Volume Title: Annals of Economic and Social Measurement, Volume 1, number 3

Volume Author/Editor: NBER

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm72-3

Publication Date: July 1972

Chapter Title: Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File

Chapter Author: Benjamin Okner

Chapter URL: http://www.nber.org/chapters/c9435

Chapter pages in book: (p. 325 - 362)

Annals of Economic and Social Measurement, 1/3, 1972

CONSTRUCTING A NEW DATA BASE FROM EXISTING MICRODATA SETS: THE 1966 MERGE FILE

BY BENJAMIN A. OKNER*

The 1966 MERGE File is a new microdata source which contains information from the 1967 Survey of Economic Opportunity (SEO) and the 1966 Tax File. For most units, the file includes SEO demographic asset/liability information plus detailed income data from the tax return(s) filed by the family in 1966. This article contains a detailed explanation of the procedures used to construct this research tool.

Economists' data needs have changed dramatically during the last twenty to thirty years. In the past, published tables and summary statistics were generally sufficient to meet most researchers' requirements. But as a result of the widespread availability of electronic computers and an increased interest in social problems at the microeconomic level, there is now an effective demand for large amounts of disaggregated economic and demographic information. Unfortunately, the supply of usable microdata is still far short of both the quantity and the quality demanded.

I. THE NEED FOR A NEW DATA FILE

Despite the ease with which one can obtain a current estimate of total U.S. personal income, there are no official statistics on the size distribution of such income or any cross-classifications of personal income by typical demographic characteristics of the population.[1]

Annual information from the U.S. Internal Revenue Service exists on *income subject to tax* from *individual tax returns*, but the omission of data for people not required to file distorts the distribution for those at the low end of the income scale. The Census Bureau also collects income information in its Current Population Survey each year from a sample of about 30,000 *households*. But, in addition to using a different analysis unit, Census employs a *total money income* concept (which includes nontaxable transfer payments but excludes taxable realized capital gains). If used carefully, together these two data sets may be helpful to the researcher investigating questions regarding the distribution of total money income. However, neither the Internal Revenue nor the Census data contain any information on the distribution of nonmoney income and therefore cannot be linked with the personal income or other aggregative statistics.

The lack of a consistent and comprehensive set of household income data prompted the construction of the new microanalytic data base discussed in this

* The author is a member of the Economic Studies Staff of the Brookings Institution. The views expressed are his own and do not purport to represent the views of the other staff members, officers, or trustees of the Brookings Institution. The study was financed under a research grant to the Brookings Institution from the U.S. Office of Economic Opportunity, and was presented at the NBER Workshop on the Use of Microdata in Economic Analysis, October 22, 1970.

[1] Estimates of the size distribution of personal family income were prepared by the Office of Business Economics (OBE), U.S. Department of Commerce, between 1944 and 1963 (when they were discontinued). Work needed to resurrect the series is currently underway at OBE, but it is likely to be some time before such data are again regularly published.

paper. This new file is an indispensable part of an empirical study of the distribution of federal, state, and local taxes among U.S. families now underway at Brookings. When we began the study, two possible sources for 1966 microlevel income information existed: (1) the Internal Revenue Service Tax File containing information from federal individual income tax returns filed for 1966; and (2) the 1967 Survey of Economic Opportunity (SEO) data file based on field interviews for a sample of the total population in early 1967.[2] Although neither of these data files, by itself, was adequate for estimating the size distribution of income needed for the tax burden study, they each contained important information which, if combined, would provide a suitable basis for estimating the required distribution.

For several reasons, the SEO population was chosen as the basis of the new data file. It contained a stratified representation of the total U.S. population collected on a family basis[3] which seemed to be the most useful unit for our analysis. The income information collected in the SEO includes receipts from nontaxable as well as taxable sources and is therefore much more comprehensive than a concept which includes only income subject to tax. In addition, the demographic data available for each family are much richer than can be obtained from tax returns. However, there is a serious disadvantage to using the SEO because the income data are known to be understated (especially among higher-income families) since capital gains were not included in the survey income concept and also because of the well-known phenomenon of income underreporting in sample surveys.

In creating the MERGE data file, we selected and combined the best information available from both the 1966 Tax File and the 1967 SEO File. The SEO family record was used as our base and selected information was imputed to each family record on a systematic basis from the 1966 Tax File. Thus, the newly created MERGE File contains demographic and income information for low-income SEO families who are not in the tax-filing population as well as the more complete—and we believe, more accurate—income tax information for higher income individuals.

In the remainder of this paper, the detailed steps involved in actually constructing this new and unique set of household microdata are described.

II. Creating the Merge Data File

The 1967 SEO File contains data from a stratified sample of all U.S. families and individuals. The 1966 Tax File population consists of a subset of this same population, viz., those individuals who filed income tax returns. On the basis of

[2] The 1966 Tax File contains a stratified sample of data from close to 87,000 individual tax returns. For a detailed description of the file, see "The Brookings 1966 Federal Individual Income Tax File," Brookings Computer Center Memorandum No. 42, June 30, 1968 (mimeo), which is available on request.

The 1967 Survey of Economic Opportunity was conducted by the U.S. Bureau of the Census for the U.S. Office of Economic Opportunity. The SEO File data are derived from interviews with about 30,000 households and include income information for 1966 and supplemental financial and demographic data as of the date of interview. A full description of the survey may be obtained from the Office of Planning, Research, and Evaluation, U.S. Office of Economic Opportunity.

[3] Even though this differs from Census Bureau practice, throughout this paper the term "family" is used to include both families of two or more persons and unrelated individuals (families of size one).

income and other information reported, we were able to estimate which families in the SEO population would not have been expected to file for 1966. And after the nonfilers were excluded, the remaining SEO units represent families who were in the population from which the 1966 Tax File sample of returns was drawn. For these families, it was possible to estimate the kind of tax return(s) filed using reported SEO information.

Once the SEO tax-filing group was determined, it would have been ideal to obtain the corresponding tax information directly from the Internal Revenue Service. However, this was precluded because both Census and Internal Revenue have very stringent policies with respect to maintaining the confidentiality of their data. In place of an exact one-to-one matching by name for each family, a less satisfactory—but feasible—means of stochastically simulating this matching procedure was developed. This process involved two major parts: (1) on the basis of information available in the SEO file, we estimated whether any members of each SEO family would be expected to file in 1966 and if so, the number and kinds of such returns; and (2) for filers, "SEO tax units" were created and actual returns from the Tax File "similar" to the SEO tax returns were randomly selected. The actual tax return data were then merged with the existing information in each SEO family record. Since there were close to 30,000 matches to be made, the selection and linking of returns was all done by computer.[4]

For most families, the final MERGE data file contains all the demographic information and data concerning receipts of nontaxable income from the SEO File plus figures from a tax return assigned to it from the 1966 Tax File. For SEO families who were deemed to be nonfilers, the MERGE File includes no tax return information. And for a small number of very high-income units, there exists no SEO demographic data.

Creating SEO Tax Units

Income Allocation

For each interview unit, certain income information—wages, nonfarm business income, and farm income—was collected in the SEO for each person 14 years and over. However, for the other income components only the total amount received by all members of the interview unit was available from the SEO. Of these other items, the following are taxable under the federal individual income tax: rent, interest, dividends, government pensions, private pensions, and other regular income.

Because the tax-filing status of an individual is largely dependent on the total amount of taxable income received, the first step in forming SEO tax units was to allocate the taxable income components among eligible members in each family. This was done for the various items using the following procedures.

Rent, interest, dividends, and other regular income. These items were allocated equally between the head and wife of the interview unit, or to the head of the interview unit if no spouse existed.

[4] All programming and computer operations for the project were performed at the Brookings Computer Center. Jon K. Peck was primarily responsible for devising as well as programming the many intricate operations required for the study. The project would have been impossible without his ingenious and dedicated efforts and grateful acknowledgement is given for his help.

Government and private pensions. If the interview unit contained any persons aged 45 or over who worked less than 50–52 weeks in 1966 because they were retired, pension income was allocated equally among such retirees. In order to avoid allocating an unreasonably small amount of income to any individual, a minimum amount criterion ($1,800 for government pensions and $1,000 for private pensions) was imposed if the number of eligible recipients was greater than one.[5] If the test was not met or if the interview unit contained no eligible recipients, the total amount of government and private pensions was allocated to the interview unit head and wife. In estimating a person's taxable pension income, the total was reduced by 10 percent to adjust for some nontaxable return of invested capital.

Tax-filing Criteria

After taxable income was allocated among individuals, SEO tax units were created using the filing requirements of the Internal Revenue Code applicable for 1966. Individuals who had earnings subject to wage withholding were assumed to file returns in order to obtain tax refunds. In order to link the persons in each interview unit with the proper tax units, it was also necessary to determine whether they would have qualified as a dependent in some tax unit. As described below, the income and demographic information for each person in the SEO interview unit were used to determine his tax-filing status.

Income Criteria

A person was assumed to file a tax return if any of the following conditions were met:

(a) He had total taxable income of at least $600 ($1,200 if age 65 or over);
(b) He had absolute farm income plus absolute nonfarm business income of at least $400;
(c) He had wage income between $200 and $600 ($1,200 if aged) and his occupation was *not* newsboy, baby sitter, private household keeper or laundress, bootblack, charwoman and cleaner, farm worker, or gardener and grounds keeper.[6]

Using these income criteria, it was first determined if any family member other than the head and wife would file a return and if so, tax units were created for them. It was then determined whether the family head and spouse (if she existed) would file a tax return. If so, a tax unit was created for them and a record was made of any dependents to be associated with their return. Thus, it was possible for a person in an SEO family to file his own return and also to be claimed as a dependent on another tax return filed in the family—as is often the case in the real world.

[5] Thus, if the total amount of pension income was "T" and there were "n" potential recipients, the income was distributed equally among the "n" eligible persons only if T/n was equal to $1,800 for government pensions. If T/n was less than the $1,800 minimum, $T/(n-1)$ was computed and if the results passed the $1,800 minimum test, that amount of government pensions was distributed equally among $(n-1)$ of the eligible recipients chosen randomly from among the total.

