Volume Title: Studies in Income and Wealth

Volume Author/Editor: Conference on Research in Income and Wealth

Volume Publisher: NBER

Volume ISBN: 0-870-14168-6

Volume URL: http://www.nber.org/books/unkn51-1

Publication Date: 1951

Chapter Title: An Income Size Distribution from Income Tax and Survey Data, 1944

Chapter Author: Maurice Liebenberg, Hyman Kaitz

Chapter URL: http://www.nber.org/chapters/c5728

Chapter pages in book: (p. 378 - 462)

Part VII

An Income Size

The Interagency Technical Committee On Income Distribution, made up of representatives from the Office of Business Economics, Bureaus of Agriculture Economics, of Labor Statistics, of the Census, and the Federal Reserve Board, chaired by the Budget Bureau, procured many of the tabulations indispensable for constructing the estimates in this report.

Distribution from Income Tax and Survey Data, 1944

Maurice Liebenberg
and
Hyman Kaitz
NATIONAL INCOME DIVISION, DEPARTMENT OF COMMERCE

As noted in other papers, e.g., Part VI, size distributions from sample field surveys are deficient in many respects and account for only a portion of total consumer income as independently estimated. Bureau of Internal Revenue tax returns are likewise inadequate; since they are classified by adjusted gross income and since the tax return is the basic unit they do not constitute size distributions that are meaningful for some economic analyses. These two sets of data, the main sources from which size distributions can be constructed, can be used singly or to complement each other. Apart from the difficulties of converting to common units and income concept, complementary use is preferable because BIR data can be assumed to be most reliable at the

higher income levels, precisely where surveys are most inadequate. Conversely, because of legal filing requirements tax data are deficient at low levels of income where surveys are more satisfactory.

The purpose of this paper is to discuss briefly some methods of adjusting income size distributions and to present a size distribution for 1944 based upon the joint use of tax returns and field surveys. As pointed out below, the distribution, though superior to either source in accounting for aggregate income, must be considered as merely a first approximation to the desired distribution.

A Methods of Deriving Adjusted Income Size Distributions

The primary reason for adjustment is that size distributions of income from surveys or income tax returns do not completely account for aggregate income as estimated by methods which are possible when size distributions are not an objective. Broadly speaking, this deficiency can be removed or mitigated by three main lines of approach:[1] (1) simple transformations of given distributions—methods that assume either a constant Lorenz curve or some specific change in the relative distribution based on a minimum of usable information; (2) the use of income tax data to supplement survey results, primarily to achieve more adequate reporting at high income levels; and (3) the segregation, separate adjustment, and eventual recombination of relevant component groups of reporting units which require and are susceptible to separate treatment, e.g., recipients of wages and entrepreneurial income.

The simple transformation based on the assumption of a constant Lorenz curve has been fully treated in the literature.[2] Trans-

[1] As noted below, these three lines of approach are not mutually exclusive and can be used, to some extent, to supplement one another. Yet they are distinct in the method that is given primary emphasis in dealing with the problem of missing income.

[2] Edward Ames, 'A Method for Estimating the Size Distribution of a Given Aggregate Income', *Review of Economic Statistics*, XXIV (1942); David Durand, 'A Simple Method for Estimating the Size Distribution of a Given Aggregate Income', ibid., XXV (1943); and 'An Appraisal of the Errors Involved in Estimating the Size Distribution of a Given Aggregate Income', *Review of Economics and Statistics*, XXX (1948).

formations involving changes in the Lorenz curve were used in the estimates of 1941–42 incomes by the Office of Price Administration; see Neal Potter and David Rosenblatt, 'Method of Estimating the Distribution of Civilian Money Income in 1942', *Studies in Income and Wealth, Volume Eight.* The second method is exemplified by the National Resources Committee study for 1935–36; see *Consumer Incomes in the United States* (Washington, D.C., 1938). Estimates have not yet been based on the third method but some experimental work has been done in the Department of Commerce in segregating earner groups, the first step. The estimates in Section C are best classified under the second method although transformations are also used. Before entering upon an account of the methodology, the three approaches are briefly explored.

1 *Simple Transformations*

a *Constant Lorenz curve*

Transformations that assume a constant Lorenz curve have the great merit of simplicity.[8] Moreover, they can be used when lack of data preclude more refined adjustment. No recourse has to be made to any data other than the distribution requiring adjustment and the aggregate to be achieved. Since the assumption implies that each income be multiplied by a constant, the problem is merely one of changing interval limits and interpolating for the desired original limits. Little else can be said for this method. And if other usable data and sufficient resources are available it is open to serious objections.

First, consider the very magnitude of the adjustment required. In 1944 the discrepancy between Census survey results and an estimate of aggregate income was about $30 billion, almost a fourth of total income. To maintain the same Lorenz curve while

[8] The adjustment of distributions to achieve any given aggregate under the condition that the relative distribution be maintained can be accomplished in many ways (see the articles by Ames and Durand). The method suggested by Ames depends upon known properties of the Lorenz curve. His 'derivative' curve depends upon the property that the ratio of any given income to the average of the distribution is equal to the derivative of the Lorenz curve evaluated at the point of the given income. A method based upon such properties is not necessarily superior. Though it has a certain elegance, it is inferior in practice to some interpolation formulas that can be applied directly to the frequency distribution itself.

achieving the desired aggregate puts undue stress on the assumption of equal percentage underreporting of income.[4] Moreover, the method does not allow for known inadequacies of the initial relative distribution, e.g., survey understatement of incomes at the higher levels. This bias suggests prior correction which militates against the direct application of the method.

Furthermore, component sources of income, such as wages and salaries, dividends and interest, are subject to varying degrees of underreporting (see Part VI). Since such components account for unequal proportions of total income at various income levels, correction for missing income may affect different income levels unequally.

Arguments in favor of the constant Lorenz curve assumption that point out the approximate stability of the relative distribution over time are often specious. Of course, only small changes in the distributions of survey data between years not far apart or in which economic conditions do not change markedly are noticeable, as can readily be seen by examining the Lorenz curves for survey results for 1944–46 (ibid.). However, near coincidence of Lorenz curves over time is perfectly consistent with marked

[4] Although the application of constant multipliers to every income assures the maintenance of the relative distribution, identical Lorenz curves do not necessarily imply such multipliers. Only when the additional condition of maintaining rank is imposed can the latter be assumed. Thus, such multipliers may be regarded simply as a device to achieve the same relative distribution. The maintenance of Lorenz curves can be regarded as a special case of the general transformation $y = h(x)$ in the expression

$$\int_0^x f(x)dx = \int_0^{y=h(x)} g(y)dy$$

where $f(x)$ and $g(y)$ are the two density functions before and after transformation and $h(x) = ax$. This transformation, of course, does not imply that every $y_i = ax_i$. The more general interpretation, however, involves a certain methodological obscurity which can be justified only by empirical verification of the final distribution or by empirical evidence of the constancy of the Lorenz curve of income before and after adjustment. (See note 8 for a similar interpretation that can be made with respect to the source pattern method.) But until a sufficient basis for the constancy of the relative distribution is obtained, and as a necessary step in such verification, we must concern ourselves with known *deficiencies* in the basic data and with the *direct* implications of procedures designed to correct them. It is in this latter context that a constant Lorenz curve implies the assumption of equal percentage underreporting of income.

differences between an adjusted and unadjusted distribution and is therefore not a valid argument for adjusting distributions by means of the constant Lorenz curve assumption. The stability noted over time may reflect simply a stability of bias. Conversely, comparisons of income inequality based upon unadjusted distributions are themselves dangerous since apparent differences may not be due to changes in the underlying relative distribution but may reflect changes in reporting bias.

The decision to embark upon a more elaborate adjustment must depend, of course, upon the purposes for which the distributions are intended. Whether slight changes in the relative distribution would significantly modify many conclusions based upon them must be considered since adjustments other than those based on the simplest assumptions require rather elaborate and time-consuming computations. Frequently, however, the particular use of the data, e.g., inequality or welfare comparisons, makes a careful assignment of missing income necessary if inferences are to be valid.[5]

b *Source pattern transformation*

Another transformation by which a size distribution can be adjusted to achieve a desired aggregate allows for 'source patterns', that is, it takes into account differences in the shares of the major types of income at the various levels of total income. Consisting essentially in adjusting a relative distribution by assigning the missing income of each type in proportion to the reported amounts at each level of total income, its chief merit is that ad-

[5] Many comparisons of income inequality are too crude to be worth while. The question what constitutes a small or insignificant change is too often determined by a visual impression of plotted Lorenz curves or by uncritical examination of coefficients of inequality. Because of the insensitivity of the Lorenz curve, changes in distribution must be marked before differences between two curves become apparent. The associated coefficient of concentration is similarly insensitive; moreover, it hides even obvious changes in portions of the distribution. The use of the entire area under the line of equal distribution as the basis for an index of inequality is bound to throw into seeming insignificance all except tremendous changes in the relative distribution. The needed change in perspective is acquired by relating such measures to a norm established by distributions for many years. Needless to say, changes in the relative distribution as revealed by unadjusted sample data should be appraised as significant in the statistical sense, i.e., in the light of sampling fluctuations.

ditional data on underreported income in each major category are used separately instead of total underreported income alone. Although it assigns missing income from each source proportionately to the reported amounts, it changes the over-all relative distribution because it allows explicitly for differences in reporting among the various income shares as well as in the amounts from each source at various levels of total income. Another merit is that it is only somewhat more complex than maintaining the Lorenz curve of total income, and at least in projections over considerable periods probably yields superior results.

Devised initially for projecting income size distributions over time, the method yielded rather good results in the case of the Office of Price Administration estimates.[6] Formally, of course, adjusting distributions for any one period to achieve a desired aggregate can be considered as a problem in projection, since in both a given distribution must be transformed into another which will yield the correct aggregate income and at the same time reflect the causes that may account for a shift in the relative distribution.

In practice, the method consists, briefly, of 7 steps: (1) dividing the initial distribution into segments, say, deciles, for which associated income source patterns are available; (2) determining, for each segment, the percentage distribution of income by source; (3) multiplying the percentage for each source in the various segments by the ratio of the total amount from that source, as estimated independently, to the total amount in the initial distribution; (4) adding the adjusted percentages within each segment; (5) adjusting their sums, provisional multipliers, proportionately so that when applied to total income in each segment from the initial distribution they will yield the desired total income for the entire distribution; (6) connecting these adjusted multipliers with a continuous curve to permit their

[6] The source pattern method as outlined here is a specific procedure used in the projections by the Office of Price Administration and is not meant to refer to other methods which may utilize source pattern information. For a full outline of methods used in constructing the OPA estimates see Potter and Rosenblatt, op. cit.

application to the segment limits; and (7) applying the multipliers and interpolating for the original interval limits.[7]

The source pattern method is based upon assumptions only slightly more acceptable than those underlying the method that maintains the relative distribution of total income. The reasons given above for the failure of the constant Lorenz curve technique apply here with only slightly less cogency. There are, for example, good reasons why an addition required in wages and salaries should not be made proportionate to reported wages and salaries. In the case of income tax data, almost all the missing wages and salaries can be assumed to be received by the lowest income levels. For missing wages of recipients not included in the distribution who are assignable to a specific portion of the distribution the method must lead to error. Even when all recipients of a particular type of income are considered to be present in the distribution the method leads to similar error if underreporting is not proportionate at all levels of the distribution (see Chart 4 for results of applying the method to the 1944 Census survey distribution).

Another assumption underlying the method, the maintenance of rank, is not valid. Every unit classified at a specific total income level is assumed to have the associated source pattern whether it has or not. This assumption, together with the continuous multiplier curve which gives the increases in income for

[7] In an unpublished paper, 'On Some Structural Properties of Distributions of Income', David Rosenblatt explores fully the nature of the source pattern transformation together with its underlying assumptions. The approach, strictly formal, utilizes matrix algebra to represent the transformation in symbolic form. Thus the provisional multipliers m_i for any i-th segment can be given by

$$\begin{pmatrix} m_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ m_k \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1r} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{k1} & & a_{kr} \end{pmatrix} \begin{pmatrix} T'_1/T_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ T'_r/T_r \end{pmatrix}.$$

where (a_{ij}) is the source pattern matrix for k segments and r sources, T_i the total amount of the i-th source in the entire initial distribution, and T_i' the corresponding amount of the source desired.

each level of total income, maintains rank.[8] Thus, the basic problem of changes in rank is not dealt with under the source pattern method. The need for such changes will be further discussed below; it suffices to say here that full adjustment requires methods that permit units to change their relative positions in a distribution if missing incomes are to be properly assigned. In practice, of course, income cannot be assigned to individuals but may be approximated by segregating groups requiring different treatment. The essence of the source pattern method is that *all* units are shifted to a degree dictated by the average composition of income at their level.

In summary, the source pattern method is not only relatively easy to apply but also has merit because it gives some weight to the varying degrees of completeness with which income sources are reported. However, it is in error in relying upon an assignment of missing sources proportionate to amounts reported and in failing to effect needed changes in rank that should follow upon full adjustment.[9]

[8] It is the continuity noted in *average* source patterns that permits an assumption of continuity in constructing the multiplier curve. If the segments could be increased arbitrarily so that they equaled the number of units in the population, such patterns would not be averaged and the patterns noted would be discontinuous, as expected. Application of the source pattern method under such conditions of discontinuity would necessarily involve changes in rank. Even apart from this extreme assumption the final results of adjustment are not independent of the size of the segment chosen.

If the data are in appropriate form, it is quite feasible to adjust individual sample schedules for a proportionate increase in each type of income reported. Tabulations based upon the adjusted total incomes would then achieve the desired changes in rank (Rosenblatt, op. cit.). The method would be in error to the extent that proportional adjustments are inapplicable, not to mention possible difficulties in working with large samples.

As in the case of the simple transformation that maintains the same Lorenz curve (see note 4) the varying multipliers of the source pattern method can be interpreted as not involving any assumptions concerning the maintenance of rank, i.e., the multipliers are regarded merely as operators that achieve the desired final distribution. Except for transformations based upon known mathematical equivalence, however, it is impossible, given this interpretation, to predetermine the probable effectiveness of any method to bring about a desired result. In every case, resort must be had to empirical verification of the final product since the usual criteria of reasonableness and consistency cannot be applied to the method as a progression of meaningful steps. In addition, the source pattern method is especially sensitive to such decisional matters as the size of segment on which the multipliers are based, and, as just said, it is impossible to determine the degree of segmentation that will yield the most accurate results.

[9] By applying the source pattern method to subgroups, rank can be changed.

2 *Supplementing Survey Results by Means of Tax Data*

The second procedure, the one adopted in constructing our estimates, relies upon supplementing survey data by tax data to get at missing income. If, e.g., survey results are deficient at the higher income levels, a distribution 'tail' obtained by adjusting tax data is substituted.[10] This method is superior to those discussed so far because it utilizes additional reliable data and necessitates fewer crude assumptions. Since it best describes the procedures followed in constructing the estimates in Section C details are postponed. Here it is described in broad outline and some of its deficiencies mentioned.

a *Reasons for partial results*

The chief limitation is that the adjustment is only partial, i.e., supplementing survey results by tax data will not necessarily account for all the income independently estimated. In the ideal case it would of course. If survey results were complete below the $3,000 level, say, and tax data equally complete above that level all would be well if the two sources were spliced at that level. In practice, however, this is far from the case and the final total differs considerably from the desired aggregate.

The investigator has several choices when confronted with the still inadequate results of combined survey and tax data. The final adjustment to achieve the desired aggregate can be made by regarding the tax data, modified on the basis of audit information, as correct and throwing the burden of additional assessment on the income levels taken directly from the survey. However, audits probably do not bring to light large amounts of unreported income.

Thus, units receiving only wages or only wages and nonfarm business income can be isolated and the method applied. The difficulty is that even within any subgroup rank positions are maintained by the arbitrary proportional assignment of the missing sources. There can be no doubt, however, that such subdivision would be a net improvement since changes in rank are permitted at least between subgroups. When carried to its logical extreme this modified approach is an improvement in that it avoids the assumption that all units at a specific level of total income take on the source pattern at that level. When such subdivision is made and empirical data are used to correct the component distributions, the method is similar to method (3) outlined below.

[10] This method was used in the National Resources Committee study of 1935–36 incomes.

One alternative is to regard survey data as the more reliable and adjust the tax data to get at the missing income. This decision must rest ultimately upon confidence in the completeness of a survey and interview techniques.

Another alternative, to correct the entire distribution for the missing income, is perhaps the most attractive but confronts us with the nature of the final correction. We are faced again with the choice of resorting to transformations such as those mentioned above or possibly to other independent data. With the exception of the transformation that maintains the relative distribution of total income, which can always be made no matter what the starting point, adjustments are exceedingly difficult at this stage. Thus, if the investigator tries the source pattern approach, he will find that suitable patterns are not readily available for income levels after the rather elaborate adjustment of returns to achieve the desired economic unit and appropriate income concept.

In view of these difficulties the partial result may well be left unaltered. It is definitely superior to unadjusted survey data since at least one obvious area of deficiency, the upper income levels, was considered. In our estimates, for example, survey results yielded only $111 billion of the desired $140 billion. Supplemented by tax data, a total of $125 billion was achieved including approximately 48 percent of the originally missing income. Needless to say, omission of the final adjustment means that only a first approximation to the desired distribution has been attained. How close the approximation is, and whether the results, despite their incompleteness, can be useful, depend of course upon the amount and distribution of the missing income still to be accounted for.

b *Changing the unit of classification and the income concept*

The method that incorporates income tax data assumes that the data can be adjusted adequately to effect necessary changes in income concept and unit—the major task in constructing the estimates.

As will be seen in Section C, tax returns require rather elabo-

rate adjustment before they can be used to supplement survey data. Selma Fine and Enid Baird mention the need for such adjustment and give an example of procedure in 'The Use of Income Tax Data in the National Resources Committee Estimate of the Distribution of Income by Size', *Studies in Income and Wealth, Volume Three.* The problems encountered in constructing a distribution for 1944, however, are somewhat different from those found in using 1935–36 data. Our task was facilitated by a more appropriate income concept for classifying 1944 tax returns, by additional data from *Statistics of Income Supplement Compiled from Income Tax Returns for 1936* on matching incomes of husbands and wives, and by more information on capital gains and losses.

The tabulation of tax distributions by size of adjusted gross rather than net income is, of course, a substantial gain and obviates the need for any adjustment for deductions. But the adjusted gross income concept, though far better than net income, still requires adjustment to remove such items as net taxable capital gains and losses, and to add tax exempt interest, military income, etc. before the survey income concept can be approximated. With respect to these additional adjustments the task has been made more manageable only in the case of capital gains and losses where new information has become available. Military income adds both a new problem and a relatively large amount of income not covered on tax returns.

The 1936 cross-tabulation of husbands and wives filing separate returns has added valuable information for converting income tax returns into family units. But it contributes to only one part of the problem of combining tax returns into families.[11]

[11] The use of the family as the desired unit depends, of course, upon the ultimate use of the distributions. Because of intrafamily economic relationships the usual sociological concept also defines a family unit in an economic sense. But the family is merely one of many possible classifications and from some viewpoints unsatisfactory; for instance, welfare comparisons cannot easily be made. The units comprising a distribution are not identical in all relevant respects other than income, the criterion for classification. In 'Resource Distribution Patterns and the Classification of Families', *Studies in Income and Wealth, Volume Ten,* William Vickrey attempts to remove some of these differences by introducing the family size and adult-child composition of its members. Needless to say, the unit chosen can materially affect its distribution by size of income (see Part 1).

