

1979 SOLE PROPRIETORSHIP EMPLOYMENT AND PAYROLL: PROCESSING METHODOLOGY

Nick Greenia, Internal Revenue Service

I. BACKGROUND

As the result of an interagency agreement between the Internal Revenue Service (IRS) and the Small Business Administration (SBA), IRS Statistics of Income (SOI) Division is augmenting its tabulations of business financial data (income statement, and balance sheet, when possible) with two additional data items, payroll and employment, from employment tax returns, Form 941 and Form 943. Employment is also to be used as an additional table classifier. The Small Business Administration (SBA) expects that the tabulations will prove useful in the continuing development of its Small Business Data Base in fulfillment of its Congressional mandate (P.L. 96-302 Title IV) to evaluate public policy and economic trends that affect small businesses without thereby placing any additional data collection burden on small businesses [1].

To produce these enhanced data, SOI is linking its perfected [2] sample files of business information and tax records for corporations (Form 1120 series), partnerships (Form 1065), and sole proprietorships (Schedules C, F, or Form 4835 appended to Form 1040) to their respective Forms 941 (Employer's Quarterly Federal Tax Return) and/or Forms 943 (Employer's Annual Tax Return for Agricultural Employees) in order to abstract employment and payroll from the latter two types of records. The linkage is effected through the Employer Identification Number (EIN).

These studies commence with Tax Year 1979 and will be repeated for all three types of business entity for Tax Year 1982 to coincide with the Economic Censuses. Thereafter, they will be undertaken annually for corporations and quinquennially for partnerships and sole proprietorships [3].

For the Tax Year 1979 Sole Proprietorship Employment and Payroll Study, the process entailed attempting to (a) link the 108,335 business Schedules C and F and Forms 4835 appended to Forms 1040 on the SOI Individual sample file to possible counterpart employment and payroll records in the population files of some 5 million Forms 941 and 943 for all types of business entity; (b) resolve multiple matches and mismatches for matched sole proprietorship/employment and payroll records; and (c) reweight for false unmatched sole proprietorship records.

II. SOURCE FILES

Each of the business employment and payroll studies will add employment and payroll data to the financial data already available from the IRS SOI business statistics series by matching SOI sample files of business income and tax returns with the corresponding quarterly or annual Employer's Tax Returns reporting Federal income tax withheld and Social Security (FICA) taxes (Forms 941 and Forms 943).

Processing for the 1979 Sole Proprietorship Study consisted of linking by EIN sole proprietorship business records associated with the SOI-perfected Tax Year 1979 Form 1040 sample file [4] to Census-perfected extracts of their corresponding Form 941 (Employer's Quarterly Federal Tax Return) and Form 943 (Employer's Annual Tax Return for Agricultural Employees) records. Sole proprietorship business records were appended to the sole proprietor's Form 1040 and for this study were one of the following three types:

- (1) Schedule C (Profit or Loss from Business or Profession),
- (2) Schedule F (Farm Income and Expenses), and
- (3) Form 4835 (Farm Rental Income and Expenses and Summary of Gross Income from Farming or Fishing).

File extracts containing EIN, payroll, and employment were provided by Census for the population of some 5 million Forms 941 and 943 (Census deleted Form 943 employment due to its unreliability as a consequence of the March 12 reporting requirement, seasonality of farm employment, and exclusion of certain employee groups not under Social Security) for Calendar Years 1978, 1979, and 1980. The Census-perfected extracts of Form 941 and Form 943 data were themselves derived from tape extracts originally produced on a contractual basis by IRS (initial processor of the complete data set for tax administration purposes) as authorized by Internal Revenue Code section 6103 for Census as part of Census' ongoing effort to update annually its Standard Statistical Establishment List (SSEL).

Generally, problems of access to data were minor for SOI since all source documents were IRS-related and originally filed with IRS. While data access posed little difficulty for SOI, however, SBA could receive only tabulations of aggregated data--no files of microdata records--due to the restrictions IRS places on the disclosure of confidential taxpayer data under sections 7213 and 7431 of the Internal Revenue Code.

III. MATCH/MERGE METHODOLOGY

Foremost among the challenges presented by the 1979 Sole Proprietorship Study were those relating to the matching variable itself, the EIN, and the sole proprietorship's filing period. Each of these factors directly affected linking procedures and strategies regarding the Form 941 and Form 943 data.

