertaken by this group, report on their activities, and highlight some of the findings.

THE ERUMS PROJECT

The new workgroup was originally comprised of representatives from the Bureau of Labor Statistics (BLS), the Social Security Administration (SSA), the Internal Revenue Service (IRS), the Bureau of Economic Analysis (BEA) and the Office of Management and Budget (OMB). Later a representative from the Committee on National Statistics and observers from the Census Bureau joined the group.

This group knew that an important factor contributing to its predecessor's inability to fully accomplish what it set out to was that many of the activities planned by the group were beyond the control of its members. In light of the difficulties encountered by the previous Work Group, the new group agreed that the objectives and tasks to be undertaken should be items that the members felt were achievable and could be controlled by the group.

The group determined that its primary objective would be to conduct a pilot study designed to match information, at a microdata level, from employer wage reporting and establishment reporting systems of BLS, SSA and IRS. Thus the name Employer Reporting Unit Match Study (ERUMS) Workgroup emerged. This micro study would differ from that proposed by the previous work group in that it would focus on the reporting unit relationships between the BLS and SSA systems, supplemented with information from IRS at the employer level. This type of study would allow the group to carefully examine and gain insight into the differences and similarities of the three systems so that recommendations could be made regarding: (1) the development of a system that uses common identifiers for collecting informat on pertaining to the same organization; and (2) developing consistent procedures to be used in administrative and statistical data collection efforts for defining reporting units and identifying and coding reporting unit characteristics. In addition, conducting the microdata match study would provide valuable experience in learning how to accomplish interagency exchanges of data and files within the framework of current regulatory constraints. Also, the group hoped to learn from the pilot study how a cooperative interagency data exchange can be used to identify and correct errors, deficiencies and shortcomings in the systems of the participating agencies.

In the first several meetings of the ERUMS Work Group, the members concentrated on outlining plans for the study. These early deliberations resulted in the following:

1) Scope of Study.--Considering the resources available to the work group, it was decided that a sample of records should be selected from one State for the pilot study. This would make it possible to do a thorough review and analysis of matched and unmatched cases.

2) Data Access.--It was clearly recognized from the beginning that, because of current restrictions on the release of identifiable information, careful consideration must be given to the steps to be taken in order for the group to gain access to and use the required microdata records. Instead of the approach outlined by the previous group, this group felt

that it would be necessary to conduct the study under interagency agreements between the participating agencies. Extensive discussion centered around what the terms of such agreements would be. It was very important that these agreements contain well defined statements as to the purpose of the pilot study and assurances of the protection of confidentiality of identifiable information.

3) Data Sources.--Each participating agency identified the files from their system that could be made available for the match study, given a satisfactory outcome of the access issue. The SSA would provide records from its master file of employers which contains information used to code geography and for workers in statistical files and from records of employer wage reports furnished on Form W-3 (Transmittal of Income and Tax statements). The IRS would provide records of information from Forms 940 (Employer's Annual Federal Unemployment Tax Return), Form 941 (Employer's Quarterly Federal Tax Return) and Form W-2 (Wage and Tax Statement). BLS would furnish employer information from reports that States are required to file under the Unemployment Insurance program and summarized in their ES-202 report.

4) Data Processing.--The BEA personnel offered their services in performing the computer processing required for the microdata match. An appropriate sample of records from the BLS, SSA and IRS systems for the one State was to be selected and matched based on a specific set of variables. The group recognized that a substantial amount of manual data processing would be required after the electronic match was done. An examination and analysis of the matched and unmatched records would be a key part of the pilot study and should provide the work group with much of the information needed to meet its objectives.

Description of Activities

The group then set out to define some specific tasks that needed to be done. These were:

- Develop a project description;
- Select a State and obtain their permission to use their records for the match study;
- Document the data files to be used in the match;
- Develop sampling criteria;
- Develop matching criteria;
- Draft interagency agreements to cover data exchanges and the work to be done; and
- Develop a timetable for accomplishing specific objectives.

