ERS Staff Paper

Economic
Research
Service

Food and
Rural
Economics
Division

Number 9904

Topcodes and the Great U-Turn in Nonmetro/ Metro Wage and Salary Inequality

John Angle
Charles M. Tolbert

Topcodes and the Great U-Turn in Nonmetro/Metro Wage and Salary Inequality. John Angle and Charles M. Tolbert. Food and Rural Economics Division. Economic Research Service. U.S. Department of Agriculture. Staff Paper No. 9904.

Abstract

Part of the perceived increase in wage and salary inequality in the early 1980's may be due to social scientists using Bureau of the Census topcodes in Current Population Survey (CPS) data as if they were valid incomes. A topcode is the number that the Bureau of the Census substitutes for a reported income bigger than the maximum disclosable income in CPS public use sample files. Large incomes are rare and, consequently if disclosed, might allow the respondent to be identified, thus breaching the pledge of confidentiality from the Bureau of the Census to the respondent. In the 1960's, 1970's, and 1980's, the Bureau of the Census used the maximum disclosable income itself as the topcode. Estimating a measure of inequality using a topcode as if it were a valid observation on an income yields an underestimate. This downward bias was acute in the late 1970's when the Bureau of the Census did not raise its maximum disclosable income in a time of rapid inflation. It did though in the early 1980's, making some of the measured increase of income inequality in the early 1980's artificial. This downward bias on estimates of inequality in the late 1970's affected metro income inequality more than nonmetro because metro areas have a higher proportion of large incomes. Nonmetro inequality is historically higher than metro inequality. The nonmetro/metro gap in inequality was overestimated in the late 1970's.

Keywords: Great U-Turn, income, inequality, metro, nonmetro, topcodes

Acknowledgments

We are indebted to Dr. David McGranahan, former Chief of the Population, Labor, and Income Branch of the Rural Economy Division of the Economic Research Service for recommending Tolbert and Lyson (1992) as a landmark paper in the field of nonmetropolitan income distribution. Tolbert and Lyson (1992) deal with many issues beyond those examined here.

Washington, DC 20036-5831

August 1999

Contents

Summary

Many scholars have reported the reversal of a trend toward lower inequality in U.S. income distribution around 1980. This reversal has been called the Great U-Turn. Tolbert and Lyson (1992) examine inequality in nonmetro and metro earned income from the 1960's through 1990 using three measures of inequality: the variance of the logarithms of income, Theil's measure, and the coefficient of variation. The latter two measures support the Great U-Turn scenario circa 1980; the first measure supports a Great U-Turn scenario around 1970. This difference in timing is a consequence of Tolbert and Lyson's use of the Bureau of the Census' income topcodes. The Bureau of the Census substitutes a topcode for a valid income when that valid income exceeds the maximum disclosable income in its public use micro-data files. Tolbert and Lyson's research is based on the March Current Population Survey (CPS). In the CPS data that Tolbert and Lyson analyze, the topcode used by the Bureau of the Census is the maximum disclosable income itself. Thus, for example, a wage and salary income of $200,000 reported for calendar year 1977 in the March 1978, CPS would appear in the CPS micro-data file as $50,000, the maximum disclosable wage and salary income in that year.

While using topcodes as if they are valid income observations is not usually a problem, this practice became problematic in the late 1970's because the Bureau of the Census did not increase the topcode during a period of rapid inflation, thus lowering it in constant dollar terms. In the late 1970's, so many incomes were topcoded that estimates of measures of inequality were biased downward. Some inequality measures are more sensitive to this bias than others. This bias distorted the measured gap between nonmetro and metro inequality in the late 1970's since a larger proportion of metro incomes were topcoded than nonmetro. The measured gap between the inequality of nonmetro and metro incomes was enlarged by the artificially low estimates of metro income inequality. Nonmetro inequality is historically higher than metro inequality. The nonmetro/metro gap in inequality was overestimated in the late 1970's.

Topcodes and the Great U-Turn in Nonmetro/Metro Wage and Salary Inequality

John Angle
Charles M. Tolbert

Introduction

There has been great concern among social scientists about a major change in income inequality patterns in the United States around 1980. Much of the social science literature has concluded that a long period of decreasing income inequality came to an end during this time to be followed by a decade of increasing inequality. This turnaround was termed the "Great U-Turn" by Harrison, Tilly, and Bluestone (1986a, b, c). See also Kuttner (1983), Lawrence (1984), Thurow (1984, 1987), Harrison and Bluestone (1988), Levy (1988), Burtless (1990), the *Quarterly Journal of Economics* (February, 1992), Levy and Murnane (1992), Danziger and Gottschalk (1993), and Frank and Cook (1995). These and other researchers concluded that the 1980's saw more than just a moderate increase in inequality, equivalent to giving back a few year's progress earlier in decreased inequality, but rather a fundamental transformation of the distribution of U.S. personal income toward much greater inequality.

The Great U-Turn and Nonmetro Income Inequality

What happened around 1980 to nonmetro income inequality and the gap in the United States between nonmetro and metro income inequality? The landmark paper on this subject is Tolbert and Lyson (1992), "Earnings Inequality in the Nonmetropolitan United States: 1967-1990," published in *Rural Sociology*. Their paper is based on the March Current Population Survey (CPS), a large household sample survey conducted in March each year by the U.S. Bureau of the Census. Tolbert and Lyson use three different measures of earnings inequality: (1) the coefficient of variation, (2) Theil's measure, and (3) the variance of the logarithms of income. They do not employ the most widely used measure of inequality in social science, the Gini concentration ratio. Figure 1 of their article shows a 3-year moving average of three measures of inequality against time and is reproduced here as figure 1. *(See figures following the References section of this report.)* This figure shows the variance of the logarithms of income bottoming around 1970 and increasing from there, while the other two measures of inequality (Theil's measure and the coefficient of variation) decline steeply through the late 1970's and then rise steeply in the early 1980's. Tolbert and Lyson defined income as the sum of wage and salary income, self-employed income, and farm income for residents of nonmetro areas. Their analysis was based on the earned income of full-time, year-round employed workers, excluding individuals with missing data on any of the variables, for example, occupation and industry.