While the EIN was a required entry for a Form 4835 if Form 943 was filed, it was required for a Schedule C or Schedule F if the sole proprietor had a Keogh plan (self-employed deferred compensation plan) or was required to file an employment (Form 941 or Form 943), excise, or alcohol, tobacco, and firearms tax return. Matters were complicated for Schedule C and Schedule F, however, by the Keogh plan provision

as follows. Prior to 1978, employers maintaining Keogh plans were required to have an EIN in order to complete Form 5500-K (Annual Return/Report of Employee Pension Benefit Plan for Sole Proprietorships and Partnerships with Fewer than 100 Participants and At Least One Owner-Employee), even if the only participants were owner-employees (sole proprietors and certain partners). In 1978 and 1979, owner-employee Keogh plans without common-law employee participants (i.e., with only owner-employee participants) were no longer required to file Form 5500-K, but Schedule C and Schedule F instructions for EIN completion still read as described above; that is, Keogh plans without common-law employees were not excluded explicitly. Of the more than 650,000 Forms 5500-K filed for Plan Year 1977, some 450,000 were for plans without common-law employees. Therefore, while it is unclear what the impact of such a situation was for 1979 Schedules C and F, it is apparent that the potential for problems in the 1979 Sole Proprietorship Employment and Payroll Study (false matches to Forms 941 and Forms 943) was considerable.

The EIN potential problem was compounded by the fact that while sole proprietorship Forms 941 and 943 were processed by IRS and posted by EIN to the IRS Business Master File (the computer data storage system from which the original Form 941 and Form 943 file extracts were produced for Census processing/perfection), the sole proprietorship records (Schedules C and F, Form 4835) were processed with the appropriate Forms 1040 and posted to the IRS Individual Master File (IMF) by the Form 1040's Social Security Number (SSN). Little testing or perfection was performed for the sole proprietorship's EIN, and thus, the potential for false matches as well as false non-matches--due to incorrect and even missing EIN's on the IMF side--was significant.

If the sole proprietorship's EIN posed a problem for the link operation, so did its filing or accounting period. Since (a) no such item existed on the business records themselves (it was abstracted from the one Form 1040 to which multiple sole proprietorship records could be appended), (b) a Form 1040 whose accounting period ended in other than December was presumed to have a full-year fiscal accounting period, and (c) 98.6 percent of the 92,694,302 Forms 1040 processed for Tax Year 1979 had Calendar Year 1979 accounting periods, SOI decided that part-year records and other possibly out-of-scope records (e.g., certain prior-year returns) would not be excluded from processing. Instead, the assumption was made that all sole proprietorship records should be treated as full-year calendar 1979 accounting period records. Accordingly, significant savings of both time and money were realized by disregarding the accounting period from the SOI Form 1040 sample file and using only the 1979/1980 Census Form 941/943 file for this study (instead of both the 1978/1979 and 1979/1980 files, as was done for the 1979 Partnership Employment and Payroll Study).

Since EIN generally was required as an entry on the business schedule only in the event of payroll taxes (Forms 941 and 943) or a Keogh plan, EIN-linkages could be contemplated for just a subset of the sole proprietorship sample.

In fact, of the 108,335 Schedules C and F and Forms 4835 on the SOI Sole Proprietorship sample file, only 31,008 had EIN's and, therefore, could be viewed as potential initial matches with the Forms 941 and 943. By type of record, the sample counts were the following.

Form 4835:	40
Schedule F:	2,612
Schedule C:	28,356

IV. PROBLEMS AND RESOLUTIONS

Of the 31,008 records with EIN's (see Figure 1), 24,153 matched on EIN with Forms 941 and/or Forms 943 on the 1979/80 Census extract (EIN was unique for each Form 941 or Form 943 but could have been shared by a Form 941 and a Form 943). Of these 24,153 matches, 4,503 were multiple matches, meaning an SOI sole proprietorship record matched to a Form 941 or Form 943 matching either another SOI sole proprietorship record, an SOI partnership record, or an SOI corporation record. Of the inter-business entity (instead of intra-business entity) multiple matches, 117 were for sole proprietorships matching Forms 941/943 with records on either the SOI Partnership sample file or the SOI Corporation sample file. Consequences would have been dire indeed had all these multiple matches not been individually reviewed (an operation to be treated as obligatory, given the size of the largest possible sole proprietorship weight--over 2,000--and the simply astronomical amounts of payroll, hundreds of millions of dollars per Form 941 for a number of cases, reported for what were probably large corporations).