[6] Individuals with low wages are not required to file returns but it was assumed that they would do so in order to obtain refunds of income tax withheld by employers. Exceptions were made for those employed in the occupations listed since wage withholding is not typical for such jobs.

Dependency Test

A person qualified as a dependent of the family head if he was in the same interview unit and met any of the following conditions:

(a) He was under 14 years old;
(b) He was under 19 years old, a child of the head and passed the support test;
(c) He was a child of the head, a student, and passed the support test;
(d) His total taxable income was less than $600 and he passed the support test;
(e) He was a relative of the head residing in the household and passed the support test.

A person passed the support test if either of the following two conditions were met:

(a) His total income was less than $250; or
(b) His total income was less than half the total income of the family head and wife and was less than $3,000.

Type of Return

For each SEO tax unit, the type of return filed was determined on the basis of the following criteria.

Single individual return. All (primary and secondary) individuals were assumed to file single individual returns. In addition, all persons who passed the filer test and had no dependents were assumed to file single individual returns.

Joint return. All families with both spouses present were assumed to file joint returns. Such returns were also created for families with only one spouse present and whose reported marital status was "married, spouse absent." There were no returns created for married persons filing separately.

Surviving spouse. If the head of the family did not have a spouse, it was determined if he or she was a widower whose spouse had died within the two preceding years and if there were any children in the family who qualified as dependents. If so, a surviving spouse tax return was created.

Head-of-household. If a widowed head of a family had no children but there were other persons in the family who qualified as dependents, then he or she was assumed to have filed a head-of-household return. Also there was created a head-of-household return for the head of a family whose marital status was divorced, separated, "never married, but has child," or "never married, other" if there were other family members present who qualified as dependents.

Tax Unit Creation Results

As shown in Table 1, the total number of SEO tax units created compares quite closely with the actual number of tax returns filed for 1966. The total of 67.3 million created units is just 1.4 million, or 2 percent, less than the 68.7 million returns actually filed. The accuracy of the results varies by marital status, but among joint returns, which comprise 60 percent of the total number filed, the estimated number of SEO tax units differs from the actual number by only 138,000. The

TABLE I

CREATED SEO TAX UNITS AND ACTUAL TAX RETURNS FILED, 1966, BY MARITAL STATUS
(In thousands)

Marital status	SEO tax units	Actual tax returns filed
Single individual	22,322	25,182
Married filing joint return	41,511	41,373[a]
Unmarried head-of-household and surviving spouse	3,469	2,164
Total	67,302	68,719[b]

[a] Since there were no returns created for married persons filing separately, this is the sum of returns for married couples filing jointly and one-half the number of returns of married couples filing separate returns. This adjustment is needed to make the figures for created and actual returns comparable.

[b] The total shown is less than the published figure for the number of 1966 returns filed because one-half the numbers of married couples filing separate returns is deducted. See note (a).

largest discrepancy is among single individual returns where the procedure resulted in an underestimate of the total by 2.9 million. But this is somewhat offset by the 1.3 million over-estimate of the number of surviving spouse and head-of-household returns. Since our filing criteria follow the statutory requirements very strictly, it is possible that part of the discrepancy is derived from individuals who filed single individual returns even though qualified to file as surviving spouse or head-of-household. In addition, we may have been too stringent in setting the income filing requirement at $200 if there are a substantial number of children with earnings below that level who filed for a refund of withheld taxes. Since all such returns would have been nontaxable, the omission will not have any impact on the final tax burden figures and the discrepancy in the number of single individual returns is not considered serious.

Selecting the Match Return

Once the SEO tax units were created, the next part of the MERGE File creation involved selecting actual returns from the Tax File to be linked with each SEO family unit.[7] In effect, for this process, each created SEO tax unit was used primarily as a vehicle for deriving the information now needed to select an actual return from the Tax File to be linked with the corresponding individual(s) in the SEO family.

Computer Matching

The initial step in the linking process was to group the tax units in each file into "equivalence classes" defined by comparable characteristics available in both the SEO File and Tax File. The characteristics used were (1) marital status under which the return was filed; (2) whether the head (or spouse) of the tax unit

[7] Of course, the procedure outlined did not apply to those SEO families in which we did not expect any member to have filed a tax return for 1966. There were about 7 million SEO families, of the 62 million, in which the combination of reported income, family, size, and other characteristics were such that the family was excluded from the tax-filing population.

was age 65 or over; (3) the number of dependent's exemptions in the unit; and (4) the reported pattern of income.

The income pattern variable was specially constructed for the linking process. First, each tax unit was classified into one of the four possible major income source categories—wages, business, farm, or property income. The categories were defined to be exhaustive and the largest income source value(s) (in absolute amount) was used to determine the major source group. Thus, if a tax unit reported a business loss of $10,000 and dividend income of $500, it would have been classified into the business major source category even though the property income (dividends) of $500 is algebraically greater.

Once the major source of income was determined, each return was further classified into a minor income source category (within each major source group). Tax units which did not meet any minor source criteria were put into either a (1) "major source is sole income source" category; or (2) "minor income exists, but is negligible" category (i.e., is not sufficiently large to meet the minor income source criteria). Since the significance of any given dollar amount of income differs depending upon the source from which derived, the criteria for the existence of a qualified minor source varied by income type.[8] Altogether, there were thirty-five possible income pattern categories (thirty-two plus three special categories for tax units in which the major income source was negative).

The initial class definitions would have resulted in more than 1,000 different equivalence classes,[9] many of which would have been empty or would have contained very few units. The number of equivalence classes was reduced by eliminating and/or combining a large number of previously defined categories and creating a single marital status and age variable. As finally used, the actual number of equivalence classes was equal to 74, used to effect a total of 28,643 tax unit matches.[10]

In almost all cases, the actual selection of a Tax File return was done by computer using tightly prescribed rules for defining an acceptable match. The general procedure was to consider all Tax File returns within the equivalence class of the SEO tax unit as comprising the population of potentially acceptable matches.

Since we did not expect to find an actual Tax File return with exactly the same amount and pattern of income as reported in the SEO tax unit, the first matching rule established an acceptable range of major source income from which a Tax File return could be selected. This was initially set equal to the major source income of the SEO tax unit plus or minus 2 percent of that amount. In addition, to insure that we were not overly restrictive for low-income units and not overly generous for high-income ones, the band of acceptable major source amounts

[8] See the Appendix for detailed definitions of the major and minor income source categories.

[9] This is the product of two age groups (under and over age 65); three marital status categories; five classes for number of dependent's exemptions (one through four plus five or more); and the 35 income pattern categories.

[10] The reader should not infer that there was an average of 387 matches per class, as would be derived simply by dividing the total number of matches by the total number of equivalence classes. There was a very wide variation in the cell counts in each class which, of course, reflects the prevalence of different income patterns among different kinds and sizes of families. Selected summary statistics on the matches by equivalence class are given in the Appendix.

had to equal at least the major source income amount plus or minus $50 and could not exceed the major source income amount plus or minus $500 (resulting in minimum and maximum bands of $100 and $1,000 respectively). Thus, if reported major source income was $6,000, the 2 percent criterion would establish an acceptable income band ranging between $5,880 and $6,120. Since the band is equal to $240 ($6,120 less $5,880), neither the $100 minimum nor $1,000 maximum band size criteria would be applicable (in fact, these restrictions were operative only for units with income below $2,500 or above $25,000, respectively).

For all returns in the acceptable income range, within each equivalence class, a "consistency score" was then defined and computed to take account of hitherto unused information for effecting a suitable tax return match. For each of the factors entering the consistency score, tax return data from each potential match were compared with information in the SEO family record and if the items were "consistent" (in terms of joint presence or absence of items), the return was given consistency score points. The six factors used for consistency scoring purposes were:

(a) Home mortgage interest deduction or property tax deduction on tax return vs. home ownership or debt (or house value included in farm value) in SEO—*12 points*;

(b) Interest or dividend income on tax return vs. interest or dividend income or ownership of stocks, bonds, or other interest-bearing assets in SEO—*8 points*;

(c) Farm income on tax return vs. farm income or farm assets or debt in SEO—*10 points*;

(d) Business income on tax return vs. business income or business assets or debt in SEO—*10 points*;

(e) Rental income or real estate property tax deduction on tax return vs. rental income or real estate assets or debt in SEO—*9 points*;

(f) Nonzero capital gains income on tax return vs. dividends or interest on stocks, bonds, etc. in SEO. Also, capital gains equal to zero on tax return vs. earnings from property in SEO is consistent—*8 points*.

As can be noted from the listing, the maximum possible consistency score was 57. However, only Tax File returns in the top 25 percent of the initial group when ranked by consistency score were eligible for matching with an SEO tax unit. An additional constraint was imposed: the minimum consistency score in the top quartile had to equal at least 25 points out of the possible 57. Because of the way the points were awarded, this meant that a tax return had to meet at least half of the consistency tests in order to be assigned to an SEO tax unit.

All tax returns which were within the acceptable income range for the SEO tax unit we were attempting to match and which also passed the consistency score test were eligible for selection and linking. From the eligible returns within the group, the return assigned was randomly selected with a probability of being chosen proportional to the weight of the return in the Tax File.[11]

Almost all the matches were made using the procedure just described. However, there were instances in which the initially defined income band contained

[11] This procedure guarantees random selection since the Tax File weights are equal to the inverse of the probability of selection from the total universe of tax returns filed during 1966.

no tax returns or the consistency scores of returns in the initial income range did not meet the top quartile minimum of 25. When this occurred, the initial income range was increased by an additional 1 percent, plus and minus, and the minimum and maximum dollar amount constraints on the size of the income range were increased somewhat.[12] Consistency scores were computed for all new returns in the wider income range and if a suitable match was found (still using the same criteria as above), the computer program assigned the selected return to the proper SEO tax unit and proceeded to the next unit.

