THE CREATION OF A MICRODATA FILE FOR ESTIMATING THE SIZE DISTRIBUTION OF INCOME*

BY EDWARD C. BUDD†

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

This paper describes the methods used by the Office of Business Economics, U.S. Department of Commerce, in creating a microdata file for use in estimating the size distribution of income. It explains the techniques of statistical matching involved in merging microdata files from various sources to correct and supplement income estimates in the original field survey (The Current Population Survey) and to incorporate additional information that can be used to estimate items and types of income not contained in the original file.

An important development in recent years in research in the area of the personal sector of the economy in the United States has been the creation of data files for samples of individuals and households, and their increasing availability to both government and private research workers. These include such files as the Internal Revenue Service's Tax Model of individual returns for 1964 and later years (TM); the condensed version of the Bureau of Labor Statistics' Survey of Consumer Expenditure for 1960–1961 (CES); the Federal Reserve Board's Survey of Financial Characteristics of Consumers (SFCC) (for income year 1962) and Survey of Changes in Family Finances (for 1963); the 1960 Census one in a thousand sample; the March income supplement of the Current Population Survey (CPS) and the Survey of Economic Opportunity (SEO) for income years 1965 and 1966. Others, such as the Office of Economic Opportunity's Longitudinal Study, might also be mentioned.

This development has several advantages from the user's standpoint. No longer is it necessary for him to rely on the tabulations published by the agency in sole possession of the tape—tabulations which may or may not correspond with his particular requirements. While he is always free to request the agency to run special tabulations, manpower and data processing limitations often produce extended delays, for it is unlikely that the agency will view the request with the same high priority as does the user. Furthermore, user's needs, running beyond ordinary tabulations and requiring editing or modifying a file or performing simulations, are greatly expedited, if not made possible in the first place, if the user has direct access to the tape.

For the suppliers there are, of course, certain problems to tape release. For the sample survey or tax data, information in the record may permit the identification of the individual respondent, and certain fields may need to be deleted for the sake of confidentiality. Another problem concerns the condition

*Prepared for the Workshop on the Use of Microdata Sets in Economic Analysis, sponsored by the National Bureau of Economic Research, October 22–23, 1970. I am indebted to Daniel Radner for correcting some errors in an earlier draft of this paper.

†The views expressed here are the author's and do not necessarily reflect those of the Department of Commerce.

of the tape when it is released. If, as McClung[1] suggests from his experience with the SEO, tape files are analogous to agency worksheets rather than to tables the agency publishes, an additional workload may be placed on the producer in documenting the file—explaining how it was generated and how the tape is to be read. If, on the other hand, the agency attempts a better job of cleaning up the tape before release, processing costs and time delays may be significantly increased. My own experience (which, however, does not encompass the SEO) has been that, with some exceptions, tapes could be used with relatively little consultation with the producing agency and without substantial input of resources devoted to editing and improving the file.

It needs to be, and perhaps is, recognized that many of the deficiencies in tape files now available simply reflect the limitations of the basic data sources from which they are derived. Errors in coding, editing, and transcribing information onto tape from the schedule or return will always be with us, and are presumably reflected in the products derived from tape files, such as published tabulations, even if less apparent in the aggregated data. Use of micro files has simply made us more aware of a pre-existing condition.

In sample surveys, however, the problems go well beyond errors in processing, to errors and oversights by enumerators in the interview process (witness the some 300 "not available" responses on race in the SFCC, over 11 percent of the sample!), to respondents' errors in filling out questionnaires or in answering enumerators' questions, and to the failure of some respondents to answer certain questions at all, particularly those on income and assets. Even the failure of the enumerator to obtain an interview can produce bias in the data, especially on income. For example, the noncontact rate in the SFCC for those with incomes below $15,000 was less than 2 percent; for those over $15,000, about 6 percent.[2]

The deficiencies are perhaps most serious for the income data. Underreporting by those who do report their incomes runs around 9 to 15 percent on the average—much greater for certain income types, such as interest and dividends. The receipt of some income types, particularly property income and certain transfers, is often not reported at all; indeed, one often suspects that a "none" is a substitute for a refusal to answer the question. And the nonresponse rate is, of course, highest for the income questions themselves.

There is some controversy on the importance of these deficiencies in the basic data sources. Published tables, which present size distributions of income and medians and/or means for one or a few economic and demographic variables at a time, may not, it is true, be too sensitive to some of these errors or to methods for mitigating their effect. More sophisticated methods for allocating incomes to nonrespondents, for example, do not seem to produce higher median incomes

[1]Nelson McClung, John Moeller, and Eduardo Sigel, "Transfer Income Program Evaluation" (mimeo), paper presented at the Workshop on the Use of Microdata Sets in Economic Analysis, National Bureau of Economic Research, October 1970.

