STATISTICAL RESEARCH DIVISION REPORT SERIES
Statistical Research Report Number:
CENSUS/SRD/RR-85/13

Processing Research Study for the 1982 Economic Censuses

by

T. Christopher Dyke
Statistical Research Division
U.S. Bureau of the Census
Rm. 3554, F.O.B. #3
Washington, D.C. 20233

(301) - 763 - 5909

Report Issued: January 27, 1986

1. Executive Summary

The processing research study was intended to measure the effect on originally reported data of each stage of the census processing. The measures obtained demonstrate the effectiveness of each stage at reducing bias, the potential increase in bias due to each stage, and the interactions between stages. This may indicate where savings can be obtained by reduction or elimination of some activities, or where improvements to activities or additional stages may be required.

An economic census record from a census mail questionnaire passed through several distinct stages during the census processing, including keying, complex editing, problem solving, and analyst review following preliminary SIC and area tabulations. This study followed a sample of establishments through each processing stage, generating files which contained establishment records corresponding to a particular processing stage. Participating divisions included Economic Surveys (ESD), Construction Statistics (CSD), Industry (IND) and Business (BUS). Industry Division conducted the Censuses of Manufactures and Minerals. Business Division conducted the Censuses of Wholesale Trade, Retail Trade and Service Industries.

Some general observations evident from the study are:

1. Keying errors, while not large in number, tend to introduce large positive values for a data item and therefore lead to large negative changes when the item is corrected. These errors are of two types: a) values keyed in units instead of thousands and b) two key code-data value pairs being combined so that one variable has a huge value inserted, while the second variable becomes a non-response item. Commonly these keying errors required an analyst review of the record to correct the error, thus errors of this type became quite costly.

2. The execution of this study, as well as the data processing between divisions, would become less complicated if some common rules were applied. As the Bureau moves towards a complete ASCII environment, those divisions which continue to use the FIELDATA character set will change to ASCII. While three of the participating divisions produced files mixing both character and integer binary fields, Construction Statistics Division's format was packed binary, which allows no use of letters or symbols, only numbers. Business Division was the only division which did not use CENIO as the basic file format. In effect, this myriad of processing characteristics requires great programming detail to be attended to when records are transferred from one division to another, commonly, necessitating translation from one language to another, one character set to another or one file structure to another. All of these factors require programming time, effort, and expense.
3. Each of the four participating divisions used a markedly different flag system to describe the basic characteristics attendant to a data variable e.g. its reporting status, impute status, and indication of how a change occurred. One common phenomenon of these flag systems was that they were not used with great accuracy. It was not uncommon, throughout all the divisions, to note a change in the data with the appropriate flag saying no change had occurred and vice-versa. Also, where a change had occurred, and was duly noted, the reason for the change was often obscure.
4. The data generated for this study represents a valuable source of information concerning the 1982 Economic Censuses which can be used not only for additional research but also for testing hypotheses concerning Economic data without incurring the cost of creating new data for such a

purpose, since the data was collected for a sample of establishments that included all SIC's and was stratified by single units (companies with one establishment) and multiunits (companies with more than one establishment).

5. Based on the sample selected for this study, 6.5% of the establishments from the two basic mailout files had no record returned to SRD from any participating division. While the original mailout file had a return rate to SRD of 94.8%, the third quarter birth file had only a 73.9% return rate. Minerals had the lowest return rate at 86.7%.

6. The general trend of census processing is to reduce the data values throughout the entire procedure. Besides the large negative changes that occurred from the correction of keying errors, sizeable changes at the national, trade area level were noted throughout the processing. For example, the drop of 19.5% in the annual payroll figure for wholesale trade from the next-to-last stage to the published data showed the following changes at the SIC level: SIC 5052, - 74.4%; SIC 50-5052 (i.e. SIC 50 except SIC 5052), - 24.3%; SIC 51, -10.0%. Moreover, at the state level for SIC 5052, a 97.4% drop in Ohio was observed, while a 47.4% increase was shown for Louisiana. It was not uncommon for recent changes of this magnitude to be the result of changes in 5 or fewer establishments, especially at smaller geographic areas. Thus it behooves an analyst to be aware of the sensitive nature of values in published tables to one or two changes at the establishment level.

7. Business Division publication data was received in January, 1985, Constructon Statistics publication data in May, 1985, and Industry publication data in June, 1985. Economic Surveys Division will not have their final data until late 1985 and that portion of the study will be published at a later date.

8. For the basic data items included in the study, the general level of the processing, as measured, either by the percent changes in the volume of the data or by the percent of establishments changed was substantial. As might be anticipated, a larger percentage of establishments experienced changes earlier in the processing than later. For each basic data item studied, Table 1.1 summarizes the percent of establishments changed by the first two passes of the data through each division's processing system. These changes reflect the status of the basic data items after the appropriate division receives the data from ESD. Therefore, any changes that occured in an establishment's record during the Jeffersonville processing from the receipt of a form to the record's transmittal to the divisions are not represented. As the extent of the processing at Jeffersonville is far from cursory, the value of records requiring further processing by the divisions is considerable.

Table 1.1 Summary of Percent of Establishments Changed by Division over all variables between the initial three processing stages

	Business		Construction		Industry	
Percent	1st Pass	2nd Pass	1st Pass	2nd Pass	1st Pass	2nd Pass
0- 9	2	17	0	14	6	5
10-19	6	0	1	0	4	0
20-29	6	0	1	0	1	2
30-39	3	0	2	0	3	3
40-49	0	0	5	0	2	3
above 50	0	0	5	0	2	4
Total Basic Items	17	17	14	14	18	18
Average Percent	21.4	2.4	44.4	4.0	25.6	32.0
Median Percent	20.6	1.9	41.7	3.8	15.5	35.6

9. For minerals and manufactures basic data items, changes affecting the national level were small after the final industry review. This occurs nearly a year prior to the creation of the final historical file. Evidence from the special SIC's in Manufactures suggests that changes affecting the state level were made during the area reviews but that these changes were not frequent. Indeed, six of the 10 special SIC's had fewer than three states affected.
10. For Construction SIC 1794, Excavating and Foundation work, totals at the national level were greatly affected by one large keying error for payroll in one establishment in Colorado. Subsequent errors were propagated by the edit through this establishment's record and all were large enough to affect the national totals for this SIC.
11. During Minerals processing, pseudo-establishments were created by amending the state and county codes to the Census File Number. These establishments represented "well-heads" connected to the establishment that required tabulation at more exact geographic levels. For the value of shipments and cost of materials, this process introduced large errors that required correction at the next stage.