If no suitable match was found on the second try, the class limits were again expanded (by plus and minus one percent each time) and consistency scores were computed for the new tax returns included in the enlarged set of eligible returns. The computer program terminated after seven unsuccessful automatic assignment attempts and then executed a hand-matching procedure. As shown in Table 2 statistics, the class expansions and hand-matching were rarely needed. Of the 28,643 tax unit matches made, 27,912, or 97 percent were accomplished using the initial criteria.[13]

TABLE 2

NUMBER OF COMPUTER TAX UNIT MATCHES BY NUMBER OF BAND EXPANSIONS REQUIRED FOR MATCH

Band expansions	Number of matches
Match found on initial attempt	27,912
Match found after 1 expansion	271
Match found after 2 expansions	129
Match found after 3 expansions	81
Match found after 4 expansions	50
Match found after 5–7 expansions	49
Computer match impossible after 7 attempts, returns hand-matched	151
Total	28,643

[12] Each time the percentage range size was increased, the minimum band size was increased by plus and minus $10 and the maximum was increased by plus and minus $125. In effect, these merely compensated for the percentage changes in the income range and continued to be relevant only for returns with very low or very high incomes.

[13] The term "hand matching" which describes the assignment procedure used for 151 of the SEO tax units is misleading since the process was highly computer dependent. In fact, the procedure is probably unique since a batch-environment computer was used interactively to select returns to be matched from the computer console. After seven expansions of the acceptable income range, a list of all possible tax returns which might be selected was written on the computer printer. The list included pertinent information about the SEO tax unit as well as the income and consistency score data for all Tax File returns that had been located for possible matching with the SEO tax unit. By means of the computer sense switches, it was possible for the analyst to continue expanding the income range for eligible returns for as long as he desired in an attempt to find an acceptable Tax File match. After each successive hand-determined expansion, there would be listed on the line printer all the information concerning new tax returns which became eligible for matching as a result of expanding the income class boundaries. Of course, as in the case of the computer-selected matches, all potentially eligible returns had to be selected from the same equivalence class as the SEO tax unit.

As would be expected, the increased income range expansions made eligible for assignment tax returns with incomes increasingly divergent from those of the SEO tax unit. In some instances, the income divergence was accompanied by substantial increases in the consistency scores of the returns

Weight Adjustment for High-income Units

After completing the match-merge process, we discovered substantial differences in the derived amount of income reported by high-income SEO families and the total published by the Internal Revenue Service. This discrepancy was primarily the result of the different methods used for drawing the Tax File sample and the sample of SEO households which produced a large disparity between the actual number of cases at the high end of the income scale in each file.

The differential sampling, in fact, nullified the original match-merge procedure for high-income families. For example, in the SEO sample, high-income families were generally chosen at a sampling rate of 1/3,000 and the data for each family are therefore multiplied by a weight of about 3,000 to obtain population estimates. In the Tax File sample, the returns are grouped into 13 strata, depending primarily on income level. The sampling rates for the strata differ and range between 1/4,000 for low-income returns to 1/1 for those with very high incomes. As a result of the different sampling schemes, the SEO File contains very few cases of high-income families (each of which has a weight of about 3,000), while the Tax File contains thousands of high-income returns (each of which has a very low population weight). Thus, when a tax return was matched with one of the high-income SEO families, all the tax data originally associated with a very low Tax File weight were multiplied by the much larger SEO family weight. And, as might be expected, the estimated aggregate amount of income on such returns became vastly overstated.

While this problem could have been solved either by aggregation of Tax File returns or duplication of SEO family units, we were unable to devise a satisfactory, practical procedure for doing this. Aggregation of high-income tax returns into a smaller number of "representative units" was rejected because we were unwilling to accept the concept of the "typical millionaire." On the other hand, duplicating even a small proportion of the SEO records 3,000 times was impractical in terms of the computer capability.

The problem was finally resolved by splitting the MERGE File into two parts. For all families with positive income of under $30,000, the SEO and tax return data as derived from the match-merge process as described were included. For all SEO families with $30,000 or more income or with negative income, the

---

made available. But this was not necessarily the case. And in many instances, increased class expansions did not increase the number of available returns.

Since it would have been exceedingly difficult to formulate and program acceptance rules for the tax units that fell into the hand-matching group, the final decision in selecting a match was left to the analyst. We made each decision after subjectively weighing the evidence (and considering the trade-offs) concerning the alternative returns made available in the process. The actual match was then effected by entering into the computer console switches the number of the Tax File return selected as the best possible match. The computer program used this information and then treated the selected return in the same manner as one that had been chosen automatically with the programmed matching algorithm.

Whenever a satisfactory assignment was made, data from the selected tax return were incorporated into the SEO family record and income information from the return (or multiple returns) was transferred into (and replaced the initial information in) the SEO family record. At this stage, adjustments were made to the Tax File data to correct for known underreporting of certain income components. These were based on unpublished information from the Office of Tax Analysis, U.S. Treasury Department. Since 97 percent of all matches were made on the initial attempt, in most instances the differences between the two sets of data were quite small.

original SEO records were deleted and replaced with Tax File returns. (The $30,000 income level was chosen for splitting the file since it is at that point that the SEO and Tax File weights start to diverge widely.)

Thus, there are no SEO data for high-income tax returns or returns on which reported income was negative.[14] Such returns were merely appended to the MERGE File without any of the SEO demographic data. These two special groups of tax returns accounted for less than 2 percent of the 70.6 million returns filed in 1966.

Income Adjustments

Since the total income recorded in the MERGE File was less than the aggregate "adjusted family income" (AFI) computed for 1966,[15] the final step in creating the MERGE File involved adjusting the SEO and Tax File income components to correspond with national aggregate figures.

In Table 4, we present a comparison of the MERGE File income data and the AFI control totals. As can be seen, the total income accounted for in the

TABLE 4

COMPARISON OF ADJUSTED FAMILY INCOME AND MERGE FILE INCOME BEFORE ADJUSTMENT, BY SOURCE OF INCOME, 1966
(In billions)

Source of income	Adjusted family income	MERGE File income[a]	Difference	MERGE File income as percent of adjusted family income
Wages, salaries, and other labor	$423	$415	$ 8	98%
Nonfarm proprietors'	45	46	−1	100
Farm proprietors'	14	6	8	43
Rents	19	18	1	95
Royalties	1	1	—	100
Dividends	15	12	3	80
Personal interest	24	21	3	88
Transfer payments	34	25	9	74
Other accrued gains on assets	84	84	—	100
Total	$659	$628	$31	95%

[a] MERGE File income excludes adjustments for nonreporting and underreporting of income.

[14] In fact, these two groups of returns are not really different since returns with substantial negative income are usually filed by wealthy families. Hereinafter, both the high-income and negative income returns are referred to as the "high-income portion" of the MERGE File.

[15] The adjusted family income (AFI) concept was developed for, and is unique to the tax burden study. These data were derived primarily from the Office of Business Economics (OBE) personal income figures and individual income tax information from the Internal Revenue Service, both of which were adjusted to take account of differences in income concepts and the populations covered. The AFI concept is intended to correspond as closely as practicable to an economic concept of income, viz., consumption plus tax payments plus (or minus) the net increase (or decrease) in the value of assets during the year. AFI is defined to include only income which accrues *directly* to individuals and families; it does not include earned, but undistributed, income. For a detailed description of how the adjusted family income figures were derived, see Benjamin A. Okner, "Adjusted Family Income: Concept and Derivation," Brookings Technical Working Paper II, for the Distribution of Federal, State, and Local Taxes Research Program, March 1971 (rev.) (mimeo), which is available on request.

MERGE File was 95 percent of the AFI estimate for 1966. However, the degree of agreement between the two sets of figures varied considerably by income source. The data for wages were very close. But before adjustment, the reported farm proprietors' income amounted to only 43 percent of the expected AFI amount. In addition, there were less serious, but significant discrepancies between the expected and reported amounts of interest, rent and transfer payments.

The income reported in the MERGE File was less than the AFI estimate because income information was collected using the Census money income concept; and there was both nonreporting and underreporting of income by the survey respondents.

Although nonreporting and underreporting are conceptually separable, in practice it is very difficult to distinguish between these two kinds of response error. On the basis of various data from outside sources, we believe that most of the difference between the reported and AFI aggregate amounts for factor payment components resulted from underreporting.[16] And for similar reasons, we concluded that transfer payments in the MERGE File were understated primarily because of respondent nonreporting.[17]

For components where we believed the differences were due to underreporting, the MERGE File data were adjusted to the AFI aggregates under the assumption that such underreporting was not related to other characteristics of the survey unit. This was done by applying a single ratio to the reported income of all units to increase it to the aggregate adjusted family income amount. In the case of nonreporting, the adjustment procedure was more complex. In these instances, we imputed missing amounts stochastically to MERGE File units based on various other characteristics of the survey unit.

In addition to the adjustments to correct the survey data for underreporting and nonreporting, there were several imputations to add information to the MERGE File which was not available (because it was not collected) in the SEO or the Tax File. These included such items as net imputed rent on owner-occupied homes, tax exempt interest on state and local bonds, and employer supplements to wage and salary income.[18]

III. Concluding Remarks

Creating the MERGE data file was a costly and time-consuming operation. It took well over a year and involved several man-years of labor input and computer time. Although it involved a tremendous investment of resources, we feel that the effort was worthwhile and that the file is an extremely useful analytical tool. While it was constructed primarily for use in our tax burden study, the

[16] For example, a recent study of farm income reporting indicates that there are large "differences in concepts used in accounting for expenses" of farmers which lead to substantial underreporting of farm income on tax returns. See Edward I. Reinsel, *Farm and Off-farm Income Reported on Federal Tax Returns* (U.S. Department of Agriculture, Economic Research Service, 1968), pp. 27–33.

[17] Comparison of the MERGE File data on social security and public assistance benefits recipients with Social Security Administration program statistics indicated that there was a large amount of nonreporting for these two income components.

