BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-86/25

ADMINISTRATIVE DATA STUDY FOR THE 1982 ECONOMIC CENSUSES: PART II

by

Robert T. O'Reagan
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

(301) - 763-5350

With the assistance of:

Paul Hanczaryk
Economic Surveys Division

Recommended by: Nash Monsour

Report completed: September 30, 1986

Report issued: November 11, 1986

1. Executive Summary

The Study of Administrative Data was designed to examine the differences between reported 1982 Economic Census data and administrative record data for economic censuses tabulations. A report by Burns (1986) treats that comparison in a wider context; the part of the study covered in this document deals only with examination of photocopies of Census questionnaires and photocopies of the associated administrative records (Forms 941, 1040C, 1065, 1120, 1120S) for the same establishments, all of which were in certain Standard Industrial Classifications (SICs) within the state of Nebraska.

The possibility existed that complete paper documents could show handwritten notes or scratchwork which might shed light on reporting discrepancies. Keying errors might also have had some impact.

Statistical probability sampling was not used to select the 306 establishments examined. Staff members from the economic areas supplied Statistical Research Division (SRD) with a list of SICs that were of particular interest; these were SICs 50-Wholesale Durable Goods, 52-Retail Building Materials and Mobile Homes, 422-Public Warehousing Service, and 862-Professional Organizations Service. For convenience, all establishments were selected from one state, Nebraska, and thus their IRS forms were filed at a single IRS regional office. Cost dictated the number of establishments for this exploratory venture at about 300.

Absence of a probability sample restricts the inferences to a non-statistical type.

Based on observation, the author's conclusions are as follows:

1. No clear patterns of differences were detected for the fields "employment", "first quarter payroll", nor "annual payroll."

2. The "receipts" field showed two apparent reasons for notable discrepancies. First, about 45 percent of the IRS records were for fiscal rather than calendar periods (where that was determinable). Very few Census reports were for fiscal periods, so that in itself could explain many of the differences. The second explanation is to some extent intersected with the first; 30 percent of the IRS records show zero in the unedited tape version of "receipts". This is generally not the figure shown on the IRS photocopy. There are various reasons that the IRS photocopy would show receipts while the unedited IRS tape file would not, primarily that the tax return records from IRS on tape were often received after our closeout (i.e. September 15, 1983) for establishments reporting to IRS on a fiscal year basis.

   Additionally, there is some evidence that "returns" are not always deducted from receipts on the Census questionnaires.

3. Keying errors are not a major problem in terms of frequency, though two of these notable cases are stunning in terms of dollar magnitude. Keying errors seem slightly more common on the Census side than on the administrative records but this is not conclusive.

4. Discernible reporting errors on Census questionnaires appear to be perhaps more frequent than keying errors. Some differences look suspiciously like reporting errors but that cannot be nailed down.

5. Notes, scratchwork or other items to which the Census Bureau was not entitled were almost never visible in the masked photocopies we received from IRS.

The Bureau need not request additional photocopies. Had there been promising avenues for further pursuit, the potential existed for ordering up to 10,000 more establishments, but the cost is around ten dollars per establishment and though our original request was in November of 1984, the final copies were not received until July, 1986.

Recommendations all relate to conclusion number 2. Many of the tax returns not received by closeout (and therefore showing zero in the receipts field on tape) were received by early January 1984. As an example, only 72 percent of the corporations had receipts data available from IRS by mid-September 1983, but that increased to 92 percent by early January 1984. If operational constraints can be overcome, these late receipts might in the future be provided to the divisions.

Alternatively, it would usually be feasible to use the prior (fiscal) year receipts figure if the current (fiscal) year figure is not available. Either would overlap the nominal calendar year by approximately six months. A much more extensive study would be necessary to determine overall whether a fiscal year figure is an acceptable proxy for a calendar year figure.

2. Detailed Findings of this Study

The observations discussed in this report are based on 195 establishments from SIC 52 (Retail Building Materials and Mobile Homes), 52 establishments from SIC 50 (Wholesale Durable Goods), 52 establishments in SIC 422 (Public Warehousing Establishments), and 7 establishments from SIC 862 (Professional Organizations Service). These 306 establishments are all single-units, all in Nebraska.