As the group got into these items, a number of obstacles began surfacing that were inhibiting the smooth and orderly progress toward attaining the primary objective. The major problem area centered around confidentiality issues; such as, assurances that can be given to the State selected for the pilot study about protection of information from their records, decisions on who will have access to the identified microdata, and requirements for the protection of tax return information. There were also problems with the multi-agency type interagency agreement that had been drafted to cover the pilot study work. It was evident that modifications had to be made to the original plans.

The decision was made that BLS would perform the computerized match operations. It was felt that confidentiality problems could be substantially reduced if access to confidential records was limited to those with a "need to know," in this case, BLS and SSA personnel. The issue of expanding this "special" group was to be addressed as the study proceeded and if it was determined to be desirable to do so. Since then there have been a number of occasions when it seemed desireable to expand the special group to include additional members of the ERUMS Workgroup, but in each case concerns over confidentiality assurances and the acompanying need to renegotiate the agreements that had been executed as discussed below, won out over the benefits of expansion.

Proposals for the content and format of interagency agreements were revised several times. Intense discussions focused on problems each agency had with issues that were under consideration for inclusion or exclusion in the agreement. The group finally decided that an agreement between IRS and BLS covering the use of tax information, the work to be performed and the products to be obtained would be the best way to proceed. An agreement covering the conditions of use of SSA data by BLS would be handled through a separate document. Subsequently both of these agreements were adopted by the Workgroup and formally accepted and approved by officials of the agencies involved.

A draft was prepared which described the ERUMS project in terms of some of the specifics of the matching operations as well as a statement of the purpose. This description was later revised to include the statistical products to be obtained (Exhibit A). The group then accepted BLS' recommendation that records for the State of Texas be used for the microdata match. BLS obtained permission from Texas for their records to be used in the study. All of the files under consideration for use in the study were documented. As work on the project proceeded, slight modifications were made to some of the specifics of the project description and in the contents of the files that were available at the time the actual matching operations were performed. The substantive contents of the files used are shown in Exhibit B.

Sample Design and Selection; Electronic Matching Operation

To assist the group in designing the sample of records to be used in the match, BLS obtained a

set of universe counts from their UI Name and Address File for the State of Texas. These counts provided the number of single and multi unit records in the file in terms of the Employer Identification Number (EIN) and are summarized as follows:

Total Number of Employers = 270,612
Single-unit Employers = 267,487
Multi-unit Employers = 3,125

Based on these counts a sampling rate of 6 in 100 was selected for the initial stage sample that yielded 201 multi-unit EINs and 16,135 single unit EINs from the BLS UI file. The selection was based on the occurences of six pairs of randomly selected digits in the 7th and 8th positions of the EIN.

The EINs selected in this first stage were also matched to SSA files and records that matched on EIN were selected. In addition, records with EINs not in the Texas UI file but having the same pattern of selection digits and at least one Texas establishment in SSA files were selected. These operations resulted in the selection of 16,734 EINs that were classified as single unit in SSA files and 491 EINs that were classified as multi-unit.

(Exhibit C provides a description of the structure of the BLS and SSA files from which the records were selected, defining single unit and multi-unit concepts.)

The records selected from the BLS and SSA files were matched on EIN and classified in terms of match status and whether they were single or multi-unit in their respective files. During this match, for employers that were single unit in both BLS and SSA files, an additional comparison was made to determine if there was agreement on county and 2-digit SIC. Also, records in the BLS multi-unit category that had 20 or more establishments were identified separately. The results of those operations were as follows:

Status		# of EINs
BLS	SSA	
single	single 1/	8,689
single	single 2/	4,392
single	none 3/	2,698
none 3/	single	3,559
multi	single 4/	88
multi	single 5/	6
single	multi	356
multi	none 3/	41
none 3/	multi	69
multi	multi 4/	60
multi	multi 5/	6
	Total	19,964

See footnotes on figure below.