The variance of the logarithms in figure 1 shows the Great U-Turn occurring circa 1970. The two other measures of inequality (Theil's measure and the coefficient of variation) indicate a Great U-Turn around 1980. When did inequality stop decreasing and start increasing? Is this difference in time simply a matter of which measure of inequality one chooses to use or is something else producing this divergence? I address this question by examining changing metro and nonmetro inequality in wage and salary income between 1967 and 1986.

Note: Charles Tolbert is a professor at Louisiana State University in Baton Rouge, LA, where he is the Chair of the Departments of Sociology and Rural Sociology.

The Topcode Hypothesis

This paper hypothesizes that the different timing of the Great U-Turn, depending on the choice of measure of inequality, can be explained by how Tolbert and Lyson and many other social scientists use the code the Bureau of the Census substitutes for large incomes to protect the identity of people with large incomes. This code is called a "topcode" and is used in Bureau of the Census data products such as the March CPS.

The Census Bureau releases detailed data from the CPS. The proportion of people with large incomes, say $100,000 (in 1989 dollars) a year, is quite small, and the release of information on those with large incomes could lead to the identification of respondents. Thus, the Census Bureau does not publish an income component, such as wage and salary income, over a certain specified amount, the maximum disclosable income. Instead, any income over the maximum disclosable income is replaced with the maximum disclosable income itself, known as the topcode. A data file with identifying information deleted and large incomes topcoded conceals the identity of respondents receiving incomes over the maximum disclosable income. The Bureau of the Census publishes public use microdata samples of the CPS. These samples contain data on individuals but the individuals cannot be identified from the data in the file.

In a few instances the maximum disclosable income might have been what a respondent reported. However, most records that report the maximum disclosable income as the income amount were topcoded, and the topcode is likely to be an underestimate of the income amount it replaces. Nevertheless, most social scientists use topcoded income amounts as if they are valid observations. This practice is generally sound since the Bureau of the Census usually chooses a maximum disclosable income amount high enough that few incomes are underestimated by their topcodes. However, the late 1970's was a time of rapid inflation (see fig. 2) and the Bureau of the Census kept the maximum disclosable income at $50,000 (in nominal, current dollars) from the March 1968 CPS, which asked about 1967 income, through the March 1981 CPS, which asked about 1980 income (see fig. 3). Meanwhile inflation lowered the real (adjusted for inflation) value of $50,000 (see fig. 4) and an increasingly large proportion of incomes was topcoded by the Bureau of the Census (see fig. 5). See table 1 for the maximum disclosable income in nominal and constant 1989 dollars in each year between 1967 and 1987.

This paper examines wage and salary income alone rather than the sum of income types that Tolbert and Lyson chose to analyze. The Bureau of the Census separately topcodes wage and salary income from self-employment income, which, in turn, is separately topcoded from farm income. Wage and salary income is the largest component of earned income for most people. Also, this paper examines individuals age 25 to 65 with at least $1 of wage and salary income. Tolbert and Lyson focus on those individuals employed full-time with year-round employment. This paper looks at a more inclusive group.

Measures of income inequality are especially sensitive to the proportion of the population that is very poor or very rich. An increased relative frequency at either end of the distribution can increase the measure of inequality, which is devised so it becomes smaller when incomes are more equal, and bigger when incomes are less equal. If a researcher estimates a measure of income inequality with CPS data in 1967 to 1986 using topcodes "as is," that is, as if they were valid income observations, the measure would be estimated with no income bigger than the maximum disclosable income. Most researchers who use the public use micro-data samples from the March CPS compute measures of inequality by using topcoded incomes "as is." In most years, the maximum disclosable income, the topcode, is sufficiently high that only a few incomes are topcoded, and these very few cases are not influential enough by themselves to seriously bias the estimate of an income inequality statistic.

Blackburn and Bloom (1987) recognized, however, that, since the largest incomes make the largest contributions to the Gini concentration ratio and most other measures of inequality, using topcodes "as is" can bias estimates of inequality downward since this practice underestimates very large incomes. Further, Blackburn and Bloom recognized that a change in the topcode will affect this bias and complicate making comparisons over time. To make their results comparable over time, they selected a single topcode in constant dollars. They used an even lower maximum disclosable income level than the Bureau of the Census uses in every year, except the year with the lowest inflation-adjusted maximum disclosable income level. Thus, they topcode valid inflation-adjusted incomes in years when their topcode is lower than the current Bureau of the Census topcode in constant dollar terms.

A better alternative is to use the best point estimator of incomes exceeding the maximum disclosable income instead of the maximum disclosable income itself. The best point estimator of topcoded incomes is the expectation of incomes at and in excess of the maximum disclosable income, i.e., the mean of these incomes.[1] Call this expectation, x^*:

$$x^* \equiv \frac{1}{\delta}\int_{\hat{x}}^{\infty} x\, f(x)\, dx, \tag{1}$$

where:

$x \equiv$ income

$\hat{x} \equiv$ maximum disclosable income

$\delta \equiv \int_{\hat{x}}^{\infty} f(x)\, dx$

$f(x) \equiv$ the functional form of the tail of the income distribution.