The more general study (see Burns, 1986) of which this is an offshoot omitted cases which had a blank or zero for a field under examination; this study included all cases. Also, for many tables and comparisons in that parent study, outliers were excluded; this study included all such cases.

No means nor standard deviations were computed since this was not a statistical sample. Patterns which were recognized in the earlier broader study--such as the overall downward bias in the administrative data--were not retested here. There was, however, no evidence to contradict any of those earlier findings.

Differences between administrative unedited tapes and 1982 Census unedited data were defined in the following ways:

- For the field March 1982 employment a discrepancy of two or more employees was called a "difference".
- For the field first quarter of 1982 payroll a discrepancy of $2,000 or more (to eliminate rounding differences) was called a "difference".
- For the field annual payroll for 1982 a discrepancy of $2,000 or more (to eliminate rounding differences) was treated as a "difference" for the tables so labelled, but other tables labelled accordingly used $100,000 as a view into significant "differences".

Categories are usually defined as based on the unedited version received on tape from IRS. Comparably, the Census data are unedited original tape versions. The determining values or "true" values are taken to be the entries seen on the photocopies of the response forms, and where these values are being discussed, that will be made explicit.

Tables 1-5 are counts of establishments in categories as just described.

2.1 First Quarter Employment

Overall, 26 percent of the first quarter employment figures differ by two or more employees between unedited administrative data tapes and Census unedited original data tapes. One establishment differed by 26 employees (for unknown reasons). Ten of the 79 difference cases are due to incomplete Census response, two were cases not keyed by Census, one was clear Census response error, one each keypunch errors by Census and by IRS and two are possibly multi-units. No real pattern is visible.

2.2 First Quarter Payroll

First quarter payroll differs by two thousand dollars or more in 25 percent of the establishments. Of these 75 cases, 25 are due to incomplete Census reports, two are clear Census misreporting, three were for some reason not keyed by Census, and two are Census punch errors. One of the misreporting differences was colossal, showing $46 million on the Census form rather than $46 thousand; no, it was not a keying error. No consistent pattern could be seen.

2.3 Annual Payroll

As to annual payroll, 31 percent of the establishments exhibited

differences of two thousand dollars or greater between Census unedited and IRS unedited tape files. Comparisons with the photocopies provided some explanations. Of these 95 cases, 17 were Census partial non-responses, four were Census key errors (one colossal), three were Census reporting errors clearly, two were not keyed by Census for some reason. No striking patterns are discernible.

2.4 Annual Receipts

This is another story. There do seem to be some patterned explanations for the 71 percent difference rate (if the $2,000 figure is used to define difference) or the 32 percent difference rate (if $100,000 is defined as different). Only 5 of the total 306 establishments are on a fiscal year (non-calendar year) basis as the Census photocopies document it; at least 112 of the IRS photocopies indicate a fiscal year reporting period. This in itself would be expected to cause differences. Of these 112, 84 percent exhibit ($\geq$ $2,000) differences, but 45 percent show large ($\geq$ $100,000) differences.

Another notable group contributing to the difference column is the 59 establishments for which no IRS photocopy was received. In this category 83 percent have differences and 54 percent display large differences. The similarity between these proportions and those for the fiscal year group are supported in another way. Both groups (fiscal year and no tax photocopy) have a high proportion of zeroes in the receipts field on the unedited IRS tape. The fiscal reporters have a zero in the receipts field 27 percent of the time and the no-tax-photocopy group has a zero in the receipts field 80 percent of the time. Calendar year tax photocopies have only 11 percent zeroes in the receipts field.

Zero in the receipts field on the IRS unedited tape is only in agreement with available IRS photocopies 20 percent of the time. In the other 80 percent of the cases, the author assumes that zero is present because the tax form was not available for keying. Remember that fiscal year 1982 might not end until June 30, 1983. Add to that a minimum of four months processing turnaround plus possible filing extensions granted (two of those were seen) and the IRS receipts figure might not be available on tape for "calendar 1982" until sometime in spring 1984. The high proportion of zeroes in the receipts field for establishments for which there was no photocopy makes that group look more like fiscal year reporters than calendar year reporters.