Figure 1. 1979 Sole Proprietorship Employment and Payroll
Preliminary Unweighted Processing Counts
(Pre-Reweightings)

Item	Number of Businesses (Schedule C and F, Form 4835)
Statistics of Income	
Sample.....	108,335
Without EIN.....	77,327
With EIN.....	31,008
Initially matched on EIN to 1979/80 Form 941 and/or Form 943.....	24,153
Initially unmatched on EIN to 1979/80 Form 941 and/or Form 943.....	6,855

All multiple matches were manually reviewed using one-line record listings containing the following data items: EIN; sole proprietorship industry code; sole proprietorship code (to distinguish between Schedules C and F and Form 4835); Form 1120/1065 code (to identify inter-

business multiple matches, but only those from SOI sample files); sole proprietorship business receipts, business deductions, and proxy payroll (salaries and wages plus cost of labor); Form 941 calendar 1979 payroll; Form 941 calendar 1980 payroll; Form 943 calendar 1979 payroll; and Form 943 calendar 1980 payroll.

At least two factors (other than the questionability of the sole proprietorship's EIN) are responsible for exacerbating the multiple match (as well as the false non-match) situation. The first is the sole proprietorship/corporation "connection" and helps explain at least some of the sole proprietorship/corporation multiple matches and mismatches. Apparently, sometimes a corporation such as a large department store will subcontract work to a sole proprietorship, say, for appliance repair or upholstery cleaning, and the sole proprietorship will incorrectly report the corporation's EIN instead of its own. The second factor concerns multiple sole proprietorships run by the same sole proprietor, even in different business activities. The sole proprietor might legitimately file several different business returns--each with the same EIN (when EIN is necessary)--and either one Form 941 or Form 943 for all businesses or one for each (also using only one EIN). Regardless, IRS would end up processing several business returns but only one consolidated (by either the proprietor or IRS) Form 941/943 containing all employment and payroll data for the sole proprietor. This latter consideration turned out to be quite significant due to the high number of "multiple matches" which were of this variety.

Resolution of multiple matches was accomplished first by "transcribing to unmatched status" sole proprietorship records with non-zero proxy payroll (the sum of salaries and wages plus cost of labor) which matched to a Form 941 or Form 943 whose payroll was egregiously greater than the sole proprietorship's proxy payroll (often sole proprietorship/corporation matches probably). Second, the assumption was made that for purposes of this processing stage, records with zero proxy payroll generally should become unmatched records. Finally, within each group of both like SSN's and EIN's (to ensure that "like" sole proprietorships also belonged to the same sole proprietor or Form 1040), the remaining matches of sole proprietorship records with non-zero proxy payroll were "perfected" by reapportioning the Form 941/943 payroll and employment data among the sole proprietorship records based on their share of the like group's total proxy payroll. When possible, this reapportionment scheme was applied according to the type of sole proprietorship record best corresponding to the Form 941 or Form 943. For example, if a Form 941 and a Form 943 matched a Schedule C and a Schedule F, the Form 941 data were accorded to the Schedule C and those of the Form 943 to the Schedule F. If a Form 941 or a Form 943 matched both a Schedule C and a Schedule F, the Form 941 or Form 943 was reapportioned among both schedules.

Comparison listings were used after resolution to ensure that all problem matches had, in fact, been remedied. Subsequent to multiple

match processing, the final stage in mismatch or false match testing was performed: scrutiny and resolution of matches in which Form 941 or Form 943 payroll exceeded the business record payroll or proxy payroll by at least \$1,000 (see Figure 2). Manual review of one-line listings for these records identified only 45 matches worth retaining; the remainder were dispatched to unmatched status via an algorithm which required Form 941/943 payroll to be strictly less (no tolerance) than the sole proprietorship's business deductions (business deductions was chosen in case proxy payroll had been reported or was "hidden" in deduction items other than cost of labor and salaries and wages) in order for the match to be kept. (The tolerance was dropped for this resolution process due to the large weights observed for a number of sole proprietorships and also because business deductions was sometimes zero.) Comparison listings were again used to verify that no anomalies slipped through processing [5].

Figure 2. 1979 Sole Proprietorship Employment and Payroll
Unweighted Match-Processing Counts
(Pre-Reweightings)

Category	Sole Proprietorship Records		
	Initial EIN Matches to Form 941/943	Retained as Match	Rejected as Match
TOTAL.....	24,153	22,279	1,874
Multiple business record matches.....	4,503	3,612	891
Form 941/943 payroll exceeded business deductions by \$1,000*....	737	45	692
Records with zero 1979 Form 941/943 employment and payroll*	291	0	291
Other matches.....	18,622	18,622	0

* NOTE: Matched records meeting this condition but resolved as unmatched during other processing stage are excluded from this count.