Information about f(x) in the sample has been lost by topcoding. However, knowledge about the shape of the right tails of income distributions can be used to form an estimate of f(x) and consequently of x^*. Vilfredo Pareto (1897), discussed in Arnold (1985), introduced a functional model for the right tail of income and asset distributions at the end of the 19th century. This functional model, the Pareto probability density function (pdf), is known to be inexact (Henson, 1967). Instead of estimating x^* as the mean of the fitted Pareto pdf from the maximum disclosable income, $\hat{x}$, to infinity, x^* can be more directly estimated from March CPS data where the maximum disclosable income was so high that almost no incomes were topcoded. The few incomes topcoded can be estimated by the x^* of the Pareto pdf. These topcoded incomes are so few that estimating them by $\hat{x}$ has a negligible effect on estimates of measures of inequality. With these few topcoded cases thus estimated, the sample $x^*/\hat{x}$ ratio is calculated as the simulated maximum disclosable income. $\hat{x}$, is ratcheted lower \$1,000 at a time. The resulting observations on the ratio $x^*/\hat{x}$ measure the effect of lowering the topcode. They can then be fitted by a smooth curve and the smooth curve can be used to estimate the ratio at a given $\hat{x}$ in terms of constant dollars.

[1]The Bureau of the Census began assigning the mean of incomes at and in excess of the maximum disclosable income as the topcode of the public use sample of the March CPS beginning with the March 1996 CPS. These means are separately calculated by gender, race, and whether the person was employed year-round, full time.

This procedure has two advantages over the use of the Pareto pdf alone to estimate x^*. First, it is based on the sample right tail rather than the Pareto pdf, known to be an inexact model of the right tails of income distributions. Secondly, the procedure does not make the false assumption, as estimating x^* by integrating the Pareto pdf must, that there is no frequency spike (pile up of frequencies) at the maximum disclosable income. The Bureau of the Census chooses large round numbers as maximum disclosable incomes, and larger incomes tend to be reported as rounder numbers, i.e., exactly divisible by $10,000 (Angle, 1994). So one could expect a substantial fraction of incomes will be reported as this income. When $x^*/\hat{x}$ is estimated via a lowering of a simulated topcode, frequency spiking affects the estimate and thus is taken directly into account.

This paper first shows that using topcodes as if they were valid income observations became problematic in the late 1970's because these were years of rapid inflation. The Bureau of the Census did not raise the maximum disclosable income and topcode of $50,000 in these years. Consequently, in terms of constant dollars, the topcode was lowered, and an increasing proportion of incomes was topcoded. Second, this paper replaces the information deleted by topcoding with x^*, the mean of incomes equal to and in excess of the maximum disclosable income. Finally, this paper demonstates that the three measures of inequality that Tolbert and Lyson use, as well as the most widely used measure of income inequality, the Gini concentration ratio, tell a consistent story about the timing of the Great U-Turn. When x^* is substituted for $\hat{x}$ as the topcode, all measures of inequality indicate that the Great U-Turn occurred closer to 1970 than 1980, and that there was no unusual divergence between nonmetro and metro inequality around 1980.

Inflation and the Nominal Topcode

Figure 3 and table 1 show that for wage and salary income from 1967 through 1980 the U.S. Bureau of the Census employed $50,000 as the maximum disclosable income and topcode. During this period, inflation reduced the purchasing power of the maximum disclosable income. Figure 2 shows inflation ran at high levels throughout the 1970's and 1980's. The cumulative effect of this inflation was a rapid lowering of the maximum disclosable income in terms of inflation-adjusted, or real dollars (fig. 4 and table 1). Figure 4 and table 1 show the maximum disclosable income plunging from $170,000 in constant 1989 dollars in 1967 to a low of $75,556 in constant 1989 dollars in 1980. The consequence of this lowering of the maximum disclosable income was a surge in the proportion of incomes topcoded, as shown in figure 5.

Figure 5 shows that the proportion of incomes topcoded peaked in 1980. The rise in the proportion topcoded was swift in the late 1970's, the obverse image of the plunging real value of the maximum disclosable income. Figure 5 shows that the proportion of incomes topcoded is consistently higher in metro areas than nonmetro areas. Also, there are more large incomes in metro than nonmetro areas. Figure 6 shows that the right tails of wage and salary income distributions, the proportion of larger incomes, in metro areas has been above the right tails of the distribution in nonmetro areas.

Simulating the Lowering of the Maximum Disclosable Income

During the 1970's, the U.S. Bureau of the Census held the nominal maximum disclosable income constant during a period of rapid inflation. Consequently, each year more and more incomes were topcoded, almost 2 percent by 1980 in metro areas. But the downward biasing effect of topcoding on a measure of income inequality has yet to be demonstrated. How serious is topcoding 2 percent of

Table 1 -- Nominal and real topcodes of wage and salary incomes in the March Current Population Survey

Year income received	Year of March CPS	Nominal topcode for wage and salary income	Topcode for wage and salary income in terms of 1989
			Dollars
1967	1968	50,000	170,000
1968	1969	50,000	163,704
1969	1970	50,000	156,738
1970	1971	50,000	149,831
1971	1972	50,000	143,506
1972	1973	50,000	138,558
1973	1974	50,000	131,548
1974	1975	50,000	119,459
1975	1976	50,000	110,500
1976	1977	50,000	104,492
1977	1978	50,000	98,004
1978	1979	50,000	91,322
1979	1980	50,000	83,712
1980	1981	50,000	75,556
1981	1982	75,000	104,082
1982	1983	75,000	98,368
1983	1984	75,000	94,043
1984	1985	99,999	120,929
1985	1986	99,999	116,622
1986	1987	99,999	113,332