2. Detailed Findings of the Study

Table 2.1 summarizes, at the national trade area level, the changes experienced by each of the variables studied. The table shows five stages of processing for Business and Contruction data and ten stages for Industry data. In all cases, stage one represents the data as it was received by the divisions from ESD and the last stage represents the publication data. No "outliers" i.e. extremely large changes in one establishment's record were corrected. In addition to the volume percent change, the table gives the

percent of establishments which exhibited any change for an item from one stage to the next. This table is an aggregate of the tables produced at the SIC and type of unit levels. These more detailed tables, available upon request, display the changes from stage-to-stage by two basic components: the changes from establishments as they move from one SIC to another and the changes made in individual records within an SIC. The description of the various stages for the trade areas can be found in Section 4.2 for Business, in Section 4.3 for Contruction, and in Section 4.4 for Industry.

TABLE 2.1 Summary of Weighted Percent Change from Stage-to-Stage, by Variable

Trade Area Variable	PERCENT CHANGE - VOLUME					PERCENT OF ESTABLISHMENTS CHANGED			
	1st to last	1 - 2	2 - 3	3 - 4	4 - 5	1 - 2	2 - 3	3 - 4	4 - 5
WHOLESALE TRADE									
Sales	+3.9	+8.6	+2.2	-5.5	-1.0	21.0	4.9	12.0	8.1
Annual Payroll	-78.5	-26.9	-37.7	-41.2	-19.5	39.1	3.5	17.8	11.4
First Quarter Payroll	-48.0	-5.8	-17.2	-14.1	-22.4	37.1	3.0	8.5	6.4
First Quarter Employment	-49.2	-39.4	-2.0	-15.0	+0.6	36.0	2.9	8.1	6.4
Operating Expenses	-69.9	-46.8	-9.7	-35.3	-3.0	27.4	3.9	10.6	8.1
End of 1982 Inventories	-82.0	-51.5	-9.1	-58.2	-2.7	18.1	3.3	8.4	6.6
End of 1981 Inventories	-77.5	-45.6	-10.9	-53.0	-1.5	20.6	3.4	9.1	7.7
RETAIL TRADE									
Sales	-28.0	-5.6	-0.5	-21.2	-2.6	21.5	1.9	8.0	3.7
Annual Payroll	-53.8	-12.6	-2.0	-34.0	-18.2	19.9	1.6	5.9	3.5
First Quarter Payroll	-51.9	+4.3	-2.8	-31.4	-30.8	19.5	1.6	5.7	1.8
First Quarter Employment	-49.9	-4.3	-6.0	-42.1	-1.7	18.5	1.5	5.6	1.6
SERVICES INDUSTRIES									
Receipts	-93.0	-84.3	-7.5	-46.2	-10.5	23.7	2.0	10.7	5.4
Annual Payroll	-79.9	-20.2	-16.5	-65.2	-13.2	20.7	1.8	8.7	4.9
First Quarter Payroll	-71.6	-5.7	-11.0	-61.1	-12.9	19.5	1.7	8.1	4.7
First Quarter Employment	+3.9	+29.6	-5.7	-9.0	-6.5	18.2	1.7	7.9	4.7
Operating Expenses	-72.3	-28.8	-4.1	-57.4	-4.8	3.2	1.1	5.2	4.2
Tax-Exempt Receipts	-80.1	0.0	-7.8	-77.5	-4.4	0.0	0.9	5.2	4.1
CONSTRUCTION INDUSTRIES									
Total Employment	+31.1	+527.0	-54.3	-16.2	-45.4	52.3	2.7	0.8	0.4
Construction Workers	-64.5	+83.8	-72.7	-28.3	-1.3	70.1	2.8	0.7	0.4
Total Payroll	-97.0	-94.5	-3.2	-13.0	-35.0	31.4	3.4	1.0	0.5
Construction Workers Wages	-65.5	-59.5	+7.9	-20.1	-1.2	40.7	3.6	1.0	0.4
Construction Workers Hours	+38.3	+57.5	-2.7	-3.2	-6.8	64.0	8.3	4.1	29.0
Total Construction Receipts	-99.1	-99.0	-2.8	-1.0	-5.4	41.0	4.0	1.2	0.5
Net Construction Receipts	-99.4	-99.4	-5.6	-1.6	+7.5	45.1	4.3	1.2	0.5
Value Added	320.6	+521.0	-11.3	-0.8	-23.0	59.1	4.7	1.3	0.5
Cost of Materials & Fuels	-77.7	-69.7	-0.8	-2.6	-23.9	54.8	4.4	1.3	0.5
Payments for Work Subcontracted out	-36.5	-12.7	+3.0	+0.0	-29.4	29.2	3.3	0.9	0.4
Rental Payments	-87.6	-86.9	-1.5	-0.2	-0.4	14.5	2.9	0.1	0.4
Total Business Receipts	-44.7	-21.6	-4.9	-1.1	-24.9	41.8	4.1	1.1	0.5
Cost of Materials	-77.0	-68.5	-2.4	-0.5	-25.1	41.5	4.3	1.2	0.4
Cost of Fuels	-84.2	-81.9	+27.1	-30.6	-0.9	36.7	3.5	0.2	0.6
MANUFACTURES									
Total Employment	-49.3	+6.9	-51.8	-0.7	-0.6	31.5	7.0	1.5	1.6
Salaries & Wages	-98.0	-98.0	+0.9	-0.8	0.0	35.3	1.8	1.3	1.5
Production Workers	+3.7	-10.2	+17.8	-0.6	-1.1	11.6	29.9	1.4	1.6
Production Workers Hours	-99.0	-93.6	-20.8	-0.3	-73.3	9.4	49.8	1.2	1.5
Production Workers Wages	-96.3	-24.9	-95.0	-0.7	-0.3	9.7	53.2	1.4	1.5
Value Added	+3.5	-44.5	+84.9	+3.2	-1.1	83.4	60.7	5.5	3.2
Cost of Materials	-106.0	-106.5	-5.6	-2.3	-0.2	10.8	32.4	3.3	2.1
Value of Shipments	-35.9	-34.0	-1.4	-0.3	-0.6	46.7	7.3	3.4	3.5
New Capital Expenditures	-23.1	0.4	10.9	-1.5	-3.0	19.2	29.4	1.3	1.2
MINERALS									
Total Employment	-85.8	+3.5	-86.1	+1.6	+3.2	29.0	10.9	8.0	12.7
Salaries & Wages	+28.0	+20.8	-0.2	+2.1	+3.9	38.1	2.6	7.3	15.3
Production Workers	+27.6	+2.3	+35.4	+4.0	-5.6	3.2	38.9	8.2	13.4
Production Workers Hours	-36.3	0.0	-32.2	+2.0	-2.8	2.6	52.9	9.0	14.3
Production Workers Wages	-2.2	-21.0	+25.7	+5.2	-4.4	2.4	47.5	8.2	14.9
Value Added	+22.1	+40.5	+13.3	-12.0	+72.9	66.9	58.8	20.9	25.4
Cost of Materials	+16.3	+8.9	+10.8	-4.3	+42.8	11.7	45.7	15.7	21.2
Value of Shipments	+22.4	+38.0	+6.1	-5.8	+79.3	43.6	7.2	11.7	23.1
Total Capital Expenditures	+12.6	+1.0	+38.9	-21.6	+2.2	5.8	39.5	12.0	16.7