The various types of return tabulated by the Bureau of Internal Revenue for 1944, for example, are: (1) joint returns of husbands and wives; (2) separate returns of husbands and of wives (husbands and wives with community property income tabulated separately); and (3) returns of single persons and married persons not living with spouse. This classification is far indeed from the desired family unit. Type (3) requires division into its component groups: (a) single individuals (one-person families); (b) heads of broken families; and (c) family supplementary income recipients other than spouse. Moreover, all these types of return except 3a and 3c must be further divided into 'main' and 'subfamilies'. The Census Bureau definition of family includes members of subfamilies whereas subfamily members may file their own tax returns. After separation all these types of return must be combined into the desired family units. The steps are many and because information is lacking entail rather complex and not too satisfactory procedures. The techniques chosen must be appraised entirely in the light of possible manipulative error; see Section B for some examples of error that may be introduced.

In summary, despite the difficulties, the appeal of this method lies mainly in its use of reliable supplementary data. The assumptions required are many but involve primarily questions of combination and manipulation as opposed to those of a more sweeping character underlying methods that maintain the relative distribution.[12] Even if procedure materially affects the distribution it is reasonable to believe that the error will prove

[12] Another technique used to supplement survey material in the absence of actual conversion of tax data is the Pareto extrapolation in the 1941 BLS-BHNHE survey. If the more elaborate adjustment cannot be undertaken, this method can be recommended. It was especially heartening to find that the results of the extrapolation used in that survey closely approximated those obtained by the more elaborate adjustments made to 1941 tax data by Mr. Pechman (see Part IV). The method must remain somewhat precarious, however, and cannot be offered as an adequate substitute for the longer task, since despite the enormously useful fact that the Pareto function fits the upper levels of an income distribution rather well, the slope of the line is sensitive to the character of the survey distribution at the point of joining, and excellent results in any one year may prove mainly fortuitous.

small if the adjustments are carried through in sufficient detail and the scope of the assumptions is limited.

3 *Method Involving the Segregation of Earner Groups*

This method can be said to be an attempt to improve the assignment of missing income by proper segregation of the groups receiving it. Correct assignment to the composite of heterogeneous units in the total family distribution is difficult even if distributions of missing income by type are known; knowledge of missing income of a particular kind, or even of missing income recipients, is not directly translatable into procedures that make proper assignment possible. The reasons are, of course, the complexity of income interrelations and the multi-income recipients in the distribution. The purpose of segregating relevant groups is simply to render the assignment of missing income more precise. The problem of adjustment becomes essentially one of constructing more adequate component distributions from independent sources. Upon recombination into family units these corrected component distributions will affect all levels of total family income in a manner consistent with the nature of the corrections.

The component distributions required must be determined by the kind of income for which correction is to be made. In addition, the type of income receipt to be corrected for must itself be the criterion of classification. If we segregate the group receiving wages but the data are tabulated by total income it is doubtful whether wages can be adequately corrected for.

Distributions by size of source are always easy to construct from survey material. Difficulty arises when more than one source is to be corrected for. Certain units are members of more than one group as defined by income from a specific source because they contain multi-income recipients. Their position in the final distribution is determined as much by these associated incomes as by the source used to classify them. Correction of any or all component distributions by the size of each source is insufficient in itself to permit correct assignment of units by their combined sources. This can be done only if the interrelations of

each type of income with other types are made explicit and used to effect the required combination. Cross-tabulations that make explicit the interrelations among types of receipt for units having more than one type are needed if a method involving the segregation of earner groups is to be used to adjust distributions.

For units receiving income from only two sources the cross-tabulations required are the familiar 2-dimensional ones relating two variables. Receipts from more than two sources suggest multi-dimensional cross-tabulations, which are too formidable to be a workable tool for size distribution adjustment.[13] The necessary simplification can be achieved mainly by redefining the unit and by grouping some sources. If the individual and not the family is the unit, the number of recipients of multi-source income, especially from earnings, is greatly reduced. Classification of individuals instead of families has the additional advantage that most empirical data that can be used to correct component distributions pertain to individuals; e.g., OASI information on wage earners is for individuals. Furthermore, BIR tabulations of returns for years prior to 1948 can be converted more easily into individual than into family distributions. Forthcoming BIR tabulations of unincorporated business returns will permit similar estimates of entrepreneurial income. In fact, most of the data by which size distributions are corrected can be utilized more directly when component distributions are for individuals and this method, which requires an arduous division and recombination of family units, is considered only because these additional sources of information are available.

The combination of several sources into a single group is another simplification. Since the main sources are earnings it is proposed that they be kept distinct and the grouping confined to nonearnings items. The loss of accuracy by such a grouping is, in the main, illusory since adequate correction of certain nonearn-

13 The multi-dimensional case can be approximated by a series of cross-tabulations; e.g., for recipients who receive wages, entrepreneurial income, and interest, wages can first be related to entrepreneurial income, then wages plus entrepreneurial income to interest. This 'ladder' approach has promise but is not treated here partly because it is too complex and partly because, given present sources, many of the marginal distributions cannot be adequately corrected.

ings distributions by their own size is exceedingly difficult.[14]

The method proposed, therefore, consists of four steps: (a) deriving from survey material distributions of individuals for each of the three main earnings sources by its own size (wages, farm, and nonfarm entrepreneurial income); (b) cross-tabulating these sources for multi-source earners;[15] (c) correcting the several earnings distributions by their own size by using independent data whenever possible; (d) using the cross-tabulations, appropriately modified, to combine incomes for these recipients.

The result, assuming adequate adjustment of the size distributions by size of earnings source, will be a corrected earnings distribution for individuals; these earners must be combined into families and the income from nonearnings sources added.[16] Both steps can be accomplished by two more cross-tabulations both of which can be made from survey data: individual earnings by family earnings and total family earnings by total family income.

The method cannot be expected to be without difficulties. Some are inherent in the method itself. For example, the division of units into component groups necessitates recombination on the basis of a cross-tabulation known to be in error; for two related sources it is believed that this difficulty can, in the main, be overcome. The cross-tabulation of individual by total family earnings is more difficult since it involves the number of earners per family and their combination into family units. The computation of an adequate nonearnings margin presents another difficulty. Nevertheless, earnings constitute such a large proportion

[14] The problem of estimating the size distributions of some items included in 'income other than earnings' will perhaps remain an impediment to a full adjustment of income distributions unless new sources of data are discovered. At present it is exceedingly difficult to estimate the number of recipients of some types of income in this group, not to mention the larger problem of estimating their distribution.

[15] Individuals with three earnings sources are not treated here since they are relatively few. For example, the inflated Census survey for 1945 yielded 43.4 million individuals with some civilian earnings (excluding incomplete schedules) and only 13,000 individuals with three civilian earnings sources.

[16] For the sake of simplicity the nonearnings correction is postponed until earners have been combined into family units. Because the data are poor, correct allocation of such income to the several individual earnings groups would be exceedingly difficult and would, moreover, multiply the cross-tabulations necessary.

of total consumer income that distributions for earner groups adequately corrected and combined into family units will be a tremendous advance.

B Major Differences in Methodology Between Present and Earlier Studies

The outstanding previous attempt to supplement survey results by income distributions based on tax returns was the National Resources Committee study under the direction of Hildegarde Kneeland. Subsequent attempts, including our own, followed the broad plan underlying that pioneering study and benefited greatly from it. When it was being carried on, data were even scarcer than they are now. The improvement in basic data has facilitated somewhat the task of constructing our estimates and, at the same time, permitted re-examination of some assumptions it was forced to adopt.

Before describing our procedures in detail we shall discuss the methods on a somewhat more general level, and show how the adjustments differ from those made previously.

1 *Cross-tabulations*

In adjusting income size distributions we are continually concerned with problems of either combining the incomes of separate units or adding or subtracting income from the distributions of certain units. These two types of adjustment are required because the family is frequently a multi-income unit and incomes differ in concept or are deficient in coverage.[17] They are easiest made by cross-tabulations which give explicitly the relations among units or among income components. Such relations are seldom simple; that is, one income characteristic is almost never uniquely associated with another. In the cross-tabulation between husbands' and wives' incomes, for example, at each level of husbands' incomes there is a distribution of wives' incomes covering a wide range. One advantage of cross-tabulations is that

[17] In practice the two types of adjustment reduce to one—adding or subtracting incomes.

they show these conditional distributions explicitly. In adjustments involving relations among more than two characteristics the cross-tabulations revealing the conditional distributions are, of course, multi-dimensional, but in practice only 2-dimensional cases are considered in any single operation.

Though the need for cross-tabulations has long been recognized it has been insufficiently emphasized that adequate adjustment of size distributions requires their frequent and exhaustive use. Generally, units at a particular level of income in a given classification are widely dispersed among the size classes of any other classification. Failure to recognize this fact leads to inadequate adjustment. In the adjustment of income tax data neglect of such cross-tabulations leads to an improper assignment of units by income level.

Some attempts to adjust tax data, the National Resources Committee study of 1935–36 incomes, for instance, have been handicapped by the absence of such tabulations. The dubious expedients adopted when they are lacking will now be briefly examined.

The most common makeshift is to use average relationships, that is, to base the adjustment upon the average correction necessary at given levels of income. In a sense, such methods entail the substitution of a single statistic for the entire conditional distribution required. The underlying relationships are 'condensed', i.e., all dispersion is removed from given rows or columns.

The violence done by ignoring dispersion depends upon many factors. If the data do not reveal marked dispersion, or if the adjustment concerns few income recipients or minor amounts of income, the distortion is not serious. Indeed, in such cases, the adjustment may well be dispensed with altogether. As will be seen below, the adjustment to remove capital gains and losses constitutes, at least for 1944, just such a minor change in the over-all distribution. Since the cross-tabulation between adjusted gross income and capital gains (or losses) shows marked dispersion this relatively minor effect is due entirely to the small number of units having such items and the related fact that only a negligible amount of income is involved. Thus, the effect of the

adjustment is slight on all portions of the distribution except the top levels of income. When detail is required at these levels, however, accurate adjustment is, of course, necessary.

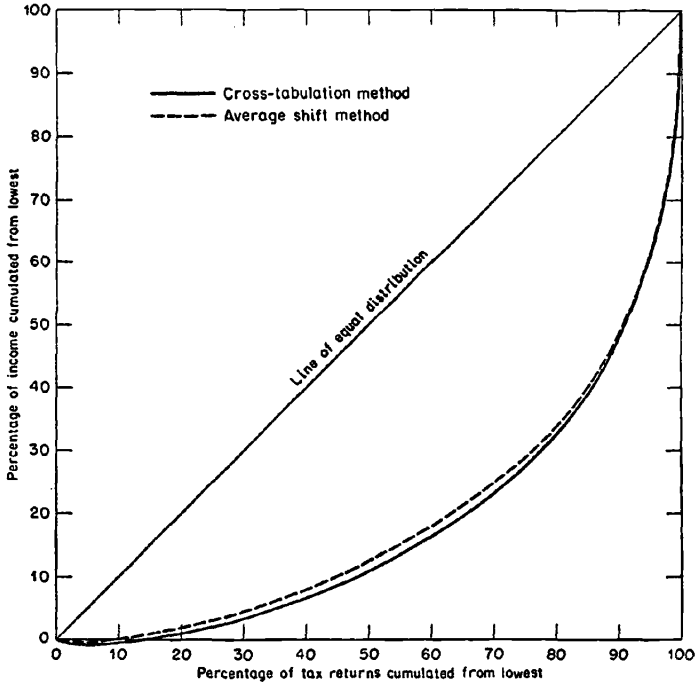
The effect of ignoring dispersion in treating capital gains and losses can best be seen when attention is confined to the units reporting them (Chart 1). One Lorenz curve is based upon adjustment by average relationship, the other upon adjustment by means of a cross-tabulation between total income and capital gains. The marked differences between the two Lorenz curves show how different the distributions are.

Methods that ignore dispersion, though they may cause only slight error in some instances, are inherently faulty, and in other instances may cause substantial distortion. If husbands' and

Chart 1

Lorenz Curves for the Results of Excluding Capital Gains by Various Methods

Returns Reporting Capital Gains



wives' incomes were matched by 'averages' we could be certain that great violence would be done the distribution. The outstanding instance of complete failure, in the authors' experience, was an attempt to construct distributions by size of net entrepreneurial income for various industrial groups when only distributions by size of gross income and the average net income at each level of gross were given. Net-gross income ratios are ordinarily much lower at high than at low levels of gross income, particularly in such industries as retail trade, e.g., the average net income at a high level of gross was as low as 2 percent of gross income. Giving all members of the high gross class such a low net income yielded a distribution with practically no 'tail'. Though this is a rather special case, distortion is always present to some degree whenever the method is used. Instances where this distortion can be substantial are so common that when dispersion is believed to be wide and the total impress large it is better to assign the dispersion arbitrarily on the basis of, say, a constant coefficient of variation for some kind of skew distribution.[18]

The tendency to make what are mainly nominal adjustments is not confined to methods that rely on average corrections. The possible effects of 'averaging' techniques are relatively easy to recognize although they have been minimized or disregarded in practice. Another type of transformation appears attractive at first but may lead to large errors. It relies on data for past years showing distributions both before and after a given combination or transformation. These 'before' and 'after' distributions can be directly related by means of a simple transformation curve which is then applied directly to current data. An example is the matching of the incomes of husbands and wives filing separate returns. The cross-tabulation for 1936 cannot, because of changes in the number of returns filed and marked differences in income as well as in several other factors, be directly applied in 1944. This tabulation, however, is the only basis for current matching of separate returns. The given cross-tabulation can be

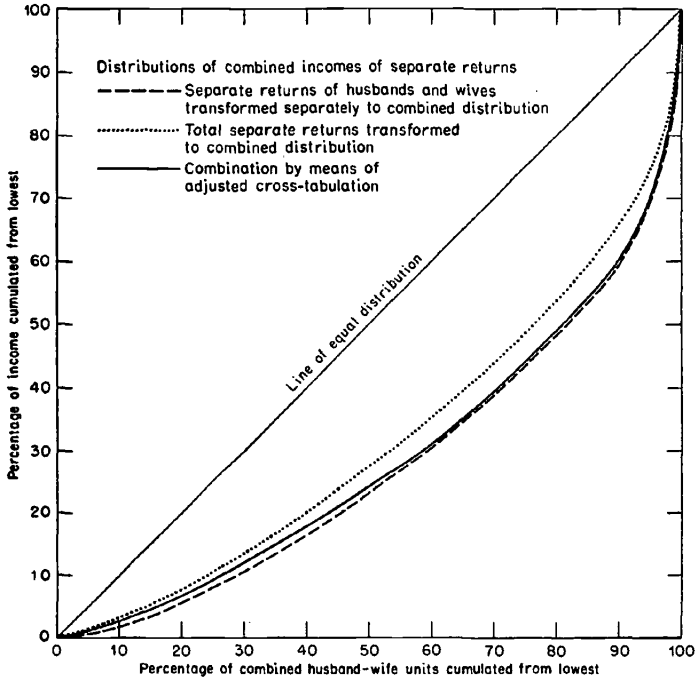
[18] We experimented with various artificially constructed cross-tabulations in connection with our work on entrepreneurial income. The results were promising and, given the opportunity, we shall try to develop them further.

ignored and the income levels in the separate and combined distributions related at points of equal cumulated percentages. The transformation curve derived by this process can then be used in the current situation to transform a distribution of separate returns into a combined distribution.[19]

What is accomplished by such a transformation can be seen from Chart 2. The solid line is the result of a more detailed method using the cross-tabulation explained at some length in

Chart 2

Lorenz Curves for the Results of Matching Separate Returns by Various Methods



[19] The error in using the transformation curve based upon data for past years follows from the assumption that the transformation function is constant over time and independent of the form of the initial distributions. There always exists a transformation $y = f(x)$ which will correctly change one distribution into the other. Adequate transformations of this direct kind are exceedingly attractive in size distribution work but are difficult to obtain. Even if the transformation curve type suggested by experience is taken over, the parameters of the function may remain undetermined.

Section C. The dotted line pertains to the transformation referred to above relating total separate returns to the combined distribution. The dash line is the result of still another transformation of a similar kind. Relating the husbands' and wives' distributions separately to the combined distribution and adjusting the result to the correct combined income, it approximates more closely the distribution obtained by the more detailed procedure. This comparison is not intended to suggest that the solid line is necessarily the true Lorenz curve. Because of the adjustments required to adapt the original cross-tabulation for current use, the more detailed procedure underlying it also involves assumptions that may be questioned. The main purpose is to stress the differences between the curves. All curves in Chart 2 achieve the same aggregate yet the differences in the relative distributions are substantial. It is in this sense that procedures are fundamental, since the changes in the relative distribution in this illustration are actually larger than those between survey distributions in recent years.

It may be argued that this illustration exaggerates the error and that when these results are incorporated in a distribution for all families the total error will not be large. In the illustration, however, the separate return segment is by no means a minor component of the total distribution. Moreover, adjustments of similar nature made to other portions of the distribution may accentuate rather than hide the faulty treatment.

The above illustrations of the use of averages and direct transformations on the basis of past data stress the need for alternative procedures. On each occasion when a similar adjustment was required in constructing our estimates cross-tabulations were used if possible. In explaining the general approach, therefore, it may be said that one important difference between our methods and those used formerly is the emphasis on cross-tabulations.

2 *Supplementing Tax Returns before Combination into Family Units*

Our methods differ substantially in another way from former procedures for adjusting tax distributions. The 1944 tax dis-

tributions are complete to an unprecedented degree. The low filing requirements together with the withholding provision and the high level of income brought a tremendous number of taxpayers into the distribution. It is no longer true, therefore, that adjustments to tax data serve merely to augment the survey distribution by providing substitute frequencies for the upper income levels; it is now possible, assuming certain problems in adjusting for income concept are solved, to get a substitute distribution for all except the lowest levels of income. Since for 1944 the income tax data yielded more income than the Census survey despite partial coverage due to the statutory filing requirement and the exclusion of nontaxable sources, there is a definite advantage in including converted income tax units as far down the income scale as possible. Other reasons for dealing with the entire tax distribution are the desirability of level by level comparison with survey results and the fact that an exact point of splicing to the survey distribution need not be decided upon in advance.

Associated with the availability of a relatively complete distribution of tax returns and the initial indeterminacy of the exact point of splicing to survey results is the requirement that the tax return distribution, when converted into a family distribution, be as complete as possible above any given income point. Even if the eventual point of splicing is predetermined, this requirement of completeness compels the inclusion of data for all levels of income recipients because of the process of combining tax returns into family units where units may combine irrespective of individual level; units classified at even the lowest levels in a distribution by size of their own incomes can conceivably join with other units far up the income scale. In the National Resources Committee study, for example, the modified tax distribution was appended at the $7,500 level of total income. To obtain a family distribution starting at that level adjustment had to begin at a lower point in the tax distribution because units would be shifted upward. Our procedures can be considered, from one viewpoint, as carrying this precaution to an extreme. It is perhaps best exemplified in the case of supplementary recipients where the incomes of husband-wife units at

the $10,000 level, say, may be increased by the small incomes of additional recipients.