Note: The times series does not go past 1986 income because, beginning with the March 1988 Current Population Survey, the wage and salary variable was split into two variables, wage and salary income in the longest job and next longest held job, each separately topcoded. Dollar values were adjusted to constant 1989 dollars using total personal consumption expenditure index numbers. Table B-3 (Council of Economic Advisers, 1996).
Source: March Current Population Survey.

incomes? The best way to find out is to simulate the lowering of the maximum disclosable income in a March CPS that has few topcoded incomes. The Pareto pdf can be used to infer the mean of the few cases at and above the maximum disclosable income, $\hat{x}$. As the lowering of the maximum disclosable income is simulated, an inequality statistic will be calculated two ways, with $\hat{x}$ as the topcode and with x^* as the topcode. The best estimate of the inequality statistic, with the original data and estimates of x^* based on the Pareto pdf substituted for the high $\hat{x}$, will also be plotted. The Gini concentration ratio, by far the most widely used income inequality statistic, will be estimated in this simulation of the lowering of the maximum disclosable income.

Gini's original formulation of the concentration ratio was as the normalized version (i.e., forced into the interval [0,1)) of an earlier statistic Gini had devised, the mean difference (David, 1983). In a sample of size n of observations on income, Gini's mean difference statistic is:

$$\text{Gini's mean difference} \equiv \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} \left| y_i - y_j \right|, \tag{2}$$

the sum of the absolute difference of all ordered pairs of observations, (y_i, y_j), divided by the number of ways it is possible to draw ordered pairs of incomes from the sample. Division by twice the mean of the sample normalizes this statistic:

$$\text{Gini concentration ratio} \equiv G \equiv$$

$$\frac{1}{2n(n-1)\,\bar{y}} \sum_{i=1}^{n}\sum_{j=1}^{n} \left| y_i - y_j \right|. \tag{3}$$

This statistic can function as a summary statistic of the Lorenz curve, a generalization of the measurement of inequality by saying how much of total income, the richest 1, 2, 5 percent, etc., individuals receive. Income distribution is concentrated when the top tiny percentage receives a disproportionate share of total income. This close relationship between the Gini ratio and the Lorenz curve means that the Gini ratio is very sensitive to how well total income is measured. Because income is concentrated, a good estimate of total income requires good estimates of the largest incomes.

The Gini concentration ratio is sensitive to the largest incomes in the sample. An examination of the numerator of the definition of the Gini ratio, G, (equation 3), illustrates this point. Assume that y_i is the largest income in the sample. Most of the differences whose absolute value will be summed to make y_i's contribution to G will be, since an income distribution is right skewed, by far the largest contributions of any observation to G.

Four consecutive March CPS's, from 1968 through 1971, have maximum disclosable incomes in constant 1989 dollars that are quite high: $170,000, $163,704, $156,738, and $149,831. The nominal dollar maximum disclosable amount in all 4 years is $50,000. The few cases topcoded in each year can be estimated as the mean of the Pareto pdf in excess of the maximum disclosable income amount. Pooled together, these 4 years of CPS wage and salary income data present an opportunity for a simulation experiment of the effect of using topcodes "as is" on the Gini concentration ratio.

The simulation experiment of using topcodes "as is" begins with a maximum disclosable income of $150,000 in 1989 dollars. Wage and salary incomes have been adjusted to 1989 dollars using the personal consumption expenditure deflators (Council of Economic Advisers, 1996, table B-3). Then the simulated maximum disclosable income is ratcheted down $1,000 at a time to $40,000 in terms of 1989 dollars. At each income level, the Gini concentration ratio is then estimated two ways: (1) with all incomes above this level replaced with that level, i.e., like using the maximum disclosable income as the topcode, and (2) with all incomes above the topcode level replaced with the sample mean of all incomes at and above that level. This simulation experiment was done separately for metro and nonmetro incomes. At every income level, a greater proportion of the metro sample was above the topcode. Figure 7 displays the simulation for metro wage and salary incomes, figure 8 shows the simulation for nonmetro incomes.

Figures 7 and 8 show that: (1) using topcodes "as is" underestimates the true Gini concentration ratio; (2) that this degree of underestimation depends in an almost linear way on the proportion of cases topcoded; and (3) that substituting the mean of the sample at and above the topcode, x^*, for the Bureau of the Census topcode, neutralizes the problem of topcoding--at least up to about 7 percent of the sample topcoded. If more than 7 percent of the sample is topcoded, then the Gini concentration ratio estimated with x^* as the estimate of the topcoded incomes begins to drift downward, but much less rapidly than estimates of the Gini concentration ratio estimated using Bureau of the Census topcode "as is." Clearly one obtains a better estimate by using x^* for the Bureau of the Census topcode, the maximum disclosable income, instead of $\hat{x}$.

How to Estimate x^* Once Incomes Have Been Topcoded

Once income data have been topcoded, the sample x^*, is unknown to the user of a CPS public use micro-data sample file since the topcoding has eliminated sample information about incomes in excess of the maximum disclosable income amount, $\hat{x}$. This part of the distribution is the right tail, the relative frequency distribution of large incomes. The right tail of the income distribution is quite regular (Pareto, 1897). In fact, in unconditional income distributions of substantial populations defined geographically (what most social scientists mean by the term "income distribution") a gently tapering right tail is universal. This knowledge of the regularity of the right tail can be used to estimate x^*.