INDUSTRY : LATER STAGES

Census Variable	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10	5 - 6	6 - 7	7 - 8	8 - 9	9 - 10
MANUFACTURES										
Total Employment	-0.1	-0.1	0.0	0.0	0.0	1.0	0.1	0.0	0.1	0.0
Salaries & Wages	-0.1	0.0	-0.1	0.0	0.0	1.0	0.0	0.0	0.1	0.0
Production Workers	-0.1	-0.1	0.0	0.0	0.0	1.0	0.0	0.0	0.1	0.0
Production Workers Hours	-23.8	-0.1	-0.1	-0.1	0.0	1.0	0.1	0.0	0.1	0.0
Production Workers Wages	-0.1	0.0	0.0	0.0	0.0	1.0	0.0	0.1	0.1	0.0
Value Added	-0.5	-0.8	0.0	0.0	0.0	15.1	0.1	0.2	0.3	0.0
Cost of Materials	-0.4	+0.7	0.0	-0.2	0.0	1.4	0.1	0.1	0.1	0.0
Value of Shipments	-0.4	0.0	0.0	-0.1	0.0	1.8	0.0	0.1	0.1	0.0
New Capital Expenditures	-27.3	-0.2	-0.3	-0.1	0.0	1.0	0.0	0.1	0.1	0.0
MINERALS										
Total Employment	+1.3	-6.5	-0.2	0.0	0.0	1.7	0.2	0.1	0.0	0.0
Salaries & Wages	+1.8	-1.4	-0.2	0.0	0.0	1.7	0.6	0.4	0.0	0.0
Production Workers	-0.6	-7.7	-0.2	0.0	0.0	0.9	0.7	0.1	0.0	0.0
Production Workers Hours	-0.2	-4.9	-0.2	0.0	0.0	1.4	0.6	0.1	0.0	0.0
Production Workers Wages	-0.1	-1.7	-0.2	0.0	0.0	1.0	0.6	0.1	0.0	0.0
Value Added	-48.8	-0.4	-1.1	0.0	0.0	1.8	2.4	0.9	0.9	0.3
Cost of Materials	-27.0	-3.4	-0.1	0.0	0.0	1.8	1.4	0.3	0.1	0.1
Value of Shipments	-48.9	-2.3	-0.9	0.0	0.0	1.0	1.2	0.7	0.1	0.0
Total Capital Expenditures	-1.0	+1.1	0.0	0.0	0.0	0.9	1.7	0.2	0.8	0.2

All records whose SIC code was 5052, Wholesale Distributors of Coal and other Minerals, were transmitted to SRD whether or not they were included in the original sample. When these records were tabulated with a weight equal to unity, the data represented the actual values for this SIC at each stage of processing and, moreover, allowed for the exact breakdown of the data at the state level since all records for SIC 5052 were present. Table 2.2 shows the percent change, at the state level, for this SIC from the 4th stage of processing to the 5th or publication stage. This table exemplifies the large changes that occurred late in the processing in this SIC and the corresponding larger changes at the state level with respect to the national level.