[2]The "incomes" used are those that persons in the sample reported in the 1960 Census, or for those over $50,000, adjusted gross income reported on 1960 tax returns, obviously not 1962 incomes "reported" to an interviewer. The percentages were computed from unweighted observations and may tend to overstate nonresponse rates for those over $15,000. *Survey of Financial Characteristics of Consumers*, pp. 50–52.

for such records or result in a perceptible change in the distribution. Even here the effect of errors cannot be overlooked entirely—certainly the effect of underreporting. With the increasing availability of microdata files, however, and the consequent opportunity for obtaining smaller cells than have heretofore been published and for exploring the relationships among fields different from those embodied in current tables, problems of data quality become increasingly serious.

What I am arguing for, I should add, is the development of methods for improving quality in basic sources, not restricting the use of the files by producing agencies to whomever they may judge to be a "knowledgeable" user, or in other ways. It is after all the user's reputation that is at stake in the use he makes of the file, not the producer's, and it is the former's responsibility to exercise the required professional caution in not pressing for answers or results beyond the capability of the underlying data base to supply them. Users themselves need to become experts on the limitations of the data in micro files as well as on their potentialities. It might also be noted that this problem arises with the release of any data, not just micro data. For example, researchers investigating the cyclical variability of income size distribution make use of annual size distributions obtained from the CPS with little regard for the data problems that may produce much of the variability they seek to explain.[3]

The one ongoing sample survey under U.S. government auspices is the CPS, and the Census Bureau has over the years introduced a number of changes that have improved the quality of the annual income supplements. To mention just a few, a method of assigning income to nonrespondents to income questions was introduced in 1961; another income question was added in 1966; improved editing and field procedures were introduced in 1969 and an experimental program was undertaken in order to improve the quality of income response and reduce nonresponse rates. One of the outcomes of this program has been the collection of income and work experience data in the same rather than successive months, which should eliminate "nonmatches" and errors in matching data from two different surveys, and improve the consistency between earnings and work experience for individual records. Under current discussion is a proposal for an "Experimental Spring Supplement," involving a small-scale research sample in addition to the regular CPS income supplement; I earnestly hope this project will be given the priority it deserves. It is doubtful, however, whether improved survey methods will meet all our needs for expanded and better quality data; they are obviously of little help for files generated before such improvements are introduced. There is still a need to explore other methods.

Limited use for quality improvement purposes may be made of editing procedures, some of which are designed to make the information in a record consistent; others, to replace missing information. Consistency edits are useful,

[3]T. Paul Schultz, "Secular Trends and Cyclical Behavior of Income Distribution in the United States: 1944–1965," *Six Papers on the Size Distribution of Wealth and Income*, Lee Soltow, ed., Studies in Income and Wealth, No. 33; Charles E. Metcalf, "The Size Distribution of Personal Income During the Business Cycle," *Am. Econ. Rev.*, v. 59 (September 1969), pp. 657–668.

but there is often some question as to which is the "right" and which is the "wrong" answer. If a person reports previous year's work experience, but no earnings, do we eliminate the experience, or allocate the earnings? What if the reverse occurs? Similar problems arise in connection with income and asset information, although the latter case is complicated by the fact that "none" for one and an amount for the other could be consistent answers. The preferable thing to do may be to leave the inconsistency and let the user determine the best procedure for his objective.

A somewhat different case arises where the information is actually missing, either because certain questions were missed in the interview or because the respondent refused to provide an answer, the latter a particularly serious problem for income questions, although many of the "no income" responses, as I have already noted, look suspiciously like nonresponses (NA's). Unless one is willing to make the assumption that the distribution of NA's is the same as the distribution of respondents for a particular field—an assumption inconsistent with the fact that NA's and respondents differ for other characteristics on which both groups provide information—the missing information for each record must be filled in before that field can be used. Assignment or allocation methods have become more sophisticated since they were first introduced into the 1961 CPS for income NA's, but it would still be safe to say that they have not been able to fully take account of the presumed relation between nonresponse and income; indeed, it is difficult to devise methods that will. For the March 1965 CPS, for example, we experimented at the Office of Business Economics with a considerably more detailed matrix of characteristics employing more than five times as many cells as were used in the Census matrix at that time. While the relationship between the economic and demographic characteristics of NA's and the incomes assigned to them was undoubtedly improved, the income total allocated was, if anything, less than that originally allocated by Census.

Another line of attack in improving microdata files obtained from field surveys, which can also be used to add information not originally a part of the file, is to combine them, on a record-by-record basis, with information derived from other microdata sources, particularly administrative records such as tax returns or social security earnings or benefit records. Tax returns are particularly useful in correcting or supplementing income data from field surveys, especially for the upper tail of the distribution. Certain income types, notably property income, are better reported, and information not normally included in field surveys, such as taxes paid, can be incorporated. If tax return and sample survey files can be merged, some of the limitations associated with using a tax return data file, such as IRS's Tax Model, can be partially overcome. For one thing, one is not restricted to the use of the tax return unit as a recipient unit, since the merging process combines the tax returns into consumer units (families and unrelated individuals). Further, the merging process has implicit in it an estimate of those who were assumed not to have filed a tax return; an estimate of their incomes is, of course, available from the field survey. For some purposes it is important to have the file cover the entire population, not just those who filed tax returns. The absence of nonfilers in the tax return file is one reason why it is difficult to manufacture a pseudo-family file from a tax return file by using

methods to combine into family units the returns that are filed. Third, the sample survey file can often provide estimates of tax-exempt income which are, unfortunately, not obtainable from the tax return. Finally, I am tempted to mention the demographic and economic characteristics available from the survey file that can be identified with the tax unit, but it is just this absence of information on the tax return that makes the statistical merging of the two files so difficult.