From these counts subsample ratios were determined in order to yield about 200 multi-unit employers and 200 single unit employers, with approximately equal numbers of cases selected from each of these groups. For the multi/multi and multi/single categories, all EINs with 20 or more reporting units in the BLS file were selected with certainty. The results of the final sample selection were:

Status		Ratios	# of EINs Selected
BLS	SSA		
single	single 1/	173.78	50
single	single 2/	87.84	50
single	none 3/	53.96	50
none 3/	single	71.18	50
multi	single 4/	2.59	34
multi	single 5/	1.00	6
single	multi	8.90	40
multi	none 3/	1.00	41
none 3/	multi	1.73	40
multi	multi 4/	1.76	34
multi	multi 5/	1.00	6
	Total		401

1/ Same county and 2-digit SIC.
2/ Different county or 2-digit SIC.
3/ "none" means having no 1982 wage report.
4/ Employers with less than 20 establishments in BLS file.
5/ Employers with 20 or more establisments in BLS file.

Information from BLS and SSA records that was needed for the manual match operations was prepared in separate listings of each source file for the final sample of 401 cases. At the same time IRS began the task of arranging for the extraction of information from their master files for the final sample, which would be added later to the BLS and SSA data in order to produce the descriptive and analytical tabulations called for in the project description and the BLS/IRS agreement.

Manual Matching

Procedure.--BLS and SSA staff met a number of times to conduct the manual matching operations using the listings that were prepared from the final sample of 401 employers. A worksheet was prepared upon which the group recorded information about the relationship and comparability of geographic and industry codes between the BLS and SSA files for each employer and of the units of those employers that were multi-unit. An electronic data record was prepared from the worksheet and tabulations describing these results were prepared for use in analysis and investigation.

A different version of SSA's wage report file, which was not available for the electronic match, was used for the manual matching operations. This file provided additional information that was helpful in determining the status of records in SSA files. Reports for employers whose workers were not subject to social security taxes, delinquent employer

reports, and reports of household employers, all of which were not part of the file that was used in the electronic match, were now available.

Further investigation revealed that a number of employers listed in SSA's multi-unit file were filing wage reports as single unit entities. There were also some cases where single unit employers were entering information in the establishment number field of the wage report, which caused them to be erroneously accreted to the multi unit file.

The knowledge gained through the preliminary investigation enabled the group to reclassify the records of the employers selected in the final sample. The SSA records were reclassified based on what was learned in this preliminary investigation and numbers were assigned to the final grouping of records that would be examined in the next stage of the manual review and matching operations. The group numbers are primarily used for reference purposes when reporting on the results of those operations. The reclassification efforts resulted in the following distribution of EINs in the final sample:

Status		Group	# of
BLS	SSA	No.	EINs
single	single	1	149
single	no wage rpt.	2	35
no wg.rpt	single	3	87
multi	single	4	93
single	multi	5	6
multi	no wage rpt.	6	12
no wg.rpt	multi	7	3
multi	multi	8	16
	Total		401

Each group of records were reanalyzed with respect to geographic and industry comparability as before, with appropriate changes made to the electronic data record.

The final sample cases were weighted to the first stage sample and to the universe. Attachment 1 of Exhibit D shows the distribution of these counts in terms of number and percent of the total.

Tabulations were prepared descibing the results of the matching operations for certain key classifications of records. These are shown as Attachment 2 of Exhibit D. The observable results were recorded for the other classifications that were not described in tabular form.

Highlights of Results.--The highlights of the results of the manual matching are shown below. A more detailed report on the results of the manual matching operations is available from the author upon request.