Lumping the known fiscal reporters and the no-photocopies-available into one group accounts for 56 percent of the establishments, 83 percent of the zero receipts on the unedited file. IRS Calendar year reporters account for 44 percent of the establishments reviewed, but only 17 percent of the large differences. Among these calendar year cases, the difference was on 4 occasions seen to be "returns" which were deducted properly on the 1040c but not on the Census questionnaire apparently. Returns were not separately visible on other kinds of IRS forms; about 12 percent of the 1040c reporters failed to deduct returns from their Census receipts figure.

In the face of a non-statistical sample and large uncertainties or gaps in the data, a statistical conclusion cannot be drawn. But the two problem groups--fiscal year reporters when categorizing from the photocopies, or zero receipts cases when categorizing from the computer readable file--seem to be related and definitely account for most of the discrepancies.

The one percent or so keying/reporting errors which account for enormous numerical differences should be detectable by rather simple edits.

3. Sample Design and Selection

Statistical sampling was not applied in this study nor the broader study which preceded it (see Burns, 1986). The establishments examined here were in fact a subset of those from the earlier study. Staff members from the economic areas supplied Statistical Research Division (SRD) with a list of SICs that were of particular interest. For convenience, all establishments were from one state; this reasonably guaranteed that only one IRS regional office needed to become involved in the process of retrieving documents and photocopying them.

The SICs were:

50	Nebraska	Wholesale Durable Goods
52	Nebraska	Retail Building Materials & Mobile Homes
422	Nebraska	Public Warehousing Service
862	Nebraska	Professional Organizations Service

These SICs in Nebraska contained approximately 2,867 establishments in 1982, and include both single-unit and multi-unit establishments, though the cases under investigation were all single units. Single units were chosen because consideration was focused on administrative record data that might be used in lieu of response data for economic census tabulations. Because IRS is our source for the list of such establishments, no confidential Census information was released in transmitting those employer identification numbers.

The size of the sample (306 establishments) was determined solely by cost. If promising results were observed within this small test group, the budget and the sample could have been expanded.

The data fields which were visible on the IRS photocopies are detailed in attachment 1; all other fields were suppressed before delivery. Census made photocopies of the report forms from the 1982 Economic Censuses where such forms were available.

Census found documents for almost 99 percent of the establishments; IRS supplied copies of 941s relating to 97 percent of the establishments and copies of the tax forms for about 81 percent of the cases.

Delivery of the IRS photocopies was slow. Though the request was initiated in November of 1984, the final batch was not received until late July 1986. This is an observation rather than a complaint. Census should recognize that future requests will probably be of low priority compared to IRS's necessary regular work.

4. Processing the Data

The overall study of which this was the second part involved two different research approaches. For Part I, which was reported on by Burns (1986), comparisons were made between computerized data files containing administrative and economic censuses response data sets. Part II, reported on here, examined photocopies of census questionnaires and photocopies of associated IRS records (1040C, 1065, 1120, 1120S, 941) for the same small subset of establishments. The possibility existed that original paper documents would show handwritten notes or scratchwork which might shed light on some discrepancies. Keying errors might also have had some impact. It must be recognized that discrepancies or keying errors had to be defined in relation to the computerized version of the data.

Discrepancies then, were defined with respect to a computer printout showing both the original census data as keyed and the unedited administrative data. As with Part I of the study, the basic data items compared were employment, first quarter payroll, annual payroll, and receipts. Unlike Part I, establishments were not excluded from consideration when one or even all of the fields were blank. Rather, those blanks or zeroes were researched against whatever paper copies were available.

Since means and variances were computed in Part I, and meaningful regressions were attempted, clear outliers were usually omitted from that study. In Part II, outliers were of interest and were always included.

As a consequence of the non-probability sample, the small number of observations, and the inclusion of zeroes and outliers, no statistical estimates are attempted. The reader is referred to Part I (Burns, 1986) for such measures.

Cell counts only are recorded (see tables 1-5). Column categories are defined as regards computer printouts of unedited data. Row categories are defined by reference to administrative record photocopies.

TABLE 1.

Differences (>2 Employees) in First Quarter Employment
Between Unedited 941 Data Tape and Census Unedited
Original Data Tape
Nebraska--Single Units--1982

SIC Group	Total Establishments	Differences > 2 Employees
All	306	79
52	195	48
50	52	11
422	52	19
862	7	1

TABLE 2.