One method of merging survey and administrative records is the exact match: starting with the sample survey records, an attempt is made to locate the tax return(s) filed by the survey unit and/or its corresponding social security records. If the sample survey covers a sufficiently large proportion of the population, the matching process can be done the opposite way: starting from a sample of tax returns or social security records, an attempt is made to locate the corresponding units in the population survey. This latter method in fact was used in the 1960 Census-IRS match, although in retrospect it seems to have raised as many difficult problems as the 1950 Census-IRS match, which went from a sample of Census units to their corresponding tax returns. For small sample surveys the latter method is not feasible in any case.

The Social Security Administration has been working on a series of three way link projects, involving the matching of a sample of records from the March 1964, 1965, and 1966 CPS, and the 1966 and 1967 SEO, with tax returns and social security records. This approach is not without its problems, either, as the amount of time consumed in this project would suggest. One problem, so far as the field survey-IRS part of the link is concerned, is determining precisely for which records a tax return has been filed but cannot be located, and for which a tax return has not been filed at all. The exact matches do not provide an independent estimate of the number of (legal or illegal) nonfilers or their characteristics.

A more serious problem is the inaccessibility of the files to those outside the Social Security Administration (SAA). Whatever the reason offered—confidentiality, insufficient quality of the match, desire to control use made of the data—it means that the kinds of research that for practical purposes require access to the file by the user can never be carried out. SSA could, of course, mitigate the effect of this decision to some extent by preparing studies, such as Census has carried out for the Census-IRS matches, on relationships among data contained in the CPS, tax returns and social security records. For example, it would be useful to know something about the characteristics of, and CPS incomes of, persons filing various types of returns, e.g., how many joint returns are filed by other than "married, spouse present" in the CPS, or about relationships among self employment incomes for each of the three sources. The agency has, however, given high priority to another, quite independent question: deciding by a series of elaborate rules what is the "best" estimate of the particular income type from each of the three sources. While work on what SSA calls "calibration" proceeds, other equally important aspects of the files remain unutilized.

There is therefore a need for other methods for linking microdata files containing both survey and administrative records, with reliance on statistical

rather than exact matching of records. If it is a CPS-IRS link that is desired, the trick is to find returns similar to, rather than identical with, those filed by a CPS family, or to decide whether any returns, indeed, might have been filed by that unit. There are, of course, a number of different ways of doing it. The paper by Okner[4] describes the method used by the Brookings Institution to carry out a statistical match between the 1967 SEO and the 1966 Tax Model. I will give you a nontechnical description of the one used at OBE for matching the March 1965 CPS and the 1964 Tax Model.

At this point it might be useful to comment on the reasons for undertaking the project. It grew out of the need to establish a new methodology for estimating OBE's old income size distribution series. One of the major purposes of this series was to account for all of family personal income as estimated by OBE. It required what might be described as a "synthetic" methodology—estimating a size distribution by using data from a variety of different sources, including field surveys, tax returns, business and administrative records of one sort or another, and the aggregate income types as estimated in the national accounts. The old series did not, however, meet the need for micro distributions; the only breakdowns available were for nonfarm families, farm operator families, and unattached individuals. Indeed, it would have been impossible to do so, since the underlying microdata tapes either had not been created or were not available for use outside the agency creating them. Heavy reliance had to be placed on published tabulations and cross-tabulations, with the necessity for interpolating within class intervals as items of income were added or deducted.

In developing the new methodology we have relied almost entirely on the use of microdata files—including the March 1965 CPS, the 1964 Tax Model, the SFCC, and the IRS audit study for 1963—rather than published tabulations. The old series depended on tabulations from the 1950 Census-IRS match to provide a link between family units and tax return units; our first step was to statistically match, record by record, the CPS and the Tax Model. At the same time we developed methods for correcting the Tax Model income types by use of the audit file. After assembling the matched file into families and carrying out the audit correction of the Tax Model, we next executed a record-for-record match of our merged file with the SFCC. This step was necessary in order to provide additional information for estimating the distribution of certain types of imputed income, as well as state and local bond interest and accrued savings bond interest, which are not subject to income tax. In the remaining space, I will try to give an abbreviated description of each of these steps.

It might be well to point out certain assumptions made and guidelines developed for the match. First, common information between the two files was to be used as much as possible so as to minimize the need for random matching of records. Second, it was assumed that the CPS correctly represented the (domestic noninstitutional) population universe, and the Tax Model (TM), the universe of tax filers; reweighting of records was to be avoided where possible so that we could come up to the income or other aggregates implied in either

[4]Benjamin Okner, "Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File," forthcoming, *The Annals of Economic and Social Measurement.*

file.[5] Third, the methods used were to preserve as nearly as possible the relationship between the CPS demographic and economic (e.g., work experience) characteristics and the CPS income types and amounts.