The intent underlying both multiple match and mismatch processing was that only matches with almost certain probabilities of being "good" were to remain as matches. That is, the assumption was that possibly marginal matches were to be treated during these processing phases as "truly false" matches. The goal was to produce a solid reweighting base of good matches so that

reweighting for false non-matches based on their characteristics would be as accurate as possible. It was thought that any marginal cases would be more suitably accounted for later by those characteristics which allied them more closely with either true matches or true non-matches as a result of reweighting analysis.

V. REWEIGHTING

On a weighted basis, only 11.1 percent of the 12,329,982 sole proprietorships in the SOI 1979 population matched a Form 941/943 after resolution of multiple matches and mismatches. Since 82.3 percent of sole proprietorships did not have an EIN and only 7.4 percent of all unmatched records had EIN's, however, this statistic is not as discouraging as it might first appear. In fact, the match rate was 63.0 percent when only records with EIN's are considered.

Final problem adjustments consisted of reweighting for false non-matches [6], based on analytical tables of matched and unmatched frequencies classified by industry, Form 1040 adjusted gross income, business receipts, and proxy payroll (cost of labor plus salaries and wages). Unmatched frequencies were further broken down according to whether sole proprietorship records were with or without EIN, since imputation factors might differ considerably for these two sets.

Reweighting was more significant in terms of impact for the 1979 Sole Proprietorship Study than the 1979 Partnership Employment and Payroll Study [7] largely due to the sole proprietorship EIN problem (the EIN's potential absence and other complications as discussed above) and the distribution of unmatched proxy payroll. Of the \$42.4 billion reported as proxy payroll by all sole proprietorship records (matched and unmatched), only \$28.8 billion or 67.9 percent was accounted for by matched records. If proxy payroll is a good indicator of "true matchability" (97.7 percent of matched records also reported proxy payroll), it seemed that a significant portion of true matches remained to be "found," given that 27.6 percent of unmatched records with EIN's and 22.2 percent of unmatched records without EIN's also reported proxy payroll. Of course, to the extent that proxy payroll consists of contract labor or other "non-true" payroll components, it might not be such a good indicator for certain sole proprietorships--especially for proprietorships filing Schedules F but not required to file Form 943 for employees not under Social Security (see Data Limitations below). Imputation for "missing" data rather than reweighting for false non-matches might be more the issue then.

Reweighting was based upon a file of data defined differently in terms of matched and unmatched status from that of the 1979 Partnership Employment and Payroll Study. For the 1979 Partnership Study, a matched record was defined, primarily for reasons of simplicity and expediency (it was also the first of the business employment and payroll studies to be undertaken and, consequently, the first to encounter new obstacles and the attendant deadlines and cost restrictions in surmounting them), as any Form 1065 matching on EIN with a

1978, 1979, or 1980 Form 941 or Form 943 containing either employment or payroll for 1978, 1979, or 1980. This definition unfortunately allowed into tabulations some records with both zero employment and zero payroll for 1979, since they contained data for either 1978 or 1980. While this definition is being discontinued for future business employment and payroll studies, it also was not used for the 1979 Sole Proprietorship Study, even though a file containing two years (1979 and 1980) of Census Form 941/943 data was used for matching purposes. In fact, only records matching on EIN to a 1979 Form 941 or Form 943 containing employment or payroll data are considered matches--and these criteria must have been met even after multiple match and mismatch problem resolution. That is, records initially "matched" but later transformed to unmatched status as a result of resolution processing are not considered matched for reweighting and table purposes.

VI. DATA LIMITATIONS

Following are qualifications necessary to better understand the data in terms of conceptual limitations posed by slightly different terminologies employed across return forms as well as differences in data reporting requirements:

(a) Sole proprietorship proxy payroll was defined as the sum of salaries and wages plus cost of labor in order to be consistent with the definition of proxy payroll used for the 1979 Partnership Employment and Payroll Study. While this item was used primarily for purposes of comparison with Form 941/943 payroll during multiple match and mismatch processing, definitional differences between these two versions of payroll also warrant aggregate comparisons to ascertain what effect not only actual but also perceived differences had on the data.

Salaries and wages and cost of labor were available from Schedule C as the items wages (form instructions required the reporting of both salaries and wages) and cost of labor but from Schedule F and Form 4835 only as the item labor hired. All of these items should have excluded compensation of the proprietor, but since the Sole Proprietorship Study required gross payroll, they included amounts deducted for jobs or WIN credits.