Table 2.2 Wholesale SIC 5052 Percent Changes from Stage 4 to Stage 5

	Counts	Sales	Annual Payroll	First Quarter Payroll	First Quarter Employment	Operating Expenses	Ending 1981 Inventory	Ending 1982 Inventory
United States	-5.5	-4.0	-75.5	-28.9	-7.2	-7.7	-.2	-1.0
Alabama	-5.9	-10.6	-27.0	-24.8	-30.3	-23.6	-12.4	-25.5
Alaska	.0	.0	.0	.0	.0	.0	.0	.0
Arizona	.0	.0	.0	.0	.0	.0	.0	.0
Arkansas	.0	.0	.0	.0	.0	.0	.0	.0
California	-8.3	.0	.2	.1	.0	-1.4	.2	.2
Colorado	.0	.0	.0	.0	.0	.0	.0	.0
Connecticut	-15.4	-38.8	-1.3	-.4	2.3	-64.5	.1	4.9
Delaware	.0	.0	.0	.0	.0	.0	.0	.0
Wash., D.C.	-28.6	-8.6	-63.0	-62.3	-50.0	-84.3	.0	.0
Florida	-5.9	.0	.0	.0	.0	-1.9	.0	.0
Georgia	.0	.0	.0	.0	.0	.0	.0	.0
Hawaii	.0	.0	.0	.0	.0	.0	.0	.0
Idaho	.0	.0	.0	.0	.0	.0	.0	.0
Illinois	-2.9	.0	.0	.0	.0	.0	.0	.0
Indiana	-13.3	-11.4	-23.7	-39.9	-30.1	-20.3	.0	.0
Iowa	-50.0	-5.5	.0	.0	.0	.0	.0	.0
Kansas	.0	.0	.0	.0	.0	.0	.0	.0
Kentucky	-7.4	.3	-.1	-1.2	-3.3	.3	7.4	2.5
Louisiana	.0	7.4	47.4	184.6	300.0	44.9	17.3	.0
Maine	.0	.0	.0	.0	.0	.0	.0	.0
Maryland	.0	.1	1.9	.0	.0	.6	19.7	6.8
Massachusetts	.0	.0	.0	.0	.0	.0	.0	.0
Michigan	-4.8	.0	.0	.1	.9	-.1	.1	.0
Minnesota	.0	.0	.0	.0	.0	.0	.0	.0
Mississippi	.0	*******	*******	.0	.0	*******	*******	*******
Missouri	.0	.1	.6	.5	1.1	.7	.0	.0
Montana	.0	.0	.0	.0	.0	.0	.0	.0
Nebraska	.0	.0	.0	.0	.0	.0	.0	.0
Nevada	.0	.0	.0	.0	.0	.0	.0	.0
New Hampshire	.0	.0	.0	.0	.0	.0	.0	.0
New Jersey	-3.7	-4.2	.3	.0	.0	.0	.1	.1
New Mexico	.0	.0	.0	.0	.0	.0	.0	.0
New York	-2.2	.0	.0	.0	.0	.0	.0	.0
North Carolina	-20.0	-.2	-10.9	-11.5	-10.3	-11.5	.0	.0
North Dakota	.0	.0	.0	.0	.0	.0	.0	.0
Ohio	-5.9	-8.2	-97.4	-80.4	-13.9	-23.0	-5.0	-4.9
Oklahoma	.0	.0	-.3	.0	.0	.0	.0	.0
Oregon	.0	.0	.0	.0	.0	.0	.0	.0
Pennsylvania	-1.3	-.8	-1.4	-1.8	-2.8	-3.2	1.9	-1.0
Rhode Island	.0	.0	.0	.0	.0	.0	.0	.0
South Carolina	.0	.0	.0	.0	.0	.0	.0	.0
South Dakota	.0	.0	.0	.0	.0	.0	.0	.0
Tennessee	-5.7	-.3	-3.5	-3.2	-6.1	-2.5	.0	.0
Texas	-5.6	-27.8	-36.5	-32.8	-34.9	-25.6	.0	.0
Utah	.0	.0	.0	.0	.0	.0	.0	.0
Vermont	.0	.0	.0	.0	.0	.0	.0	.0
Virginia	-3.6	-.1	-.5	-1.1	-1.2	-2.2	-.9	-1.2
Washington	.0	.0	.0	.0	.0	.0	.0	.0
West Virginia	-13.6	-3.5	-10.3	-9.2	-10.6	-14.6	.0	.0
Wisconsin	.0	.0	.0	.0	.0	.0	.0	.0
Wyoming	.0	.0	.0	.0	.0	.0	.0	.0

******* indicates a value that changed to a positive number from zero.

Table 2.3 is another example of a type of detail table produced from this study. Using the state breakdown tables for Construction SIC 1794, Excavating and Foundation Work, the following computation was completed for each variable and state. The following example uses the variable construction workers' hours for all establishments in Alabama:

Volume change from stage 1 to stage 2	=	915
Volume change from stage 2 to stage 3	=	58
Volume change from stage 3 to stage 4	=	-51
Volume change from stage 4 to stage 5	=	-100
Total positive volume change from stage 1 to stage 5	=	973
Total negative volume change from stage 1 to stage 5	=	-151
Net volume change from stage 1 to stage 5	=	822
Gross volume change from stage 1 to stage 5	=	1124
Net change/gross change	=	73.29%
Gross change/pub. value (2715)	=	41.40%

Thus, gross changes equaling 41.4% of the final tabulated value occurred for the establishments in Alabama. Table 2.3 was constructed by computing the gross change/published value ratio for each of the 51 states. Then, for each of 15 variables the distribution and certain averages were computed. This table indicates the large amounts of change to which the volume of the data are subjected.

TABLE 2.3. DISTRIBUTION AND MEAN OF THE GROSS CHANGE IN A VARIABLE AS A PERCENT OF THE FINAL TABULATED VALUE OVER THE STATES FOR SIC 1794

	0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-99%	100+%	Average % for all States	Average % for States w/o 100+% Changes	U.S. %
ESTAB	38	13	0	0	0	0	0	0	0	0	0	7.2	7.2	6.4
TE	2	3	9	13	12	2	4	2	0	1	3	11053.0	39.0	16328.0
CW	1	0	0	2	2	4	3	4	1	4	30	14526.9	63.9	21660.9
PRTE	3	8	13	9	4	1	3	1	1	0	8	62693.6	31.1	94367.9
PRCW	2	10	11	8	7	2	0	2	1	0	8	1149.7	31.4	1361.9
HOURS	0	1	8	8	7	13	6	4	1	0	3	61.7	47.1	67.6
TCR	4	5	9	14	7	1	1	1	0	0	9	276.1	30.7	286.8
NCR	2	6	9	11	6	3	1	1	1	0	11	308.4	33.1	331.4
VA	2	2	4	12	5	5	4	2	1	0	14	455.6	40.8	576.5
CM&TF	2	5	10	13	4	2	1	1	0	1	12	528.9	33.5	427.6
SO	7	9	7	10	4	4	4	1	0	0	5	140.4	31.9	117.9
RENT	3	8	6	6	2	7	1	2	2	2	12	478.3	38.3	678.4
TR	2	6	9	13	5	6	2	1	0	0	7	369.8	34.5	471.8
CM	3	3	10	11	8	3	3	0	0	0	10	644.9	34.3	590.9
TF	5	8	8	9	4	4	2	0	0	0	11	896.8	29.2	1003.7