One of the major problems in meeting the first condition is that there is so little demographic information on the tax return; marital status of the taxpayer and the number of taxpayer and dependency exemptions is about the extent of it. We used marital status and age—whether the taxpayer (and/or spouse, if a joint return) is 65 and over—together with the existence and size of the following income types: wage and salary income, business (schedule C) and partnership income, farm (schedule F) income, and property income (interest, gross dividends, and rent). The records in the CPS and the TM can be classified into six groups, based on the analogy of marital status in the two files and on age: (1) age under 65: married couples living together (inclusive of couples in subfamilies, but exclusive of couples married in 1965)—joint and separate returns of husbands and wives; (2) same as (1) but age 65 and over; (3) age under 65: other family heads (including heads of subfamilies)—heads of households, surviving spouses, and single returns with one or more dependents; (4) same as (3) but age 65 and over; (5) age under 65: other persons (i.e., other relatives and unrelated individuals), including as individuals persons married in 1965—single returns with no dependents; (6) same as (5) but age 65 and over.

The correspondence of persons and tax returns within each of these groups is certainly not exact; for example, married couples may not be living together; a single person may be supporting relatives living elsewhere. Indeed, we have little exact knowledge of the kinds of returns filed by various persons or couples. Certain modifications in the above classification were therefore introduced in order to increase the comparability of records between the two files in each group. One consideration was the number of couples or persons relative to the number of returns in the group. An insufficient number of returns relative to the persons in one group would leave too many units without a tax return, and the unmatched units might have CPS taxable incomes well above IRS filing requirements. The opposite case would leave CPS units with returns even though their CPS incomes were well below filing requirements. Thus, in view of the number of joint and separate returns we had relative to the number of married couples, we decided to allocate separate returns with one exemption among groups (3), (5), and (6) in accordance with the proportion of other heads to other persons in the subgroup of persons—"married-spouse absent" and separated—who might have filed such returns, instead of matching up such returns to convert them to "pseudo joint" returns to be assigned to groups (1) and (2).

Another test we used was the correspondence between the distributions of wage and salary income in each group in the two files, since it is the best and most fully reported type in either file. In general, the CPS distributions were lower than those in the TM, with fewer frequencies in the upper brackets and

[5]Since we used only those CPS records which contained income information—three quarters of the sample in income year 1964—whereas Census weights were computed on the basis of the full sample—records in the three-quarter sample were reweighted so that tabulations of income data would come up to the CPS universe. The reweighting scheme controlled for family relationship, farm-nonfarm residence, age, race, and sex.

more in the lower. Most of the difference can be accounted for by the fact that wage income is more completely reported on tax returns and consequently the mean wage income of the tax distribution is higher. If an adjustment is made for this difference, the groups in general did meet this test. In the last two groups, however, there were more frequencies in the upper brackets for other persons than for single returns. Some returns from the first four groups were therefore allocated to the latter two to improve the correspondence between upper parts of the two wage distributions. Our work in this area, I might add, leads me to suspect that "too many" higher income joint returns are being filed, and "too few" other types of returns, judging at least by the CPS data.

The next step was to determine which CPS records in each group were not to be given a tax return, i.e., were to be classified as nonfilers, since the weighted number of records in each group had to be the same for matching purposes. Consideration was given to the existence and size of self-employment income, as well as to CPS "taxable" income, which we defined to include earnings, property income, and "all other," including pensions and annuities, veterans' payments, workmen's compensation, royalties, and interpersonal transfer payments. (Perhaps only half of CPS "all other" income is subject to tax.) While space is lacking to describe the procedure in detail, the net result was that all of those classified as nonfilers reported taxable incomes in the CPS below IRS filing requirements. The criteria were worked out in terms of legal filing requirements; no attempt was made to estimate the existence or number of illegal nonfilers.

The matching of returns within a group was then carried out on the basis of the size and amount of wage (and salary) income, self-employment income, and property income. We gave primacy to wage income as a link, because it is more consistently and more accurately reported in both sources than is self-employment income. Link studies, such as the IRS-Census match, certainly point to this conclusion. Further, the correspondence in the two sources among the self-employment distributions is certainly less close than for wage income. The tax return distributions for nonfarm self-employment, for example, show considerably more dispersion than the CPS ones, with more frequencies at the top, and many more with low, especially negative, incomes. It would be an interesting question—one that cannot be examined here—to account for these differences.

These are reasons for giving self-employment a secondary role in the match, not for putting it aside entirely. The matching studies do indicate a high degree of consistency in reporting its existence, if not its amount, in both sources. Furthermore, it is a particularly important source of income for those with no wage income: about half of those with self-employment have no other earnings types, and a high proportion of filers with no wage income have self-employment; four fifths of all joint returns with no wage income, for example, had business, farm, or partnership income.

