

## Linking Administrative Records Over Time: Lessons from the Panels of Tax Returns

*John L. Czajka, Mathematica Policy Research, Inc.*

---

### *Abstract*

*In 1985 and again in 1987, the Statistics of Income (SOI) Division of the Internal Revenue Service initiated panel studies of taxpayers. Taxpayer identification numbers (TINs) reported on a sample of tax returns from the 1985 and 1987 filing years were used to identify panel members and search for their returns in subsequent years. The 1987 panel also included efforts to capture dependents, based on the TINs reported on Aparents@ and dependents' returns. This paper describes and assesses the strategy used to identify panel members and then capture and link their returns. While the availability of a unique identifier greatly simplifies data capture and record linkage and, as in this case, may determine whether or not a record linkage project is operationally feasible, imperfections in the identifiers generate a range of problems. Issues addressed in this paper include elements of operational performance, validation, and measuring the completeness of matching or data capture. Recommendations for improving the success of such efforts are presented, and implications for linkage across administrative records systems are discussed.*

### Introduction

How often, when confronted with a task requiring the linkage of records with imperfectly listed names and addresses, recorded in nonstandard formats, do we long for a unique identifier? This paper addresses some of the problems that analysts may face when they perform exact matches using a unique identifier. The paper deals, specifically, with records that have been linked by an exact match on social security number (SSN). The question it poses is, when is an exact match not an exact match? The paper is more about “unlinkage” than linkage per se. The linkages created by exact matches on SSNs represent the starting point. The work that ensues involves breaking some of these linkages as well as creating additional ones. The findings reported here may be relevant to any effort to link administrative records by SSN, whether longitudinally or cross-sectionally.

### Overview of the Statistics of Income (SOI) Panel Studies

Over the years, the SOI Division of the Internal Revenue Service (IRS) has conducted a number of panel studies of individual (1040) tax returns. These studies employ a common methodology, for the most part. A base year panel sample is selected from the annual SOI cross-sectional sample, which provides a large and readily available sampling frame for such studies. Panel members are identified by their SSNs, as reported on their base year sample returns. The IRS searches for and captures all returns that list panel SSNs as filers in subsequent years. The returns captured by this procedure are then linked longitudinally. In reality, what are linked over time are persons, and these person linkages imply linkages between tax returns. In the two most recent panel studies, described below, the SSNs were edited, after this initial linkage, to correct errors and fill in missing values. After the editing was completed, the linkages were

---

re-established. As a result of this process, some of the original links were eliminated while others were added.

The 1985-based Sales of Capital Assets (SOCA) Panel began with about 13,000 base year returns. All filers on these returns were initially designated as panel members. Joint returns, which can be filed only by married couples, have two filers. Returns with other filing statuses have one filer. A SOCA Panel file covering the years 1985 to 1991 has been completed.

The 1987-based Family Panel began with about 90,000 base year returns. Not only filers but also their dependents (as claimed on base year returns) were defined as panel members. Returns filed by separately filing spouses, whether panel members or not, are to be captured and linked to the returns filed by their panel spouses. Returns filed by the dependents who are claimed in any year after the base year, whether they are original panel members or not, are to be captured and linked as well -- but only for the years in which they are claimed. Work to implement and review the SOI edits and prepare a panel file is only beginning; further editing will take place over the next few months.

## Problems Created by Incorrect SSNs

**I**ncorrect SSNs create a number of problems affecting not only record linkage and data capture but subsequent analysis of the data. In describing these problems, it is helpful to distinguish between incorrect SSNs on base year returns, which by definition include only panel returns, and incorrect SSNs on out-year returns, which include both panel and nonpanel returns.

Incorrect SSNs reported on base year returns have two types of consequences. Both stem from the fact that base year panel SSNs provide the means for identifying and capturing out-year panel returns. First, incorrect base year SSNs produce pseudo-attrition. Individuals whose SSNs were listed incorrectly in the base year will drop out of the panel when they file with correct SSNs. If these individuals are married to other panel members, they will remain in the database, but unless their base year SSNs are corrected their later data will not be associated with their earlier data. These missed linkages lead to incorrect weight assignments, which have a downward bias. A second consequence of incorrect base year SSNs is that the IRS will look for and may link the out-year returns of the wrong individuals to the base year records of panel members. The editing of SSNs is intended to eliminate both kinds of linkage errors.