For deriving a correct distribution it makes little difference whether all the incomes to be combined are in the initial tax data or not since in either case they can augment incomes at all levels. If we confine ourselves to tax returns, or to returns above an arbitrary level of income, deficiencies may appear throughout the distribution. It is in this sense that the relatively complete tax distributions available in recent years are still insufficient to assure an accurate family distribution above any arbitrary point. The initial tax distributions must be augmented to include *all* income recipients before combination can yield an adequate family distribution.[20] We therefore regard all the given tax distributions as deficient, requiring correction for the total number of recipients.

This supplementation of tax returns before combination into family units, made possible by the approximate completeness of the tax data for 1944, is consistent with using cross-tabulations treated in the preceding section where an entire (conditional) distribution of recipients classified by the size of their own income is associated with a unique level of income for related units. From this viewpoint the supplementation of tax returns can be interpreted as an attempt to obtain as complete a cross-tabulation as possible.

3 *Segregation and Combination of Subfamilies*

Another area of difference between the present and former methods of converting tax data into family units is our concern with subfamilies. The family as defined in the Census survey consisted of two or more persons living in the same household related by blood, marriage, or adoption. Inherent in this concept are parent or parent-child units (other than the family head and his wife) which for tax purposes may constitute groups separable

[20] If the splicing point on the family income distribution is high, augmentations to the tax distributions below the level of legal filing requirement may be disregarded because the effect above this point can not be large. Since we adopted procedures permitting as low a splicing point as possible, augmented distributions can have appreciable effect.

from the 'main' family. In the Census survey the incomes of these subfamilies are added to those of other family members in computing total family income. In the tax distributions, however, these subfamilies are tabulated as separate units and are indistinguishable from the 'main' family. Adoption of the Census definition, therefore, compels an adjustment for these subfamilies.

This adjustment consists in segregating the tax returns of such subfamilies and combining their incomes with those of the 'main' family members. Since in 1944 large numbers of the population were in the armed forces and there was a marked housing shortage, persons living in subfamily status were numerous. Adjustment for these families was deemed important, therefore, and a substantial part of the labor of converting the tax distribution was expended on segregating and dealing effectively with them.

4 *Family Supplementary Income Recipients*

Another difference concerns the treatment of family supplementary income recipients. If we confine ourselves to tax returns no hint is given of how these units can be segregated or combined with other units to constitute families as we define them. Survey relationships must, therefore, be exploited to effect this adjustment. The segregation and combination of these units is undoubtedly the weakest link in converting the tax data. Since the segregation can be reasonably accurate, the problem centers mainly upon combination with other units in the distribution.[21]

In the NRC study supplementary recipients were combined by extrapolating both the average number per family and the average amounts of their incomes beyond the survey levels. When applied to the husband-wife units from tax returns, these averages permitted estimates of both the number of recipients and

21 The authors believe that the main reason so-called 'matching studies' of data from surveys and tax returns should be encouraged is to supply information on this problem. The problem of supplementary recipients is part of the general one of converting tax returns into family units which could be aided materially by examining such returns under the correct family classification. It is hoped that such studies may provide the basic relationships for future combination of tax units. At the very least, they will throw some light on the acceptability of assumptions such as those used in making the estimates in this report (see Sec. C).

their incomes at each family level.[22] The adjustment related supplementary income to the total income at each level of husband-wife income to obtain the basis for shifting units up the income scale. As no dispersion was assumed, the units were not distributed properly.

Here we inquire first into the nature of the relation between the incomes of husband-wife units and of supplementary recipients. As will be seen below, some evidence, based upon survey material, suggests there is little relation, as Dorothy Brady suggested in conversation with one of the authors, between the incomes received by these two groups. Other evidence suggests a small positive correlation. In Section C5, however, the slight correlation is shown to be of little consequence; independence can safely be assumed with only minor effect on the results. The assumption that these two groups of units were independent in the statistical sense immediately provided the statistical bridge whereby combination could be effected.

C Description of Procedures Used in Constructing a Family Income Size Distribution for 1944

1 *Adjustment to Remove Net Taxable Capital Gains and Losses*

The income concept used to classify tax data differs in several ways from consumer money income, the concept used in the surveys. The chief component of this conceptual difference is military income, which is included in the latter but largely excluded from the former. Other areas of difference are 'other income', e.g., social security payments, pensions, as defined in the surveys but excluded from tax returns, and net taxable capital gains and losses which are included in the tax return concept of adjusted gross income but not in the survey income concept. Adjustment for capital gains and losses changes the aggregate

[22] Extrapolation to high levels is, of course, hazardous, but in some form is unavoidable in any treatment of survey data. In a strict sense, the averages extrapolated in the NRC study should have been related to family incomes *excluding* such supplementary income because the husband-wife 'nuclei' obtained from the distribution of tax returns to which the averages were applied excluded such income.

relatively little. Military income in the 1944 survey totaled approximately $6.0 billion and 'other income' about $2.1 billion; net capital gains and losses on tax returns were only $0.9 billion.

The effect of the capital gains adjustment is pronounced at the upper income levels, however, because a relatively large percentage of the units at these levels have such gains. We decided to make a careful adjustment for capital gains and losses primarily because detail at high levels of income is required if the need for data throughout the income scale or a small 'and over' class at very high levels is to be satisfied.[23] Moreover, a complete distribution of adjusted gross income corrected for such gains and losses is interesting in its own right, revealing as it does the degree to which income distributions are modified by the exclusion of such items.

Accurate adjustment for capital gains and losses required, as did almost all others, a cross-tabulation of adjusted gross income by capital gains (or losses) for units receiving income from this source. Since the BIR did not provide such a cross-tabulation for 1944, we had to construct one. Tabulations for preceding years could not be used directly because they were truncated at $5,000 net income and in some years, e.g., 1942, the last year for which there were such tabulations, net gains were classified as either short or long term and no hint was given concerning the identity of units that may have had both. In 1944 both long and short term gains were netted and the 1942 cross-tabulations were therefore not directly applicable. Even apart from these differences in classification, the direct use of data for a prior year is hazardous without considerable adjustment.

The BIR did present a distribution of returns by size of capital gain as well as the number of returns reporting and the average net gain for each level of adjusted gross income for 1944. Hence the marginal distributions of the desired cross-tabulation were known, as well as the average net gain for each column (con-

[23] The 'and over' class in current surveys is still either '$7,500 and over' or '$10,000 and over' although incomes have risen markedly since these limits were set. Data at higher levels are difficult to obtain, but an 'and over' class beginning at $10,000 in 1947, say, is roughly comparable with one beginning at $5,000 in 1935–36 as far as the percentage of families above that level is concerned.

ditional distributions). Our problem was to construct suitable column distributions to yield the given averages and at the same time meet the marginal distribution of capital gains by size.

We constructed these conditional distributions from two sources: cross-tabulations for preceding years and BIR data for adjusted gross income levels above $100,000, recorded on its transcript cards. The conditional distributions of capital gains obtained from the latter source can be accepted as precise, but were confined to adjusted gross income levels above $100,000. The latter data were drawn upon also in determining the form of the size distribution of capital gains for levels below that extremely high income, where such information was not available. They, together with cross-tabulations for earlier years (suitably modified), were the base for most of the required column distributions for income levels below $100,000.

a *Constructing a cross-tabulation for capital gains by income level*

Cross-tabulations were available for several years before 1943, but only for incomes above $5,000; moreover, they were presented separately for long and short term net gains. The size distributions and average net capital gain for each net income level differed markedly from year to year. A cross-tabulation for a particular year was chosen after comparing the average capital gains for each level of net income for 1935–41 with the figures for 1944.

In view of the later adjustments, the fact that net income was the basis of classification in the former years while the 1944 data were based upon the adjusted gross income concept was deemed of little moment. Furthermore, the separation of long and short term gains in the earlier years implied that the conditional distributions would not be strictly applicable, but the error introduced by our decision to confine attention to long term gains also could be expected to be of little consequence.[24]

24 We decided to base provisional conditional distributions upon tabulations for preceding years after initial attempts to base them on some specific distribution function had failed when tested against data for preceding years. We discovered,

The 1940 cross-tabulation yielded average net gains most closely approximating those for 1944. The row distributions of this tabulation were therefore taken as first approximations to the desired conditional size distributions of gains for the various income levels below $100,000 for 1944. They were actually approximations in two respects: the averages, although close, differed from those yielded by the 1944 data; and there was no assurance that they would yield the marginal distribution of capital gains by size of capital gains. Further adjustment was therefore required. We adjusted first to the given averages, then modified the distributions to agree with the marginal distribution. It would have been better to adjust simultaneously to meet both conditions, but no convenient procedure was at hand.

Examination of the conditional size distributions for earlier years gave ample evidence that they conformed fairly well to lognormal distributions for all except the extremely high levels of capital gains. More specifically, they resembled lognormals somewhat below a point where gains about equaled the upper limit of each net income class associated with a distribution. We could assume, therefore, that in adjusting for average capital gains we could use a transformation that would change a lognormal distribution into another of the same form and at the same time leave the top levels almost untouched. Actual distributions from 1940 tabulations were modified to achieve the desired 1944 average gain for the given income class by means of a transformation suggested by the lognormal case (see App. A). The distributions obtained by means of this transformation, together with the data for adjusted gross income classes above $100,000, constituted column distributions for each level of adjusted gross income above $5,000. Frequencies for the cells at levels of ad-

for example, that although the lower portions of the distributions by size of gain could be closely approximated by lognormal distribution functions, the fit was unsatisfactory at higher levels. The curve type suggested was that of a composite function which would permit a more rapid falling off of frequencies at the high levels than in the lognormal. Several composite types proved unsatisfactory when tried. The fact that lognormal functions fit major portions of the distributions reasonably well, however, did provide the basis for the transformation below, as well as suggest the treatment of capital losses where the distribution was limited to a range up to $1,000.

justed gross income below $5,000 were derived by extrapolating each row distribution.[25] The row and column distributions of these extrapolated values were then smoothed and the cross-tabulation finally adjusted so that all rows and columns added to both marginal distributions.[26]

b *Constructing a cross-tabulation for capital losses by income level*

The adjustment for capital losses differed from that for gains primarily in that conditional distributions from a preceding year did not enter into the calculation. Since the law limited net capital loss allowed as a deduction from adjusted gross income to $1,000, this amount determined the maximum shift of any unit to be effected by the adjustment and immediately narrowed the range of possible error. Moreover, as in the case of capital gains, a lognormal curve type could be assumed to apply over this narrow range. The theoretical functions that had proved unsatisfactory in the treatment of capital gains could be used in constructing conditional distributions. This more synthetic approach had additional merit in that less difficulty was encountered than in transforming empirical curves.

From the correct conditional distributions for adjusted gross income above $100,000 we found that the logarithmic coefficients of variation were approximately constant. The coefficients were computed from distributions of total realized net capital losses without the statutory limitations on their deduction for tax purposes. On the assumption that the same coefficient would hold for lower income and given the average loss at each level of adjusted gross income for returns having such losses, lognormal distributions were constructed for all levels of adjusted gross

[25] Extrapolation was done graphically on lognormal paper. Many of the empirical distributions dealt with in size distribution work are approximately of lognormal form, at least for portions of their range.

[26] W. E. Deming's iterative method was used to effect this adjustment; see *Statistical Adjustment of Data* (Wiley, 1943), p. 115. When the final column distributions were checked for discrepancies from desired averages caused by the adjustment, it was found that most differences were 3 or 4 percent with occasional extreme discrepancies ranging up to 21 percent. No further adjustment was made to remove these discrepancies as the total effect could be assumed to be minor.

income below $100,000 (see App. B). The conditional distributions, like those for capital gains, were adjusted to add to the known margins.

c *Results*

The cross-tabulations of adjusted gross income and capital gains or losses for units having net gains or losses permitted the relatively easy adjustment for removing such items from adjusted gross income. In adding net capital losses or subtracting net capital gains for levels of adjusted gross income below $100,000, each cell of the cross-tabulation was taken separately; addition or subtraction was confined to frequencies within the cell and carried through on the assumption of uniform density. For levels of adjusted gross income above $100,000, the amounts exclusive of capital gains and losses were computed from the transcript cards. To all frequencies with $1,000 or more of realized capital losses $1,000 was added because of the statutory limitation (Table 1). The effect on all except the highest income levels in the distribution of all returns is slight indeed—ample evidence that when detail is not required at very high levels, careful adjustment for capital gains and losses is superfluous. At the very high levels, however, differences are substantial.

Despite the extreme income inequality noted in distributions of capital gains by their own size, removing them made the distribution by adjusted gross income level of returns having them more unequal. As is evident from Chart 3, the distribution by adjusted gross income minus capital gains and losses for returns reporting them is more unequal than the unadjusted distribution. This effect, which is surprising in view of the known inequality of capital gains by size, is evident also from the 1936 data in the *Statistics of Income Supplement*. It arises from the capital gains adjustment alone and cannot be explained by the adjustment for capital losses.

The change in income inequality noted is that measured by the Lorenz curve and its associated index, the coefficient of concentration. Actually, despite an increase in the area under the line of equal distribution after adjustment for the removal of capital

items, the two Lorenz curves in Chart 3 cross at a very high level of income.[27] Pareto curves drawn for high levels of income re-

Table 1
Capital Gains and Loss Adjustment
Individual Income Tax Returns for 1944

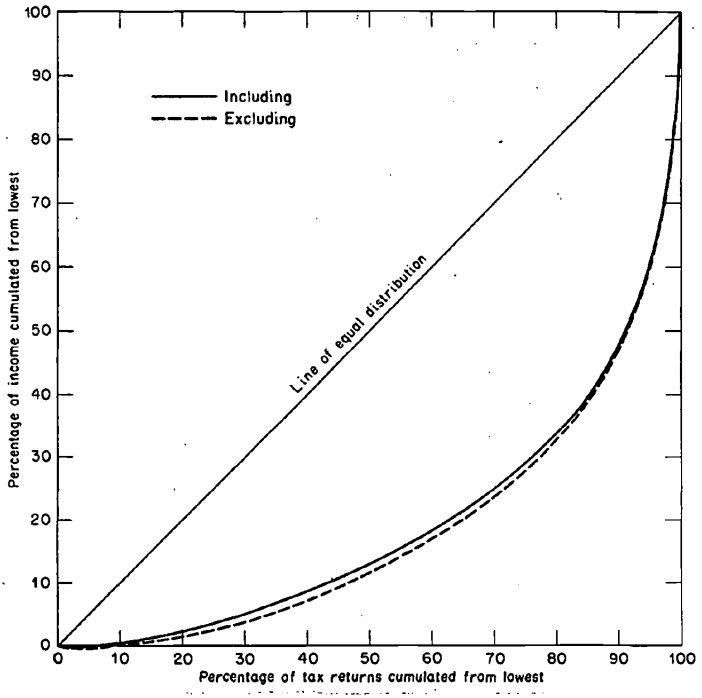
Income Class ($000)	Returns with Capital Gains or Losses by: Adjusted gross income	Returns with Capital Gains or Losses by: Adjusted gross income excl. capital gains & losses	All Returns by: Adjusted gross income	All Returns by: Adjusted gross income excl. capital gains & losses
Deficit	23,594	35,388	191,905	203,699
0– .5	39,494	82,628	3,260,590	3,303,724
.5– 1	93,347	112,328	6,068,006	6.086,987
1– 2	232,256	252,989	14,066,411	14,087,147
2– 3	253,787	235,676	11,319,701	11,301,590
3– 4	206,155	186,665	6,920,298	6,900,808
4– 5	139,987	122,407	2,817,030	2,799,450
5– 10	258,185	231,132	1,834,433	1,807,380
10– 15	86,264	82,613	298,478	294,827
15– 20	42,442	40,828	129,466	127,852
20– 25	27,727	26,145	67,537	65,955
25– 30	16,709	15,774	38,435	37,500
30– 40	19,575	18,341	41,610	40,376
40– 50	10,659	9,952	0,422	19,715
50– 75	12,721	11,324	21,669	20,272
75– 100	4,353	3,971	7,294	6,912
100– 150	3,367	2,986	4,873	4,492
150– 200	1,112	910	1,565	1,363
200– 250	517	415	665	563
250– 300	268	219	351	302
300– 400	256	198	318	260
400– 500	136	98	155	117
500– 750	137	92	159	114
750–1,000	55	39	62	46
1,000–2,000	48	36	50	38
2,000–3,000	4	1	6	3
3,000–4,000	2	2	2	2
4,000–5,000	2	2	3	3
5,000 & over	1	1	1	1
Total	1,473,160	1,473,160	47,111,495	47,111,495

[27] The two curves cross in 1944 at approximately the $40,000 level of adjusted gross income or at a point below which lie approximately 98 percent of the returns and 73 percent of the income. We found double probability paper useful in revealing details of Lorenz curve plottings for areas usually obscured when curves are drawn on the customary arithmetic scale. This paper, which permits the plotting of data in terms of normal deviates, reveals detail at all levels with equal clarity. Since the area between the line of equal distribution and that of any Lorenz curve drawn on this paper is, of course, infinite, it cannot be used as it stands to measure over-all inequality.

Chart 3

Lorenz Curves for the Distributions of Adjusted Gross Income Including and Excluding Capital Gains and Losses

Returns with Capital Gains and Losses



vealed differences in slope toward more equality. In short, the effect of removing capital gains and losses from the distribution by adjusted gross income is somewhat complex and cannot be adequately described by over-all measures of inequality.

The reasons for the particular effect on income inequality of the capital gain and loss adjustment are best seen by examining the full cross-tabulations relating them to adjusted gross income. The amount of capital gains at each level of adjusted gross income also suggests the results. Although average capital gains increase with each level of adjusted gross income, the total amounts expressed as percentages of total adjusted gross income at each level do not trace any such pattern. The pattern for 1944 was definitely U-shaped, the minimum occurring at approximately the point of

crossover mentioned above. In other words, for returns reporting them, capital gains are relatively more important at low and extremely high levels of income than in the mid-range.

If dispersion and changes in rank are ignored, the U-shaped curve can be regarded as a multiplier curve to effect a transformation based on an average correction at various income levels.[28] In fact, this multiplier curve was the basis for the Lorenz curve in Chart 1. It immediately suggests the increase in over-all inequality as well as the differential effects noted after capital gains and losses have been removed.

d *Removing capital gains and losses from different types of return*

The last step was to carry through the adjustment to remove capital gains and losses from each category of return: joint, separate, and returns of single persons and married persons not living with husbands and wives. Attempts to construct cross-tabulations of adjusted gross income and capital gains and losses for each type of return would be extremely arduous and, in view of the minor differences expected, probably lead only to spurious accuracy. We therefore used the cross-tabulation based upon all returns. We assumed that the distribution of capital gains and losses at each level of adjusted gross income was the same for all types of return having such gains or losses at all levels of adjusted gross income except those above $100,000 where exact information was taken from the transcript cards for each type of return.