In each group we ranked the units by size of wage income and separated them into a number of wage rank classes, with an equal number of frequencies from the two files in each class, although frequencies varied from one wage

class to the next. For example, a bracket might encompass all records in each file lying between the 4th and 5th percentiles (from the top), even though this might include all records in the CPS lying between $19,000 and $17,500, and in the TM from $22,500 to $20,000. The number of classes had to be large enough so that there was not too much disparity between the top income in the TM and the bottom income in the CPS (a disparity of $5,000 in the above illustration), and yet small enough so that some weight would be given to self-employment and property incomes in the matching process. For example, it would be possible to make each class coextensive with a CPS record, which would mean that records would be matched solely by their rank in the wage distribution without regard to other income types. The number of wage classes ranged from 37 for husband-wife couples under 65 to 16 for other heads 65 and over.

Each wage class was then broken up into 4 subclasses, based on whether the record in either file had nonfarm self-employment income (NFSE), farm self-employment (FSE), both, or neither. For the first three subclasses, the records were reranked by absolute size of such income, from highest to lowest. Since the number with, say, NFSE in the CPS (or TM) in subclass 1 could exceed that in the TM (or CPS), the excess, and hence those with the lowest NFSE amount (in absolute value) had to be transferred to the fourth subclass so that we would be assured of the same weighted number of records in each of these four subclasses for matching purposes. A consequence of this transfer was that not every record with NFSE and/or FSE in one file could be matched with a record with the corresponding income type in the other—although roughly 90 percent of those with NFSE, and 76 percent of those with FSE, in the CPS were matched with a return containing Schedule C, or Schedule F, income respectively. The relatively poorer showing for farm income was due to the overall shortage of tax returns with farm income compared with those reporting farm income in the CPS.[6] It might be noted that these percentages are similar to percentages shown by exact matches for those who are found to have reported a particular type of self-employment income in either or both sources.

For the first three subclasses records were initially matched in each cell on the basis of their rank in self-employment income, from highest to lowest in algebraic value. This method gave no weight in matching to size of property income and for the records involved resulted in almost a random relation between CPS and IRS property income. For each of those cells based on wage rank classes with no, or a small amount of, wage income—comprising only a sixth of the cells, but encompassing almost three quarters of the weighted records with self-employment income—we therefore established a small number of additional classes based on size of self-employment income. These were determined by ranking the records by size of the latter income type from highest to lowest, with the same weighted number in the two files in each subclass—a method identical with that used in setting up the wage income brackets themselves. The purpose was to give some weight to size of property income in matching those self-employment records for which wage income was small or absent entirely.

[6]Part of this difference may be attributable to the fact that farm partnership returns had to be included in NFSE, since there is no industry breakdown of partnership income in the TM for individual returns.

Within each of these size and type subclasses—with the exception noted above of approximately a quarter of the records with self-employment income which were matched by size of self-employment and wage income alone—the records were reranked by size of property income from highest to lowest; those with no such income were ranked in random order. The records were then matched by their rank. The assumption in this final step of the matching procedure is that those with property income in the CPS, as compared with those reporting a "none," are more likely to have reported it on their tax return and to have reported a larger amount. The procedure itself implies that *within* one of our subclasses a "none" in CPS could not get a tax return with a property income amount greater than the tax return amount assigned to a CPS unit reporting a positive property amount. We know, of course, that property income is more fully reported on tax returns, and reported by a larger proportion of the recipients, than is true of the CPS. In any case, the net result of the matching procedure was that those CPS units reporting property income tended to receive a tax return with a larger property income than the amount they reported in CPS, and many more reporting "none" in the CPS received a tax return with a property income amount than was true in the opposite case. The percent of filers reporting property income in the CPS, and the corresponding tax return percentages, are shown in Table 1.

TABLE 1

PERCENT OF RECORDS WITH PROPERTY INCOME IN CPS BEFORE AND AFTER MATCHING THE CPS WITH THE TAX MODEL

Group	Number in Group (1,000's)	Percent of Records with Property Income	
		CPS*	TM
Married couples:			
(1) Under 65	36,046	32	46
(2) 65 and over	3,368	54	80
Other heads:			
(3) Under 65	3,810	23	28
(4) 65 and over	381	54	87
Other persons:			
(5) Under 65	19,452	18	30
(6) 65 and over	2,145	57	83
All groups:			
Under 65	59,308	27	40
65 and over	5,894	55	81

*Exclusive of nonfilers (those not assigned a tax return).

It is apparent that the method described did meet the first guideline mentioned above, that of minimizing reliance on random draw techniques for carrying out the match. In order to meet the second guideline, it was necessary to devise some method for handling the different weighting systems in the two files. While the weight for each person in the CPS is not the same, it is only indirectly related

to income, if at all. The TM, on the other hand, is stratified by type of return (1040A's, business and nonbusiness 1040's) and by size of adjusted gross income. The solution finally adopted was to split the records in each file so that the weights for two records with the same rank in the particular subclass would be the same. To illustrate, suppose in a given subclass with a weighted frequency of 5,000 (defined, of course, to be the same in both files), there are 2 CPS records with a weight of 2,500 and 5 TM records, each with a weight of 1,000. The matching procedure then created 6 merged records, 4 with a weight of 1,000 each and 2 with a weight of 500 each. A little reflection will show that the number of matched records in a cell by this method is equal to one less than the sum of the unweighted number of CPS and TM records in that cell. For the merged file as a whole the number of unweighted records is approximately equal to the sum of the unweighted records in the two separate files.