The probability density function (pdf) that Pareto suggested as a model of the right tail of income distributions (Evans, Hastings, and Peacock, 1993) is:

$$h(x) = ca^{c}x^{-c-1}, \qquad (4)$$

where:

$$a \equiv \text{smallest income fitted}$$

$$c > 1,$$

implies that the mean of incomes in excess of $\hat{x}$, x^*, is:

$$x^* = \left(\frac{c}{c-1}\right)\hat{x}, \qquad (5)$$

which in turn implies that the ratio, $x^*/\hat{x}$, is a constant:

$$\frac{x^*}{\hat{x}} = \frac{c}{c-1},$$

because c is constant.

While the Pareto pdf has been widely used to model the right tails of income distributions, one of its drawbacks is that, empirically, the ratio, $x^*/\hat{x}$, has been found not to be constant except possibly in the extreme right tail of the distribution (Henson, 1967), i.e., $x^*/\hat{x} \rightarrow$ a constant as $\hat{x} \rightarrow \infty$. Figure 9 shows that the ratio $x^*/\hat{x}$ is not constant over $\hat{x}$. Figure 9 is based on the March 1968 through the March 1971 CPS files, covering wage and salary incomes in 1967 through 1970. In all 4 years, the maximum disclosable income in terms of nominal dollars was \$50,000. In terms of the value of dollars in 1989, the maximum disclosable incomes were quite high: \$170,000, \$163,704, \$156,738, and \$149,831, respectively. The few incomes that exceeded the maximum disclosable income and were topcoded with the maximum disclosable income were estimated by the mean of the Pareto pdf in excess of the maximum disclosable income. The Pareto parameters of the fits to the nonmetro and metro wage and salary distributions were estimated by the Henson method (1967). However, there are so few cases above the maximum disclosable income that the treatment of these topcoded cases has only a negligible effect on the estimate of the ratio $x^*/\hat{x}$. While it is possible that the ratio $x^*/\hat{x}$ changes in the years after 1970, it is unlikely that this change is large relative to the difference between the nonmetro and metro income distributions, and, as figure 9 illustrates, the $x^*/\hat{x}$ ratios of the nonmetro and metro income distributions are quite similar.

Henson (1967; cited in Shryock and Siegel, 1973), Spiers, 1977, and Parker and Fenwick, 1983) suggests estimating the c parameter of the Pareto pdf, using only data from the extreme right tail of an income sample. However, since empirical estimates of the ratio $x^*/\hat{x}$ are available, it is better to recognize that $x^*/\hat{x}$ decreases as $\hat{x}$ becomes large by smoothing the estimate of the ratio over $\hat{x}$. The smoothing is done by multiple regression. The ratio $x^*/\hat{x}$ is regressed on $\hat{x}$ and $\hat{x}^2$. The fitted values of the regression are taken as the smoothed values of the ratio (fig. 9). The nonmetro and metro ratios are separately smoothed.

Measures of Inequality Estimated with x^* as Topcode and $\hat{x}$ as Topcode

With an estimate of x^*, the mean of wage and salary incomes in excess of the maximum disclosable income, $\hat{x}$, in hand, I can estimate how much better the use of x^* is as a topcode over $\hat{x}$ in estimating inequality from March CPS data that has been topcoded. The method has three steps. In step 1, I estimate the three measures of inequality employed by Tolbert and Lyson (1992) plus the more commonly used Gini concentration ratio, using the topcodes of the Census Bureau as if they were valid observations. These topcodes are the maximum disclosable income amount, $\hat{x}$. Step 2 is to replace each occurrence of $\hat{x}$ with x^* and re-estimate the measures of inequality. Step 3 is to graph the two estimates to examine the difference.

Figure 10 displays estimates of four measures of inequality in nonmetro and metro wage and salary income estimated from the March 1968 through the March 1987 Current Population Surveys. These measures are estimated from the wage and salary incomes of people 25 to 65 years of age with at least $1 of wage and salary incomes. The estimates in figure 10 use Bureau of the Census topcodes as if they are valid income observations. Tolbert and Lyson based estimates on people employed full time, year round without missing data on selected variables such as occupation and industry and then smoothed their estimated measures of inequality with a 3-year moving average.

Despite the differences in the definitions of income and population used to estimate the measures of inequality, figures 1 and 10 display many commonalities. First, all three measures in figure 1 (variance of logarithms of income, Theil's measure, and the coefficient of variation) show roughly the same patterns as the graphs of the same statistics in figure 10. Thus, the comparison shows that nonmetro wage and salary inequality is higher than metro, the variance of the logarithm of income indicates both started rising in the early 1970's, and the metro Theil's measure and coefficient of variation dip downward just before 1980 while the nonmetro statistics do not. All three measures show that inequality peaked in the mid-1980's and was in decline at the end of the 1980's. Figure 10 departs from figure 1 in the graph of Theil's measure and the coefficient of variation. In figure 1, both statistics fall during the 1970's, then rise sharply during the 1980's, to about where they began in the 1970's only to fall again in the late 1980's. In figure 10, there is a distinct dip, particularly in the metro statistic just prior to 1980, but the general trend, both metro and nonmetro, was up during both the 1970's and 1980's, with a downturn in the late 1980's. The drop in these statistics just prior to 1980 is less prominent in figure 10 than in the Tolbert and Lyson data. The Gini concentration ratio of wage and salary income shows a profile over time much like those of Theil's measure and the coefficient of variation, except that there is less of a downward dip just before 1980.