[18] Details concerning all imputations are reported in another paper. See Benjamin A. Okner, "The Imputation of Missing Income Information," Brookings Technical Working Paper III, for the Distribution of Federal, State, and Local Taxes Research Program, April 1971 (mimeo).

research value of the MERGE File certainly is not limited to this work. We expect to distribute copies to other researchers in the future and believe that the file will be useful in a large variety of other research projects.

Of course, procedures similar to those used to create the MERGE File could be used to construct other microdata sets. However, I would not overemphasize the desirability of building new files in this way. We had to use a less-than-optimum strategy in order to proceed with our tax study. But the only correct way to construct a merged data file is by one-to-one direct linking of information from different microdata sources.[19]

The feasibility of doing this is now quite limited. And current prospects for rapid progress in direct linking are not encouraging. Nevertheless, there is little doubt of the high expected return from a substantial investment of resources in this area. Increased efforts by researchers to obtain access to and use of various existing microdata sources are clearly needed and warranted.

The Brookings Institution

Appendix

The match-merge process required the calculation of the major and minor income source for each tax unit and the use of this information to group tax returns into equivalence classes from which the actual matches were selected. The details concerning the definitions of these concepts are given below.

Major and Minor Income Source

Each tax unit in both the Tax File and the SEO File was classified into a single major income source category on the basis of the amounts and types reported. The criteria for grouping were similar for both files, but involved slight differences because the data available were not identical in both data sets. Because of the differences that existed, it was necessary to aggregate the amounts for several income sources in order to obtain total business income in the Tax File and in both files it was necessary to sum several income components to obtain the total property income. Wage and salary income and farm income were reported individually in both the SEO and Tax Files.

Tax File. In the Tax File, capital gain income and royalty income were not used for determining the major income source since comparable information on these receipts was not available in the SEO.[20] Wage and farm income information was collected separately and, hence, involved no aggregation.

Business income was reported under several different categories: income from business or profession, partnership income, and small business corporation income. The total business income for the return was calculated as the sum of the absolute amounts of each of these components.

[19] Work of this kind is now underway in the Office of Research and Statistics, at the Social Security Administration. Their project involves direct linking of micro-unit data from a subsample of SEO units with information from the Social Security Master Earnings Records and tax returns.

[20] Capital gains are specifically excluded from the income concept used in the SEO. While royalty income should have been reported, it was included in the same category as "other regular income" on the SEO questionnaire and we believe that little, if any, of such income was actually reported.

Property income was also reported under several different categories in each tax return. The total was computed as the sum of taxable dividends and dividend exclusion; interest income from banks, savings and loan associations, and all other sources; net rental income; income from pensions and annuities; and income from estates and trusts.

SEO File. Determining the amounts of the various major types of income was much simpler in the SEO tax units since information there was collected in much less detail than was the case for tax returns. The only category which involved aggregation was property income. Total property income was computed as the sum of dividends, interest, rental income, and pensions for each tax unit.

Determining major income source. For each tax unit in both files, the largest income source (in absolute value) was deemed to be the major income source. The various possibilities were considered in the following sequence: (1) wages; (2) nonfarm business income; (3) farm income; and (4) property income. In each case, the test made was whether the absolute value of the source being considered was greater than or equal to the sum of the absolute values of the other three possibilities. Thus, in the unlikely event that the amount of wage income was exactly equal to the sum of all other income sources, the unit would be classified as having wages as the major income source.

Minor income source. In order to refine the pattern of income variable, we next defined a minor income source classification within each of the major income source categories. The individual components for each of the four minor source categories were defined as they were for the major income source calculation.

To qualify as a minor source, the amount of income in the category had to be at least $50 and had to equal at least 20 percent of the major income source amount for wages; 20 percent for nonfarm business income; 15 percent for farm income; and 2 percent of the major source amount for property income. In order to relax the stringency of the criteria for higher-income units, a dollar amount floor was also established. If the amount of income in a category was equal to or greater than the following minima, it was automatically accepted as a minor source: for wages, $3,000; for nonfarm business income, $3,000; for farm income, $2,000; and for property income, $400.

Of course, whenever a unit received income from only a single source, none of the minor source requirements were met and in the equivalence class definition table below, these units are designated by only one letter indicating the major income source (e.g., "*W*" indicates that wages were the major and sole income source). Another possibility was that the unit reported income from some source other than the major one, but that the secondary income source did not meet any of the minor income source criteria. In this case, "e" is used to indicate the presence of a small amount of income other than the major source. In addition, since we did not allow for other than a major and minor income source, if these two categories did not exhaust the total income reported, we indicate the presence of additional income by "e". Thus, "*W* + *P* + e" indicates that the major income source was wages, the minor income source was property, and that there was some additional income reported in the tax unit. For almost all units, the pattern described by the major and minor income source categories was found to be more than adequate.

TABLE A-1

EQUIVALENCE CLASS DEFINITIONS FOR MATCH-MERGE PROCESS

Class Number	Income Pattern[a]	Marital Status and Age[b]	Number of Exemptions
1	W	Single Under 65	1
2	W	Married Under 65	1, 2
3	W	Married Under 65	3
4	W	Married Under 65	4
5	W	Married Under 65	5+
6	W	Single Under 65	2–5+
7	W	All marital status 65 +	1–5+
8	W + e	Single Under 65	1
9	W + e	Single Under 65	2–5+
10	W + e	Single 65+	1, 2
11	W + e	Single 65+	3
12	W + e	Single 65+	4
13	W + e	Single 65+	5+
14	W + e	Married All ages	1–5+
15	W + B or W + B + e	Single All ages	1–5+
16	W + B or W + B + e	Married All ages	1, 2
17	W + B or W + B + e	Married All ages	3
18	W + B or W + B + e	Married All ages	4, 5+
19	W + F or W + F + e	All marital status All ages	1–5+
20	W + P	Single Under 65	1–5+
21	W + P	Married Under 65	1, 2
22	W + P	Married Under 65	3
23	W + P	Married Under 65	4, 5+
24	W + P or W + P + e	All marital status, 65+ Single 65+	1–5+ 1–5+
25	W + P + e	Single Under 65	1–5+
26	W + P + e	Married Under 65	1, 2
27	W + P + e	Married Under 65	3
28	W + P + e	Married Under 65	4, 5+

TABLE A-1 (continued)

Class Number	Income Pattern[a]	Marital Status and Age[b]	Number of Exemptions
29	W + P + e	Married 65+	1–5+
30	B	Married All ages	1–3
31	B	Married All ages	4, 5+
32	B	Single All ages	1–5+
33	B + e	Single All ages	1–5+
34	B + e	Married Under 65	1, 2
35	B + e	Married Under 65	3
36	B + e	Married Under 65	4
37	B + e	Married Under 65	5+
38	B + W or B + W + e	Single All ages	1–5+
39	B + W or B + W + e	Married Under 65	1, 2
40	B + W or B + W + e	Married Under 65	3
41	B + W or B + W + e	Married Under 65	4, 5+
42	B + W or B + W + e	Married 65+	1–5+
43	B + F or B + F + e	All marital status All ages	1–5+
44	B + P	Single All ages	1–5+
45	B + P	Married Under 65	1, 2
46	B + P	Married Under 65	3
47	B + P	Married Under 65	4, 5+
48	B + P	Married 65+	1–5+
49	B + P + e	Single All ages	1–5+
50	B + P + e	Married Under 65	1, 2
51	B + P + e	Married Under 65	3
52	B + P + e	Married Under 65	4, 5+
53	B + P + e	Married 65+	1–5+
54	F	All marital status All ages	1–5+
55	F + e	All marital status All ages	1–5+
56	F + W or F + W + e	All marital status All ages	1–5+

TABLE A-1 (continued)

Class Number	Income Pattern[a]	Marital Status and Age[b]	Number of Exemptions
57	F + B or F + B + e	All marital status All ages	1–5+
58	F + P or F + P + e	All marital status All ages	1–5+
59	P or P + e	Single Under 65	1–5+
60	P or P + e	Married Under 65	1–5+
61	P or P + e	Single 65+	1–5+
62	P or P + e	Married 65+	1–5+
63	P + W or P + W + e	Single All ages	1–5+
64	P + W or P + W + e	Married Under 65	1–5+
65	P + W or P + W + e	Married 65+	1–5+
66	P + B or P + B + e	Single All ages	1–5+
67	P + B or P + B + e	Married Under 65	1, 2
68	P + B or P + B + e	Married Under 65	3–5+
69	P + F or P + F + e	All marital status All ages	1–5+
70	P + B or P + B + e	Married 65+	1–5+
71	B + e	Married 65+	1–5+
72	Negative B	All marital status All ages	1–5+
73	Negative F	All marital status All ages	1–5+
74	Negative P	All marital status All ages	1–5+

[a] In designating the source of income, the letters have the following meaning: "W," wages; "B," business; "F," farm; "P," property; and "e," epsilon (small amount of income from sources other than those designated). The first letter given always denotes the major income source and if it is not followed by other letters, it is the only income source (e.g., "W" means wages are major and only source). The second letter designates the minor income source and when two capital letters appear the sum of the major and minor income is equal to the total. If "e" appears in the income pattern list, it means that the sum of the major and minor source income was less than total income; or if no minor income source existed, the major income source was less than total income.

[b] In this table, "single" marital status includes head-of-household and surviving spouse returns as well as single individual returns. "Married" refers to joint returns filed by married couples. Age "under 65" or "65+" is determined on the basis of the age of the taxpayer or spouse. If either spouse is 65 or over, the return is in the latter category.

Equivalence Classes

After the major and minor income source pattern for each return was determined, this plus other information from each return was used to group the tax units in each file into equivalence classes for matching. Although the number of possible classes was very large, there were 74 equivalence classes actually used in the match-merge process. The definitions of characteristics used for forming the classes are presented in Table A-1.