Incorrect SSNs on out-year returns, as was stated, may involve both panel and nonpanel returns. If a panel member's SSN is misreported on an out-year return, after having been reported correctly in the base year, the out-year SSN will not be identified as panel, which may prevent the panel member's return from being captured at all. This is true if the panel member whose SSN is incorrect is the only panel member to appear on the return. While many panel returns continue to be selected for the annual cross-sectional sample in the years immediately following the base year, such that a panel return may still be captured despite the absence of a panel SSN, the incorrect SSN will prevent the panel member's being linked to the earlier returns. If a *nonpanel* return incorrectly includes a panel SSN, this error will result in, first, the return's being captured for the panel and, second, the wrong individual's data being linked to the panel member's base year record.

The bias that may be introduced by incorrect SSNs is distributed unevenly. Certain types of returns appear to be more prone to erroneous SSNs than others. Clearly, error rates are higher among lower income returns than among higher income returns. They may be higher as well among joint returns filed by couples who have a better than average chance of divorce in the next few years, although this observation is more speculative.

The dollar costs of incorrect SSNs cannot be overlooked either. In addition to the editing costs, there is

a cost to collecting and processing excess returns.

## Identifying Incorrect SSNs

The SSN lacks a check digit. The SSN was established long before it became commonplace to include in identification numbers an extra digit or set of digits that can be used in an arithmetic operation to verify that the digits of the number “add up” right. As a result, there is no quick test to establish that a reported SSN was recorded incorrectly. Instead, it is necessary to make use of a number of other techniques to validate and correct the SSNs that are reported on tax returns or other administrative records.

Range checks are an important tool in screening out incorrect SSNs early in processing. Range checks of SSNs build on what is known and knowable about the distribution of numbers that have been issued by the Social Security Administration (SSA). A very limited range check can be based on the fact that the first three digits of the nine-digit number must fall into either of the ranges 001-626 or 700-728. SSNs with lead digits that fall outside these ranges must be incorrect. (The IRS uses an additional range to assign taxpayer identification numbers to persons who cannot obtain SSNs; these numbers are valid for IRS purposes but cannot be linked to other data.) More elaborate tests may utilize the fact that the 4th and 5th digits of the SSN have been assigned in a set sequence, historically. For each set of first three digits, SSA can report what 4th and 5th digits have been assigned to date or through a specific date. Most of the nine-digit numbers that have never been issued -- and, therefore, are incorrect -- can be identified in this manner. In addition, the SSNs that were assigned to persons who have since died can be obtained from SSA. Brief records for most SSA decedents can be accessed via the Internet.

The IRS maintains a validation file, using data obtained from SSA, to verify not only that particular numbers have ever been issued, but that they were issued to the persons who report them. The validation file contains up to 10 “name controls” for each SSN, where a name control consists of the first four characters of an individual’s surname. If an individual changes his or her name numerous times and registers these changes with SSA, the different name controls will be present on the validation file, sorted from the latest to the earliest. The name control is a relic of period of much more limited computing capacity and less powerful software. The inability of name controls to differentiate among members of the same family, for example, restricts their utility for the editing of tax panel data, since misreporting among family members is a common type of error.

SSA maintains much more extensive data for its own validation purposes as well as other uses. Essentially all of the information collected on applications for new or replacement social security cards is retained electronically. The SSA will also perform validation exercises for other agencies. This was not an option for the IRS data, which could not be shared with SSA, but it may be a viable path for other users to take. In performing its validation and other matching exercises, SSA relies heavily on exact matches on multiple characteristics. SSA utilizes partial matches as well but without the framework of a probabilistic matching algorithm. As a result, SSA’s validation tends to be conservative, erring on the side of making too few matches rather than making false matches.