Figure 11 shows the same four measures of inequality of metro wage and salary income as figure 10 estimated with Bureau of the Census topcodes as if they were valid observations, and with those topcodes replaced with x^*, the estimate of the mean of topcoded incomes. The difference between these two estimates of a single statistic provides an estimate of the biasing effect of using Bureau of the Census topcodes "as is." In figure 11, the two estimates of the variance of the logarithms overlap, indicating that the variance of the logarithms as an estimator of inequality is relatively insensitive to the extent of topcoding in the data, almost 2 percent of the metro sample in 1980. Theil's measure and the coefficient of variation are much more vulnerable to topcoding, with the Gini concentration being somewhat vulnerable but less so than these two statistics.

Figure 12 graphs the inequality measures of nonmetro wage and salary income. In any given year, there are fewer than half as many nonmetro wage and salary incomes large enough to be topcoded as metro wage and salary incomes (fig. 6). Thus, one would expect topcoding to be a less serious problem in the estimation of nonmetro inequality measures. Figure 12 shows that the downward bias of using Bureau of the Census topcodes "as is" is barely noticeable only in 1980 and a few years just prior in the estimation of the Gini concentration ratio, and somewhat larger in the same years for Theil's measure and the coefficient of variation. Nonmetro wage and salary income is potentially as vulnerable as metro wage and salary income to the downward bias in estimating inequality due to using topcodes "as is," but the proportion of nonmetro incomes large enough to be topcoded is small enough to keep the bias small.

Figure 13 shows the corrected estimates of the four measures of inequality of wage and salary income, nonmetro and metro, estimated with x^* substituted for the Bureau of the Census topcode.

It is straightforward to demonstrate that the larger the proportion of Bureau of the Census topcodes used "as is," the bigger the downward bias on estimates of Theil's measure, the coefficient of variation, and the Gini concentration ratio. Figure 14, for example, plots the difference (Theil's measure estimated with x^* used as the topcode minus Theil's measure estimated with Bureau of the Census topcodes "as is") against the proportion of the sample topcoded. The straight line is the fitted linear model of the regression of the difference on the proportion. The observed differences cluster closely around the line. This simple ordinary least squares (OLS) regression has an r^2 of 0.917. Its intercept and slope are statistically significant beyond the 0.001 level. The equation of the line is $y = 0.000862 + 0.689685$. This means that an increase in the proportion of cases topcoded from virtually 0 to 0.02, i.e., 2 percent, would account for a downward bias of 0.0147 in the estimated Theil's measure. How big is that? The metro maximum Theil's measure estimated with x^* was 0.2846 in 1986 and the minimum 0.2416 in 1968. The maximum minus the minimum is 0.0430. So the bias induced by using the 2 percent of the sample topcoded by the Bureau of the Census as if they were valid observations is as large as 0.34, over a third, of the difference between the maximum and minimum attained by the Theil's measure in this 24-year period. Now, imagine that downward bias disappearing suddenly in the early 1980's, when there actually was an increase in inequality, probably related to the deep recessions of 1980 and 1981-82. The combination of the two events in measured income inequality might well look like a surge of income inequality, leading many analysts to see a Great U-Turn around 1980.

But Theil's measure, clearly affected by using Bureau of the Census topcodes "as is," is not used as frequently as other measures of inequality. More frequently used than Theil's measure is the coefficient of variation, and the use of either is, by far, exceeded by the use of the Gini concentration ratio. Figure 15 shows that the differences between the coefficient of variation estimated with Bureau of the Census topcodes used "as is" and the coefficient of variation estimated with these topcodes replaced by x^*, resemble a linear function of the proportion of the sample topcoded. As in figure 14, there is evident heteroskedasticity: more scatter among the smaller proportions topcoded. That is hardly surprising, however, since more noise in estimating a relationship is to be expected from a small sample.

Figure 16 shows that the differences between Gini concentration ratio estimates using Census Bureau topcodes "as is" and the Gini concentration ratio estimates that replace Bureau of the Census topcodes with x^*, also fall close to a line. The line of figure 16 was estimated by OLS and has the form $y = 0.000541 + 0.306289x$, where x is the proportion topcoded. The r^2 of the regression is 0.465. The range of proportions topcoded seen in the sample is from 0 to 0.02 (2 percent). The difference between the high (0.403) and low (0.369) Gini concentration ratios (estimated with x^*) of metro wage and salary income in 1967 through 1990 is 0.034. The downward bias on estimates of the Gini concentration ratio resulting from using Census Bureau topcodes for 2 percent of the sample, about the maximum percentage topcoded in 1980 from the regression line of figure 16, is 0.00667. This coefficient represents 20 percent of the difference between the maximum Gini and minimum Gini of metro wage and salary income from 1967 through 1990.

Conclusions

This paper demonstrates that use of the Bureau of the Census wage and salary income topcodes as if they were valid data introduces a downward bias into measures of inequality estimated from such a data set. Theil's measure and the coefficient of variation are sensitive to this bias, the Gini concentration is somewhat so, and the variance of the logarithms of income is less sensitive than the other measures of inequality. This bias becomes unacceptably large in the case of the first three measures of income inequality if more than 1 or 2 percent of the sample is topcoded, since the bias induced becomes a substantial fraction of the true swings in these statistics over the course of several decades. Also, the Bureau of the Census had no standard in the 1960's, 1970's, and 1980's for determining size of the smallest income to be topcoded or what percentage of incomes to topcode. The maximum disclosable income, which was used as a topcode, was increased from time to time in an *ad hoc* way. These adjustments distort comparisons over time.

Nonmetro wage and salary income is less subject to this bias than metro wage and salary income because there are substantially fewer incomes large enough to be topcoded in nonmetro than in metro areas. When the Bureau of the Census topcodes more than 1 or 2 percent of the largest incomes, metro income inequality (as measured by Theil's measure, and the coefficient of variation or the Gini concentration ratio) will appear to be substantially lower than it really is while nonmetro income inequality will be less affected by Bureau of the Census topcoding. The result is, circa 1980, an artificial widening of the gap between nonmetro and metro income inequality. Nonmetro income inequality has been historically greater than metro income inequality, not because of a proportionally greater number of large incomes in nonmetro areas but because of a proportionally greater number of small incomes (see fig. 6).