Annals of Economic and Social Measurement, 1/3, 1972

COMMENTS

BY CHRISTOPHER A. SIMS

Okner in this paper describes an ingenious procedure without discussing its theoretical basis. A little thought on the structure of the practical problem Okner is trying to solve shows that his method produces biases which could be avoided by the use of different, but no more complicated, procedures.

Okner is confronted with two samples, one roughly three times the size of the other, with the overlap (the number of individuals appearing in both samples) clearly negligible. Certain variables, which we shall call X, appear in both samples.[1] Other variables, Y, appear only in the larger (IRS) sample, and still others, Z, appear only in the smaller (SEO) sample. Okner would prefer to have a single sample with information on X, Y, and Z. Since this does not exist, he proceeds to generate from the available data an artificial sample which he hopes has the same properties as the ideal sample.

Okner's problem is a special case of the following general problem: Given samples from two marginal distributions of a joint distribution, estimate the joint distribution and generate a sample from it. The hard part of this problem is estimating the joint distribution. Okner has handled this part badly by failing to separate it from the trivial problem of sample generation.

Once we have framed the problem this way, it is clear first of all that there is no information in the sample on the joint distribution of Z, Y conditional on X. To get the joint distribution of X, Y, and Z, we need an *a priori* assumption about the joint distribution of Y, Z conditional on X. The one Okner implicitly makes is that Y, Z are independent for given X. This may be a reasonable way to proceed, but the assumption ought to be stated explicitly and discussed. There are possible objections for the procedure. Apparently the most important component of Y is capital gains income. Is it true that, given other major components of the tax return, capital gains are independent of sociological categories? It seems to me possible that ownership of stocks might be more important for older people, more *or* less educated people (I can think of arguments both ways), and urban area residents. To the extent that capital gains make an important contribution to income distribution, the artificial sample Okner has generated will give artificial answers to questions about income distribution by the categories just named.

A more important criticism of Okner's procedure is that, once we have accepted the assumption of Z, Y independence for given X, a procedure which simply matches observations from the two samples is likely to generate biases. To see this, let us first consider what a good procedure might be. Ideally, one

[1] Though one of the major components of Okner's work involved generating X in a form comparable to that in the IRS sample from the SEO sample, I have no objection to the procedure by which Okner did this. Also, one of Okner's purposes was to correct for suspected bias in the SEO sample data for X. I think he could have done this better by a different procedure but leave this issue to the side for now.

would formulate a prior distribution on parameters of the conditional distributions of Y and Z given X, then use the sample information to estimate those distributions. The statistical technique required to implement this approach lies outside the standard econometric repertoire, but it could be handled along lines formally similar to those used by Robert Shiller[2] in putting Bayesian prior distributions on lag distributions. Short of the ideal procedure, one could devise an approximately Bayes procedure which avoids any serious bias along lines in many ways similar to what Okner does.

Suppose X has dimension k and that in every region of the X space we can specify a k-dimensional interval $I(X)$ such that for values of X separated by less than $I(X)$ the distribution of Y conditional on X is independent of X. We could then partition X-space into cells, none of which was larger than the local $I(X)$. In cells densely filled with observations, we could estimate the conditional distribution of Y directly within the cell by standard methods. Where cells were sparsely populated, we would need some smoothness assumptions on the parameters of the conditional distribution of Y, and these would lead us to some kind of regression technique—possibly as simple as just interpolating values from nearby, more densely populated cells.

We could do as Okner does and avoid the problem of estimating the X, Z distribution by using the original sample points for X, Z in our artificial sample. Furthermore, there would be little harm in matching Y values with X, Z values directly, without explicitly estimating the conditional distribution of Y, so long as all matches stayed within cells. Where Okner's procedure results in bias is where he makes matches which run across cells.

As it stands, Okner's procedure selects a "cell" corresponding to each point in the SEO sample, then looks for matches within it. His cells are defined by age, family status, income pattern, and income level. Okner's approach led him to want cells which would all contain small non-zero numbers of sample points, without explicit concern for the criteria developed in this comment. Hence he aggregated across age and marital status classifications to obtain cells which are clearly broader in that dimension than would be consistent with nearly constant parameters for the Y distribution, yet he chose narrow bands for the level of major source income. In any case it is clear that some of his matches run across not only his own cells (the "hand matches"), but also the cells which would arise from the criteria of this comment.

Two kinds of bias are likely. First, where cross-cell matches occur randomly, the variance of the conditional distribution of Y given X will clearly be upward biased. This will tend, e.g., to impart upward bias to measures of income inequality conditional on Z. In Okner's problem, that would be an upward bias in income inequality within sociological categories. More important, wherever the density of the X, Y sample varies systematically with X, there will be a systematic tendency for cross-cell matches to be found on the side of X where density is higher. This effect explains the bias Okner found at the upper end of the income distribution. For high incomes, sampling density in the IRS sample is increasing with total income. Thus in searching for a match to a high-income SEO return, Okner was

[2] "A Distributed Lag Estimator Derived From Smoothing Priors," Paper presented at the 1971 Meetings of the Econometric Society.

more likely to find the match on the high-income side because there were more IRS returns on that side.[3] As far as I can see, the disparity in sampling densities between SEO and IRS samples by which Okner purports to explain the bias is irrelevant. The bias comes from the *slope* in the IRS sampling rate in a region of the distribution where cells are sparsely populated.

Finally a word about Okner's correction for reporting bias in the SEO income items. The way to correct for that would be to compare the distributions of items in the SEO "tax returns" with the corresponding distributions from the IRS sample. Where disparities occurred, they could be corrected by rescaling the SEO distribution so it matched the IRS distribution. All of this could be and ought to be done *before* the matching procedure, since otherwise the reporting bias infects the matching. As it stands, Okner corrects the artificial joint distribution to match the IRS distribution, after the matching. I strongly suspect that some of the bias being corrected for in this way comes out of the matching procedure itself. If reporting bias had been taken care of at the start, statistically significant biases in the artificial joint distribution would provide a warning signal for some kinds of bias in estimation of the joint distribution.

University of Minnesota
National Bureau of Economic Research

[3] Okner does describe use of relative sample weights in his random selection procedure to assure truly random choice. But this applies only to *within* cell matches. In areas of the distribution where cells were sparsely populated, the random selection procedure was never invoked.

Annals of Economic and Social Measurement, 1/3, 1972

COMMENTS

BY JON K. PECK

Sims' comments on the Okner matching procedure provide a much-needed theoretical analysis, but some of the implications of his analysis for the validity of the match require further consideration.

It is clearly true that the two samples available provide no information on the joint distribution of Y and Z given X when one sample contains observations on the sets of variables X and Z and the other on X and Y. This necessitates some assumption such as independence. But this may not be as bad an assumption as Sims supposes, since X, the group of variables on which both samples contain data, includes not only certain kinds of income but significant amounts of demographic data as well (even though the data are not always exactly comparable between the two samples). Whether this is adequate will depend on the dimensions which are of interest in the matched population.

The ideal procedure which Sims sketches would make explicit many assumptions which are implicit in the Okner procedure, but the explicit specification of the joint prior distribution would be an enormous task because of the very large number of parameters to be specified and the interdependencies in the prior.

There is a choice to be made between matching procedures such as that used by Okner and more explicit estimation techniques for the joint distribution such as regression analysis or averaging or interpolation schemes. If one could confidently specify the functional relationships involved, this could be a satisfactory approach. But the problem is made more complex by the need to preserve many nonlinear relationships, some of which are stochastic, among predicted variables. For example, the IRS code prescribes numerous relationships among the variables in a tax return so that the average of n valid tax returns is not in general a valid tax return. While violating these relationships may do little damage in some aggregate dimensions, for many purposes these relationships are very important. For example, for calculating the effects of a change in the standard deduction percentage on the income distribution, the answers calculated from averaged and unaveraged returns are likely to differ significantly.

Another difficulty with a regression approach is the danger of introducing systematic relationships among variables, or strengthening existing ones, which will lead a researcher unaware of the manner in which the data were constructed to find statistically significant but artificial relationships among the sample variables. Of course, this is a danger with any imputation procedure, but it is particularly acute in the regression case. Ideally, the researcher should not have to know that he is dealing with an artificial sample.

Sims asserts that biases arise in the matching process when matches occur across the cells within which the distribution has constant parameters.[1] The biases

[1] Rosenblatt [1] has shown that there does not exist any unbiased, nonnegative, and symmetric estimator of a univariate continuous density function. This is not to imply, of course, that identifiable biases cannot be reduced or eliminated.

arise both in the conditional variance of Y given X and in the mean of the Y, Z distribution. It is difficult to assess the importance of these biases. Firstly, the cell definitions in the matching were dynamic and "fuzzy-edged". The set of potential IRS tax return matches was strictly restricted to a subset of the original cell grouping. The subset consisted of those returns lying within a nonsymmetric and variable width income interval containing the SEO tax return income; the width of the interval depended on the density of returns at that point and the SEO income amount. This set was restricted further by a set of other criteria (the consistency scores) which were used in a more flexible and probabilistic way. If this resulted in an unsatisfactory set of returns, the set was enlarged by widening its boundaries successively until an acceptable choice was available or the procedure gave up. Then a match was chosen by hand. Of the 28,643 matches, 97 percent were accomplished using the first set found. Therefore, the seventy-four cells listed by Okner should be regarded as very loose upper bounds on the eligible population.[2] Further, although some biases may be induced by this classification system, it is not obvious that in terms of, say, the mean squared error of the matches, the procedure will not do as well as the method Sims suggests. There is clearly a tradeoff between larger cell populations and thus greater heterogeneity within cells, and small homogeneous cells which are sparse in some dimensions. But it is not obvious that the particular cells which Okner chooses are optimal. If one is really to assess the quality of the matching process, the uses to be made of the result must be considered, and a loss function should be specified.