In editing the SSNs reported on tax panel records, the IRS staff employed a number of evaluation strategies. These are discussed below.

## The SOI Editing Strategy

The editing strategy employed by SOI staff for the two panel databases included several key elements. The first was the use of automated procedures to flag probable errors. The second was the reliance on manual or clerical review to evaluate the cases that were flagged as containing probable errors. Automated validation tests were not always definitive in identifying false matches, so expert review was

often necessary. Furthermore, there was no attempt to automate the identification of the appropriate corrections. The clerical review was responsible, then, for determining if an SSN was indeed incorrect, identifying the correct SSN or an appropriate substitute, and then implementing the needed corrections. The third element of the editing strategy was to correct the base year panel SSNs to the fullest extent possible. This is an important task because the corrected SSNs identify panel members in future years. The fourth element was to eliminate cross-sectional “violations” in the out-years -- that is, instances where particular SSNs appeared as filers multiple times in the same tax year, or where the SSNs listed as dependents matched to filers who were not the dependents being claimed. The last element of the editing strategy was to use automated procedures to apply SSN corrections to other years, where errors might exist but may not have been flagged. These corrections are directed at situations where a taxpayer continues to report an incorrect SSN for a filer, a separately filing spouse, or a dependent, year after year or at least for multiple years. These misreported SSNs may not always be flagged as probable errors. Furthermore, it is highly inefficient to rely on independent identification and correction of these errors.

### **Limitations of the Editing Strategy**

The overall strategy has two notable limitations. First, the sheer number of cases that could be flagged as probable errors in a panel database containing nearly a million records, as the Family Panel file does, is very imposing. The obvious response is to limit clerical review to cases whose probabilities of error are judged to be very high. The SOI Division designed a number of validation tests. Certain tests were considered to be fatal; all violations had to be corrected. For other tests, multiple failures or specific combinations of failures were necessary in order to trigger a review. If a test is associated with a low probability of error, clearly it is inefficient to review all cases. But if there is no other test that in combination with this one can identify true errors with a high enough probability to warrant review, then errors will be missed. Below we discuss some of the problems associated with identifying incorrect secondary SSNs.

Another limitation is that cross-sectional error detection strategies have been favored over longitudinal strategies. This can be attributed to two things. First, some of the desired linkages are cross-sectional in nature, and cross-sectional tests have a direct impact on the quality of these matches. Second, it is difficult to define longitudinal tests that identify cases with high probabilities of error. The kinds of longitudinal conditions that suggest errors in SSNs involve breaks in continuity -- for example, changes in the SSN of a spouse or in some aspect of filing behavior. While incorrect SSNs will produce such breaks, most of the occurrences are attributable to genuine change.

### **Validating SSNs Against IRS/SSA Records**

In editing the SOCA and Family Panel files, SOI staff used an IRS validation file that contained fields obtained, ultimately, from SSA. These fields were the SSN, up to 10 name controls, and the date of birth. Identifying variables that were present on the panel records included:

- SSNs (primary, secondary, and dependent);
- Return name control (derived from surname of first-listed filer);
- City and state;
- Full name line -- starting in 1988; and
- Name of separately filing spouse -- starting in 1988.

That the SOI Division did not begin to obtain full names until 1988 proved to be unfortunate for both panels. Having full names for the base year would have allowed panel members to be identified by both name and SSN. Some of the problems of validation that grew out of the limited identifying information that was present for the base year returns in both panels are discussed below.