2 *Matching and Combining Separate Returns of Husbands and Wives*

In 1944 5.1 million tax returns, approximately 11 percent of all returns filed, were separate returns. Thus, husbands' and wives' incomes for a substantial number of returns had to be matched

[28] As noted above, ignoring dispersion and changes in rank simplifies at the cost of considerable error. The simplification is resorted to here only to facilitate exposition. The Lorenz curve based upon the average changes in income reveals a crossing with the initial distribution by adjusted gross income at approximately 78 percent of the returns and 32 percent of income; that based on the full cross-tabulation, a crossing at the 98 and 73 percent points.

as one step toward achieving the combined income of the family unit. As mentioned above, no hint is given in *Statistics of Income* how separate returns combine.[29] The only data are an actual matching study of separate returns of husbands and wives published in the *1936 Statistics of Income Supplement.* This study, which unfortunately was not available to the National Resources Committee in constructing the 1935–36 distribution, is sadly out of date because of the large rise in the level of incomes and possible changes in filing practices.

Conceivably, survey relationships might yield data for such a combination. But the special tabulations required were not available when we made our estimates. The value of such data would be limited, moreover, because of the assumed identity of incomes reported in surveys and on tax returns and, more important, because survey data are deficient at the very income levels where separate returns predominate. If the cells in the cross-tabulation relating husbands' and wives' incomes are to have sufficient entries, surveys must be heavily supplemented at the higher income levels. Furthermore, if survey material is to be useful for this purpose it is necessary, either by adding suitable questions to the schedule or by analyzing the interview results, to determine which individuals filed jointly or separately. Only when they filed separately could the survey cross-tabulations of husbands' and wives' incomes be assumed to apply to the matching of separate returns.

For the above reasons, the investigator must rely upon the cross-tabulation in the *1936 Statistics of Income Supplement,* at least for the time being. The problem is mainly the best possible use of the now antiquated cross-tabulation so as to avoid, if possible, limitations imposed by the span of time. Simply forcing the entries in the 1936 cross-tabulation (see, e.g., the methods outlined in Deming, op. cit.) so that the sums of both rows and

[29] It is regrettable that the tax form, no matter how the tax is imposed, does not list additional questions which would permit easy combination into family units. This would be desirable for both separate returns and those of supplementary family income recipients. At present a matching study would prove enormously difficult because of the large number of returns and the fact that the sample used as the basis for the *Statistics of Income* tabulations could be expected to contain relatively few instances where all returns for a family were included.

columns agree with the given 1944 marginal distributions of husbands and wives by the size of their own income was not considered satisfactory. This method is warranted only when there is reason to believe that departures from the desired margins are due to random factors. In the case of the 1936 cross-tabulation one would expect a shift upward of frequencies in all the row and column distributions which could not be accomplished adequately by methods dependent upon assumptions of random error.

We based upward shifts in both row and column distributions upon the known changes in the marginal distributions. The sums of both row and column distributions were made to agree with the marginal frequencies directly and further adjustment was unnecessary (see App. C). The procedure modified the 1936 cross-tabulation in accordance with the known changes in the marginal distributions and altered the underlying relations revealed in the 1936 cross-tabulation only slightly.

The size distributions of the separate returns of both husbands and wives were adjusted before they were used as marginal distributions in the cross-tabulation. First, separate returns filed by husbands and wives were not equal in number, although their filing implied similar action by both husband and wife. In 1944 there were over 200,000 more returns in the wife distribution. This excess can be attributed to several factors. At least a portion is due to sampling error. Second, difficulties arise in classifying returns by sex when information is incomplete and assignments cannot be made accurately from the name. There are also the possibilities that returns are improperly filled out by the taxpayer and that husbands in the armed services stationed overseas postponed filing until they came back whereas their wives filed on the customary dates. In the absence of data, the excess returns of wives were arbitrarily assigned to the category 'Returns of single persons and persons not living with spouse'. As will be seen, this category is itself later divided so that this allocation affected, in some small degree, all returns except separate and joint. The 200,000 excess was removed from the 1944 distribution of separate returns of wives on a proportional basis.

Once the adjusted marginal distributions were derived, husbands' and wives' incomes were added by means of the modified 1936 cross-tabulation on the assumption of uniform density in each cell, permitting the use of rather simple formulas. However, further adjustment was required to compensate for the error introduced by this assumption. Formulas that assume uniform density in each cell are subject to bias and tend to yield distributions with insufficient frequencies in intervals at or near the modal class. The combined distribution contained approximately 2 percent more income than the separate distributions. This error was corrected for by shifting frequencies from all classes toward the mode.[30] A similar error probably occurred in adding capital gains and losses but was ignored because of the minor amounts involved.

3 *Separating Returns into Categories of Family Membership*

As mentioned above, the types of return classified by the BIR must be subdivided so that they can be appropriately combined into family units as defined in field surveys. The BIR classification is only partly suitable for this purpose. Joint returns and combined separate returns of husbands and wives are roughly equivalent to head-wife families in the surveys except as far as they include husband-wife units of subfamilies, e.g., a married son and his wife living with his parents. But single returns, including returns of married persons not living with spouse, comprise various returns, some of which represent heads of families, others, single individuals,[31] and still others, supplementary income recipients in families.

The choice of subgroups required for later combination into family units is dictated by the role of family members as taxpayers (Table 2).

30 See Appendix D for an account of the method of shifting the frequencies. After shifting frequencies, interpolation for the desired class limits was carried through by formula (see App. E).

31 'Single individual' does not refer to marital status; it differentiates the person not living with relatives from the family of two or more related persons. Recent Census Bureau releases have called this group 'individuals not in families'.

Table 2
Income Tax Returns Allocated to Family Membership Groups

BIR Types of Return	Role in Survey Family	Membership in Main or Subfamily	
		Main	Subfamily
1 Joint 2 Separate	Husband-wife units in normal families	X_1	X_3
3 Single	Heads of 'broken' families	X_2	X_4
	Supp. income recipients in families	X_5	X_6
	Single individuals	X_7	..

The 'main' family comprises persons in the family other than in subfamilies. The 'subfamily' is a married couple with or without children, or one parent with one or more children under 18, living in the household and related to, but not including, the head of the household or his wife.

From the viewpoint of their status in the family, types 1 and 2 are identical and are taken directly from joint returns and from separate returns after matching. Thus, as can be seen from Table 2, six subgroups are segregated for families of two or more persons, and one for single individuals. These subgroups must be singled out in the BIR tabulations so that they can be appropriately combined. More specifically, the combinations require matching and combining persons in groups X_3 through X_6 with the appropriate 'main' families in groups X_1 and X_2 to determine families (as defined in field surveys) comprising all related persons living in a household.[32]

Although the desired subdivisions of single returns could be approximated from BIR data on exemption status and some tabulations for prior years of heads and nonheads of families classified under single returns, returns in that category were separated on the basis of survey material.[33]

[32] Six component distributions assignable to families of 2 or more are mentioned mainly for the sake of completeness. In practice, this number was reduced to 5 on the assumption that there were no supplementary income recipients in subfamilies, i.e., group X_6 was ignored.

[33] Difficulties in using BIR data to suggest the subdivision were due mainly to the fact that exemption status alone is insufficient to determine family status as given in Table 2. Moreover, information on heads and nonheads of families classified under single returns was not obtainable for a year close enough to 1944 to be representative of that year with respect to family composition and level of income. Efforts to isolate the desired groups on the basis of these data were rather unsatisfactory. The distributions were inconsistent in that some were pitched higher than the survey material and some were not. In any case, there were no BIR data to suggest how subfamilies could be segregated.

a *Separation of income tax returns*

The general procedure was to base percentage patterns at each income level on Census survey distributions for each group in Table 2. The pattern for a given level was then applied to tax returns at that level after capital gains and losses had been removed and after incomes of husbands and wives filing separate returns had been combined.

As Table 2 suggests, 2 sets of percentage patterns are needed to achieve the desired division of tax returns: first, 4 distributions to divide single returns into groups X_2, X_4, X_5, and X_7 (group X_6 was ignored); second, 2 distributions to divide husband-wife units into those in main families, group X_1, and in subfamilies, group X_3.[34] The first group of 4 distributions requires distributions of individuals' incomes; the second consists of distributions of combined husband-wife incomes.

The basic data for constructing the distributions were taken from a Census tabulation of survey data for 1944 giving, by civilian earnings levels of individuals, distributions of (a) heads of 'broken' families, (b) heads of normal (husband-wife) families, (c) wives of heads of normal families, (d) other relatives of heads, and (e) single individuals. Under the BIR classification distributions b and c would be in the joint and separate return categories; distributions a and e in the single return category; distribution d would contain units that may be in any of the 3 categories since it contained family supplementary income recipients as well as all members of subfamilies.

The Census tabulation did not, therefore, obviate the need for considerable manipulation. The advantage, however, was considerable, because single individuals and heads of broken 'main' families were already segregated. Distribution d, other relatives of heads, had to be subdivided, it is true, but subfamily members were combined only with family supplementary income recipients. As shown below, the distribution of subfamily members was taken from another Census Bureau tabulation and

[34] The 'husband-wife' unit as used here and elsewhere in this report is a single unit consisting of the combined incomes of the husband and wife within the family.

the distribution of family supplementary recipients obtained by subtraction.

The Census tabulation giving the various distributions by size of individuals' civilian earnings had to be adjusted to the BIR concept of adjusted gross income. A second adjustment was necessary since the Census distribution of individuals, although appropriate for the separation of single returns, was not suitable for classifying husband-wife units in the tax distribution. The first limitation required adjustment for income other than earnings, while the second required the combination of appropriate distributions of individuals in Census data.

Six income size distributions were derived from Census data corresponding to groups X_1–X_5 and X_7 in Table 2 (see App. F). All distributions were by size of consumer money income minus military income, a classification that roughly approximated adjusted gross income minus net taxable capital gains and losses. They ended with a '$10,000 and over' class, the limit for detailed classification in the survey data. To get percentage patterns for all levels of income, each of the 6 distributions therefore had to be extrapolated to higher income levels. Extrapolation up to the $100,000 level was by means of Pareto curves for each distribution based on parameters computed from data in the income class immediately below $10,000. No extrapolation was made beyond $100,000.

Percentages were then computed at each income level for the distribution between main and subfamily husband-wife units (to be applied to BIR income size distributions of joint returns and combined separate returns) and among heads of broken main families, heads of broken subfamilies, supplementary income recipients, and single individuals (to be applied to the BIR income size distribution of single returns).

b *Census Bureau and derived BIR distributions for similar units*
At this stage certain BIR data for comparable units and income classifications can be compared directly with estimated distributions based upon Census data (Table 3).

Table 3
Income Size Distributions Based upon Census and BIR Data, 1944 (thousands)

Income Level[a]	Husband-Wife Units[b]		Individual & Single Returns[c]	
	Census	BIR	Census	BIR
Loss	127	126	86	74
Zero	505	...	13,770	...
0–.5	2,127	605	6,509	2,520
.5– 1	2,387	1,858	4,013	3,801
1– 2	6,077	5,907	6,345	6,947
2– 3	7,615	7,015	2,378	3,096
3– 4	5,412	5,618	608	961
4– 5	2,363	2,895	161	294
5–10	1,689	2,031	158	228
10 & over	464	571	71	100
Total	28,766	26,625	34,099	18,021

[a]Income level is that of adjusted gross income after capital gains and losses have been removed or, approximately, Census survey consumer money income minus military income.
[b]Husband-wife units contain the combined incomes of husbands and wives in either main or subfamilies and are the sum of joint and combined separate returns in the tax data or of Census survey 'heads of primary families', 'wives of heads', and normal subfamilies.
[c]The constituent distributions in this group are heads of broken main families, heads of broken subfamilies, family supplementary income recipients, and single individuals, as taken from BIR tabulations (adjusted to exclude capital gains and losses), and from estimates based on Census survey data as described above.

The number of husband-wife units from Census surveys exceeds that from tax returns at all income levels up to approximately $3,000 (Table 3). Above that point the number from tax returns is larger. The cross-over point for single returns is lower, at approximately $1,000. The excess of Census survey over BIR data is, of course, to be expected at levels below $1,000 for husband-wife units and below $500 for single returns because of legal filing requirements. The excess above those levels is less easily explained.

The comparison does not yield evidence of superior reporting in the survey for the low income levels. If survey and BIR distributions were spliced at the cross-over point, too many husband-wife units would be included in the distribution. If we take Census husband-wife units for all levels below $3,000 and BIR units for those above, the total would far exceed such units in

the population, even if the units in the Census zero income class are disregarded.[35]

The most plausible explanation for the heavy representation of husband-wife units at levels below $3,000 in the Census data is that Census survey data are biased relative to BIR data.[36] This bias, which might be deduced from the much lower aggregate income accounted for in the survey, is evident here from examination of comparable units (cf. Part VI). Furthermore, it was apparently not caused exclusively by the failure of the survey to include some family supplementary income recipients, for the low pitch of the distribution is obvious for husband-wife units alone.

A similar failure to achieve adequate representation at the higher levels can be noted in subsequent income surveys by both the Census and the Federal Reserve Board. In some of the later surveys, however, particularly those of the FRB, the percentages of units at the higher income levels are substantially larger, suggesting a marked diminution in the bias due to reporting (cf. Part IX).

Ultimately, the cause for such bias must be determined by examining the survey procedures themselves including various controls in the 'blow-up' of results, corrections for refusals, and analysis of population groups that may have been missed in the sampling.[37]

From the viewpoint of constructing an income size distribu-

[35] A similar statement cannot be made for the single return distribution because the Census distribution includes supplementary family members 14 and older of whom many are in the zero class.

[36] One possible qualification of the above explanation concerns the nonremoval of 'other income' (see App. F). It may be argued that once account is taken of this item the excess of Census survey data in the range below $3,000 would be removed. However, rough estimates, based on Census source pattern tabulations, revealed that even extreme assumptions regarding the downward shift of recipients of 'other income' would be insufficient to remove the excess.

[37] Actually, both BIR and Census survey data are biased. The failure of many farm families to file tax returns is reflected in the absence of such units in the (probably) lower part of the BIR distribution. It is questionable, however, whether the inclusion of the farmers who do not file returns would materially affect the conclusions because similar groups were missed in survey data and the increase in BIR units due to the inclusion of such units would not reach the Census survey frequencies below $3,000 because of the limitation on the number of husband-wife units referred to above.

tion from tax data to supplement survey findings, the above considerations present many difficulties. If the survey is regarded as representative in that it gives the reported incomes of the entire population, identical units must occur at different income levels in the two sources. Splicing BIR results to survey material, therefore, requires care and cannot be done at any point where the distributions are approximately the same.

c *Splicing derived BIR to Census distributions*

Because of the above considerations regarding bias in survey results and the difficulty of determining the point of splicing, we kept the BIR distributions, derived by applying the percentage patterns, down to the $1,000 income level for husband-wife units. The chief reasons were the expected collapse of the BIR data below that point due to the legal filing requirements, the minimizing of the area of possible duplication of units in survey and tax data, and the consistency of the independent estimate of the number of normal families with the distribution thus obtained.

The decision to splice to Census survey data at the $1,000 level for husband-wife units resembled that to retain the BIR distribution for all income levels above $500 for individuals. As mentioned in note 35, a population control similar to that for husband-wife units was not available for this group.[38]

The derived BIR distributions were spliced to the Census distributions directly at the specified levels of income and no attempt was made to smooth. The frequency in the zero income class for each distribution was the difference between the Census estimates of the total units in each category and the number of units with positive and negative incomes.

[38] Splicing husband-wife units at $1,000 was found in subsequent tests to be consistent with the alternative of using Census frequencies below the legal filing requirement of $500 for husbands and wives separately. To get from BIR data the estimated distributions of husbands and wives classified by their own incomes, distributions of husbands and wives filing joint returns with two incomes were combined with distributions of husbands and wives filing joint returns with one income and those filing separate returns. When Census frequencies were substituted below $500 and incomes of husbands and wives recombined, the results closely approximated the size distribution of Census husband-wife units below the $1,000 level.

Essentially, the splicing to survey data at such low income levels implies that the final income size distribution of family units, constructed at a later stage, is based primarily upon BIR data.[39] The need for completeness in each size distribution at this stage was indicated in Section B where it was pointed out that recipients at all levels of income will, after being combined into family units, affect the frequencies of family units throughout the distribution.

4 *Allocating Population Groups to Categories of Family Attachment*

The income size distributions obtained by applying percentage patterns to BIR data, and supplemented by survey frequencies below the point of filing requirement, accounted for all individuals 14 and older in 1944. That is, each distribution, when augmented by the Census frequencies, contained also a class of zero income recipients which, when added to units with positive or negative incomes, accounted for the entire population in the specified group.

At this stage the income distributions had to be assigned properly to the family nuclei for later combination. Though the number and income size distribution of all supplementary family members had already been estimated, it was necessary to determine how many were associated with normal families, with broken families, and with other supplementary family members in the same family unit, etc. The income size distributions of normal and broken subfamilies required similar allocation.

a *Population in categories of family attachment*

The basic Census survey tabulation did not give any hint except the total number of relatives of family heads, a category that included subfamily members as well as supplementary income recipients. Many other Census survey tabulations gave relevant

[39] It should not be concluded that survey material contributed little to our estimates. The above outline of methodology is ample evidence of the frequent and exhaustive use of survey data to effect the required classifications and as controls. Survey material was used later also to suggest how units combine. Relations based on survey data were indispensable in constructing our estimates.

information but none was directly applicable. Therefore, a sequence of tables or 'bridges' incorporating the various sources of information was constructed, thereby permitting approximation to the desired result. Many of these tables required adjustment because they were for survey years after 1944. As indicated in Appendix G, estimates for 1944 were derived for (a) the number of normal and of broken families distributed by size classes of the number of persons 14 and older, and (b) the number of normal families 'with normal subfamilies' and 'with broken subfamilies' distributed by size classes of the number of persons 14 and older. By subtracting normal families in (b) from those in (a) the number 'without subfamilies' was estimated by size classes of the number 14 and older.

For broken families, distributions by size classes of the number 14 and older were needed separately for those 'with normal subfamilies', 'with broken subfamilies', and 'without subfamilies'. It was assumed that for any given size class of the number 14 and older, the percentage division among the three categories was approximately the same for broken as for normal families.

These distributions provided all the relationships for later combination.[40] Since the total number of combinations was unmanageable we arbitrarily limited possible combinations made with any family unit (husband-wife unit or head of broken family) to two, e.g., one supplementary family member and one subfamily, or two supplementary family members. At most, one subfamily was permitted in any one combination.

With this limitation and with the classification of normal and

[40] A test was made at this point by comparing the number of families classified by the number of civilian earners from the 1944 Census survey with the number classified by the number of supplementary family members (i.e., the number of supplementary family income recipients if $0 income is included as a class) obtained above. The comparison can only be approximate since the Census data classify families as civilian earners and the BIR estimates are for total income recipients.

FAMILIES CLASSIFIED BY NUMBER OF EARNERS OR INCOME RECIPIENTS
(000)

	No. of Earners or Income Recipients in Family			
	0	1	2	3 or more
Census	2,578	18,806	9,360	2,570
BIR composite	1,093	17,667	11,382	3,172

broken families derived above, we determined the population in the various groups shown in Table 4; e.g., normal (husband-wife) families of 3 persons without subfamilies would have 1 supplementary family member, whereas broken families of the same type would have 2.[41]

b *Income size distributions for categories of family attachment*
As can be seen from the stubs of Table 4 the 14 distinct family types exhaust the list of permitted combinations of major family membership groups. All the persons listed in any one of the columns 3–7 belong to a distinct family membership group.