While this method is necessary to assure that tabulations of any income type in the matched file will come up to the aggregate for that type implicit in the relevant file before matching, it has the disadvantage of producing a rather unwieldy file in terms of length. We have experimented with other techniques for matching within cells that do not involve splitting records, including various sorts of random draw methods, but they have all produced estimates of particular income aggregates which are hardly within an acceptable range of the corresponding control. For example, a method which picks that IRS record which falls at the midpoint of the CPS record, in terms of the latter's weight in the ranking, when both CPS and IRS records are ranked from high to low, produced a 24 percent underestimate of TM dividend income and a 19 percent underestimate of partnership income.

The third guideline—preserving the CPS relationship among demographic and economic characteristics and income amounts and types—can be effectively followed only insofar as these characteristics are related to the existence and size of the various CPS income types used in the match with the TM, in view of the virtual absence of demographic data in the latter. If nonwhites, for example, generally have lower (but nonzero) property incomes than whites, they should end up with tax returns having lower property incomes than returns assigned to whites. Difficulties arise where a part of the match uses random assignment, or the demographic characteristic is related to an income type not used in the match and not correlated with those that are. In our match, for example, the CPS records without property income in the cell were ranked in random order to be matched with tax returns, some of which contained property income and some of which did not. If there is a lower probability of nonwhites having property income for any given wage and self-employment income than whites, the procedure will assign too many returns with property income to nonwhites, too few to whites. While the OBE match does not seem to have done violence to the relation between race and total money income as computed from the TM, there is no definite assurance that the detail on the income types will be free of this defect. Similarly, the relation between age and money income computed from the tax return follows the original CPS pattern, although this would, of course, be expected for the age group 65 and over, since that was the one age bracket that could be treated separately in the match.

Since the file had been split into many pieces in the match—married couples, other relatives, and unrelated individuals were all treated as separate units—it was necessary to reassemble the file in terms of family units. While this might seem to be a relatively simply matching operation, it was complicated by the fact that split weights for other relatives in the family no longer necessarily corresponded with the differently split weights of the head. One method, to have computed a weighted average for each income type from the various split records for each other relative, would have given biased results; in particular, it would have overstated the number of families with any particular income type. We finally used a procedure similar to that employed in the match—a further splitting of records for heads and family members, with the number of split records for any one family being equal to the sum of split records for head (and spouse, if any) and other relatives, minus one. The resulting file of married couples, other family members, and unrelated individuals contains close to 200,000 unweighted records. The use of random draw techniques for combining other relatives into families would again serve to reduce the length of the file created, and may well be a more feasible method than it is for the original match itself.

After completion of the CPS-TM match, the next step was to correct the TM income types for that part of underreporting on tax returns which would have been eliminated had each return in the file been subject to audit. For this purpose we used the half sample of about 50,000 tax returns from the 1963 audit study of individual returns (Tax Compliance Measurement Program or TCMP). This file contained, for each income type, the amount of income reported by the taxpayer, and the amount as corrected by the auditor. The correction methods developed were applied separately for eight groups of returns; joint returns and all other returns, for under 65 and 65 and over, short and long forms.

The rationale for the method that was finally adopted can be illustrated by asking, if we had had an audit study for 1964, would we have preferred to use it rather than the 1964 TM for matching with the CPS, assuming both files represented the tax return universe for 1964? (Indeed, both the 1963 audit study and the 1964 TM were subsamples of *Statistics of Income* samples for their respective years.) Since our answer to this hypothetical question was in the affirmative, our purpose became one of capturing the relationship between the before and after audit distributions for each income type for 1963 and applying those relationships to the corresponding income type in the TM. The correction of dividend income may be taken as an example. Returns containing dividend income both before and after audit were first ranked from highest to lowest by size of dividend income as reported by the taxpayer and the aggregate amount of dividends reported by each percentile of this distribution was determined. Next, the same returns were *re*ranked from highest to lowest by size of dividend income as corrected by the auditors and the aggregate amount of dividends after audit was computed for the same percentiles. The ratio of aggregate dividends after audit to dividends before audit, computed separately for each percentile, provided the required correction ratios. Returns with dividend income in the TM were then ranked from highest to lowest and grouped into the same percentiles used in the TCMP. The appropriate TCMP correction ratios for dividend

income were then applied to the dividend income reported by the returns in each of these percentile groups.[7]

While the method used gave us the appropriate after audit distribution for each income type, there was no assurance that it would give us the same after audit distribution of total adjusted gross income or AGI (exclusive of capital gains) contained in the TCMP. To test this possibility we applied our percentile correction ratios to each income type in the TCMP as reported by the taxpayer, added up the "corrected" income types for each return to obtain a "corrected" AGI for that return, and compared the size distribution of such "corrected" AGI's with the size distribution of AGI's in the TCMP as actually corrected by the auditors. The two size distributions were virtually identical, indicating that correcting each income type separately without regard to the corrections the auditors made in other income types on the same return did not distort the final corrected distribution of AGI.