During the 1970's, the Bureau of the Census kept the topcode at a fixed nominal $50,000 for wage and salary income. Any wage and salary income in excess of this maximum disclosable income was replaced by $50,000. While inflation raged during the 1970's, $50,000 rapidly became a smaller income and more people were having their wage and salary income topcoded. The percentage of incomes topcoded reached a maximum for income received in 1980, and as many as 2 percent of the metro wage and salary incomes and 1 percent of the nonmetro wage and salary incomes were topcoded. Then, just as income inequality really did increase in the early 1980's, probably because of the deep recession of 1981-82, the Bureau of the Census substantially raised the topcode of wage and salary income, canceling much of the downward bias. So the apparent increase in measured inequality (using the Gini concentration ratio, Theil's measure, or the coefficient of variation) included both the real increase and the canceling of the downward bias due to the raising of the topcode. The canceling of the downward bias exaggerated an actual sharp increase in inequality.

The percentage of wage and salary incomes topcoded has been much smaller since the topcode was raised from $50,000. Were topcoding to again become a problem, there are two adaptive responses. One is to use the variance of the logarithms as the measure of inequality. The logarithmic transformation disproportionately decreases the contribution of large incomes, and this statistic is relatively insensitive to using the Bureau of the Census topcodes "as is." The other response is to substitute an estimate of the mean of incomes in excess of the topcode for the Bureau of the Census topcode, which is the maximum disclosable income.

References

Angle, John. 1994. "Frequency Spiking in U.S. Income Distributions," *Proceedings of the Social Statistics Section of the American Statistical Association.* Alexandria, VA: American Statistical Association.

Arnold, B.C. 1985. "Pareto Distribution," *The Encyclopedia of Statistical Sciences.* Kotz and N.L. Johnson (eds.). Vol. 6: pp. 568-574.

Blackburn, McKinley, and David Bloom. 1987. "Earnings and Income Inequality in the United States," *Population and Development Review.* Vol. 13: pp. 575-609.

Burtless, Gary (ed.). 1990. *A Future of Lousy Jobs? The Changing Structure of U.S. Wages.* Washington, DC: Brookings Institution.

Council of Economic Advisers. 1996. *Economic Report of the President.*

Danziger, Sheldon, and Peter Gottschalk (eds.). 1993. *Uneven Tides: Rising Inequality in America.* New York: Russell Sage Foundation.

David, H.A. 1983. "Gini's Mean Difference," *The Encyclopedia of Statistical Science.* S. Kotz and N.L. Johnson (eds.). New York: Wiley. Pp. 436-437.

Evans, Merran, Nicholas Hastings, and Brian Peacock. 1993. *Statistical Distributions.* New York: Wiley.

Frank, Robert, and Philip Cook. 1995. *The Winner-Take-All-Society.* New York: Simon and Schuster.

Harrison, Bennett, and Barry Bluestone. 1988. *The Great U-Turn: Corporate Restructuring and the Polarization of American.* New York: Basic.

Harrison, Bennett, Chris Tilly, and Barry Bluestone. 1986a. "The Great U-Turn: Increasing Inequality in Wage and Salary Income in the U.S." Study prepared for the Joint Economic Committee, U.S. Congress.

______________. 1986b. "Wage Inequality Takes a Great U-Turn," *Challenge.* March/April 29, No. 1: pp. 26-32.

______________. 1986c. "Rising Inequality," *The Changing American Economy.* David Obey and Paul Sarbanes (eds.). New York: Basil Blackwell. Pp. 111-134.

Henson, Mary. 1967. *Trends in the Income of Families and Persons in the United States: 1947 to 1960.* Technical Paper no. 17. U.S. Department of Commerce, Bureau of the Census.

Kuttner, Robert. 1983. "The Declining Middle," *Atlantic Monthly.* July: pp. 60-72.

Lawrence, Robert. 1984. "Sectoral Shifts in the Size of the Middle Class," *Brookings Review.* Vol. 3: pp. 3-11.

Levy, Frank. 1988. *Dollars and Dreams: The Changing American Income Distribution.* New York: Norton.

Levy, Frank, and Richard Murnane. 1992. "U.S. Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations," *Journal of Economic Literature.* Vol. 30: pp. 1,333-1,381.

Miller, Herman. 1966. *Income Distribution in the United States.* U.S. Department of Commerce, Bureau of the Census.

McGranahan, David. 1980. "The Spatial Structure of Income Distribution in Rural Regions," *American Sociological Review.* Vol. 45: pp. 313-324.

Pareto, Vilfredo. 1897. *Cours d'Economie Politique.* Vol. 2. Lausanne: Rouge.

Parker, Robert, and Rudy Fenwick. 1983. "The Pareto Curve and its Utility for Open-Ended Income Distributions in Survey Research," *Social Forces.* Vol. 61: pp. 872-885.

Shryock, Henry, Jacob Siegel, and associates. 1973. *The Methods and Materials of Demography.* U.S. Department of Commerce, Bureau of the Census. Vol. 1.

Spiers, Emmett. 1977. "Estimation of Summary Measures of Income Size Distribution from Grouped Data," *Proceedings of the Social Statistics Section of the American Statistical Association.* Pp. 252-277.

Thurow, Lester. 1984. "The Disappearance of the Middle Class," *New York Times.* Vol. 133 (February 5): p. F3.