CONSTRUCTION LEGEND
ESTAB = Number of Establishments (#)
TE = Total Employment (#)
CW = Total Construction Workers (#)
PRTE = Payroll for Total Employment ($1000/Yr)
PRCW = Payroll for Construction Workers ($1000/Yr)
HOURS = Construction Workers Hours (1000 Hrs/Yr)
TCR = Total Construction Receipts ($1000/Yr)
SO = Payments for Work Subcontracted Out to Others ($1000/Yr)
NCR = Net Construction Receipts ($1000/Yr) (NCR=TCR-SO)
RENT = Total Rental Payments ($1000/Yr)
CM = Cost of Materials ($1000/Yr)
TF = Total Cost of Fuels ($1000/Yr)
CM & TF = CM plus TF
TR = Total Receipts ($1000/Yr)
VA = Value Added ($1000/Yr) (VA = TR - SO - CM - TF)

The final table for this section, 2.4, is an example of the inconsistency in the edit flags present in the data. Minerals establishments which have a record for the data upon receipt from ESD (stage 1) and a record representing the output from the referral edit (stage 2) were tallied in a table showing how many of the establishments exhibited a change in total employment from stage 1 to stage 2. For this variable, only two flags were encountered, a report flag and a raked flag. A record flagged as reported, but not raked should not exhibit any change, however this is clearly not the case.

Table 2.4. Example of Edit Flag Inconsistency for Mineral Establishments for the Variable Total Employment from Stage 1 to Stage 2

Number of Minerals Establishments	Not Reported Not Raked	Not Reported But Raked	Reported Not Raked	Reported and Raked	Total
Employment does not change	71	0	506	2	579
Employment changes	2	1	191	5	199
Total	73	1	697	7	778

3. Sample Design and Selection

Three files constituted the frame for the sample used in this study. These were the primary ESD mailout file for the 1982 Economic Censuses, a wholesale single unit "class card" mailout file and a third quarter birth mailout file for new companies not on the primary mailout file.

Discussions were held with each division to determine the nature of the sample for each census and the method by which records would be identified and transmitted to Statistical Research Division (SRD) for this study. For reasons of cost and manageability, the total sample size was 70,682 establishment records. The sample was split according to the relative sizes of the censuses and used the following stratifiers: 2 digit SIC, or type of operation code, single unit or multi-unit, and a size class indicator based

upon an annual payroll code carried on the mailout files. This indicator was not present on the third quarter birth file. Within this framework, 365 strata were identified and, within each, a systematic sample was selected.

A compact summary of this sample design is given below.

TABLE 3.1 Sample Design

FILE	CENSUS	# STRATA	UNIVERSE SIZE	SAMPLE SIZE
Primary ESD Mailout	Construction Ind.	58	172,151	9,076
Primary ESD Mailout	Minerals	20	24,745	893
Primary ESD Mailout	Manufactures	96	207,877	16,302
Primary ESD Mailout	Wholesale Trade	10	276,601	4,726
Primary ESD Mailout	Retail Trade	34	884,235	14,041
Primary ESD Mailout	Service Industries	66	924,749	17,016
Primary ESD Mailout	Central Adm. Off.	6	37,954	2,540
Primary ESD Mailout	Company	1	7,580	758
Wholesale"class card"	Wholesale Trade	5	153,180	2,634
3rd Quarter Birth	All areas except CAO and Company	69	175,305	2,696
TOTAL		365	2,864,377	70,682

The sample was designed to provide national estimates at the 2-digit SIC level, with a further breakdown by type of unit.

Another feature of the sample was the inclusion of all establishments in 14 SIC's or subsets of SIC's. Also, records for all those establishments which moved into these SIC's were obtained. This allowed for tabulations based on all of the establishments in those SIC's at a given stage. Since all of the records are present, tables for each SIC can be produced at lower geographic levels.

Ten of these SIC's were in Manufactures, with one each in Construction Statistics, Wholesale Trade, Retail Trade, and Service Industries. A special SIC was not chosen for Minerals. The SIC's for Manufactures were 2084, 2421,

2515, 2741, 2851, 3271, 3471, 3585, 3662, and 3732. For Construction Statistics, it was 1794; for Wholesale Trade, 5052; for Retail Trade, 5411 (Colorado only); and for Service Industries, 737 (California only).

4. Methodology

4.1 General

Establishments selected into the sample had a "trace sample" flag placed on the ESD control record. This flag was set to one and was to remain set throughout the entire processing. If a single unit became a splitter (i.e. was found to have more than one establishment), the flag was to be carried on all establishments connected to the original single unit. Records which were not in the sample, but entered into one of the certainty SIC's would have that record transmitted to SRD on the basis of the SIC code and not on the presence of the flag.

The mail returns were received, reviewed and keyed by the Jeffersonville, Indiana processing office, and the computer records were subsequently released to the appropriate division by ESD. As the divisions processed the data, files were transmitted to SRD. The files represented a variety of types reflecting the idiosyncracies of each division's processing system.

The initial processing of the files by SRD was the matching of the transmitted records to the sample files of mailout records. If a record did not match and the flag was set, it became a time-consuming task to tie the record to an original sample record. Over 3,300 of these unmatched records were identified and were the result of the splitter processing in Jeffersonville and of changing Census File Numbers (CFN's) for some reason, e.g. change of ownership. Of the nearly 17,000 unmatched records where the flag was not set, all but about 200 of these records were found to be transmitted by virture of their inclusion in one of the 14 speical SIC's.

These 200 records represented Censtral Administrative Office (CAO) records which were to be included in Minerals tabulations, but there was no reason for their transmittal to SRD and these records were subsequently dropped from the files.

There were also sampled cases for which no record was transmitted to SRD from any divisions. These cases were most likely establishments which went out of business and no mail return was received. It is also possible that, through a programming error, certain cases were returned with the trace sample flag changed, e.g. because of a CFN change. A total of 4009 establishments were in this category, representing 6.5% of the total sample on a weighted basis. Minerals had the lowest return rate to the sample at 86.7%.

By matching the transmitted files from the divisions to each other, it was possible to track the transfer and deletion of establishments from one census to another. This was not an overly common occurance, however, 12 establishments were identified as being tabulated in two different censuses. In general, most of the SIC-switching activity occurred within a census rather than movement from one census to another.

Table 4.1.1 provides a summary of the returns to the sample.