1) The largest single category of records were those that were classified as single unit in both BLS and SSA files (group 1). There were 149 such cases in the final sample. When inflated to the 1st stage sample (14,272 cases) and the universe (237,866 cases), this represents over 70% of the total number of cases. Of these cases, over 80% were classified in the same county by BLS and SSA. Nearly 80% of these cases that were classified in the same county were also coded to the same 2-digit industry level, with about 60% agreeing at the 4-digit level. When comparing industry classifications of these records, without regards to geography, over 70% agreed at the 2-digit level and 60% agreed at the 4-digit level. Of the nearly 13% that were classified in a different county in Texas, there was only a 23% agreement on industry codes at the 2-digit level, a sharp contrast to what was seen in other cases. Even for the nearly 7% that SSA had coded in a different State, over 67% were classified to the same 2-digit industry level. Overall the results of the matching of this category indicates a high degree of comparability between BLS and SSA records.

2) The next largest category of records (about 18% of the weighted total) were those where SSA had an indication of wage activity in Texas in 1982 for which there was no wage report in the BLS' 1982 Texas UI file (groups 3 and 7). There were a total of 90 such cases in the final sample (87 SSA single unit and 3 SSA multi-unit). These represented 3,628 1st stage sample cases and 60,466 cases in the universe. A large number of these employers (about 40%) appeared on wage reports for later years at BLS, indicating that the employer may not have begun reporting wages until the 2nd, 3rd, or 4th quarter 1982. About 26% had SSA industry codes for which UI coverage is exempt and another 16% were uncoded by SSA which could have been the same type case. There was no explanation for about 18% of the cases.

3) There were 47 final sample cases in groups 2 and 6 combined (inflated to 1,901 1st stage sample and 31,677 universe cases) for which there was no wage report for 1982 in SSA files. This was about 9.5% of the total number of weighted cases. The geographic and industry codes in SSA coding files for these cases had a relatively high degree of comparability with BLS codes at the county and 2-digit industry level.

4) A little less than 1% of the weighted total were cases that were finally classified as multi-unit in BLS files and single unit at SSA (group 4). Over 70% of these cases were coded by SSA in Texas with a relatively high degree of comparability between the county and 2-digit industry codes. For the cases

that SSA had coded in a State other than Texas, there was still a high degree of comparability of industry codes at the 2-digit level.

5) There were 0.3% of the total weighted cases that were classified as single unit in BLS files and multi-unit in SSA files (group 5). Half of these cases had comparable county and 2-digit industry codes.

6) Only 0.1% of the weighted total were classified as multi unit in both BLS and SSA files (group 8). More than half of these cases had comparable county and 2-digit industry information.

List of Accomplishments

What has been presented up to this point is a summary of the major activities and events that have been completed in the ERUMS project as of the writing of this paper. These can be condensed and restated in the following list of accomplishments:

1. development of a formal statement regarding the purpose of the Work Group;
2. outlining plans for conducting the study;
3. preparation of a project description document;
4. documentation of the data files to be used in the match;
5. preparation of an appropriate interagency agreement and conditions of use agreement that was acceptable to all concerned parties, and obtaining agency approvals for these agreements;
6. selection and obtaining approval of a State for the match study;
7. development of a sample design for the match study;
8. development of specifications for and obtaining universe counts of records in BLS files for the selected state;
9. development of specifications for the first stage sample selection ratios;
10. development of a set of criteria for performing the match operations;
11. selection of the initial sample of records from BLS and SSA files;
12. performance of the electronic matching operations;
13. preparation of counts by match category and subsampling to obtain the final sample for the manual match;
14. development of specifications for the manual match operations;
15. performance of the manual classification and matching operations;
16. preparation of tabulations summarizing the results of the manual match operations; and
17. analysis and reporting on the results of the electronic and manual matches.

Remaining Tasks

As of the writing of this paper several important tasks still remain; namely, adding IRS data to the combined data record, the preparation of additional descriptive and analytical tabulations, disclosure analysis, and the preparation of the workgroup's final report. Over the next few months BLS and SSA will be working on the preparation of the tabulations. After they have been completed, IRS, BLS and SSA will conduct a disclosure review to insure that no identifiable information is revealed in those tabulations. A considerable amount of work needs to be done in organizing the materials and drafting the Workgroup report. We hope that this will be an accomplished fact before the 1989 ASA meetings.