In a strict sense the family subgroups in a given column may be distributed differently by size of income; that is, there is no inherent reason why the distribution of supplementary family members attached to normal 'main' families with only one such member, say, should have the same distribution as supplementary members attached to broken 'main' families. In brief, because of the variety of socio-economic factors associated with the 14 family types the distributions within any family membership group may well differ.

The number of subgroups, already arbitrarily limited to have a manageable number for later combination, was obviously too formidable to estimate size distributions for each. Not only would the labor be excessive but the data are, in the main, inadequate or lacking. For supplementary family members, no data were available for 1944 to subdivide the previously estimated distribution for the major membership group. Some data for a later period (see App. F) were used to estimate the distributions of normal and broken subfamilies. They might have provided information also on differences in the size distributions of normal and broken subfamilies associated with either normal or broken main families. Similarly, they might have been used to analyze whatever differences in distribution could be noted between both types of main families associated with both kinds of subfamilies. Analyses of this sort, though of great interest, would have been too laborious.

41 In determining the population it was assumed that no members of subfamilies except the husband and/or wife were 14 or older (see App. F 2).

Table 4

Families with Various Combinations of Supplementary Family Members and Subfamilies, and Individuals 14 and Older in Each Group (thousands)

FAMILY COMPOSITION	NO. OF FAMILIES (1)	NUMBER OF INDIVIDUALS					
		Total (2)	Heads & wives of normal: Main families (3)	Heads & wives of normal: Sub-families (4)	Heads of broken: Main families (5)	Heads of broken: Sub-families (6)	Supplementary family members (7)
Normal main family with							
No subfamily with							
1) 0 supplementary member	15,183.2	30,366.4	30,366.4				
2) 1 supplementary member	6,927.9	20,783.7	13,855.8				6,927.9
3) 2 supplementary members	3,087.3	12,349.2	6,174.6				6,174.6
1 normal subfamily with							
4) 0 supplementary member	693.2	2,772.8	1,386.4	1,386.4			
5) 1 supplementary member	394.9	1,974.5	789.8	789.8			394.9
1 broken subfamily with							
6) 0 supplementary member	458.5	1,375.5	917.0			458.5	
7) 1 supplementary member	297.0	1,188.0	594.0			297.0	297.0
Broken main family with							
No subfamily with							
8) 0 supplementary member	1,281.7	1,281.7			1,281.7		
9) 1 supplementary member	2,841.4	5,682.8			2,841.4		2,841.4
10) 2 supplementary members	1,039.1	3,117.3			1,039.1		2,078.2
1 normal subfamily with							
11) 0 supplementary member	441.1	1,323.3		882.2	441.1		
12) 1 supplementary member	195.2	780.8		390.4	195.2		195.2
1 broken subfamily with							
13) 0 supplementary member	294.5	589.0			294.5	294.5	
14) 1 supplementary member	179.0	537.0			179.0	179.0	179.0
15) Total	33,314.0	84,122.0	54,084.0	3,448.8	6,272.0	1,229.0	19,088.2

Families of 2 or more; single individuals excluded.

In view of the difficulty of making adequate divisions, we maintained the same relative distribution of income in each major category of family membership irrespective of the particular family configuration entered into by the component units. For example, the 10 distributions of supplementary family members suggested in column 7 of Table 4 (2 for each entry opposite families with 2 such members) were given the percentage distribution of the major membership group comprising all such supplementary members. The same assumption was made for each of the other major family membership groups (col. 3–6). In effect, this assumption limited the number of distinct relative distributions to the 5 already estimated for the major membership groups. However, in later combinations, 32 distinct frequency distributions were used.

It is difficult to appraise the error in this assumption since the differences between the size distributions of income of the subgroups composing a major membership group have not been analyzed. But as far as these subgroups differ in relative distribution between the various family configurations it must introduce some error.

5 *Combining Income Distributions for Family Member Groups into Family Units*

At this stage of the estimates distributions were available by size of consumer money income minus military income for each of the family member groups in Table 4. Before appropriate distributions could be combined into family units we had to ascertain the nature of the relation between the various distributions of family members.

a *Income relationships among groups of family members*

Two analyses were undertaken with reference to the relationship among the distributions of family members:[42] between the in-

[42] Actually, three analyses were undertaken. In addition to those mentioned in the text, an attempt was made to determine the correlation between the incomes of family members from the Consumer Purchases Study and Minnesota data on average supplementary earnings per supplementary earner at each level of total family income. Several complicating factors, however, such as the presence of more than one supplementary earner per family, required the introduction of several assumptions in the analysis and made the interpretation of the results obscure. Little or no correlation between the incomes of family members was found but the serious qualifications on the results caused us to discard them.

comes of subfamilies and of main families to which they are attached and between the incomes of supplementary income recipients and of family heads.

Relation between the incomes of subfamilies and of main families

An unpublished Census Bureau study pertaining to incomes of main and subfamilies in 1946 gave a series of cross-tabulations in 2x2 summary form: the number of subfamilies by type (husband-wife, parent-child, other) classified in categories of under and over a specified income associated with main families in similar categories. All combinations by type were added together in 5 summary tables. All were condensations of the same underlying cross-tabulation relating the incomes of main and subfamilies, but in each a different level of income was the basis for dichotomization. Nevertheless, separate chi-square tests were made to determine whether the data were consistent with an assumption that the incomes of main and subfamilies were independent. At the 1 percent level of significance, three tests led to rejection of the independence hypothesis while two did not.

In addition to these tests for independence, tetrachoric correlation coefficients were computed for each of the 5 tables on the assumption that the cross-tabulation is a normal bivariate. More generally, if independent functions of the two variables on the margins transform the cross-tabulation into a normal bivariate, the tetrachoric correlation coefficient applies to these transformed variables. Since the logarithms of the income variables yield approximately normal distributions, the correlation coefficients can be regarded as referring to the logarithms of incomes rather than to their absolute values. Except for two negative correlations where the dichotomies had been constructed near the corner of the cross-tabulation, the coefficients lay between .15 and .20. In the strict case of a logarithmic normal surface, the correlation between absolute income values cannot exceed that between the logarithms, but because such a surface is merely approximated by a logarithmic transformation of the actual data, we can only conclude that the coefficient of cor-

relation, had it been computed for absolute values of income, would probably not have exceeded the coefficients actually found. This is illustrated by the subgroup of husband-wife main and subfamilies treated below where the correlation coefficient between the absolute values of income is .17 and the tetrachoric coefficients are higher.

The tests revealed, therefore, a slight positive correlation between the income of main and subfamilies. The effect of this mild correlation when the incomes of main and subfamilies were combined could be observed by comparing the results obtained by combining the observed data with those obtained under the hypothesis of independence between the two variables.

For the husband-wife subfamilies associated with husband-wife main families a more detailed cross-tabulation was available. First, the 2x2 tables for this subgroup of families, corresponding to those referred to above, were examined to determine whether this particular combination of main and subfamilies was representative of all combinations. The tetrachoric correlation coefficients computed for this group were slightly higher than those for all combinations, ranging from .20 to .25. Thus, any investigation, confined to this group, of the difference between results derived by combining incomes on the basis of the observed frequencies and on the basis of theoretical frequencies constructed on the assumption of independence would be expected to yield as large a discrepancy as would follow from full use of all types of main and subfamilies.

The distributions of husband-wife main and subfamily incomes were then combined using the actual cross-tabulation in one instance and one constructed on the assumption of independence in the other. The methods of adding the two variables were identical and yielded two distributions classified by total family income for the same group of families (Table 5). The closeness of the results adequately illustrates the effect of substituting the hypothesis of independence in constructing cross-tabulations for a mild relationship of approximately .20 as measured by the correlation coefficient.

Table 5
Husband-Wife Main Families Having Husband-Wife Subfamilies Derived from Observed and Hypothetical Cross-tabulations
Percentage Distribution by Combined Income Level, 1946

Combined Income Class	Observed Cross-tabulation	Assumption of Independence
$0	...	*
1– 499	.1	.2
500– 999	.6	.5
1,000–1,999	3.6	2.6
2,000–2,999	8.2	7.0
3,000–3,999	12.7	12.5
4,000–4,999	16.1	16.8
5,000 & over	58.7	60.4
Total	100.0	100.0

* Less than 0.05.

Relation between the incomes of supplementary income recipients and of family heads

The data for this test were from the BLS study for 1941.[43] They were tabulated so that the relation between the incomes of supplementary income recipients and of family heads (and wives of heads when the wife had income) could be examined. The nature of the test was similar to that outlined above; that is, the degree of correlation was determined, and, in addition, the results for the given cross-tabulation were compared with those for a cross-tabulation constructed on the assumption of independence.

In investigating the relation between the given variables only two income family units were considered. Families with two or more supplementary income recipients were tallied as many times as there were such recipients. Examination of the data indicated that this procedure would not vitiate any of the conclusions made below.

The correlation coefficient between supplementary income and the income of the head is .15, remarkably consistent with the results for 1946 noted above. Here, too, the effect of this slight correlation in the combination of the data is minor (Table 6).

[43] These data were taken from the worksheets on file in the Treasury Department, Division of Tax Research, of Albert G. Hart and Julius Lieblein for 'Family Income and the Income Tax Base', *Studies in Income and Wealth, Volume Eight* (1946).

Table 6
Two Income Families Derived from Observed and Hypothetical Cross-tabulations, Percentage Distributions by Total Income Level, 1941

Total Income Class	Observed Cross-tabulation	Assumption of Independence
$0– 499	1.1	0.9
500– 999	2.0	3.3
1,000–1,999	22.0	21.9
2,000–2,999	40.4	36.8
3,000–3,999	20.0	22.5
4,000–4,999	6.8	6.3
5,000–7,499	3.7	4.4
7,500–9,999	2.2	2.1
10,000 & over	1.8	1.8
Total	100.0	100.0

The incomes of the head and his wife were considered as one income.

b *Combining family member groups into family units*

In view of these analyses we combined the various income size distributions for the different groups of family members (as summarized in Table 4) on the assumption of independence.[44] Once the marginal distributions were known, the required cross-tabulations were easily constructed.

The actual addition based on these cross-tabulations threatened to be arduous whatever the method. The number of combinations was large and time precluded any except the simplest procedure. The many additions were carried out by first interpolating the marginal distributions to get narrower intervals and assigning an average to each derived interval. The frequencies in any cell were then assigned a place in the combined distribution by adding two averages, one on each margin corresponding to the given cell.

44 The evidence for near independence in the text must be qualified in view of the limited range of incomes covered in the survey data since combination at extremely high levels of income might manifest stronger relationships. Lack of data on combination at high income levels permits only conjecture. The great merit of undertaking studies involving the investigation of tax returns from the viewpoint of family groupings is that such relationships can be revealed. If evidence of stronger correlation is found, adequate allowance can certainly be made for it in the methodology. Because it is difficult to collect information in surveys for high income levels, tax returns constitute the only source. Until such studies are carried out, detailed results at the very high levels of family incomes are impossible.

6 *Adjusting the Income Distributions of Families and Single Individuals to Include Military Income* [45]

The income size distributions for families of 2 or more and for single individuals were now complete in that tax returns, supplemented by survey data for lower income levels, had been combined into family units. The income concept used as the basis for classifying families and single individuals, however, was adjusted gross income excluding capital gains and losses, or approximately total consumer money income minus military income. Adjustment was required, therefore, to change the basis of classification to that of total consumer money income by incorporating military income in the size distributions thus far derived. Military income in 1944 includes family allowances and allotments, and military pay and veterans' payments received by persons living in the family at the time of the survey. Allowances and allotments accounted for most military income in 1944.[46]

The data basic to the cross-tabulations required for military income were taken from a Census series of special income source pattern tables. As noted above, these tables show the number of families (or single individuals) receiving income from each of 7 sources, and the amount from each source, for each level of consumer money income. These source pattern tables were given separately for many types of earner group (e.g., families whose sole source was wages, families with wages and business incomes), as well as families of 2 or more and single individuals. Thus, we could get for single individuals with wage incomes, say, the number of persons and the amount of income from each source

[45] The sixth step, like the adjustment for capital gains and losses, is an adjustment for conceptual differences. Military income in 1944 constitutes the largest single difference between the survey concept of consumer money income and the adjusted gross income concept by which tax returns are classified. As noted below, these two adjustments for concept are only a portion of those that would be required to achieve complete comparability.

[46] As noted in Appendix F, military income had been removed from the Census survey data before they were used to distribute tax returns among the various family membership groups. The adjustment mentioned here is not strictly the reverse since the distributions to which military income is to be added are not the same ones.

at each total money income level, one of the sources being military income.

Other basic data, also from the Census Bureau, consisted of two cross-tabulations, one relating individual civilian earnings to military income, and the other, income other than civilian earnings to military income (see App. F). Briefly, the dispersion of military income at each total consumer money income level was based upon (a) the dispersion among the average amounts of military income noted among the many subgroups in the source pattern tabulations, (b) the known average from the survey for all single individuals at a given level of total consumer money income, and (c) the known distribution of military income by its own size for all single individuals. Thus, the cross-tabulation constructed to remove military income from the Census survey data was used here, in modified form, to reintroduce such income into the composite BIR-Census distribution of single individuals available at this stage of the estimates.

For families of 2 or more a similar cross-tabulation was constructed. Because of the relatively large number of separate categories in the family source pattern tabulations (14 distinct groups of families reporting military income) the estimates of the dispersion patterns are deemed even more reliable than those for single individuals.

Instead of adding the $6.0 billion reported in the 'blown-up' Census survey, we introduced $7.5 billion of military income as independently estimated for 1944.[47] Military income reported in the survey was increased approximately a fourth on the assumption that the higher amount was due to more recipients

[47] See Part VI. In a strict sense the distribution at this stage contained some military income as well as income from other sources that should have been removed to achieve complete comparability with the Census income concept. Thus, the military pay above $1,500 together with other income of officers in the armed forces stationed on posts in this country who filed income tax returns are included in the composite distributions. Our estimates are not adjusted for this income, the problem being postponed until full adjustment to independent income totals is made. The $7.5 billion of military income added at this stage consisted of military allowances and allotments, veterans' and military pay received by persons living with their families, as independently estimated for the universe of income recipients defined in the Census survey.

of such income in the population rather than to an understatement of amounts by recipients reporting in the survey.[48]

Consequently, some adjustment was required in the two basic cross-tabulations constructed for the purpose of adding military income, mainly in the distribution of units by total consumer money income within the zero military income class. For the portion of the cross-tabulations relating total consumer money income to positive amounts of military income, the conditional distributions by size of military income were retained and the frequencies increased the required 25 percent.

The composite marginal income distributions of families of 2 or more and of single individuals were then introduced into the two basic cross-tabulations. Addition of the marginal distributions in each cross-tabulation by means of the estimated conditional distributions yielded the desired distributions for families of 2 or more and for single individuals by size of consumer money income including military income or by size of adjusted gross income minus net taxable capital gains and losses, plus military income.

D Frequency Distributions of Families and Single Individuals by Income Level in 1944

The procedures outlined in the preceding section yielded income size distributions for families and single individuals based upon income tax returns supplemented by data from the Census survey. As noted in Section A, the aggregate income accounted for in these distributions exceeded those in the 'blown-up' Census survey and the initial distribution of tax returns. The Census survey yielded, after adjustment to include the estimated incomes of families and single individuals with incomes of $10,000 or more, $108–111 billion,[49] while the income tax distribution,

[48] The latter assumption was held to be more plausible because military income constitutes a unique type of receipt and therefore is more likely to be reported accurately in an interview.

[49] Minimum and maximum estimates of the income of units above $10,000 were obtained by use of a Pareto extension to the distribution below $10,000, and of BIR averages by source weighted by Census frequencies for the various sources above $10,000; see Part VI.

adjusted to remove capital gains and losses, accounted for approximately $115 billion. Because of the additions to tax return data at the lower levels of income and the inclusion of military income, the constructed distributions yielded approximately $124 billion (Table 7).

Table 7
Distributions of Families and Single Individuals and of Consumer Money Income Constructed from Income Tax and Survey Data, by Size of Consumer Money Income, 1944

Total Income Level	Families & Single Indiv.		Families of 2 or More		Single Individuals	
	Number	Amount ($000)	Number	Amount ($000)	Number	Amont ($000)
Under $500*	3,653,084	845,646	1,891,567	434,767	1,761,517	410,879
500 - 1,000	3,725,376	2,813,172	2,475,103	1,884,970	1,250,273	928,202
1,000 - 1,500	4,391,605	5,506,215	3,144,577	3,952,438	1,247,028	1,553,777
1,500 - 2,000	4,430,954	7,753,804	3,423,886	6,002,105	1,007,068	1,751,699
2,000 - 2,500	4,412,034	9,922,218	3,596,721	8,098,844	815,313	1,823,374
2,500 - 3,000	4,259,087	11,698,371	3,704,154	10,185,137	554,933	1,513,234
3,000 - 4,000	6,850,906	23,694,984	6,335,561	21,934,755	515,345	1,760,229
4,000 - 5,000	4,219,658	18,689,799	4,012,962	17,779,646	206,696	910,153
5,000 - 6,000	1,829,620	9,919,187	1,791,813	9,720,050	37,807	199,137
6,000 - 7,500	1,330,067	8,825,804	1,295,201	8,593,520	34,866	232,284
7,500 - 10,000	896,513	7,630,171	860,543	7,322,625	35,970	307,546
10,000 - 15,000	439,472	5,248,354	420,604	5,021,812	18,868	226,542
15,000 - 20,000	154,617	2,640,764	146,130	2,495,371	8,487	145,393
20,000 - 25,000	76,653	1,699,808	71,723	1,590,442	4,930	109,366
25,000 - 30,000	42,987	1,172,608	40,149	1,095,289	2,838	77,319
30,000 - 40,000	46,217	1,581,709	43,101	1,474,866	3,116	106,843
40,000 - 50,000	22,330	989,851	20,696	917,391	1,634	72,460
50,000 - 60,000	13,691	745,193	12,690	690,698	1,001	54,495
60,000 - 70,000	8,147	525,336	7,540	486,216	607	39,120
70,000 - 80,000	5,269	392,707	4,870	362,936	399	29,771
80,000 - 90,000	3,592	303,755	3,317	280,568	275	23,187
90,000 - 100,000	2,526	239,044	2,329	220,360	197	18,684
100,000 & over	9,595	1,748,749	8,763	1,584,162	832	164,587
Total	40,824,000	124,587,249	33,314,000	112,128,968	7,510,000	12,458,281

* Includes deficit class.

The distributions, therefore, still require considerable adjustment to match the independent estimate, $140.3 billion. We did not make the further adjustments and, as mentioned in Section A, our estimates must be regarded as merely a first approximation to the desired income size distribution of families and single individuals.

The estimated distributions are approximate in several respects. First, with the exception of military income we did not attempt to account for the missing income from the various sources.[50] Second, we did not adjust to render the income distri-

[50] Another correction concerns the removal of military and other income of officers who lived on posts in this country and who filed income tax returns.

butions representative of the population in 1944 rather than at the time of the Census survey, April–May 1945. Third, we did not adjust for families and single individuals in hotels, lodging houses, and other quasi-households not covered by the field survey.