Note that no attempt was made to select a return from the TM, determine the return in the TCMP most like the one selected from the TM, and apply the correction ratios obtained from that return to the return initially selected from the TM. Such a method would have been exceedingly difficult to carry out, with no particular advantages over the method actually used.

This description does not cover the complications introduced by the audit correction of negative incomes in business, partnership, farm, rent, and royalty returns, nor does it deal with the problems created by auditors' reducing to zero income types reported by certain taxpayers, and finding nonzero amounts of certain types for which the taxpayer originally reported "none." The solutions we adopted to these problems involved complications that will not be pursued further in this paper.

The last merging operation in our methodology was the combination of our matched CPS-TM file with the Survey of Financial Characteristics of Consumers (SFCC) file, a sample of 2,557 consumer units stratified by income. While the purpose of the CPS-TM match was to correct and adjust CPS income types, this match was designed to provide information by means of which we could distribute among consumer units income types not covered in our two basic files, the CPS and the TM. Our primary concern was with the following information in the SFCC: home ownership and equity in owned home (to allocate imputed rent on owner-occupied dwellings); checking and savings accounts (imputed interest on such accounts); U.S. savings bonds (accrued interest on such bonds); holdings of and interest on state and local bonds; and life insurance data (imputed interest on life insurance equity).

In contrast with the TM there is a considerable amount of information in the SFCC besides income on which a match can be carried out, and our problem was to determine which were the most appropriate characteristics to use. Because of the very small sample size in the lower income brackets in the SFCC—1 to

[7]In an attempt to keep this account abbreviated, I have had to simplify the methods employed at various steps. For example, it was necessary for some income types to smooth the correction ratios. This was done by combining percentiles into quantile groups of various sizes. In addition, the correction ratios turned out to be so small (i.e., so close to one) for wage and salary income that the latter income type was left as reported in the TM.

43,155 in the under $3,000 bracket, for example—mere multiplication of characteristics used for matching would have produced many empty cells or cells with only one or two SFCC records and would have required the consolidation of many cells by hand. A similar problem exists at the top, although in this case it is the CPS that is short of records relative to the SFCC when SFCC income size brackets are used.

We therefore concentrated on those characteristics which appear to be most relevant to home and liquid asset ownership: dollar income level (using the SFCC's definition of its income strata for income size brackets); type of consumer unit (family or unrelated individual); age (6 age groups); race (white, nonwhite); and major source of earnings, which was used only for families (wage; farm;[8] nonfarm self-employment; nonworker). Space is lacking to defend fully these choices for the cells within which matching took place. Data from the Survey of Consumer Finances convinced us that the rise in home ownership in the postwar period can largely be accounted for by the rise in real income levels. Use of a ranking technique in terms of income, as employed in the CPS-TM match, would therefore not have been appropriate, since the SFCC data were for 1962 rather than 1964. The reasons for using age and race are obvious. For major source of earnings, self-employed, at any given income level, have larger asset holdings than wage and salary workers, and the latter than nonworkers, except, perhaps, for the age 65 and over group. Even with this limited breakdown we still had 540 cells, many of which were empty. Hand consolidation of cells was therefore necessary where there were CPS records but no SFCC records in a particular cell. One of the major sufferers from this consolidation was the white-nonwhite breakdown, since there were so few nonwhite records in the SFCC file.

The next problem was the matching of records within cells. In contrast to the way we set up the CPS-TM match, the weighted number of records in any one cell from the two files did not have to be the same; it was therefore necessary to inflate (or deflate) the weights of the SFCC records in a cell to make the weighted counts in the two files the same. Within a cell, records were ranked from high to low by size of interest income; for those without such income, the ranking was in random order. The matching method used did not involve a further splitting of records, as we had done in the other match; instead, we picked that SFCC record which fell at a "selection point" defined to be one third of the way down the CPS record in terms of the latter's weight (rather than at the midpoint of the CPS record), when records in both files are ranked in terms of the classifying variables (interest income or random order in this case).

To illustrate the procedure, suppose in a given cell there are 5 CPS-TM records, each with a weight of 1,000, and 2 SFCC records, record A (with the larger interest income) with a weight of 1,200, and record B with a weight of 1,300. The first step is to multiply the weights in the SFCC records by 2, so that

[8]Because of the treatment of farm home ownership in the SFCC (a farmer's home is considered as part of his farm assets), it was necessary to include all farm operators (occupational code) in this category, irrespective of whether farm income was their major source of earnings.

the frequencies in the cell are the same (5,000). The records are then ranked from high to low (or by random order, if certain of the CPS-TM records have no interest income), cumulated, and the "selection point" for the CPS records computed. These operations are illustrated below:

CPS Record	CPS Weight	Cumulated Weight	"Selection Point"	SFCC Record	SFCC Weight	Cumulated Weight
1	1,000	1,000	333	A	2,400	2,400
2	1,000	2,000	1,333			
3	1,000	3,000	2,333	B	2,600	5,000
4	1,000	4,000	3,333			
5	1,000	5,000	4,333			

As is now evident, the first 3 CPS-TM records will be matched with SFCC record A; the last 2, with record B. Had the midpoints of the CPS records been used, the 3d CPS-TM record would have been matched with B instead of with A, since its selection point would have been changed to 2,500. On the other hand, had we matched by splitting records, the 3d CPS record would have been split into two parts, one with a weight of 400 and matched with A, the other with a weight of 600 and matched with B, for a total of 6 matched records ($5 + 2 - 1$).