Table 4.1.1 Returns to the Sample

	Universe	Sample Size	Returns	Return Rate (%)	Weighted Sample Size	Weighted Returns	Weighted Return Rate (%)	'ADDs'
Mailout File								
Minerals	24745	893	785	87.9	24769	21480	86.7	0
Manufactures	207877	16302	15572	95.5	207740	199824	96.2	10781
Construction	172151	9076	8141	89.7	172338	154117	89.4	3944
Wholesale	276601	4726	4559	96.5	276616	266597	96.4	228
Retail	884235	14041	13416	95.5	884345	837260	94.7	598
Services	924749	17016	16435	96.6	924789	886976	95.9	1200
CAO	37954	2540	2256	88.8	37957	33710	88.8	0
Company	7580	758	757	99.9	7580	7570	99.9	0
TOTAL	**2535892**	**65352**	**61921**	**94.7**	**2536134**	**2407534**	**94.9**	**16751**
Wholesale "Class Cards" File Third Qtr.	153180	2634	2421	91.9	153197	141668	92.5	0
Birth File	175305	2696	2331	86.5	175476	129620	73.9	0
TOTAL	**2864377**	**70682**	**66673**	**94.3**	**2864807**	**2678822**	**93.5**	**16751**

4.2 Processing Business Division Files

Business Division transmitted files to SRD on a flow basis. These files represented those sampled records which were included during a pass through their complex edit. All changes to Business data were to be accomplished by correction runs using the complex edit to insure that the corrections did not create an "inconsistent" record for the edit's purposes. Each record received from Business contained the basic data only, no trailer data was transmitted. Although many records were received only once in the process, it was not uncommon for several copies of a record to be transmitted. (One record was received ten times.) Many times no change to the basic data was noted in these copies as the record was re-introduced to the edit for changes in the descriptive fields or in the trailer data. All of the files received were merged into one file for input to a program designed to create the tables for this study.

This program converted the records received for each establishment into data representing four separate stages for the Business processing. Using the cycle number (i.e. the number of times this establishment had passed through the edit), the processing batch number, and the correction batch number on the records, multiple records were put into sequential order. If more than four copies of a record were present, those copies exhibiting no change in the SIC code or the basic data were eliminated.

Since each record carried the basic data as reported in separate fields from those to be tabulated, the record identified as the first pass through the complex edit provided data for two separate stages. These were stage one or reported data i.e. the data as received by Business Division from the record transmitted by ESD. Stage 2 data was outputted from the data in the tabulation fields and represented the data after the initial pass through the complex edit. Stage 3 and Stage 4 data represented the tabulation fields on any subsequent passes through the complex edit. If these records did not exist, stage 3 and/or stage 4 data was repeated from the previous record so that four stages of data would exist for each establishment. If only two records were present, the 2nd record was designated on stage 4 and stage 3 was repeated from stage 2. In essence, Stages 3 and 4 represent the interaction between the analysts and the complex edit as records were referred and corrected.

Stage 5 data was the tabulation data and its source was a pass through the entire Business Division Census file at the time the basic data was considered final. Any changes in the data between this file and the records previously collected were most likely the result of a small number of the previous correction files not being processed by SRD. Also, for any establishment which appeared for the first time in this file, mainly adds to

the special SIC's, a set of stage 1-4 records were created using the edit flags to make a complete set of data.

From these files, a set of records for four or five stages of data was created for each establishment. Each record contained the following information: Census File Number (CFN), stage number, SIC, state code, 7 basic data values, if applicable and the 7 "active" or last appropriate edit flag for a data item. The 7 data values and their appropriate trade areas were:

1. sales or receipts - WRS (Wholesale, Retail, Services)
2. annual payroll - WRS
3. first quarter payroll - WRS
4. first quarter employment - WRS
5. operating expenses - WS
6. ending inventories - W or tax exempt receipts - S
7. beginning inventories - W

The following basic data flags were defined by Business Division for each variable and were present for any record designated stage two through stage five.

Flag	Definition
A	1982 administrative data (straight substitution)
C	Corrected by clerk/analyst
D	Derived from special inquiry data (item initially blank)
H	Imputation based on the establishment's 1977 census data
I	Imputation based 1977 industry averages
J	Imputation based 1982 industry averages
K	Corrected by complex edits
P	Imputation based on the establishment's 1981 administrative data

R Reported data

S Secondary release from ESD

T Item out of tolerance with 1982 administrative counterpart

X Service only: Tax Exempt receipts forced to equal total receipts

Z Blank item set to 'Ø' by the Imputation and Tolerance Edit

4.3 Processing the Construction Statistics Division Files

Three sets of files were used to summarize the construction industries' data on which the processing study's tables were based. Two separate summary files were produced with each containing seven variables patterned after the 1982 Census of Construction Industries publications. The first summary file contained the following variables:

1. Average Number of Total Employees (TE)
2. Average Number of Construction Workers (CW)
3. Total Payroll (PRTE)
4. Construction Workers Wages (PRCW)
5. Total Construction Worker Hours (HOURS)
6. Total Construction Receipts(TCR)
7. Net Construction Receipts (NCR) (defined as total Construction receipts less payments for work subcontracted out to others)

The second summary file contained the following variables:

1. Value Added (VA) (defined as all business receipts less payments for construction work subcontracted to others and payments for materials, components and supplies and fuels)
2. Payments for materials, components, supplies, and fuels (CMTF)
3. Payments for construction work subcontracted to others (SO)
4. Rental payments for machinery, equpment, and structures (RENT)

5. Total Business Receipts (TR)

6. Payments for materials, components, and supplies (CM)

7. Payments for fuels (TF)

The first set of files was received from Construction Statistics Division (CSD) on a flow basis and represented the data as they were received by CSD from ESD. This set of files was merged into a single file to produce the stage 1 data. Each record was identified by its CFN and represented one establishment.

One record identified during this process was particularly interesting because of the keying error that produced it and the method by which it was handled in subsequent processing by the CSD complex edit. When a record is keyed, the data field is identified by a key-code which precedes that data field. In this instance, two fields were combined: total payroll (key = 115, data = 2029) and first quarter payroll (key = 116, data = 263). What was keyed and transmitted was: key = 115, data = 2029116263. Thus the payroll value was enormously inflated, while first quarter payroll would not be identified.