Associated with the first and chief deficiency is the neglect of data on additional taxes assessed after audits of tax returns. The main reason for not using what appears to be, at first glance, a means of incorporating additional income in the size distribution is that the information from audits is difficult to generalize upon.[51] There is first the difficulty of ascertaining what the total additional assessment would have been had a representative sample of tax returns been audited. The additional assessments are minimal in the sense that they pertain to only a portion of all tax returns. Since the portion selected for investigation cannot be considered representative of all returns but are rather a byproduct of administrative practice, the full amount of income that would have been discovered had the audit been more nearly representative cannot be estimated. Secondly, it is impossible to determine what proportion of the additional assessments was due to failure to report income rather than to disallowed deductions. Since the basic income concept in our estimates is adjusted gross income, only assessments due to failure to report must be taken into account in the adjustment.

The distribution by type of income incorporated in our estimates closely approximates that of the initial tax return distribution; the amounts from the various sources in the initial tax return distribution were only slightly modified by the augmentation of the data by the Census survey frequencies and income at the lowest income levels. Apart from sources such as military income the impress of the Census survey additions must be slight. The approximate pattern of sources broadly defines the areas of required adjustment but does not readily suggest procedures. Further adjustment of the distributions will clearly

[51] According to 'The Use of Audit Reports as a Means of Correcting Statistics of Income', an unpublished paper James Turner presented at the 1949 Conference on Research in Income and Wealth, the difficulties in using present audit information may be removed to a large extent in the near future.

entail considerable arbitrariness in correcting for these sources, although certain segregable groups such as farmers may be treated more adequately.

The Lorenz curve for the distribution based on income tax data (Table 8 and Chart 4) crosses the curve for the survey distribution at approximately the 75 percent point. At least some of the differences in the two Lorenz curves is attributable to the inclusion in the constructed distribution of more military income, which accrues mainly to the middle and lower portions of the distribution. But the larger number of units and larger incomes at the higher levels in the distribution based upon tax returns are also reflected in Chart 4.

Table 8

Unadjusted Census Survey Distributions and Those Constructed from BIR and Survey Data by Size of Consumer Money Income, 1944 (thousands)

Total Income Level	Families & Single Individuals		Families of 2+		Single Individuals	
	Census	BIR	Census	BIR	Census	BIR
Under $500*	4,866	3,653	2,494	1,891	2,372	1,762
500– 1,000	4,616	3,725	3,172	2,475	1,444	1,250
1,000– 1,500	4,688	4,391	3,390	3,144	1,298	1,247
1,500– 2,000	4,326	4,431	3,480	3,424	846	1,007
2,000– 2,500	4,574	4,412	3,890	3,597	684	815
2,500– 3,000	3,864	4,259	3,520	3,704	344	555
3,000– 4,000	6,608	6,851	6,298	6,336	310	515
4,000– 5,000	3,230	4,220	3,116	4,013	114	207
5,000– 6,000	1,690	1,830	1,662	1,792	28	38
6,000– 7,500	1,053	1,330	1,029	1,295	24	35
7,500–10,000	675	897	659	861	16	36
10,000 & over	634	825	604	782	30	43
Total	40,824	40,824	33,314	33,314	7,510	7,510

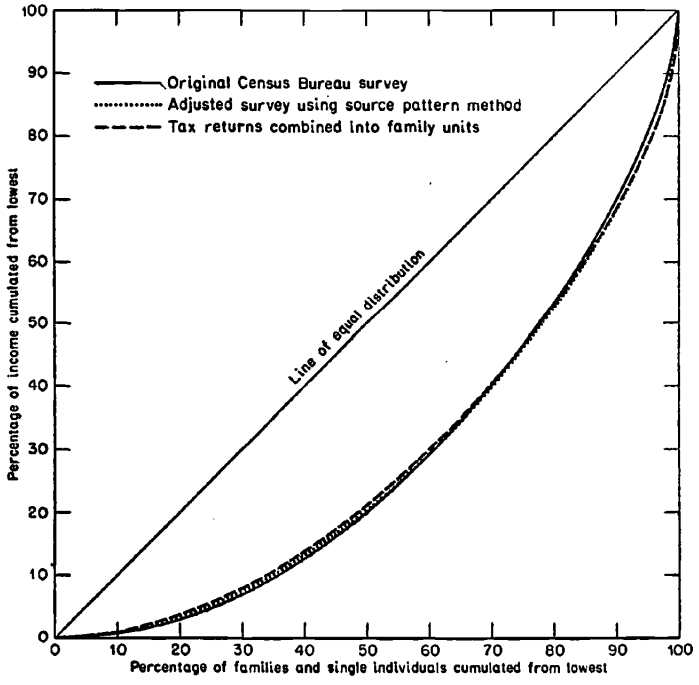
* Includes deficit class.

For purposes of comparison, a Lorenz curve of the Census survey distribution, supplemented by the assumed Pareto extension above $10,000 and adjusted by means of the source pattern transformation treated in Section A, is also shown in Chart 4.[52] Since estimates of income from the various sources included in the converted tax return distribution were not

[52] The source pattern for the $10,000 and over class, obtained by using Census survey frequencies reporting each source and BIR averages for each source to get a percentage pattern of source aggregates, was then used to distribute the income total for the class yielded by the Pareto extrapolation; see Part VI.

available, the source pattern adjustment could not be carried through to make results attain the level of total income in the converted tax return distribution and at the same time reflect a similar source composition. The source pattern adjustment, therefore, was carried through to achieve the full $140.3 billion. Hence Chart 4 may be interpreted also as a comparison of three 'blown-up' distributions, all of which achieve the $140.3 billion level of aggregate income: the original Census distribution transformed to the desired level while maintaining its original Lorenz curve; the converted income tax return distribution transformed under the same assumption; and the Census distribution adjusted by means of the source pattern transformation.

Chart 4
Lorenz Curves for the Distribution of Consumer Income



The results of the source pattern transformation are interesting in that the initial Census distribution has evidently been changed in a direction consistent with the changes brought about by the conversion of the tax data. Were the converted tax return distribution to be regarded as final and incorporating the full $140.3 billion of aggregate consumer money income, the fact that a relatively simple method, the source pattern adjustment, approximates a final distribution would be encouraging. The method requires additional empirical verification, of course, but, despite the serious limitations dealt with in Section A, it may be useful when a quick estimate is desired. Since the two Lorenz curves incorporate different aggregate incomes and different source patterns, however, we cannot conclude that the source pattern transformation would be satisfactory. The additional adjustments to the converted tax distribution to achieve the full aggregate income may, indeed, make the two curves diverge more, as is suggested by the approximate source pattern characteristics of the converted tax distribution where it can be assumed that all wages and salaries are, by and large, accounted for and that the income to be added consists largely of income normally accruing to recipients at high levels of income. The exact impress of the income items to be added cannot be assessed at this time and we can only conclude that the results obtained by converting the tax data are too partial to warrant comparisons with results of alternative methods.

By merely looking at Chart 4 we cannot, of course, judge the significance of the differences between the curves. For reasons given in Section A such curves are mainly of interest in revealing gross changes or for general illustrative purposes. We therefore examined the differences between the Census distributions and ours to determine whether they may be considered significantly larger than those arising from sampling variation.

Since the distributions derived here are based upon a complex of sample values, judgments, and statistical manipulations, the error in them is not due to sampling variation alone. The Census distributions are based upon samples, although not simple ran-

dom samples.[53] Hence in the significance tests described below the distributions derived from the tax data are regarded as populations from which the Census samples could or could not have been drawn.

Two tests were made on the distribution of all families and single individuals combined. The first was of the hypothesis that the Census distribution was a sample from the population. Chi-square, after correction for the larger standard error of the Census distribution than would have been computed for random samples of the same size, was 44.9 (for 11 degrees of freedom), a value that would occur in considerably less than 1 percent of the samples of the population. Hence this hypothesis must be rejected.

The second test was of the hypothesis that the Census distribution was a sample from a population having the same Lorenz curve as our estimates. The means of the two distributions as well as their total frequencies were equated by deflating the converted tax return distribution proportionately to achieve the same mean as the Census distribution.[54] The value of chi-square, after allowing for the increased variance as before, was 33.4 (for 10 degrees of freedom, given the additional restriction on the mean), also considerably in excess of that which would occur at the 1 percent level. Therefore, this hypothesis also must be rejected.

In view of the differences between the converted tax return and the Census distributions noted in Section C, the results of these tests are not surprising. They serve mainly to confirm the conclusions suggested by Table 3. In brief, the differences can-

53 For purposes of the tests we had to estimate the Census survey standard errors. For the 1944 survey the necessary information is not published. For 1945 published data permitted the estimate that the Census survey standard error was approximately twice that of simple random samples of the same size. The relative size of the standard errors was required as a basis for correcting the values of chi-square used in the tests. If, for example, a value of chi-square derived from a simple random sample is A, one based upon an unbiased sample that has a standard error twice as large would be A/4. This adjustment is approximate since it assumes the above factor is constant for all income levels.

54 The interpolation procedure for deflating the converted tax return distribution itself undoubtedly introduced some error. It, however, could only be small relative to any large discrepancies between the Census and the deflated converted tax distribution.

not be attributed to sampling variation but must reflect the larger questions of the reliability of survey response and general methodology.

Appendix

A Transformation Used in Estimating the Capital Gains Conditional Distributions

Let $f(x)$ represent the density function for the 1940 capital gains distribution, and $g(y)$, the transformed density function for 1944. The transformation is represented by $\log y = a + b \log x$.

We are given $M_x = \int_0^\infty x f(x)\, dx$ and $M_y = \int_0^\infty y g(y)\, dy$, where M_x and M_y are the arithmetic means of x and y respectively.

Since the distributions are dealt with in terms of class intervals, we assume that the transformation above is approximated adequately by $\log \bar{y}_i = a + b \log \bar{x}_i$, where

$$\bar{y}_i = \frac{\int_{y_i}^{y_{i+1}} y\, g(y)\, dy}{\int_{y_i}^{y_{i+1}} g(y)\, dy}, \text{ and } \bar{x}_i = \frac{\int_{x_i}^{x_{i+1}} x f(x)\, dx}{\int_{x_i}^{x_{i+1}} f(x)\, dx},$$

or $\bar{x}_i$ and $\bar{y}_i$ are the means of the i-th interval of the distribution before and after transformation. If p_i represents the proportion of the distribution in the ith interval, or the expression in the denominators of both $\bar{x}_i$ and $\bar{y}_i$ above,

$$M_x = \Sigma\, \bar{x}_i p_i \text{ and } M_y = \Sigma\, \bar{y}_i p_i \text{ added over all intervals.}$$

Since the capital gains distribution is effectively limited at the upper end by the maximum adjusted gross income of the interval, we assume that this value of capital gains has the same percentage of the distribution below it both before and after transformation, i.e., at the point x_n, $x_n = y_n$. Combining this condition with that necessary to yield the value of M_y leads to the expression

$$M_y = x_n^{1-b} \Sigma\, \bar{x}_i^b\, p_i.$$

This equation is solved for b by successive approximation.

B Estimating Capital Loss Conditional Distributions

Two assumptions are used here:
1) the conditional distribution of capital losses is lognormal between \$0 and \$1,000 (the legal limit);
2) the logarithmic coefficient of variation of the entire lognormal distribution is constant for all classes of adjusted gross income.

The distribution is given by

$$f(\log x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{\frac{-(\log x - m)^2}{2\sigma^2}\right\}$$

and the logarithmic coefficient of variation by $k = m/\sigma$
where $m = E(\log x)$, and $\sigma^2 = E[(\log x - m)^2]$.

All returns with \$1,000 or more capital loss are credited with \$1,000 only. As a result of the conditions given above, the equation for the arithmetic mean, M_x, of the distribution is derived as

$$M_x = \exp\left\{k\sigma + \frac{\sigma^2}{2}\right\} \int_{-\infty}^{t-\sigma} f(\log x)\, d(\log x) + 1{,}000 \int_{t}^{\infty} f(\log x)\, d(\log x)$$

where $t = (\log_e 1{,}000)/\sigma - k$ and k is determined from the data.

This is an equation with one unknown, σ. Given M_x, a solution can be found for σ. In practice, a table pairing values of M_x and σ was set up to facilitate solving for a large number of values of σ.

C Adjusting Cell Frequencies when a Population has Shifted

This method is designed for adjusting a 2-way quantitative classification, such as husband's income classified by wife's income, from one year to another.

Define:

$H_A(X)$ = marginal distribution function of husbands' income in year A, i.e., $H_A(X)$ yields the proportion of husbands in year A who have income less than X.

$W_A(Y)$ = marginal distribution function of wives' income in year A.

$H_A(X|Y)$ = conditional distribution function for husbands in year A, i.e., $H_A(X|Y)$ yields the proportion of husbands with incomes less than X among all husbands whose wives have incomes of exactly Y dollars.

$W_A(Y|X)$ = conditional distribution function for wives in year A.

The same notation will hold for year B, B replacing A as a subscript.

The entire cross-tabulation is available in year A and the marginal

distributions only in year B. Since all the information is given in class intervals, the conditional distributions for year A occur in the form

$$1) \qquad \int_a^b H_A(X|Y)\, d\{W_A(Y)\}; \text{ and } \int_c^d W_A(Y|X)\, d\{H_A(X)\}$$

where (a, b) and (c, d) are sets of class limits.

Transformations,

$$2) \qquad X' = \phi(X), \qquad Y' = \psi(Y),$$

may be determined empirically such that

$$3) \qquad \begin{aligned} H_B(X) &= H_A(X') \\ W_B(Y) &= W_A(Y') \end{aligned}$$

This transformation takes a marginal distribution for year A into that for year B. Assume now that this transformation holds for the entire (X, Y) plane, so that it will take a conditional distribution for year A into a corresponding conditional distribution for year B, i.e., the point (X', Y') is shifted to (X, Y). Thus

$$4) \qquad \begin{aligned} H_B(X|Y) &= H_A(X'|Y'); \\ W_B(Y|X) &= W_A(Y'|X'). \end{aligned}$$

Since the conditional distributions are actually available only in the form (1) the transformation (4) is made as

$$5) \qquad \begin{aligned} \int_{a'}^{b'} H_A(X'|Y')\, d\{W_A(Y')\} &= \int_a^b H_B(X|Y)\, d\{W_B(Y)\}; \\ \int_{c'}^{d'} W_A(Y'|X')\, d\{H_A(X')\} &= \int_c^d W_B(Y|X)\, d\{H_B(X)\} \end{aligned}$$

where (a, a'), (b, b'), (c, c') and (d, d') are all related by means of (2). Equations (5) completely determine the cross-tabulation for year B.

D Adjusting an Income Distribution for Excessive Dispersion from the Mode

When adding two variables in a cross-tabulation (see App. E2) the simplifying assumption of uniform density within each cell introduces a bias in the combined distribution. This bias is revealed by the failure to obtain the known combined aggregate due to the excessive shift of frequencies into both ends of the distribution. The effect of the assumption can be assumed to be minimal in the area of maximum

frequencies in the combined distribution. The formula given below arbitrarily assumes that the modal income does not change when the distribution is adjusted to correct for this bias. It assumes further that the transformation factor $(1 + r)$ is constant for all incomes below the mode and $(1 - r)$ for all incomes above.

Let $\bar{x}_i$ be the mean of the ith class, and f_i the frequencies in this class. Let $i = m$ be the number of the modal interval. The old aggregate is A; the new, A'.

$$1) \qquad A = \sum_{1}^{n} f_i\bar{x}_i = \sum_{1}^{m-1} f_i\bar{x}_i + f_m\bar{x}_m + \sum_{m+1}^{n} f_i\bar{x}_i$$

The transformation consists of

$$y = (1 + r)x \text{ for all } x < \text{mode},$$
$$y = x \text{ at the mode, and}$$
$$y = (1 - r)x \text{ for all } x > \text{mode}.$$

Then

$$2) \qquad A' = \sum_{1}^{n} f_i\bar{y}_i \approx \sum_{1}^{m-1} f_i(1 + r)\bar{x}_i + f_m\bar{x}_m + \sum_{m+1}^{n} f_i(1 - r)\bar{x}_i.$$

The solution for r is

$$3) \qquad r = \frac{A' - A}{\sum_{1}^{m-1} f_i\bar{x}_i - \sum_{m+1}^{n} f_i\bar{x}_i}$$

E Some Formulas Used in Constructing Our Estimates

To avoid overburdening the text and to provide a collection of useful formulas for those engaged in manipulating size distribution data, we describe some of our formulas here. Those designed for specific areas and too complex for our estimates, because of time limitations or because the reliability of the data did not warrant refinement, are omitted.

1 *Formulas to Approximate Averages in the Intervals of a Distribution*

a *First interval*

A simple formula for use in the first interval of a distribution is based on the assumption that the density function is parabolic. A function $f(x) = bx + dx^2$ is assumed such that

$$f_1 = \int_0^{x_1} f(x)\,dx, \text{ and } f_2 = \int_{x_1}^{x_2} f(x)\,dx,$$

where f_1 and f_2 are the frequencies in the first and second intervals,

and x_1 and x_2 are the corresponding upper class limits. The final formula is

1) $$\bar{x} = \frac{c(29f_1 + f_2)}{48f_1},$$

where c is the size of the first and second intervals. The formula assumes that both the density function and the first interval start at zero, and requires that both intervals be of the same size. As the value of $\bar{x}$ cannot be less than $.604c$, this formula should be used only where the mode of the distribution is above this point.

In the case of a rather large first interval, say \$500 or more, a generally superior formula may be considered worth while, one involving additional calculations and a table of areas of the normal curve. This formula is based on the fit of a lognormal curve to the bottom interval, using, as conditions to be met, the cumulative percentages under the upper limits of the first and second intervals. Thus, the initial conditions are

$$\int_{-\infty}^{\log_e x_1} f(\log_e x)d(\log_e x) = p_1, \text{ and } \int_{-\infty}^{\log_e x_2} f(\log_e x)d(\log_e x) = p_2,$$

where

$$f(\log_e x) = \frac{1}{\sqrt{2\pi}\,(b/.4343)} \exp\left\{-\left(\frac{\log_{10} x - a}{b}\right)^2\right\}$$

and p_1 and p_2 are the cumulated percentage frequencies in the first and second intervals respectively. Let $\alpha = |p - .5|$. On the basis of the values of α computed from the known values of p_1 and p_2, the table of areas of the normal curve is consulted to obtain corresponding values of z_1 and z_2, the normal deviates. The following intermediate values must be computed:

$$b = \frac{\log_{10} x_2 - \log_{10} x_1}{z_2 - z_1}; \; a = \log_{10} x_1 - bz_1;$$

$$c = b^2/.8686; \; A = \text{antilog}_{10}(a + c);$$

p' = cumulated percent corresponding to α' found in the table of areas for $z' = z_1 - b/.4343$.

The final formula becomes

2) $$\bar{x} = Ap'/p_1.$$

b *Middle intervals*

A simple and useful formula can be derived on the assumption of a straight line density function. The conditions to be met in determining the line are similar to those for the formulas above. Here the frequencies

in the two classes adjacent to the given class are used to determine the slope, b, while the constant, a, is determined independently from the frequencies in the given class. For classes of unequal size the formula becomes

$$3) \qquad \bar{x} = x_2 + c_2/2 + \frac{c_2^3(f_3/c_3 - f_1/c_1)}{6f_2(c_1 + 2c_2 + c_3)},$$

where c_1, c_2, and c_3 are the interval sizes; f_1, f_2, and f_3 the frequencies for the interval below the given interval, the given interval, and the interval above, respectively. If all intervals are of the same size and equal c, the formula reduces to *

$$4) \qquad x = x_2 + c/2 + \frac{c(f_3 - f_1)}{24f_2}.$$

In formulas 3 and 4, x_2 is the lower limit of the given interval.