In particular, Sims' explanation of the large biases at the upper end of the income distribution seems implausible. He attributes the bias to the sharply increasing IRS sampling density at the upper end of the income distribution. In fact this density increases in widely spaced jumps at \$10,000, \$30,000, \$50,000, \$100,000 and \$200,000 of Adjusted Gross Income and is constant between jumps. A typical income band for a unit with \$20,000 AGI would have a width of \$800. Therefore, it is likely that most of the weights for eligible tax returns for any single match will be equal except around these jump points.[3] In the population, of course, the density falls sharply with increasing income in this range; therefore, Sims' analysis would suggest an underestimate of high income items rather than the overestimate which was observed.

Finally, I believe that Sims has misunderstood the Okner adjustment for reporting bias in the SEO income items. The adjustment process was accomplished in approximately the manner which Sims recommends. A second correction was made after the matching. This correction needs to be carefully inspected to detect flaws in the matching process which may well be significant. But I feel these flaws, to the extent they were avoidable, are more likely to be due to an inadequate definition of a good match (i.e., a loss function) than to statistical flaws in the general approach.

Yale University

REFERENCES

[1] Rosenblatt, Murray, "Remarks on Some Nonparametric Estimates of a Density Function", *Annals of Mathematical Statistics*, Vol. 27, 1956, pp. 832–837.

[2] Even the "hand-matches", which numbered only 151, stayed within these bounds.

[3] The use of the amount of "major source" income as the income definition for matching reduces the probability that all the weights in an eligible set of returns will be equal.

Annals of Economic and Social Measurement, 1/3, 1972

COMMENTS

BY EDWARD C. BUDD*

The microdata file constructed by the Brookings Institution by statistically merging the 1967 Survey of Economic Opportunity (SEO) and the 1966 Tax Model (TM),† while one of the first of its kind, is not quite unique. The Office of Business Economics (now the Bureau of Economic Analysis) as part of its methodology for estimating the size distribution of income for 1964, has carried out a somewhat similar statistical match of the March 1965 Current Population Survey (for income year 1964) (CPS) and the 1964 Tax Model. A more complete account of that match, together with a description of the other techniques used in estimating the completed series, is contained in another article,[1] and my remarks will be confined largely to a comparison of the methods used in the two statistical link projects.

It is perhaps unnecessary to elaborate on why such links, particularly those between field surveys and administrative records such as tax returns, are needed. Information contained in one is often not available in the other (e.g., absence of demographic data on tax returns); the quality of income data in field surveys is usually inferior to that contained in administrative records; the latter, on the other hand, do not contain the information needed to assemble them into consumer units (families and unrelated individuals).

While statistical links have their limitations, there are few if any alternatives. Exact matches (matching records from different sources for the same individual or sets of individuals) are, as Okner notes, not feasible for researchers outside of those in a few selected Federal agencies. Besides criticizing Okner's match, Sims refers to an alternate procedure of statistically estimating the joint distribution of different variables in the two files (Z for exclusively SEO variables, Y for exclusively TM variables, and X for those in both files), although his description is so sketchy that it is difficult to tell what he has in mind. He correctly points out, although perhaps it is obvious, that Okner's match is based on the assumption "that Y, Z are independent conditional on X," since X is by definition all the information that the two files have in common. In effect the SEO provides a matrix of X and Z; the TM, of X and Y. A statistical match is simply a method of estimating the joint distribution of Y and Z on a micro-record basis by combining the two matrices via the X's. If Sims's suggested method avoids a record-by-record match simply by aggregating and grouping the data for the two files before carrying out the estimation of the joint distribution, the difference between his method and statistical matches seems rather trivial.

* The views expressed here are the author's and do not necessarily reflect those of the Department of Commerce. I am indebted to Daniel B. Radner for numerous discussions of the comparisons and criticisms presented in this comment.

† Editor's Note: Okner prefers to use the term "Tax File" since it suggests the possibility of using different tax *schedules*; "tax model" connotes *behavioral* relationships.

[1] Edward C. Budd, "The Creation of a Microdata File for Estimating the Size Distribution of Income," *The Review of Income and Wealth*, December 1971, pp. 317–333.

Sims lays considerable emphasis on the use of incomplete information on the relation between Y and Z obtained from outside the two files in his estimation procedure, although it would have to be somewhat more precise than the kinds of hunches that he uses for illustration. (If the outside information were complete, there would be little point to the original estimation.) While one might agree that such data ought to be used if they exist, it would certainly be helpful to those of us working on statistical matching if he would develop precise methods for doing so. Perhaps the more important role for outside information would be in determining the comparability of the X's between the two files, rather than the relation between Y and Z. In most cases, the X's are not really defined in the same way in each file, and even if they were, they would be subject to different processing and response errors. For example, external evidence on the extent of underreporting of income types in field surveys as compared with tax returns ought, as Sims notes, to be taken account of before income is used as a matching variable.

There are a number of important differences in the methods used by OBE and by Brookings in creating their respective "merged" files, several of which will be described below. Some of these differences, it is true, may be due to differences in purpose: ours was to estimate the size distribution of family personal income, with tax return incomes being used primarily to correct the amounts reported in the CPS, whereas Okner's match was designed for use in the Brookings study of tax burdens. This point should not be overemphasized, however; the most appropriate solution to a number of the problems raised in statistical matching may not be that sensitive to differences in purpose. Furthermore, both the OBE and Brookings projects require both tax information and information on income size distribution for their final results.

One major difference lies in sampling. No sampling was involved in the OBE method. Each and every CPS record was retained, on the assumption that the CPS correctly represented the domestic noninstitutional population universe; each tax return was used once and only once, on the assumption that the TM correctly reflected the universe of tax filers. Given that every return had to be assigned to some person or married couple, our problem was then to find the most likely candidates for each of those returns. Okner, on the other hand, sampled the TM to obtain the returns to be matched with the SEO units with incomes below $30,000; some returns were presumably used more than once, others not at all. Aside from the added variance resulting from sampling, this difference need not be important if the sampling can be done without introducing bias, although, for reasons cited below, his methods do not appear to have met this latter condition. SEO units above $30,000, on the other hand, were discarded and replaced with all the tax returns above that income limit. This substitution produced biases as well, although not necessarily attributable to sampling.

Another related difference lies in the handling of the different weights in the two files that were merged. One of our goals at OBE was to avoid reweighting the TM file by the substitution of CPS for TM weights after matching. (Weights differ for each person in the CPS, although they are not a function of income size; the TM, on the other hand, is stratified by type of return and by [adjusted gross] income.) We first classified our records for matching purposes into cells on the basis of information common to each file (marital status; age—under 65, and 65

and over; existence and relative size [by rank] of wage income, self-employment incomes, and property income), with each cell being defined so that it had the same weighted number of records in both files, thus assuring that the corresponding universes for the CPS and tax returns would be the same for any given cell. We then split the records in each file so that the weight of a (split) CPS record would be identical with that of the (split) TM record combined with it. To illustrate, suppose within one of our matching cells with weighted records of 5,000, there were 2 CPS records, each with a weight of 2,500, and 5 TM records, each with a weight of 1,000. Our matching procedure then created 6 merged records, 4 with a weight of 1,000 each and 2 with a weight of 500 each. Since reweighting of either file was thereby avoided, the income types (whether taken from CPS or IRS) when summed necessarily equal the corresponding income aggregate in either the CPS or TM before matching. The cost of this precision was the length of our merged file; the number of records it contains is (approximately) equal to the sum of the records in each of the two files taken separately. We could, of course, have matched CPS records with tax returns by sampling within each cell, with the probability of selection proportion to the weight of the record. This would have avoided lengthening our file, but at the cost of introducing additional sampling error, although not necessarily any bias, into the results.

Okner's procedure was quite different. For each SEO record he defined a set of criteria (e.g., marital status of return, number of dependents, presence and size of certain income types) for selecting a sample of returns eligible for matching with that record. In this process, however, he ignored the weight of the SEO records, and took account of the TM weights only insofar as the probability of selection of a return included in one of his cells for matching with an SEO record was made proportional to the former's weight. Thus there was no assurance that the population represented by the SEO record was the same as the tax universe represented by the eligible TM records. To illustrate, suppose for a particular SEO record with a weight of 3,000, there were eligible for matching with it, determined by the criteria outlined in Okner's paper, 11 TM records, one with a weight of 1,000 and 10 with a weight of 100. Assume that the draw is made and the return with the 1,000 weight is selected. It will, therefore, be reweighted by a factor of 3. Half of this reweighting (1,000) allows for the other 10 returns that had a chance to be drawn but were not, with some consequent increase in sampling error. The other half of the reweighting, however, results in increasing the weight of the returns in this part of the IRS sample over and above the weight the returns originally had, thus overstating their importance in the merged file relative to their importance in the TM. Returns in other parts of the TM file could, of course, have lost weight in the matching process. That the effect of this reweighting was not random and resulted in biases in estimating the TM income types is indicated by the fact that many of the income type aggregates in Okner's merged file differ substantially from the corresponding TM or SOI totals. A more accurate test of the resulting bias could be obtained by comparing the size distributions of income types in Okner's merged file with those in the TM; such distributions are not, however, available from the Brookings' papers. OBE's merged file, in contrast, can reproduce exactly the TM's totals and size distributions for the various TM income types.

A third major difference was the treatment of underreporting of income in the

field survey relative to tax data. (Only for farm income can the former be said to be better than the latter.) The primary purpose of our match was to adjust the CPS for income underreporting by using tax returns, and our procedures in defining cells and in matching records took explicit account of differences in income and earnings levels between the two files. In defining the wage classes used for matching records within our marital status and age groups, for example, we ranked CPS records and tax returns from highest to lowest in terms of wage income, and divided them into classes based on their percentile position in the distribution. To illustrate, one of our classes might have encompassed all records in each file lying between the 4th and 5th percentiles (from the top), including all records in the CPS between $19,000 and $17,500, and in the TM from $22,500 and $20,000. Our matching procedure in this particular (hypothetical) class would result in matching tax returns with wage income averaging about $3,000 higher than the corresponding wage incomes reported by the CPS units.[2]

Okner, on the other hand, selected returns eligible for matching with an SEO record on the basis of the dollar size of major source income (the first pass being restricted to returns with major source income within the limits of a 4 percent band of the corresponding income type reported by the SEO unit). In view of the underreporting pattern previously referred to, SEO records were undoubtedly matched with tax returns having major source income below the SEO units "true" income (or at least the income reported on the tax returns the SEO units filed in real life). Underreporting bias thus appears to pervade the Brookings match. While it is true that in a subsequent "income adjustment" stage, the various income types in the merged file were blown up or adjusted by other techniques to their corresponding control totals as defined and estimated by Okner, this latter step is not sufficient to eliminate the effect of underreporting bias from the relative distributions.