Differences (>$2,000) in First Quarter Payroll Between Unedited 941 Data Tape and Census Unedited Original Data Tape Nebraska--Single Units--1982

SIC Group	Total Establishments	Differences > $2,000
All	306	75
52	195	46
50	52	14
422	52	13
862	7	2

TABLE 3.

Differences (>$2,000) in Annual Payroll Between Summed 941 Data Tapes and Census Unedited Original Data Tape Nebraska--Single Units--1982

SIC Group	Total Establishments	Differences > $2,000
All	306	96
52	195	59
50	52	19
422	52	14
862	7	4

TABLE 4.

Differences (>$2,000) in Receipts Between Unedited Administrative (1040c, 1065, 1120, 1120s) Data Tapes and Census Unedited Original Data Tape
Nebraska--Single Units--1982--SIC Groups 50, 52, 422, 862

	All Establishments		Difference Cases		Others	
Reporting Period on IRS Document	Zero Receipts on IRS Tape	Non-Zero	Zero Receipts on IRS Tape	Non-Zero	Zero Receipts on IRS Tape	Non-Zero
All	92	214	72	146	20	68
Calendar Year 1982	15	120	8	67	7	53
Fiscal Year 1982	30	82	26	68	4	14
No IRS Document Available	47	12	38	11	9	1

TABLE 5.

Large Differences (>$100,000) in Receipts Between Unedited Administrative (1040c, 1065, 1120, 1120s) Data Tapes and Census Unedited Original Data Tape Nebraska--Single Units--1982--SIC Groups 50, 52, 422, 862

Reporting Period on IRS Document	All Establishments		Difference Cases		Others	
	Zero Receipts on IRS Tape	Non-Zero	Zero Receipts on IRS Tape	Non-Zero	Zero Receipts on IRS Tape	Non-Zero
All	92	214	58	41	34	173
Calendar Year 1982	15	120	4	13	11	107
Fiscal Year 1982	30	82	26	24	4	58
No IRS Document Available	47	12	28	4	19	8

REFERENCES

Burn, Geneva A. (1983a), "Study of Administrative Data," unpublished draft of internal Census Bureau document deated May 31, 1983.

------(1983b), "Plans for Study of Administrative Data," internal Census Bureau memorandum for Carol Corby dated July 19, 1983.

------(1984a), "Selection of Groups of Cases for the Administrative Data Study," internal Census Bureau memorandum for Carol Corby dated January 17, 1984.

------(1984b), "Documentation of Initial Plans for the Administrative Data Study," internal Census Bureau memorandum for Carol Corby dated February 3, 1984.

------(1986), "Administrative Data Study for the 1982 Economic Censuses," Statistical Research Division Report Series, RR-85/22.

SPECIFIC SERVICES AND CONDITIONS

(1) the Internal Revenue Service (Service) will provide the Bureau of the Census (Bureau) with copies of certain business income tax forms and quarterly Form 941 for 1982 for 306 EIN cases. (If additional EIN cases are requested a separate contract will be issued.) The copies will contain the following data items.

1. From Form 1120

   (a) Employer Identification Number (EIN)
   (b) Principal industrial activity (PIA) code
   (c) Reported gross receipts less returns and allowances
   (d) Reported gross royalties
   (e) Cost of goods sold
   (f) Accounting period covered

2. From Form 1120S

   (a) EIN
   (b) PIA code
   (c) Reported gross receipts less returns and allowances
   (d) Reported gross royalties
   (e) Cost of goods sold
   (f) End-of-year code
   (g) Months actively operated
   (h) Accounting period covered

3. From Form 1065

   (a) EIN
   (b) PIA code
   (c) Reported gross receipts less returns and allowances
   (d) Reported net income from royalties
   (e) Reported net farm profit
   (f) End-of-year code
   (g) Months actively operated
   (h) Accounting period covered

4. From Form 941

   (a) EIN
   (b) Total compensation paid
   (c) Tax period covered
   (d) Number of employees (for first quarter returns only)
   (e) Taxable FICA wages paid
   (f) Taxable tips reported