If we specify that the function $y = a + bx$ cannot be negative at any point in the interval, the resulting average must lie within one-sixth of the interval width from the midpoint. For the middle intervals of an income distribution this is not a serious restriction.

Additional formulas have been developed for still other assumed density functions and for the same density function with parameters determined in various ways. We have used formulas 3 and 4 for some time, however, and found them to yield good results generally.

c *'Tail' intervals and the final open-end interval*

In the intervals of a distribution well beyond the mode, the above formulas are usually unsatisfactory. For these intervals the Pareto curve has proved extremely useful. References to this function abound in the literature and formulas for averages based on it are given here only for completeness.

For a given interval with class limits x_1 and x_2, and cumulative frequencies above these limits of F_1 and F_2 respectively, the mean of the interval is given by

$$5) \qquad \bar{x} = ab/f, \text{ where}$$

$$f = F_1 - F_2 = \text{frequencies in the given interval,}$$

$$a = F_1x_1 - F_2x_2, \text{ and } b = \frac{\log (F_1/F_2)}{\log (F_1x_1/F_2x_2)}.$$

For the final open-end interval with only x_1 and the frequencies above x_1 given, the mean can be approximated by computing the inter-

* This formula for classes of the same size can be found in *Consumer Incomes in the United States*, where it was derived from equivalent though not identical assumptions. Cf. also Durand, op. cit., p. 67.

mediate value, b, for the closed interval immediately preceding the final interval, and using the formula

$$\bar{x} = x_1 b. \tag{6}$$

2 *Formulas for Adding and Subtracting Two Variables in a Cross-tabulation* It is frequently necessary to add or subtract two variables in a cross-tabulation, that is, given a cross-tabulation of variables x and y we are required to find the distribution of $x + y = z$ or of $x - y = w$.

Ideally, if we could specify the form of the distribution in mathematical terms, $f(x,y)$, the desired distribution of z or of w could be obtained from

$$g(z) = \iint_{z} f(x,y)\, dxdy, \text{ or } h(w) = \iint_{w} f(x,y)\, dxdy.$$

Since it is unlikely that cross-tabulations dealt with in income distribution work can be approximated readily as a whole by any convenient bivariate function available at present, alternatives must be considered which combine the advantages of both good approximation and a reasonably short computation.

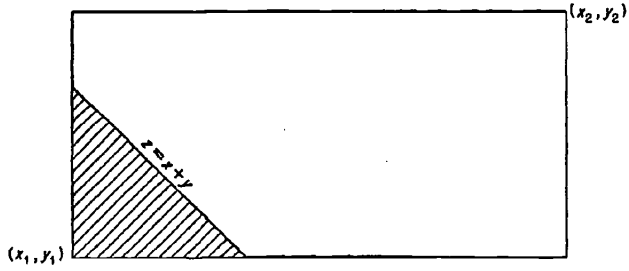
We are given the variable x with m intervals on the margin and the variable y with n intervals. The entire cross-tabulation, therefore, has mn cells. Computationally, the simplest procedure (except where all intervals have the same size; see below) is one that assigns to the frequencies, f_{ij}, in the i,jth cell, the corresponding marginal interval midpoint values of the variables, x_i and y_j. A preferable procedure, entailing little if any additional computation, is to assign the marginal interval averages, $\bar{x}_i$ and $\bar{y}_j$; that is, for the z distribution, we assign each f_{ij} at the point $\bar{x}_i + \bar{y}_j$, and for the w distribution, at the point $\bar{x}_i - \bar{y}_j$. The error lies in the assumption that neither z nor w has any dispersion within a cell. It is reasonable to expect, therefore, that the accuracy of this procedure is a function of the relative number of cells and relative cell size in the cross-tabulation. The more numerous the cells and the smaller the relative cell size, the greater the accuracy. Of course, with an increase in accuracy the method requires an increase in the amount of computation.

A fair degree of accuracy can be attained by subdividing the marginal and conditional distributions by means of interpolative methods, either graphical or by some suitable formula. A convenient procedure is to subdivide the marginal distributions by means of an interpolative formula, then obtain the frequencies for the subcells by assuming the independence of the two variables within each cell. Thus, each marginal

distribution can be divided from m x-intervals and n y-intervals into km and sn intervals respectively. The frequencies in the ks subcells within any cell can then be obtained by using the marginal distributions as controls and assuming independence, *not* uniformity. Within each subcell we may assign, as before, the marginal subinterval averages.

Another approach is to approximate mathematically the distribution surface for each cell and calculate the portions of the cell lying between specified z or w values. It permits a variety of procedures of varying complexity and computational difficulty which depend upon the form of the approximating surface assumed. The simplest among the methods assumes that all the frequencies within a given cell are uniformly distributed throughout the cell. This assumption and method, used frequently in income distribution work, have been found to be adequate in many cases and of no great computational difficulty. The nature of the surface assumed is such as to permit graphic procedures to determine the areas of each cell bounded by the lines $x + y = c$, or $x - y = c$. The rather simple formulas the assumptions yield are, perhaps, more convenient. These formulas, for both addition and subtraction, are given below.

For a given cell in the case of addition or subtraction, let the class limits of the cell be (x_1,x_2) and (y_1,y_2); the class sizes be $c_x = x_2 - x_1$ and $c_y = y_2 - y_1$; s be the smaller of c_x and c_y, and let g be the greater. For addition:



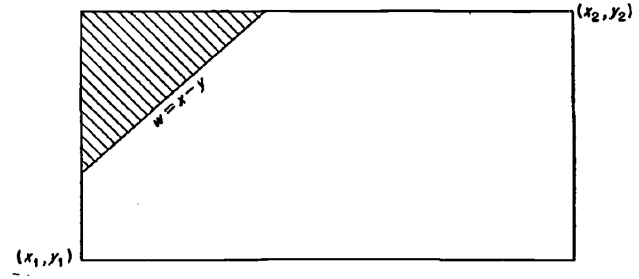
$z = x + y$. The minimum value of z in the cell is $k_1 = x_1 + y_1$. The maximum value of z is $k_2 = x_2 + y_2$. Let $P(a,b)$ represent the proportion of the total cell for which $a \leq z \leq b$. Then three formulas exist for $P(a,b)$:

7) $P(k_1,z_0) = (z_0 - k_1)^2/2c_xc_y$ when $k_1 \leq z_0 \leq k_1 + s$;

8) $P(k_1,z_0) = \dfrac{z_0 - \left(k_1 + \dfrac{s}{2}\right)}{g}$ when $k_1 + s \leq z_0 \leq k_1 + g$; and

9) $P(k_1,z_0) = 1 - (k_2 - z_0)^2/2c_xc_y$ when $k_1 + g \leq z_0 \leq k_2$.

If $c_x = c_y$, formula (8) does not apply.
For subtraction:



$w = x - y$. The minimum value of w in the cell is $k_1 = x_1 - y_2$. The maximum value of w is $k_2 = x_2 - y_1$. Let $P(a,b)$ represent the proportion of the total cell for which $a \leq w \leq b$. The three formulas for $P(a,b)$ in this case are:

10) $P(k_1,w_0) = (w_0 - k_1)^2/2c_xc_y$ when $k_1 \leq w_0 \leq k_1 + s$;

11) $P(k_1,w_0) = \dfrac{w_0 - \left(k_1 + \dfrac{s}{2}\right)}{g}$ when $k_1 + s \leq w_0 \leq k_1 + g$; and

12) $P(k_1,w_0) = 1 - (k_2 - w_0)^2/2c_xc_y$ when $k_1 + g \leq w_0 \leq k_2$.

If $c_x = c_y$, formula (11) does not apply.

Since, under the assumption of uniform density, interpolation within each cell of a cross-tabulation is dependent only upon the cell dimensions, the divisions, once obtained, can be used repeatedly for all later combinations involving the same cell configurations.

The task of adding two variables under the assumption of uniform density within each cell is considerably simplified when all the cells are of the same size. In income distribution work this condition is rarely realized, but if the same 'and over' class is desired in the combined distribution as the ones occurring in the two initial distributions (the usual case), the statement about simplification still holds. The case of subtracting one variable from another cannot be simplified in the manner discussed below unless *all* cells are of the same size. Therefore only addition will be considered here. In some instances prior interpolation on the marginal and conditional distributions is recommended to achieve this uniformity of cell size.

Given the assumption of uniform density and cells of uniform size, the operations required are simple addition and division by 2.

Let the interval size be c uniformly for both variables, except in the 'and over' class, and let a_{ij} be the frequencies in the i,jth cell. Both variables have zero origin.

variable y \ variable x	0	c	$2c$	$3c$	. . .	$(n-1)c$ and over
0	a_{11}	a_{12}	a_{13}	. . .		a_{1n}
c	a_{21}	a_{22}	a_{23}	. . .		a_{2n}
$2c$	a_{31}	a_{32}	a_{33}	. . .		a_{3n}
$3c$	.	.	.			
. . .	.	.	.	. . .		
$(n-1)c$ and over	a_{n1}	a_{n2}	a_{n3}	. . .		a_{nn}

The computational procedure is outlined in the accompanying table (the order of column elements and of columns is identical with that in the original cross-tabulation). Column F is the desired frequency distribution.

Classes of the combined variable $(x+y)$			t	F
$0-c$	a_{11}	=	t_1	$\frac{t_1}{2}$
$c-2c$	$a_{21}+a_{12}$	=	t_2	$\frac{(t_1+t_2)}{2}$
$2c-3c$	$a_{31}+a_{22}+a_{13}$	=	t_3	$\frac{(t_2+t_3)}{2}$
. . . . .	. . . . .	. . .	. . .	. . . . .
$(n-2)c-(n-1)c$	$a_{n-1,1}+a_{n-2,2}+\ldots+a_{1,n-1}$	=	t_{n-1}	$\frac{(t_{n-2}+t_{n-1})}{2}$
$(n-1)c$ & over	Sum of all remaining frequencies	=	t_n	$\frac{t_{n-1}}{2}+t_n$

We have derived formulas assuming more elaborate approximation surfaces for individual cells, such as functions of the kind, $f(x,y) = a + bx + cy$, $f(x,y) = a + bx + cx^2 + dy + ey^2$, etc., as well as functions that approximate more closely the surfaces in cross-tabulation border and modal cells. The formulas these various approximating surfaces entail all require substantially more computation than those given above which assume a uniform distribution within each cell. In addition, experience with these more elaborate formulas is still rather sketchy and incomplete. They are consequently not given here.

3 *Interpolation Formulas*

a *Interpolation for frequencies and aggregates*

i) The following procedure gives generally satisfactory results in interpolating in almost every part of an income distribution for both frequencies and aggregates. The formulas are based upon the assumption that a straight line density function adequately represents the frequencies in any interval. Thus, a function $f(x) = m_1 + m_2x$ is assumed such that

$$F = \int_{x_i}^{x_{i+1}} f(x)\,dx, \text{ and } A = \int_{x_i}^{x_{i+1}} xf(x)\,dx$$

where A is the known aggregate income and F the known frequency in the given interval. The final formulas for both frequencies and aggregate become

$$13) \qquad f = \frac{Fkz(4 - 3z) - 6z(1 - z)(A - Fx)}{k}, \text{ and}$$

$$14) \qquad a = 2Fkz^2(1 - z) - z^2(3 - 4z)(A - Fx_0) + fx_0,$$

where f = the frequencies lying between x_0 and x, a is the aggregate in this portion of the interval, F and A are defined above, x_0 is the lower limit of the interval, k the size of the interval, x the income at the point of interpolation, and $z = (x - x_0)/k$.

In addition to permitting the computation of aggregates based upon the assumptions used to interpolate for frequencies, the formulas have the merit of giving more weight to any departure from strict continuity (as revealed by the aggregates in the class) than do similar formulas based upon assumptions of continuity with the frequencies in adjacent intervals.

ii) When the first interval starts at zero, a lognormal curve may conveniently be used for interpolation. The curve (on lognormal or probit paper for quick results) is fitted to cumulative percentage fre-

quencies in the first two intervals. Arithmetically, this method involves replacing the given income level, x, by log x, and the corresponding cumulative percentage below x by z, the normal deviate read from tables of the normal curve. Linear interpolation in this series of (log x, z) values for given levels of log x yield corresponding z values, and finally, the desired cumulative percentage frequencies.

If percentages are taken of money amounts below given x levels, the procedure can be used to interpolate for income at the selected x levels.

iii) The Lagrange method of interpolation fits a polynomial to cumulative frequencies starting from any point in the distribution. Because the method involves a set of weights that are constant when the intervals and interpolation points remain the same, it is especially useful where similar interpolation must be made in several distributions.

Given a point in the distribution, the corresponding cumulative frequencies may be calculated. Formulas for fits through 2, 3, and 4 points are given below, and extensions to fits through n points will be apparent.

Let x_0 be the starting point for the cumulation of frequencies. Let F_i represent the cumulative frequencies from x_0 to x_i. $F_0 = 0$. Then, for a polynomial through two points, x_0,x_1, the formula is

$$15) \qquad F_i = F_1 \frac{(x_i - x_0)}{(x_1 - x_0)}.$$

A polynomial through three points, x_0,x_1, and x_2, yields

$$16) \qquad F_i = F_1 \frac{(x_i - x_0)(x_i - x_2)}{(x_1 - x_0)(x_1 - x_2)} + F_2 \frac{(x_i - x_0)(x_i - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

A four point fit yields

$$17) \quad F_i = F_1 \frac{(x_i - x_0)(x_i - x_2)(x_i - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} + F_2 \frac{(x_i - x_0)(x_i - x_1)(x_i - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} + F_3 \frac{(x_i - x_0)(x_i - x_1)(x_i - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)}.$$

The same weights applied to various values of F in these formulas may be applied to corresponding values of A, the cumulated aggregate amounts, to find the value of A_i at a given x_i. (In the latter case at least a four point fit should be used.)

iv) For the portion of the distribution referred to as the 'tail', the Pareto curve is a convenient base for interpolation or extrapolation. The lower limit, x, of any interval is simply replaced by log x, and the cumulative frequency F above this point by log F. Simple linear inter-

polation between adjacent points of (log x, log F) at selected values of log x yields the corresponding values of log F. Extrapolation may be carried out similarly.

We may take A as representing the cumulative aggregate above a point x and interpolate linearly between adjacent values of (log x, log A) as above.

b *Interpolation for deciles and other quantiles*

i) The assumptions in Section a (i) provide a generally satisfactory formula for decile or other quantile interpolation. Given the frequency F and the aggregate A in an interval, x_0 to x_1, we can interpolate for the aggregate up to a point x corresponding to the frequency f in the interval up to that point.

The density function constants, m_1 and m_2, are first computed in

$$m_1 = \frac{2}{k}\left(2F - \frac{3A'}{k}\right), \text{ and } m_2 = \frac{6}{k^2}\left(\frac{2A'}{k} - F\right),$$

where $k = x_1 - x_0$, and $A' = A - Fx_0$. The final formula becomes

$$18) \qquad a = m_1\frac{(x - x_0)^2}{2} + m_2\frac{(x - x_0)^3}{3} + fx_0, \text{ where}$$

$x - x_0 = -(m_1/m_2) \pm \sqrt{(m_1/m_2)^2 + 2f/m_2}$, and a is the aggregate income in the interval below the point x.

ii) As in the case of interpolation for frequencies and aggregates in a(ii) the lognormal curve can be used also for quantile interpolation. As before, the normal deviates z_1 and z_2 are calculated from p_1 and p_2, the cumulative frequency percentages below x_1 and x_2. Similar deviates, z'_1 and z'_2, are obtained from p'_1 and p'_2, the cumulative aggregate percentages below x_1 and x_2. Linear interpolation between the points (z_1,z'_1) and (z_2,z'_2) for a given value of z_i yields a corresponding value of z'_i which is then converted back to p'_i.

iii) Lagrange interpolation formulas too are convenient for quantile interpolation. The polynomial is fitted to points defined by both cumulative frequencies and aggregates. As before, let x_0 be the starting point in the distribution for cumulating both frequencies and aggregates. $F_0 = 0$ and $A_0 = 0$. At the point x_i the corresponding cumulated frequencies and aggregates are F_i and A_i. The two point formula is not included here because it is inadequate. The three point formula is written as

$$19) \qquad A_i = \frac{A_1(F_i - F_0)(F_i - F_2)}{(F_1 - F_0)(F_1 - F_2)} + \frac{A_2(F_i - F_0)(F_i - F_1)}{(F_2 - F_0)(F_2 - F_1)}.$$

The four point formula is

$$20)\quad A_i = \frac{A_1(F_i - F_0)(F_i - F_2)(F_i - F_3)}{(F_1 - F_0)(F_1 - F_2)(F_1 - F_3)} + \frac{A_2(F_i - F_0)(F_i - F_1)(F_i - F_3)}{(F_2 - F_0)(F_2 - F_1)(F_2 - F_3)} + \frac{A_3(F_i - F_0)(F_i - F_1)(F_i - F_2)}{(F_3 - F_0)(F_3 - F_1)(F_3 - F_2)}.$$

iv) Quantile interpolation in the upper intervals of a distribution, where the Pareto curve is a good fit, can be quite easily accomplished by simple linear interpolation on transformed variables. For the Pareto curve it can be shown that the relation between the logarithm of frequencies F above a given point and the logarithm of aggregate income A above this point is linear. Thus $\log A = a + b \log F$.

F Obtaining Percentage Patterns Required to Separate Income Tax Returns into Various Categories of Family Membership

1 *Adjusting Census Bureau Earnings Distributions for Differences in Income Concept*

After correction for capital gains and losses the BIR data lacked two main components of income that are included in the Census Bureau concept of consumer money income:[1] military and 'other' income as the Census tabulations call them.[2] Of the two, military income was by far the larger. Of the amount reported in the Census survey, $6 billion, only a very small portion could be expected to be in the BIR distributions, namely, military pay in excess of $1,500 paid to officers who lived with their families at the time of the Census survey and who filed income tax returns. It was assumed therefore that none was in the tax distributions. Some of the military and other income, however, of officers living on posts in this country and who filed income tax returns—income specifically excluded from the Census survey concept—was, of course, in the tax distributions, introducing an element of incomparability between Census and BIR data. Some components of 'other' income are taxable and therefore supposedly in the tax tabulations. However, social security payments

[1] See Part VI for a description of differences in income concept between survey and tax tabulations.

[2] 'Other' income as given in Census tabulations for 1944 includes social security payments, periodic life insurance payments, fiduciary income, unemployment, workmen's compensation, etc. See ibid. for independent estimates of this component as well as for amounts reported in the Census survey for 1944.

and public assistance grants, which constitute a substantial proportion of 'other' income, are expressly not taxable and therefore not in the tax data.