The second set of files was also received on a flow basis and represented the multiple passes of an establishment record through the CSD complex edit. During each pass, only those records whose trace sample flag was set to one or whose SIC code was 1794 were sent to SRD. Although a particular establishment might generate up to ten records, data records designated as stages 2, 3, and 4 were defined. Similar to Business data, stage 2 represents the 1st pass through CSD complex edit, while stages 3 and 4 represent the changes that were made in the data due to the following activities: complex edit failure, SIC review, area review, preliminary tab review as well as other analyst

activity. It was not possible to distinguish the exact reason why a change occurred in a particular data field.

From the collection of this set of files, nearly twice as many establishments were counted than were present on the stage one file. The large discrepancy between the number of establishments from these files and the stage one count of establishments is attributable to two factors: first, the movement of establishments into the special SIC 1794 where the stage one record's SIC was not 1794; and secondly, establishments which were mailed as Construction cases but were non-respondents to the Census did not have a Stage 1 record created by Construction, however for this study, an imputed record representing stage 1 data was created. This record contained zeros in the data fields

The establishment with the keying error mentioned during the stage one processing passed though the complex edit six times, however, only three versions exhibited any change for the variables in question and these were designated as stages two through four. The data for these stages showed that the inflated value for payroll was lowered by the edit by a factor slightly over 100 whereas a factor of 1,000,000 would have been correct. The edit then used the payroll data to corrupt most of the other basic items by grossly inflating them to match the payroll figure. An analyst corrected the entire record so that it matches the reported record (except for the number of contruction workers and construction workers hours) at the 4th processing stage. Besides the extreme variability in this record's individual data items, two basic ratios, the annual salary for construction workers (PRCW/CW) and the hourly wage rate for construction workers (PRCW/HOURS) also exhibited large variation. However, two accounting relationships among these variables were always satisfied, namely, NCR=TCR-SO and VA=TR-CM-TF-SO. These two

conditions imply that the complex edit does not effectively coordinate ratio-type edit tests and absoulte value-type tests over a set of several data items. See Table 2.3 for definitions of these variables.

The third file came from a pass of the entire Construction Census Data Register after all corrections were completed and represented the final census data. This file contained all the establishments that were designated in the sample or were to be tabulated in SIC 1794. The data was designated as the stage 5 or tabulation data. Any changes between the stage 4 data and the tabulation data were attributable to changes made outside the context of the edit as all of the intermediate data files were processed by SRD.

• The Census of Construction Industries is in reality, a sample not a complete census. Construction Division assigns two sample weights to their records: an establishment weight and a form weight. During the processing study, these weights were always applied to the construction data items. Any reference to a "weighted" table for construction data refers to the use of the sample weights for the processing study.

Similar to Business Division, the output for these construction files consisted of the CFN, stage number, the SIC, state code, and the seven variables for each of the two data sets (2 separate files were created). In addition for Stages 2 through 5, a hybrid edit flag was created using the impute flag fields on the construction record. This hybrid is a function of the establishment impute code with the values 0, 1, 2 and the individual item impute code with the values 0, 1, 2 and was constructed by use of the following table:

	Individual Item Impute Code		
Establishment impute code	0=total and detail reported	1=total and detail imputed	2=total reported, detail imputed
0: reported data	A	B	C
1: impute via model schedule	D	E	F
2: impute via data base procedure	G	H	I

Unfortunately there were no flags on the record specifying an analyst correction. The record with the keying error carried all "A" flags on its data items for the stage 5 record, i.e., reported data. While the data values present are indeed reported values (except for two variables) using these codes in the context of that record does not nearly indicate the rich history of processing that record.

4.4 Processing Industry Division Files

Industry Division files were transmitted to SRD using a different system from Business and Construction Statistics Divisions. Industry Division identified ten different stages and the corresponding files were received by SRD on both a flow basis and a "drop" basis (i.e. a pass through the file following the completion of a processing event). The table below describes these stages.

STAGE	Transmittal date/method	Description
1	through 12/83, flow	Data as received by Industry from ESD
2	through 12/83, flow	Output from Referral Edit
3	through 12/83, flow	Output from Correction System
4	1/84, drop	Output from Preliminary Industry Tabs
5	4/84, drop	Output from Preliminary Area Tabs
6	6/84, drop	Output from Manufactures Final Industry Tabs

7	9/84, drop	Output from Manufactures Final Area Tabs
		Output from Minerals Final Industry Tabs
8	11/84, drop	- SPECIAL TAB RUN
9	1/85, drop	Output from Minerals Final Area Tabs
10	6/85, drop	- FINAL HISTORICAL FILE

The processing of these files separated minerals from manufactures until the receipt of the final stage, when they were merged into a large industry file containing at most ten records per establishment. A record need not be present at each stage, especially those included only as a part of a special SIC, however, the large majority of establishments did have all ten stages.

Different variables were outputted for Manufactures and Minerals. These variables were the primary ones used in the Census tables published by Industry Division. Also, minerals establishments were often split into "well-heads" requiring a 15-digit CFN to be created and sent to SRD. Since the sample for minerals did not contain any special SIC's where tables by state or other geographic areas would be created, these records were merged back together into the 10-digit CFN representing the original establishment.

The files for manufactures and minerals contained nine data variables. Eight of these variables were the same for manufactures and minerals, while one differed. Besides the nine variables, the processing study record contained the CFN, stage number, SIC code, state code, and nine edit flags paired to the respective data variables. As before, there were no codes for the stage one data. The variables common to both were: total employment, total salaries and wages, number of producton workers, production workers hours, production workers wages, value added (defined differently by the two

areas), cost of materials, and total value of shipments. New capital expenditures was the ninth variables for manufactures, while minerals used total capital expenditures.

The flag paired to each variable item was derived from a much more elaborate system than Construction used. For Industry Division, each data item had its own set of flags, as well as numerous other flags carried on the records. For this study however, the edit flag was derived by using only the current year edit flags for each specific variable. The following table describes the flags as found on the processing study records.