## **Use of the Return Name Control**

Until full names became available, the only identifying information about a filer was the return-level name control, which is derived from the surname of the primary filer, which may differ from that of the secondary filer and one or more dependents. Testing for exact agreement between the return name control and any of the name controls on the validation file for the primary SSN, the secondary SSN, and any dependent SSNs could be automated easily and reliably. Exact agreement was interpreted as validating the SSN. For primary SSNs, the application of this test dispensed with well over 99 percent of the sample cases. In a clerical review of cases failing this test in the base year of the SOCA Panel, more than half were judged to be true matches. The test failures occurred in these cases because of the misspelling of a name control on either file or because the order of the SSNs on the return did not correspond to the order of the names. That is, a couple may have filed as John Smith and Mary Wesson but listed Mary's SSN in the primary position. In this case the return name control of SMIT would not have matched the name control, WESS, associated with the primary SSN in the validation file. For secondary SSNs, the application of the return-level name control test dispensed with over 90 percent of the sample cases in the base year of the SOCA Panel. Still, the remainder were too many to review. Moreover, clerical review of the cases with name control mismatches could not be expected to resolve all of these cases. A secondary filer with a different surname than the primary filer would fail the test. Without a full name line, it was not possible to establish the secondary filer's surname or even that it differed from the primary filer's surname.

## **Use of Full Name Lines**

Full name lines were not available to validate base year SSNs for either panel. From the standpoint of correctly establishing base year names, the one year lag for the Family Panel was not as bad as the three year lag for the SOCA Panel. Still, given that many erroneous SSNs are incorrect for only one year, the problem presented by changes in SSNs for secondary filers is a significant one.

The single most useful piece of information that a full name line provides is a surname for the secondary filer, from which a name control can be constructed. Basing validation tests for secondary SSNs on a secondary name control will yield substantially fewer false failures than tests that use the return level name control. With this improved targeting, clerical review of all violations becomes not only feasible but desirable.

Because the format of the name line is not exactly standard, there will be errors in constructing name controls for the secondary filer. Many of these errors, however, may occur in situations where the secondary filer has the same surname as the primary filer. For example, John and Mary Smith might list their names as John Smith and Mary. While an overly simple algorithm might yield MARY as the secondary name control, which would be incorrect and would produce a test failure, this need not undermine the validation procedures. Any strategy for using secondary name controls generated in this manner should include testing the secondary SSN against both the return name control and the secondary name control. In this example, the incorrect secondary name control would be irrelevant, as Mary Smith's SSN would be validated successfully against the return name control.

## **Strategies When Name Lines Were Not Available**

For the SOCA Panel, name lines did not become available until year four. Birth dates provided important alternative information with which to evaluate the secondary SSNs. The birth date of the primary filer implies a probability distribution of secondary filer birth years. An improbable birth year for the secondary SSN may be grounds for determining that the SSN is incorrect when it also fails a name control test based on the return name control. Birth dates proved to be particularly helpful in choosing between two

alternative secondary SSNs when the reviewer had reason to believe that they referred to the same individual.

Name lines for later years may be valid substitutes for name lines in the base year when the SSNs in question do not change. But what if the secondary SSN does change? In particular, what if the base year secondary SSN failed a validation test based on the return name control and then changed the next year? Was this a true change in spouse or was it simply the correction of an SSN? Unless the two SSNs were so similar as to leave no doubt that one of the two SSNs was in error, the editors had to consider whether the change in SSN coincided with any pronounced change in circumstances, as reflected in the data reported on the two tax returns. Did the couple move, or did the earnings change markedly? These cases reduced to judgment calls on the part of the editors. In the SOCA Panel editing, such calls appear to have favored the determination that the filer changed, not just the SSN.

### Multiple Occurrences within Filing Year

Incorrect panel SSNs may belong to other filers. If a panel member continues to use an incorrect SSN after the base year, and this SSN belongs to another filer, multiple occurrences of the SSN in question may be observed within a filing period. Such occurrences provide unambiguous evidence of the need for a correction. If the panel member does not continue to use the SSN, however, the false matches of out-year returns back to the incorrectly reported base year SSN become less easy to detect.

### Findings

Table 1 summarizes our findings with respect to the frequency of erroneous SSNs in the population of tax returns filed for 1985, based on the editing of the base year data for the SOCA Panel. Of the SSNs that were determined to be incorrect, 42 percent belonged to other persons who filed during the next six years. Thus, 58 percent of the incorrect SSNs had to be identified without the compelling evidence provided by other filers using those SSNs correctly.