Since the basic Census size distributions were by earnings classes, the adjustment to the income concept used in tax returns required changing the basis of classification in the Census tabulations to consumer money income excluding both military and 'other' income. However, because of the lack of suitable data no adjustment was made for the latter. Failure to make this adjustment introduces a certain degree of incomparability into survey and BIR tabulations. The effect would be expected to be largest at the lowest income levels both because of the heavy component of social security payments and the fact that some of the additional constituents of 'other' income are already included in the BIR tabulations.[3]

Several cross-tabulations based upon Census survey data permitted the approximate removal of military income. One consisted of a tabulation of individuals receiving military income by size of such income and by size of income other than civilian earnings (i.e., the sum of military income, interest, dividends, rents, and 'other' income). The other consisted of a tabulation of individuals receiving military income by size of such income and by size of civilian earnings (i.e., civilian wages and salaries plus entrepreneurial income). In addition, source pattern data were provided in special Census tabulations which gave for each level of consumer money income the number of families (or single individuals) receiving income from each specified source and the amounts.[4] The latter tabulation permitted the immediate allocation of consumer money income minus military income between single individuals and families of 2 or more.

For single individuals it was possible, on the basis of the above

[3] The small average amount of social security payments, together with the limitations on recipients set by law and administrative practice, suggests that the recipients of this income are mainly in the lowest income intervals, and hence below the level of filing requirement for tax purposes. This fact is borne out by Census data on family income source patterns where a strikingly large proportion of income at the lowest levels of family income comes from this source. On the basis of income size distributions of individuals or of husband-wife units rather than of families, this income could be expected to be even more highly concentrated at the lowest income levels than was evident from the Census data.

[4] The sources were civilian wages and salaries, farm entrepreneurial income, nonfarm entrepreneurial income, military income, interest and dividends, rent, and 'other' income.

tabulations, to construct a cross-tabulation relating total consumer money income as defined in the Census survey to consumer money income minus military income, which permitted shifting individuals down the income scale.[5]

Removing military income from the other distributions requiring adjustment was more difficult. The source pattern tabulations for families and for single individuals, used directly in the case of single individuals, could not be applied to the individuals composing families. Consequently, consumer money income minus military income was allocated to each distribution in the initial Census tabulations (see text, Section A3, for these distributions) and appropriate cross-tabulations relating civilian earnings to consumer money income minus military income for each were constructed.

Actually, it was not necessary to allocate civilian earnings to the various groups in the Census tabulations, since they were known for these groups, i.e., the tabulations were by size of civilian earnings. The allocation problem was thus limited to income other than civilian earnings and military income, i.e., to interest, dividends, rent, and 'other' income. Since these are relatively minor sources compared with earnings, the area of possible error involved in the allocation was narrowed.[6]

These cross-tabulations were derived from the Census tabulations previously mentioned. The assumption that persons receiving military income did not receive additional income other than civilian earnings was warranted by the Census cross-tabulation between mili-

[5] More specifically, the data used in constructing this cross-tabulation were the following, all of which were from various Census tabulations: (a) a distribution of military income by its own size, (b) the number of individuals receiving military income by level of total consumer money income, and (c) the average military income at each level of total consumer money income.

On the basis of these knowns it was possible to approximate the number of individuals having military income only (individuals who are ranked in the zero consumer money income class after the removal of military income) and to assign to the other individuals at each level of total consumer money income an appropriate dispersion pattern by size of military income consistent with the known distribution of military income by size. Since the source pattern data were given for many groups within the single individual category the varying average amounts of military income at each level of total income itself were helpful in suggesting the required dispersion.

[6] This allocation could not be made without considerable arbitrariness. It was based primarily upon average amounts of such income for single individuals which were known, the known amount for all relationship groups combined, and notions of the likely recipients of such income within the family structure. The amounts assigned agreed with the known totals and seemed reasonable.

tary income and income other than civilian earnings which showed a strong diagonal with few divergences. Subtracting the cross-tabulation between civilian earnings and military income from that between civilian earnings and consumer money income other than civilian earnings gave a cross-tabulation between civilian earnings and consumer money income excluding both civilian earnings and military income. Addition of the two variables in the last cross-tabulation permitted the construction of another: between civilian earnings and consumer money income minus military income, the goal.[7]

The last cross-tabulation, used for each of the four groups of family members in the Census tabulation (heads of normal families, wives of heads, heads of broken families, and other relatives of heads) on the assumption that the conditional distribution by size of consumer money income minus military income for each level of civilian earnings was the same for persons in each group having 'other' income,[8] gave distributions by size of consumer money income minus military income for each of the four groups.

2 *Distribution of Members of Subfamilies*

It will be recalled that individual members of subfamilies classified by earnings classes were combined with family supplementary income recipients in the Census distribution of 'other relatives of head'. To determine the proper classification of BIR returns, it was necessary to remove these members of subfamilies from the Census earnings distribution. If the Census tabulation had had an independent earnings distribution of family supplementary income recipients, the procedure would have been simple subtraction of such recipients from the given earnings distribution. Unfortunately, Census data on family supplementary income recipients were for a later year and, more important, defined such recipients as supplementary to the principal earner of the family, a classification not useful for purposes of separating different types of tax return.

[7] In a strict sense this last cross-tabulation is superfluous since the previous one was itself sufficient. The additional step was taken to minimize later computations using the cross-tabulations. Thus, the laborious task of adding distributions (done here on the assumption of uniform density throughout each cell of the cross-tabulation) was done only once rather than four times when applied to the various distributions of heads of families, etc.

[8] This did not imply, of course, that the amounts of such income were identical at each earnings level for all groups.

Since no data were available on the number of subfamilies in 1944 it had to be estimated from later data. Data for 1946–47 gave information on the characteristics of attachment to main families, and on the number of normal and broken subfamilies in 1946 and 1947; also some information on the number of such families for 1945.[9] On the basis of these figures the number of subfamilies in 1944, in both the normal and broken categories, was estimated. For the income size distribution of such families, resort was had to unpublished Census data giving such distributions for normal and broken subfamilies in 1946. In the absence of similar data for 1944, these distributions were used after adjustment for differences in the number and in income level between 1944 and 1946.[10]

These adjusted distributions of normal and broken subfamilies had to be divided into distributions of individuals composing them before they could be used to segregate supplementary family income recipients in the Census tabulation of 'other relatives of head'. Certain simplifying assumptions were necessary. First, it was assumed that no members of subfamilies except the husband and/or wife were 14 or older. This assumption, which is quite tenable in view of probable subfamily composition, permitted the broken subfamilies size distribution to be treated like individuals because by definition they contained only one spouse, i.e., only the husband or the wife could be the recipient of income. In addition, the assumption limited the number of income recipients in normal subfamilies to 2 at a maximum. Thus, for normal subfamilies there were 4 possibilities with respect to the receipt of income by the constituent members: neither husband nor wife had incomes, only the husband had income, only the wife had income, or both had income.

Under the assumption that husband-wife units in normal subfamilies in 1944 had the same proportion with zero incomes as did similar units in 1946 the number of normal subfamilies in each of the 4 categories could be determined.[11]

9 Census Bureau releases, P-S, No. 15, 'Characteristics of Secondary Families in the U.S.: Feb. 1946'; P-20, No. 17, 'Characteristics of Families and Subfamilies in the U.S.: April 1947'.

10 The adjustment for the difference in income between 1944 and 1946 was quite arbitrary. The Lorenz curve was retained for each of the two categories of subfamilies, and income was adjusted downward 10 percent on the ground that this represented the change in per capita income between 1944 and 1946.

11 Specifically, a 4 cell cross-tabulation was constructed giving the percentage distribution of individual husbands with and without income on one margin and a similar distribution of percentages for wives on the other. Since the cell consisting of zero husbands associated with zero wives was taken from the 1946 data, the entries in the remaining 3 cells were determinate.

To obtain income size distributions for the 3 groups with incomes, additional assumptions were introduced concerning how husband's and wife's incomes were related in the same subfamily. Recourse was had to a BIR cross-tabulation of joint returns with two incomes giving husbands' incomes crossed by wives' incomes. This tabulation was held to be more applicable to the case of husbands and wives in subfamilies than the previously derived cross-tabulation relating the incomes of husbands and wives filing separate returns. On the basis of the cross-tabulation and the previously obtained estimates of the number of subfamilies in each of the 4 categories, the income size distributions of individual husbands and wives could be approximated.[12]

3 *Distribution of Husbands and Wives in Normal Main Families from Census Data*

Husband-wife units represented by joint returns and combined separate returns contained units from both main and subfamilies. Subfamilies had to be removed from the income size distribution so that main and subfamilies could be combined into family units. The simplest procedure assumed that the distribution for normal, i.e., husband-wife, subfamilies estimated above was identical with that in the tax distributions. This choice, however, neglected the likelihood of a substantial difference between the income in tax tabulations and that in the Census tabulations of subfamilies. For example, the subfamily distribution from the survey may have been pitched substantially lower than the tax return distribution of identical units. If so, the husband-wife main family units remaining after the subtraction of subfamilies would be seriously in error. We therefore related the estimated distribution of husband-wife subfamilies to that of husband-wife main units from Census data to obtain percentage patterns for each income level.[13] Income size distributions of

[12] The method for obtaining the individual distributions made two assumptions: (a) the distribution of 1-income subfamilies where the husband (wife) has the income was the same as that of husbands (wives) in 2-income subfamilies, and (b) the distribution of husbands' (wives') incomes at a given level of combined income was the same as that in cross-tabulations constructed from the joint return tabulation referred to above; the constructed cross-tabulations were those of husbands versus joint incomes, and wives versus joint incomes. The first can be entertained with a fair degree of reasonableness while the second follows from the decision to use the joint return cross-tabulation referred to above. For details on the method for obtaining the individual distributions based on the data and stated assumptions, see the mathematical note in Section 4.

[13] The decision to base percentage patterns on Census data alone, made in the belief that differences between Census and BIR income size distributions are not

normal main families had therefore to be constructed from the Census data.

The initial distributions used were those of heads of families and wives of heads, i.e., husbands and wives in main families, which, together with the distribution of 'other relatives of heads', had been converted from a civilian earnings to a consumer money income minus military income classification.[14] Before the income of husbands and wives could be combined, the number of husbands and wives in each of the four categories was determined as in the case of sub-families.[15]

Individual husbands and wives in two income husband-wife units were combined on the basis of the BIR cross-tabulation of joint returns with two incomes already referred to. This cross-tabulation, because it is confined to persons filing joint returns, was not, of course, strictly applicable because the Census Bureau husband-wife units included persons filing both joint and separate returns. Ideally, a combined cross-tabulation might have been constructed incorporating both types of return. This more elaborate procedure was abandoned mainly because of lack of time. In the absence of such a composite, a cross-tabulation of joint returns (confined to noncommunity property returns) from the BIR was used because it approximated most closely the area of the distribution covered in

inconsistent with the assumption of identical percentage patterns at given income levels, is directly comparable with that concerning the treatment of the component distributions of single returns. In the latter instance it would have been possible also to subtract the estimated size distributions for single individuals and broken families (main and sub-), obtained from Census data as described above, directly from the BIR data to get the size distribution of family supplementary income recipients as a residual.

[14] The procedure for the combination of husbands' and wives' income would have been obviated had the desired tabulations been available from the survey material. Survey results, for all their deficiencies, often prove exceedingly useful if their information is fully exploited. Such exploitation, however, necessitating prior design of schedules and tabulation for the adjustment of distributions, forces a partial abandonment, if resources are limited, of the intention to gather and tabulate only information that can be directly presented to the public, since many tabulations useful in adjustments have little direct value.

[15] The margins of the cross-tabulation, consisting of the distributions of husbands and wives each by the categories 'with' and 'without' income, provided both upper and lower limits to the number of husband-wife units with zero incomes. The limits suggested, however, were too broad and a value for the cell was assigned after considerable experimentation. The number finally selected was from additional Census data on the proportion reporting zero incomes for all 2 person families combined.

detail by the Census survey.[16] Unlike the procedures for combining the incomes of husbands and wives filing separate returns the cross-tabulation of joint returns with two incomes was not modified before use.[17]

4 *Distributions of Individuals in Normal Subfamilies*

The number of intervals in the individual distributions, the same as in the combined distribution, equals k.

In the cross-tabulation of husbands in 2-income families, let h_{ij} represent the percentage, at the ith level of individual income, of all husbands at the jth level of combined income.

Hence

$$\sum_{i=1}^{k} h_{ij} = 1 \text{ for all } j.$$

For wives defined similarly and denoted by w_{ij} we have

$$\sum_{i=1}^{k} w_{ij} = 1 \text{ for all } j.$$

At family income level i, let

n_i = number of all normal subfamilies,

n_{i1} = number of 2-income normal subfamilies,

n_{i2} = number of 1-income subfamilies in which the husband is the recipient,

n_{i3} = number of 1-income subfamilies in which the wife is the recipient,

16 This noncommunity cross-tabulation of joint returns with two incomes (Preliminary Study of Individual Income Tax Returns for 1944, Oct. 1945) is itself suspect on the basis of an examination of some returns. Some returns filed for 2-income husband-wife units were, in all probability, 1-income returns, since the cross-tabulation manifests a dominant diagonal. Our estimates were not corrected for this factor although some of the procedures soften the effect. The uniform density assumption in adding the two marginal distributions, for example, distributes the effects of the prominent diagonal over a wide range. Moreover, survey results, in including data from community property states, may well improperly manifest a similar diagonal.

17 Time limitations again must be offered as the reason for neglecting the more elaborate procedure. In using the cross-tabulation in the present instance it was assumed that the percentage patterns of joint income at each level of individual income were maintained. These percentages were obtained for wives and husbands by joint income separately, with subsequent interpolation on cumulated frequency curves to achieve the correct aggregate. The method, tried experimentally on separate returns, was reasonably satisfactory.

N, N_1, N_2 and N_3 represent the appropriate totals for all income levels,

n^0_{i2} = number of husbands at individual level i from the appropriate 2-income cross-tabulations, and

n^0_{i3} = number of wives at individual level i as above.

Under the assumption in the text,

$n_{i2} = n^0_{i2} N_2/N_1$, and $n_{i3} = n^0_{i3} N_3/N_1$,

where

$n^0_{i2} = h_{i1}n_{11} + h_{i2}n_{21} + \ldots + h_{ik}n_{k1}$, and

$n^0_{i3} = w_{i1}n_{11} + w_{i2}n_{21} + \ldots + w_{ik}n_{k1}$.

In matrix notation

$(n_i)' = (n_1 \ n_2 \ n_3 \ \ldots \ n_k)$,

$(n_{i1})' = (n_{11} \ n_{21} \ n_{31} \ \ldots \ n_{k1})$,

$(n_{i2})' = (n_{12} \ n_{22} \ n_{32} \ \ldots \ n_{k2})$, etc.

Also

$$(h_{ij}) = \begin{pmatrix} h_{11} & h_{12} & \ldots & h_{1k} \\ h_{21} & h_{22} & \ldots & h_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ h_{k1} & h_{k2} & \ldots & h_{kk} \end{pmatrix}$$

and

$$(w_{ij}) = \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1k} \\ w_{21} & w_{22} & \ldots & w_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ w_{k1} & w_{k2} & \ldots & w_{kk} \end{pmatrix}$$

Then

$(n^0_{i2}) = (h_{ij})(n_{i1})$,

$(n^0_{i3}) = (w_{ij})(n_{i1})$

yielding, by simple substitution,

$$(n_{i2}) = \frac{N_2}{N_1}(h_{ij})(n_{i1}) \text{ and}$$

$$(n_{i3}) = \frac{N_3}{N_1}(w_{ij})(n_{i1}).$$

Except for the zero class, $(n_i) = (n_{i1}) + (n_{i2}) + (n_{i3})$,

or $$(n_i) = (n_{i1}) + \frac{N_2}{N_1}(h_{ij})(n_{i1}) + \frac{N_3}{N_1}(w_{ij})(n_{i1})$$

$$= \left(\delta_{ij} + \frac{N_2}{N_1}h_{ij} + \frac{N_3}{N_1}w_{ij}\right)(n_{i1}),$$

where $\delta_{ij} = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ for } i \neq j. \end{cases}$

Symbolically the solution for (n_{i1}), the set of k unknowns, is expressed as $(n_{i1}) = \left(\delta_{ij} + \frac{N_2}{N_1} h_{ij} + \frac{N_3}{N_1} w_{ij}\right)^{-1} (n_i)$.

G Allocating Population Groups to Categories of Family Attachment

Three tabulations from the Census survey for 1944, covering the population as of April–May 1945, were basic in deriving the estimates: (1) a distribution of all families of two or more persons by number of family members of all ages, (2) a distribution of normal families by the number of children under 18, and (3) a tabulation of the number of relatives of family heads (including subfamily members as well as supplementary income recipients). Other information was obtained from the Census survey for 1945 which reflected much more detail on family composition. An outline of how this information was used follows.

1 *Number of Family Members 14 and Older in Normal and Broken Families*

The general procedure was to use data from the Census survey for 1945 by means of which information in tabulations (1) and (2) above could be used to obtain required estimates for 1944, i.e., the number of persons 14 and older for both normal and broken families. This information was not given in any Census tabulation for that year but for 1945 the Census Bureau provided a fund of information: the distributions of both normal and broken families by the number of family members of all ages; a cross-tabulation for normal and broken families relating the number of family members under 18 to the number over 18; a cross-tabulation relating the number of children 14–17 to the number under 14 for both normal and broken families.[18]

The second cross-tabulation yielded by addition the estimated number of individuals of all ages combined in normal and broken families. However, since it had relatively few size classes on each margin, the first distribution was used as a control in the necessary

[18] Census release P-46, No. 8, Tables 3, 4, and 7.

extrapolation.[19] When the margin 'children under 18' was converted to 'children 14–17', by using the third cross-tabulation, the number of individuals 14 and older in normal and broken families was obtained.[20] This cross-tabulation permitted the construction of still another, classifying families by the number of members of all ages on one margin and by the number of members over 14 on the other. This table, for 1945, was used for 1944 by assuming that the conditional distributions (at each level of the number of family members of all ages) was maintained, and by introducing the given 1944 distribution of families of 2 or more by size of family into the margins. The resulting estimates of the number of individuals 14 and older in normal and broken families in 1944 were then checked with the given figure for 1944 for the number of 'other relatives of heads'. The error, approximately 2 percent, was adjusted for arbitrarily.

2 *Subfamily Groups Included in Normal Families*

Data permitting estimates of subfamily attachment could be obtained only for years after 1944. Information, given for February 1946 and for April 1947, was used after some adjustment.[21] The publication for 1946 provided a 4-cell cross-tabulation giving the number of normal and broken families with normal and broken subfamilies present. This cross-tabulation was used to associate subfamilies for 1944 as previously estimated.

The data for April 1947 were relied upon, after adjustment to 1944, to give classifications by size of family for normal and for broken families with normal and broken subfamilies present. For normal families with subfamilies, divisions by family size, confined to the number of persons 14 and older, were obtained by using the previously constructed cross-tabulation relating size of family to the number of persons 14 and older. Similar classifications by the number of persons 14 and older could not be derived for broken families with subfamilies (see text for the method used for this group).

19 Details of the construction of this cross-tabulation are not given here. In outline the method of obtaining the cells involved assumptions of independence between the categories under and over 14, given the controls furnished by the known distribution of families by the number of persons of all ages combined and the known distribution of families by number of persons under 14.

20 The margin was converted on the assumption that the distribution of families having a given number of persons 14–17 by families classified by number of persons over 18 was the same as that for families having the same number of children under 18 years of age. This assumption can be taken as approximate since the former category is a subgroup of the latter.

21 Census releases P-S, No. 15; P-20, No. 17.