One specific source of bias might be mentioned. The use of incomes as actually reported in the SEO and to IRS for defining major and minor income sources ignores the differential underreporting by income type. Thus, a "true" major source could be converted into a minor source, or a "true" minor source neglected entirely, by failing to correct for underreporting before matching, with the matches taking place more often than they should with respect to the better reported income types. Wage income, for example, is particularly well reported both in the SEO and on tax returns, at least relative to other income types. This may be one reason why so much wage income relative to Okner's control was obtained in the merged file. One way to have handled this problem would have been to blow up to control totals the income types in both files that were used for matching, before the match was carried out.

The failure to first correct for underreporting also biased the selection of nonfilers. Since nonfilers were chosen on the basis of incomes reported by SEO units and on legal filing requirements, too few SEO units would appear to have been assigned tax returns, with a bias towards nonfiler status for those SEO units having the less adequately reported income types. Indeed, the number of Okner's "SEO

[2] See Budd, *op. cit.*, pp. 324–327, for a more detailed discussion of the use of "ranking" in our matching process.

tax units" (i.e., filers) is 2.5 percent less than the actual number of returns filed, as reported in the 1966 SOI.[3]

It should be noted that the discrepancy between "SEO tax units" and actual returns filed is considerably greater for particular types of returns: SEO units assigned single returns were more than 12 percent short of the actual number of single returns filed; on the other hand, units assigned head-of-household and surviving spouse returns exceeded the actual number of such returns filed by two-thirds! This latter discrepancy should have created some suspicion that many SEO units, though perhaps eligible to file head-of-household and surviving spouse returns, were not in fact doing so, and hence some should have been assigned other types of return. Indeed, Okner's procedure might be characterized as assigning to SEO units tax returns they "ought" to have filed on the basis of information reported in the SEO, rather than the returns they *actually did* file. Indeed, the latter phrase is more descriptive of our approach to the problems of selecting nonfilers and deciding what type of return to assign to filers, although space unfortunately is lacking to justify this characterization of the differences between the two methods.

Okner's decision to discard all SEO records with incomes above $30,000 and simply substitute tax returns for that part of the SEO distribution points up some of the difficulties in the Brookings match I have already discussed. If the match is carried out by sampling, sampling error becomes a serious problem above this point because of the sharp reduction in the weight of the TM records. (This difficulty can only be resolved, as we did at OBE, by splitting records, an option rejected by Okner apparently for technical reasons.) In addition, the failure to allow for underreporting is far more serious for the upper income part of the file. It is not so much that the high income people are not in the SEO file; rather, they are being recorded at incomes below their "true" levels as reported on their corresponding tax returns. It is no wonder that there are so many more tax returns than SEO units (in terms of weighted numbers of records) at high incomes. Indeed, there are many returns with incomes well in excess of the highest income reported by any SEO unit; adherence to Okner's matching methods would, for instance, lose all of those returns lying above the income band for the highest SEO unit. In terms of my previous discussion of the reweighting problem, the population universe of the SEO and the tax return universe at higher incomes, when the comparison is made in terms of absolute income level (rather than relative income size), is simply not the same: the latter is always greater than the former.

Given these difficulties in his matching methods, it is not surprising that Okner felt forced to use a substitute procedure for the high income portion. That this path was, or had to be, chosen was indeed unfortunate for the quality of the merged file. For one thing, the upper part of the file remains on a tax return rather than a consumer unit basis; for another, no demographic information is available for high income units. Nor is this a trivial matter. While it may be true that this part of the file comprises less than 2 percent of the number of returns—presumably a

[3] See Okner's Table 1. In fact, this table underestimates by 300,000 the actual number of returns filed, since *all* (not just one half) of separate returns with *two* taxpayer exemptions must be counted. Footnotes (*a*) and (*b*) to that table can only be applied to separate returns containing *one* taxpayer exemption.

higher percent in terms of consumer units—the individuals and married couples comprising it must account for at least 10 to 12 percent of total income, even more for other income types, such as self-employment and property income.

This comment has certainly not touched on all the differences between the two merged files and the methods by which they were created. But it should be sufficient to permit the reader to judge the relative quality of the two files and their usefulness in meeting more general needs as well as the purposes for which they were created. Further, it should be kept in mind that the OBE and Brookings files are the first of their kind. Work on the statistical matching of data files is still in its infancy, and we should be able to look forward to the development of new and improved methods in subsequent work on the merging of files.

The Pennsylvania State University and
U.S. Department of Commerce

Annals of Economic and Social Measurement, 1/3, 1972

REJOINDER

BY CHRISTOPHER A. SIMS

Budd and Peck do not, as far as I can see, effectively respond to the main substance of my criticism of Okner. Perhaps, because of the way my comment was phrased, they did not understand how fundamental is my objection to creating artificial samples by matching, regardless of the details of the procedure.

To see my criticism in a different way, let us pose the question, "Is there any circumstance under which matching can be shown to lead to an artificial sample with the properties of a real sample?" That is, given the situation I set forth in my comment, with observations on X, Y from one sample and on X, Z from another sample, when will it be true that by matching observations according to X, an artificial Y, Z sample will result whose distribution is the true joint Y, Z distribution? If the joint distribution of X, Y, and Z has a probability density function, then the artificial sample will have the right distribution *only if* X, Y *and* Z *are mutually independent*. And in this case it does not matter how one performs the matching. One can match randomly and still get the appropriate distribution.

Budd, Okner, and Peck clearly do not believe their X, Y, and Z variables are independent, since otherwise they would not have bothered with their elaborate matching algorithms. I think what they are relying on is instead the assumptions (i) that Y, Z are independent given (conditional on) X and (ii) that the two samples are both very dense, in the sense that it is possible to do the matching in such a way that, if X_i is matched with X_j, the difference between the conditional distribution of Y, Z given X_i and the conditional distribution given X_j is small. The relations between the conditional distributions of Y and Z and the conditioning variable X is a regression relation.[1] If the values of the regression functions relating the parameters of the Y and Z distributions to X show very small changes between the matched pairs of X's, then the artificial sample generated by matching will have *approximately* the distribution of an actual sample.[2] But if this assumption of slowly-changing regression functions fails, the matching procedure is unworkable. Hence, it seems reasonable to suggest that anyone who uses the matching procedure present some evidence or argument that over the entire range in which he does matching, regression functions are slowly-changing relative to the gaps between matched X's. Not only do Okner (and Budd, in his *Review of Income and Wealth* paper) fail to address this question, but when we do look at the details of Okner's procedure, it seems quite likely that the regression functions are not slowly-changing across all his matches, or even across most of them.

Budd and Peck seem to think that the only way to meet my objections is to escalate the computational complexity of the procedure. It is true that I think that by a slight increase in computational complexity, a much better artificial sample

[1] By definition. A regression is such a relation. Least squares, interpolation, taking means, even matching, are different ways to estimate different kinds of regressions.

[2] Again assuming independence of Y and Z conditional on X.

could be prepared. However, if we require that there be no resort to explicit regression estimates at all, one can still improve on the Budd and Okner procedures. In the first place, one would define "cells" in such a way that within cells the regression functions can reasonably be supposed to be nearly constant. Then, in regions of the sample space where every cell has at least one X in each cell from each sample, one could match—either by Okner's procedure of random sampling within cells or by Budd's procedure of ranking all observations within these regions and matching by rank.[3] For those parts of the distribution where X's in both samples are not dense enough to allow within-cell matches, one would do what Okner already does for high-income returns: abandon the attempt to match, and list observations from both samples separately. This procedure might seem to result in a less "convenient" artificial sample. But users of the sample who are mainly interested in the dense parts of the distribution will find the inconvenience minor. Those who are interested in the less dense parts of the distribution as well will not be able to use the sample, but this is better than their being given a sample which appears to contain information about questions that interest them when in fact it does not.

Of course, many of the areas not sampled densely enough to justify the assumption that regression functions are nearly constant across all matches would be populated densely enough to justify an assumption that regression functions are, say, approximately linear across all matches. If enough computer time could be spared to estimate quite a number of local linear regressions, perhaps by rougher and quicker techniques than least squares, then the artificial sample could usefully be extended this way.

Two specific remarks by Peck deserve responses. First, he states that any use of explicit regression estimation techniques would be prohibitively difficult because of the need to preserve non-linear relations amongst, e.g., entries in the tax returns. These non-linearities seem to be of two kinds: some quantities, like numbers of dependents, have discontinuous distribution functions, and others are exact, non-linear functions of other entries on the return (like the tax given everything else). The exact dependencies can be taken care of by estimating distributions only for the independent components of the tax return, calculating the dependent components from the others after artificial values of the independent components have been generated. Entries like numbers of dependents or dollars worth of exemptions could be handled simply by not making the mistake of treating them as if they had continuous probability distributions. Thus, if for a particular observation we have estimated the conditional mean of number of dependents to be 5.3 with a conditional standard error of 0.6, we generate our observation's number of dependents from a distribution concentrated on the integers with this mean and this standard error.