Overstatement of proxy payroll may have occurred due to inclusion of payments for contract labor, such as certain janitorial, secretarial, or agricultural employees not reportable on Forms 941/943 but deducted on the business schedule, probably under cost of labor. On the other hand, understatement of payroll may have occurred if payroll were reported as commissions, legal and professional fees, repairs, other costs of sales and operations, or other business deductions. Additionally, for certain businesses in the Retail and Services industry groups, tip income would have been reportable on Form 941 but not claimed as a deduction on the Schedule C. Finally, a definition of payroll conforming more closely to the concept of total compensation might also contain contributions to both pension and profit-sharing plans and

employee benefit programs (such as health and prepaid dental insurance), though the proprietor's contributions to the latter were not specifically excluded by Schedule C instructions.

(b) For payroll, Form 941 appears to have required as reportable compensation virtually what was required in the counterpart Form W-2 and Form W-3 items; i.e., income which was taxable but not necessarily tax "withholdable." Form 943 required the reporting of all taxable cash wages to employees subject to FICA taxes, but excluded the value of non-cash items, such as food and lodging--potentially significant components of compensation for agricultural employees and also reportable on Schedule F as a deduction under labor hired. A further limitation was that reportable taxable wages were only required for workers under Social Security (thus, excluding many non-resident alien agricultural workers) and were not to exceed the FICA maximum, a little more than \$22,000 for 1979 and for purposes of this study probably not too detrimental.

In addition to taxable wages, Form 941 required the reporting of all tips and other compensation to employees even if income or FICA taxes were not withheld and specifically excluded only annuities, supplemental unemployment compensation benefits, and gambling winnings--even if income taxes were withheld on these.

(c) While the Form 941/943 March 12 reporting date for employment was an obvious data limitation, it was exacerbated by the possibility of employment double-counting due to employees who worked two or more jobs with different employers filing different employment tax returns.

(d) While testing was conducted to identify possible mismatches in which Form 941/943 payroll was abnormally high, none was attempted (primarily due to time and other cost constraints) for possible false matches or mismatches in which it was too low. For the 1982 study, it might be possible to establish acceptable ranges for payroll/proxy payroll ratios by industry, geography, and certain size classes, but any such operation should be excessively circumspect, given "hidden" proxy payroll, as well as the problem with EIN's previously discussed. (For other recommended enhancements, see also section 10, Greenia, Nick, Match Group Case Study #00002, "1979 Sole Proprietorship Employment and Payroll.")

ACKNOWLEDGMENTS

For their thoughtful review of material in this report, the author thanks Tom Jabine (National Academy of Sciences), Carol Utter (Bureau of Labor Statistics), and Doug Sater (Bureau of the Census). Appreciation is also extended to Wendy Alvey and Beth Kilss for their

help in editing the manuscript and to Dawn Nester and Rodney Turner for typing its many drafts.

NOTES AND REFERENCES

- [1] For further information on the Small Business Data Base, see Kirchhoff, Bruce A. and Hirschberg, David A., "Small Business Data Base: Progress and Potential," 1981 Proceedings: American Statistical Association, Section on Survey Research Methods; Hirschberg, David A. and Phillips, Bruce, "Using Financial Statement Data to Evaluate the Status of Small Business," 1982 Proceedings: American Statistical Association, Section on Survey Research Methods; and Rose, Paul and Taylor, Linda, "Size of Employment in Statistics of Income: A New Classifier," 1982 Proceedings: American Statistical Association, Section on Survey Research Methods.
- [2] File perfection essentially consisted of testing and resolving obvious math errors as well as data inconsistencies in each file record. Errors could have been made by the taxpayer or during a data processing stage.
- [3] A more comprehensive treatment of small business employment and payroll, forthcoming from David A. Hirschberg and Bruce Phillips of SBA, will follow the conclusion of the Tax Year 1979 corporation and sole proprietorship studies. Final tabulations for these two studies were provided to SBA in July 1985.
- [4] For a detailed account of the sampling scheme involved in selecting this sample, as well as other information--including tabulations--concerning this file, see Statistics of Income--1979/80, Sole Proprietorship Returns.
- [5] For more details on the false match resolution phase, see Problems and Resolutions, Greenia, Nick, Match Group Case Study #00002, "1979 Sole Proprietorship Employment and Payroll."
- [6] For a complete discussion of the reweighting process, including its assumptions, see Day, Charles, "Imputation Methodology, 1979 Forms 1040/941/943 Link Study," June 1985. This unpublished report is available upon request by writing to Director, Statistics of Income Division, D:R:S, Internal Revenue Service, 1111 Constitution Avenue, N.W., Washington, DC 20224.
- [7] See Greenia, Nick, Match Group Case Study #00006, "1979 Partnership Employment and Payroll."