There were two analyst correction codes, two report codes, two impute codes, two rounding codes, and two raking codes. Each of the 32 possible combinations of these binary codes was recoded using the scheme below. The Industry Division mnemonic appears in parentheses after the description and in the right most column of the table.

Analyst: 00 = No correction 10 = Correction by analyst (A)
Reported: 0 = Blank (B) 1 = Reported
Imputed: 0 = Not imputed 1 = Imputed (I)
Rounded: 0 = Not rounded 1 = Rounded (R)
Raked: 0 = Not raked 1 = Raked (S)

Analyst	Reported	Imputed	Rounded	Raked	Character Recode	Industry Mnemonic
00	0	0	0	0	A	B
00	0	0	0	1	B	BS
00	0	0	1	0	C	BR
00	0	0	1	1	D	BRS
00	0	1	0	0	E	BI
00	0	1	0	1	F	BIS
00	0	1	1	0	G	BIR
00	0	1	1	1	H	BIRS
00	1	0	0	0	I	
00	1	0	0	1	J	S
00	1	0	1	0	K	R
00	1	0	1	1	L	RS
00	1	1	0	0	M	I
00	1	1	0	1	N	IS
00	1	1	1	0	O	IR
00	1	1	1	1	P	IRS
10	0	0	0	0	Q	BA
10	0	0	0	1	R	BSA
10	0	0	1	0	S	BRA
10	0	0	1	1	T	BRSA
10	0	1	0	0	U	BIA
10	0	1	0	1	V	BISA
10	0	1	1	0	W	BIRA
10	0	1	1	1	X	BIRSA
10	1	0	0	0	Y	A
10	1	0	0	1	Z	SA
10	1	0	1	0	!	RA
10	1	0	1	1	)	RSA
10	1	1	0	0	(	IA
10	1	1	0	1	$	ISA
10	1	1	1	0	*	IRA
10	1	1	1	1	+	IRSA

4.5 Processing CAO and Company Data

As of September, 1985, the fourth and final file for this part of the study had not yet been received. A data file was created using the three available files in a format similar to the other trade areas, however tabulations were not produced awaiting the final data. The three stages for this data are:

1. Original Response Data
2. Complex Editing
3. Preliminary Company Review

The data file produced contained, as usual, the CFN, stage number, type of operation code or 9991 for company records, the state code, seven data items, along with their corresponding edit flags. The seven data variables were different for CAO and company records. These variables and the set of flags were:

CAO Data Variable	Company Data Variable	Flags (both CAO & Company)
1. Annual payroll	Total Sales/Receipts	R = Reported
2. Total Employment	Total Payroll	Z = Reported Ø
3. Total Fringe Benefits	Total Employment	X = Computer changed
4. 1982 Ending Inventories	1982 Ending Inventories	I = Computer imputed
5. Total Capital Expenditures	Total Capital Expenditures	C = Analyst correction
6. 1982 Depreciable Assets	Gross Depreciable Assets	A = SSEL (Admin.)
7. Electricity Consumed	Total Assets	N = Non-response

4.6 Sample Estimation

The sample weight, w, carried on each sampled record was the integer, k, in the expression N = nk+q for systematic sampling when N is the number of establishments in a stratum and n the sample size. Each establishment that was a response was then tabulated in the stratum indicated by its current status as to type of unit and SIC and not its sampled status. Thus the total X'_h for any variable within stratum h would be: $X'_h = \sum_{i=1}^{n'_h} w_i \, x_{ih}$ where n'_h is the count of establishments currently in stratum h and w_i is originally sampled weight carried by x_{ih}. By summing over appropriate subsets of the

total number of strata, estimates were produced for larger aggregates of the data. Constraints of time and money did not permit computation of the variances for this sampling design.

For the special SIC's, when the sample was weighted up, those cases which were included only as being in the chosen SIC and not as part of the original sample, received a weight equal to zero. Here the unweighted results are of more importance as they represent the activity of those SIC's at each stage of processing without sampling error.

4.7 Tables Constructed From the Files

Using these files, a large set of tables, available upon request, was computed for the national sample in each trade area. Corresponding to the basic stratification, each cell in the tables is identified by its SIC, whether weighted or not, data item, and type of unit. Within each of these cells, four items are present: the count of establishments, the volume of the data, the percent change from the kth stage to the last stage, and the stage-to-stage percent change.

For twelve of the special SIC's, tables were constructed showing the state breakdown for each variable and for selected ratios between basic variables. Stage-to-stage percent changes for these complete tables were also computed.

5. Relationship of this Study to Similar Studies

This was the first time a study of this nature and size was attempted. The data collected for the processing study has shown itself to be useful for three other studies conducted in conjunction with the 1982 Economic Censuses and would have been of more value if the coordination between the projects had been better.

With respect to the study, an Evaluation of Edit and Imputation Procedures used in the 1982 Economic Censuses, this processing study provided the method for capturing the data from Business Division. However, there was no overlap between the special SIC's chosen for the two studies. Even so, the two studies provided useful corroborative evidence for consistency in the types of data patterns encountered.

The construction statistics data collection methodology for this study should have been able to provide much more data for the Large Observation Study. However, the documentation concerning the processing in Construction Statistics Division provided some useful material for this study.

The timing for the selection of the samples for this study and the Content Evaluation of the 1982 Ecnonomic Census Petroleum Distributors did not permit the intentional inclusion of any cases in both studies. Some cases were included in both, but only on a chance basis, and, in hindsight, it would have been preferable for SIC 517, which contain petroleum distributors to have been chosen as a second special SIC for wholesale trade.

Of a more general nature, this study provided programming expertise within SRD to process information from the economic area. Valuable contacts within the programming branches within the divisions were established and maintained. Both of these factors facilitated the collection of economic area data for the 1982 studies and for future studies conducted by SRD in conjunction with the economic divisions. Indeed, the collection of tabulation data and detail record data for the Content Evaluation of Wholesale Trade was accomplished more easily because of the knowledge gained through the acquisition of data for the processing study.

The data base generated for the processing study should also provide a basis for testing research hypotheses in economic data without incurring the cost of creating data for such a purpose, since data was collected for all SIC's for both single units and multi-units.