Type of SSN	Percent incorrect
Primary SSN	0.57%
Secondary SSN	1.97

*Source: SOI Division SOCA Panel.*

Table 2 summarizes the findings for the 1987 filing year, based on the first year of the 1987 Family Panel. These findings include dependent SSNs, which taxpayers were required to report for the first time in that year. It is striking, first of all, how closely the estimated error rates for primary and secondary SSNs match those of the much smaller SOCA Panel. Second, the error rate for all dependent SSNs is just over twice the error rate for secondary SSNs. This is lower than pessimistic predictions would have suggested, but it could also be an understatement of the true error rate. Most dependents do not file tax returns, and so the evidence on which to base the error determinations may not be as solid as the evidence for primary and secondary filers. The other surprising feature is how the error rate for dependent SSNs takes off after the fourth listed dependent, rising to 24 percent for dependents listed in the 7th through 10th positions. It remains to be determined whether this high error rate is a phenomenon of higher order dependents or, more broadly, of all dependents on returns that report seven or more dependents. The number of sample cases involving more than five dependents is quite small, however, so the precision of these estimates for higher

order dependents is relatively low.

Type of SSN	Percent Incorrect
Primary SSN	0.49%
Secondary SSN	1.65
All dependent SSNs	3.39
1st dependent SSN	3.36
2nd dependent SSN	3.04
3rd dependent SSN	3.63
4th dependent SSN	3.56
5th dependent SSN	7.78
6th dependent SSN	13.59
7th-10th dependent SSNs	24.31

*Source: SOI Division Family Panel*

## Conclusions and Recommendations

The quality of SSNs reported on IRS records in 1985 and 1987 appears to be quite good. For primary SSNs the error rate is exceedingly low, which can be attributed in large part to the quality checks that primary SSNs must pass before the IRS will “post” their returns to its master file. Secondary SSNs have more than three times the error rate of primary SSNs, but the error rate is still low. Moreover, the IRS has increased its validation efforts with respect to secondary SSNs, so their quality should improve over time. Dependent SSNs had twice the error rate of secondary SSNs in 1987, but 1987 was the first year that dependent SSNs were required to be reported. These error rates are likely to decline as taxpayers become accustomed to the new requirements and as the cumulative effect of IRS validation efforts grows. In offering a preliminary assessment of the impact of SSN errors on data quality, I would say that, as of now, there is no evidence from the SOCA Panel that matches lost or incorrectly made due to bad SSNs will seriously compromise analytical uses of the data.

With respect to SOI editing procedures, I would make the following broad recommendations. First, the SOI Division needs to increase the amount of automation in the validation procedures and reduce the amount of unproductive clerical review time. Much of the clerical review time, currently, is spent on cases that are judged, ultimately, to be correct. The strategy that I discuss below for constructing and using secondary name controls will directly address this recommendation. In addition, the application of record linkage technology to the name control validation tests could significantly reduce the potential clerical review by allowing SSNs to pass validation when a name control contains a simple error. What I have in mind is modifying the tests so that they can take account of partial matches. Second, validation and editing must be carried out in a more timely manner. Data capture relies on an exact match to a list of panel SSNs. Unless corrected SSNs are added to the list as soon as possible, returns that could otherwise be captured will be lost.

Finally, I want to encourage the SOI Division to develop secondary name controls from the name lines

that became available in 1988 and use these name lines to edit the secondary SSNs in the Family Panel. Secondary name controls derived by even a simple algorithm from the full name line could substantially reduce the subset of cases that are flagged as possibly containing incorrect secondary SSNs. Reviewing all of the secondary SSNs that fail name control tests based on both the return name control and the secondary name control should then be feasible. Doing so will very likely prove to be an efficient way to identify virtually all cases with erroneous secondary SSNs.

## Acknowledgments

I would like to thank the SOI Division for its support of this work. I would particularly like to acknowledge Michael Weber for his efforts in designing and overseeing the editing of both panel files, and Peter Sailer for encouraging attention to data quality. Finally, I would like to thank my colleague Larry Radbill for building the data files and generating the output on which the findings presented here are based.