WHAT HAVE WE LEARNED? -- COMMENTARY

Even though the ERUMS project is not completed I feel confident in saying that we have realized many of our original objectives and have learned a number of things as we have proceeded along the long and winding path of progress which eventually leads to a successful conclusion. We have encountered many obstacles along the way which have slowed us down, halted us temporarily, or detoured us; but, in each instance we have found a way over or around whatever was in our way. Determination and dedication to this work has sustained the group members throughout the life of the project.

We have learned that conducting an interagency administrative record match is not an easy task; but it is not an impossible one either, using the fact that it has been done as evidence. We have also learned a way to accomplish exchanges of data and files of several agencies within the framework of existing regulatory constraints; that being through interagency agreements that clearly define the purpose of the exchanges, the conditions under which the exchanges are to be made, and the precautions taken to insure the protection of confidentiality of information contained in the records of the participating agencies.

As we have proceeded with the match study we have learned a great deal about the similarities and, more importantly, the differences between the relationships of reporting units in the BLS and SSA systems. During the investigation of these differences we have seen how a cooperative exchange of this type can help the participants, at least to some degree, in the identification and correction of errors, deficiencies and shortcomings that may exist in their respective systems. In fact, some corrective procedures have already been implemented at SSA as a result of things that were discovered during the investigation of the initial match results.

We have also clearly seen how a voluntary system of reporting information by establishment, with limited resources assigned the responsibility of maintaining it, in an agency where there is no direct program impact of the system, suffers greatly as compared to one with specific requirements mandated by agency regulations. The ERUMS results have pointed to the need for a full scale investigation and evaluation of SSA's Establishment Reporting Plan (ERP) system.

At this point we are not ready to make specific recommendations regarding: (1) the use of common identifiers for collecting information about the same organization; or (2) developing consistent procedures for defining reporting units and identifying and coding reporting unit characteristics. There are still a number of complicated issues surrounding these points that need to be considered, not the least of which is the development of an implementation strategy. Even at the successful completion of the ERUMS project I do not think there will be conclusive evidence that will steer us in a particular direction on how to proceed in these areas. We will have gained more knowledge and experience in an area where we might have had ideas about what things should look like but were not sure what actually existed. I believe that this work will but complete another chapter in a continuing story. From this pilot it seems likely that we will determine the need for a larger scale study to further support and extend our findings.

An enormous amount of time and effort has already gone into this project. The dedicated individuals who have been involved in this investigation are committed to contributing to efforts to improve the effective and efficient use of administrative records for statistical purposes. Yet there exists a paradox--on the one hand there is a determination and spirit of cooperation that exists among members of the workgroup, while on the other there are laws, rules and regulations, as well as possible conflicts with an agency's priorities, that hinder real sustained progress in this area.

There still remains a need for the Federal statistical community to focus attention on ways of developing a better mechanism for drawing the issues to the attention of those who are in a position to and willing to do something beyond the study and recommendation phase. It is in this area that I can report a glimmer of light on the horizon. Based on recommendations made by the Working Group on the Quality of Economic Statistics of the Economic Policy Council, legislation is currently being drafted that would designate BLS as the central collection agency for certain business identification and classification information. If enacted, the resulting system could provide the basis for a more uniform system and increased sharing of this information throughout the federal statistical community. I believe that real significant progress in this area will still depend on the combination of effective collaboration between the custodians of administrative and statistical record systems and some type of follow up legislative actions. In light of the "central collection agency" concept that has been proposed, I might suggest that effective follow up legislation would be to mandate that employers use the establishments designated in such a system when reporting information to an agency. This would be the true governmentwide establishment reporting system which may be the only certain way to insure implementation of the recommendations. I have raised this as a possibility for consideration in the past and, encouraged by these recent developments, will continue to do so. Of course, I don't know for sure that it will work, but I, for one, would like to give it a try.

(Note: References, exhibits, reports and other materials referred to in this paper are available from the author upon request.)