Peck also remarks that, because the increase in IRS sampling densities occurs in discrete jumps at points widely spaced relative to Okner's income bands, it is unlikely that my explanation for the upward bias in Okner's estimate of income

[3] Budd's procedure would make sense only within regions where we were quite sure that the X's from both samples were dense enough without any explicit division of the space into cells. Once you have the cells, defined as I am suggesting, it would be a waste of computational effort to rank observations within cells for the match.

in the upper brackets is the correct one. Here I want only to say that I think Peck effectively answers himself in his own footnote. The widely spaced jumps in sampling density are for adjusted gross income. Okner's narrow income bands are for major source income. In the upper brackets many returns probably show multiple income sources, and Okner's cell definitions make no use, as far as I can see, of amounts of minor source income. These facts are enough to make me retain my suspicion that the bias in the income attributed to high-income groups comes from the source I described.

Annals of Economic and Social Measurement, 1/3, 1972

REPLY AND COMMENTS

BY BENJAMIN A. OKNER

A theoretical mathematician and a research statistician were each situated in opposite corners of a square room. In the corner between the two men was a lovely young lady waiting to offer her charms to whichever of them reached her first. Each man had to proceed towards the young lady one step at a time. And each step taken could cover only half the remaining distance between each man and the young woman.

*Upon hearing the rules, the theoretical mathematician left the chamber because he realized that it was impossible ever to get to the lovely young damsel. The research statistician stayed and claimed the prize. While he too realized that it was impossible to reach her, he figured that he could get close enough for all practical purposes.**

Given the institutional and other constraints which now preclude exact linking of microdata records, it seems reasonable to assume that the construction of synthetic information sets—such as the MERGE File—is the only feasible way to obtain more comprehensive data than are now available from any single source. This is unfortunately likely to be the case for many more years. For most researchers who need such data now, the relevant question is not whether to construct synthetic microdata sets but how best to do so.

Since synthetic linking of different microdata files is still in its infancy, there are few generally accepted procedures or tests for determining the "correctness" of a matching procedure. Consequently, there is no objective way to decide if any given method should be labelled "unsatisfactory," "good," or "best." But given the present state of the art, surely one of the criteria used for judging the "goodness" of a synthetic microdata file is how well it fulfills the research (or other purpose) needs for which it was constructed.[1]

The primary reason for creating the MERGE File was to obtain a suitable microdata base for the overall distribution of tax burdens study currently in progress at Brookings. Clearly, the 1966 Tax File is the best data source on income for units in the upper tail of the income distribution. And if we were not concerned with the *overall distribution* of taxes, the need to create the new file would have been largely obviated. But since we obviously could not obtain information on the nonfiling population from the Tax File, it *was* necessary to add such low-income units to our data base.

Although obtaining income and other data for the low-income nonfiling population was a major reason for creating the MERGE File, there are additional benefits to be derived from using the MERGE File for the tax burden study. Among the most important are the ability to conduct the analysis on a family or consumer unit basis and the possibility of using microsimulation techniques to

* The story is attributed to Professor Leslie Kish of the University of Michigan Survey Research Center.

[1] Thus, while the methods used by Professor Budd in creating his file for the U.S. Office of Business Economics differ substantially from those I used (see pp. 325–42 above), our goals and the future uses of the files are also quite different. For a detailed explanation of his work see Edward C. Budd, "The Creation of a Microdata File for Estimating the Size Distribution of Income," *The Review of Income and Wealth*, Series 17 (December 1971).

project current data into the future. Thus, we have the ability to predict how many families not now required to file tax returns might be pulled into the filing population if the tax statutes are changed to broaden the tax base as well as to analyze how income growth over time will affect the filing status of current nonfilers.

In terms of the actual procedure used for selecting Tax File units to be attached to SEO families, there is little I can add to Professor Peck's reply to Sims' comments.[2] However, I think that it is important to underscore a very important point that Sims mentions but does not emphasize. The fact that I assumed that the Y and Z variables are independent given X (using Sims' notation, see p. 343 above) has extremely important implications for the "proper" use of the demographic categories by which families in the MERGE File should be classified. Sims illustrates the point with capital gains income; however, I do not feel this was an especially good choice since the vast bulk of such receipts are concentrated among units in the upper tail of the income distribution where the File contains no SEO demographic data. There are much more subtle things of which the researchers must beware when using the MERGE File. For example, one of the variables in the set of Z (SEO) variables is race and one in the set of Y (Tax) variables is business income. It seems reasonable to assume that receipt of business income is not independent of race. While this lack of independence may have been taken account of adequately in the matching process through the income pattern classification, there is no assurance that this is the case. Thus, while race is a variable in the MERGE File data file by which families *can* be classified, if at all possible a researcher *should not do so* without further investigation of various relationships between income and race from other sources. If the outside checks are impossible because the necessary data do not exist (which I would guess will often be the case), the analyst has the choice of either not classifying by the variable or using it at his own peril. In the latter instance, he certainly has a responsibility to explain fully what he has done and why.

Although it is clearly not a sufficient condition to "prove" that the matching procedure used was correct, a necessary one would be the reasonableness of information generated from the MERGE File. On this criterion, tabulations using the file do not suggest that the assumptions made in merging have done violence to the information. After completing the merge process, one of the first things we checked was the total federal income tax liability computed on the basis of MERGE File family records. For 1966, Internal Revenue statistics indicate that personal tax liability was $56,087 million;[3] the total calculated using the MERGE File was $54,596 million. This difference of less than three percent is certainly within sampling tolerance. In addition, both the distributions of income and taxes in the MERGE File are very close to published statistics for 1966. For those filing returns, we also found the well-established pattern of effective tax rates as income rises—moderate progression throughout most of the income scale with a regressive drop in the effective rate of taxation among those at the very top of the income distribution.

[2] One question that has been asked by several people is whether units from the Tax File were sampled with replacement in the merge process. The answer is definitely in the affirmative.

[3] U.S. Internal Revenue Service, *Statistics of Income—1966 Individual Income Tax Returns* (1968).

I want to emphasize that results such as these do not, and cannot be used to, prove that the matching procedure was "correct." Based on our work to date, however, the results we have obtained from MERGE File tabulations appear to be reasonable and consistent with other information we have concerning the distribution of income and taxes.

Even though the initial reason for creating the MERGE File was to provide the basis for estimating the distribution of federal, state, and local taxes by income levels, it has already been utilized for several other purposes. One of these was the simulation of different payroll tax changes which would help to remove the present regressiveness of this levy.

The estimates prepared indicate that the flat payroll tax now paid by the wage and salary earners could be replaced by a mildly progressive tax on total income or on earnings at reasonably moderate rates. The progressive tax would relieve those who earn less than the officially-defined "poverty lines" from making any contribution to social security out of their inadequate incomes; and it would reduce the taxes of the vast majority of income recipients, while raising taxes only for the top 10 or 15 percent of earners. The merits of these alternative methods of financing social security are just being recognized, and the public debate is already under way.

The projection capabilities noted above have already been used in research on the effects of adopting a comprehensive income tax in the United States.[4] The goal of the comprehensive tax analysis is to help to understand the large differences between the nominal and actual effective tax rates paid by U.S. families. These differences, of course, are largely due to the "erosion" of the tax base because of the numerous exclusions, exemptions, and deductions permitted under various provisions of the Internal Revenue Code. The extent of the erosion has been estimated in *aggregate terms* in the past, but reliable estimates of the differential impact of the special provisions at various income levels have never been available.

The new estimates of the yield of a comprehensive income tax involve taxation of all realized capital gains as ordinary income; taxation of capital gains transferred by gift or bequest; elimination of the exemption of interest from state and local bonds; limitation of depletion allowances to cost depletion; taxation of interest on life insurance policies; inclusion of net imputed rent in taxable income and elimination of the deductions for real property taxes and mortgage interest; taxation of transfer payments as ordinary income; elimination of most itemized deductions;[5] limitation of the standard deduction to a flat $1,300; elimination of the special exemptions for the aged and blind and the retirement income tax credit; and elimination of the rate advantages of income splitting. To make the estimates relevant to the current scene, we used projection techniques developed to raise the MERGE File income to the expected 1972 levels.

These revisions would increase the estimated tax base in calendar year 1972 (under the current tax law) from $478 billion to $644 billion, an increase of $166

[4] See Joseph A. Pechman and Benjamin A. Okner, "Individual Income Tax Erosion by Income Classes," in *The Economics of Federal Subsidy Programs*, A Compendium of Papers submitted to the Joint Economic Committee, 92 Cong. 2 sess. (1972), Pt. 1, pp. 13–40.

[5] Deductions would be allowed only for state income taxes, medical expenses in excess of 5 percent of income, charitable contributions in excess of 3 percent of income, and interest up to the amount of property income reported by the individual on his tax return.

billion, or 35 percent.[6] If the present tax rates of 14 to 70 percent were left unchanged, adoption of this comprehensive tax base would raise tax liabilities of individuals from $103 billion to $180 billion, an increase of 75 percent. This means that the tax rates could be reduced by an average of more than 40 percent and still yield the same tax as under current law.

Of course, the various features that make up the $166 billion difference between taxable income under the comprehensive income tax and the present (1972 tax) law are not evenly distributed among families at all income levels. As a matter of fact, the impact of the various changes is striking when examined by income class. For those interested, I refer you to the JEC Compendium for further results of the analysis.

I believe that the work described above illustrates well the fact that the MERGE File is an extremely valuable tool for a large variety of research purposes. I presume that the skeptic will point out that the "law of GIGO" has never been repealed, and that any results derived from the file may or may not be valid. In the final analysis, the methods used for constructing the MERGE File will undoubtedly be tested by how reliable and useful it is in serving its function(s). If "the first heavy wind to come along blows it over," obviously we'd better go back and examine the "foundation" very carefully.

While there were instances in the merging procedure when we used what many people would regard as "rough and ready" assumptions, the real issues are whether such assumptions did real violence in terms of the finished product and whether the benefit-cost ratio of attempting to improve them would be greater than one. Obviously, my answer on both these issues would be negative.

In other words, I believe we *are* close enough for all practical purposes.

The Brookings Institution

[6] The estimates for 1972 were based on projections of incomes from the 1966 base, assuming the percentage change in individual income sources would be the same as the estimated changes in personal income.