The major advantage of the method we used, in addition to being simpler to carry out, is that the length of our merged file was not increased. It should be noted, in addition, that the method in effect reweights the SFCC file, with the result that tabulating SFCC income or asset types in our merged file will not produce the same totals when those types are tabulated by weights in the SFCC file. Part of the difference, of course, was the result of the difference in income levels between 1962 and 1964, a difference our matching procedure was specifically designed to take account of. The failure to come up to the actual totals for 1964 in all cases was not, however, serious for our purposes, since the particular SFCC data merged into the file were used for allocating by size the control totals for certain income types obtained from OBE's family personal income series. Furthermore, the discrepancies can be altered if desired by varying the "selection point" used in the match, although it is virtually impossible to come up or down to every SFCC aggregate by varying that point alone.

The final step in the project involves the raising of the money income types taken from the CPS-TM file to their corresponding control totals and the assignment of both imputed income types and those money income types missing from that merged file. For those income types appearing in both the CPS and the IRS data, with few exceptions we used the tax return amount. The most important exception was farm income, which is substantially underreported in the tax data, being less than two thirds even of the amount reported in the CPS, itself subject to significant underreporting. Much of this difference is attributable to the proportion reporting a loss; a quarter of those reporting farm income on tax returns had a loss, almost double the proportion showing a loss in the CPS.

For most of the income types taken from the CPS-TM file a single ratio

technique was used for blowing up each income type to control. In the case of nonnegative income types (e.g., interest and dividends) the ratio of the OBE control total to the actual CPS or TM amount was applied. For types where loss incomes were involved (primarily self-employment income, rent, and royalties) a quadratic equation was solved for the ratio to be applied to positive amounts; the reciprocal of that ratio was applied to negative amounts. The method used assumes that losses were overreported in about the same proportion as gains were underreported; to leave the Lorenz curve for that type unchanged after blowup, the size of losses would have to be increased by the same proportion as gains, hardly a plausible procedure.

There were some exceptions to this ratio technique. For rental income, in order to reduce the percent with a loss from the 30 percent shown by tax returns after audit to the approximately 10 per cent shown by field surveys, such as the CES, SFCC, and SEO, a constant dollar amount was added to each record after using simple blowup, the sum of the two adjustments equalling the difference between the actual and the control amount. The CPS amounts of unemployment compensation and public assistance payments were brought up to control by increasing the number of recipients rather than increasing the amounts for those actually reporting in the CPS, since the number reporting receiving these types was substantially less than the number who should have reported them. The additional recipients were drawn at random and assigned mean amounts within narrowly defined cells, using such characteristics as weeks looking for work, sex, race, age, family type and size, income, and region. (Not all of these characteristics were employed for each of these transfer income types.)

Aside from imputed income, items missing from the CPS-TM file were estate and trust income, state and local bond interest, accrued interest on U.S. savings bonds, and personal contributions for social insurance. The first three mentioned types were drawn from the SFCC part of the file and adjusted to control totals by the simple ratio technique described above. Personal contributions (mostly OASDI and retirement contributions of government employees) were for the most part estimated directly from information on size and type of earnings and on class of worker (government or private) contained in the individual record.

As noted earlier, imputed income was allocated largely by information from the SFCC: imputed rent on nonfarm dwellings was distributed by size of equity in owned homes; imputed interest on checking and savings deposits, by size of such accounts (including accounts in saving and loan associations and credit unions); imputed interest on life insurance equity, by the face value of life insurance, reduced to cash surrender values by ratios reflecting the age of the family head. Farm dwelling rent and food consumed on farms, on the other hand, were distributed on the basis of data drawn from BLS's Consumer Expenditure Survey; in addition to farm residence, we took account of family size in allocating food, and of income and type of family in allocating rent. Occupation, work experience, and size of wage income were used for distributing the small amount of imputed wages in the national accounts.

This admittedly brief discussion of the final steps in the estimation of OBE's size distribution for 1964 must suffice. The major purpose of this paper,

after all, has not been to describe the complete methodology underlying the OBE estimates, but to explain the use we have made of and techniques involved in merging microdata files to correct and supplement income estimates in the original field survey (CPS) and to incorporate additional information that can be used to estimate items and types, such as imputed income, not a part of the original file. While exact matching of microdata records may be the ideal method for creating certain merged files, the day when such files will be available for general research use appears to be so far off that statistical matching of files to improve the quality of data and to extend the information found in field surveys needs to be given more attention than it has received.