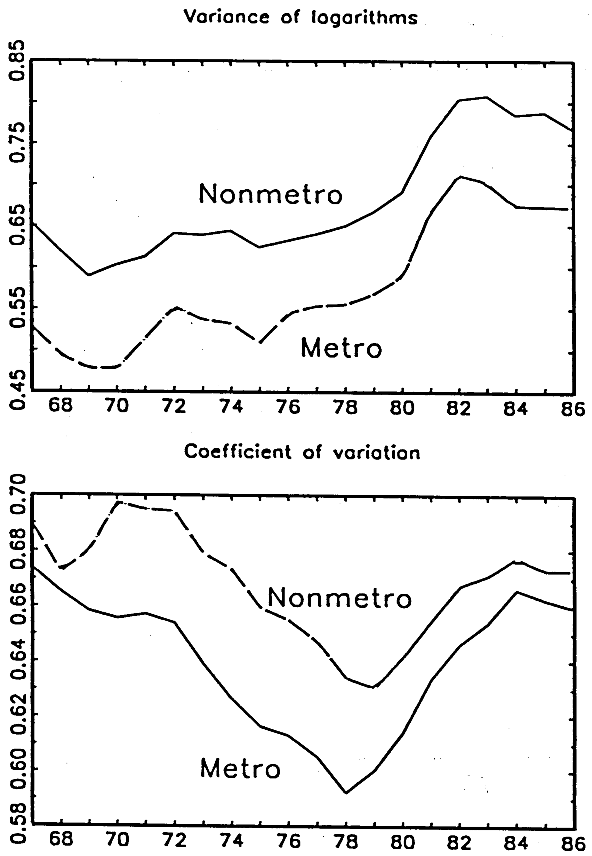
______________. 1987. "A Surge in Inequality," *Scientific American.* Vol. 256, No. 5: p. 30.

Tolbert, Charles. 1996. Personal communication.

Tolbert, Charles, and Thomas Lyson. 1992. "Earnings Inequality in the Nonmetropolitan United States: 1967-1990," *Rural Sociology.* Vol. 57: pp. 494-511.

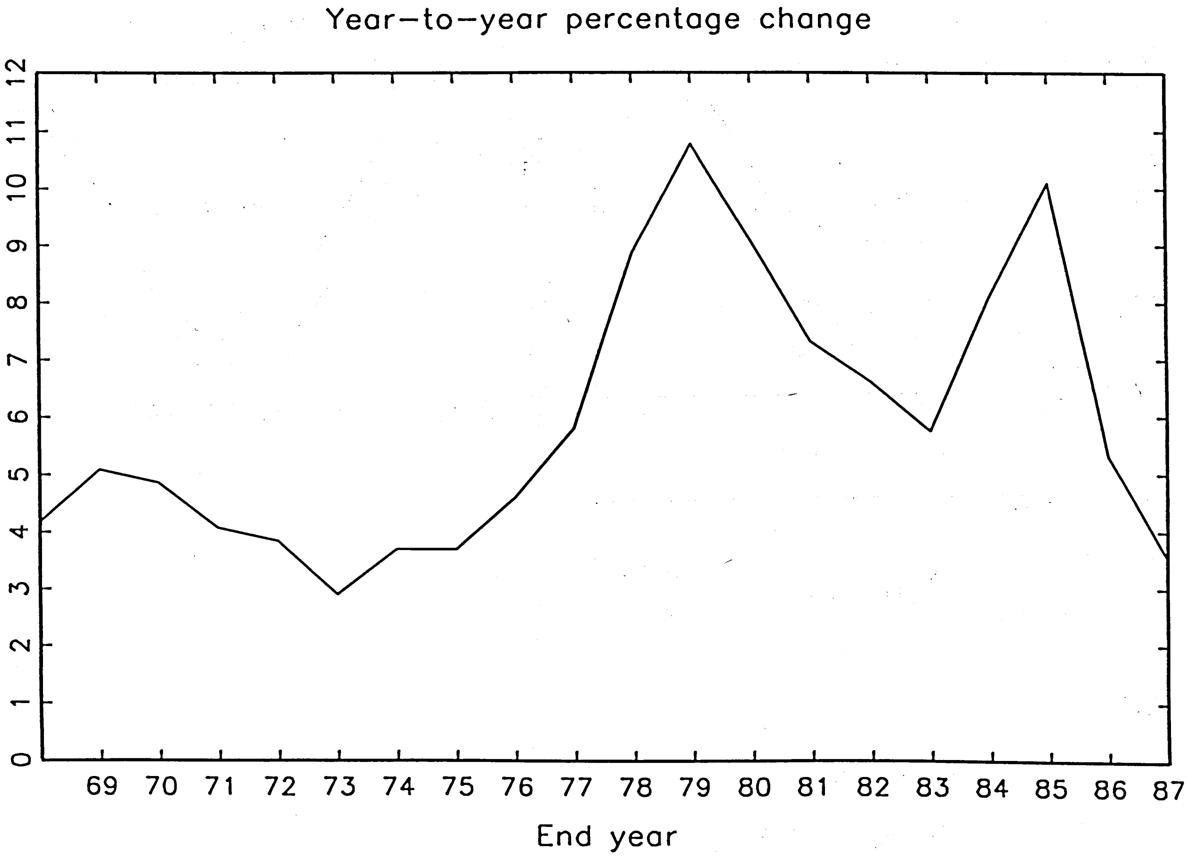
Unicon, Inc. 1997. *Manual for March, Current Population Surveys.* Santa Monica, CA: Unicon.

Figure 1. Three measures of inequality of earned income, 1967-86



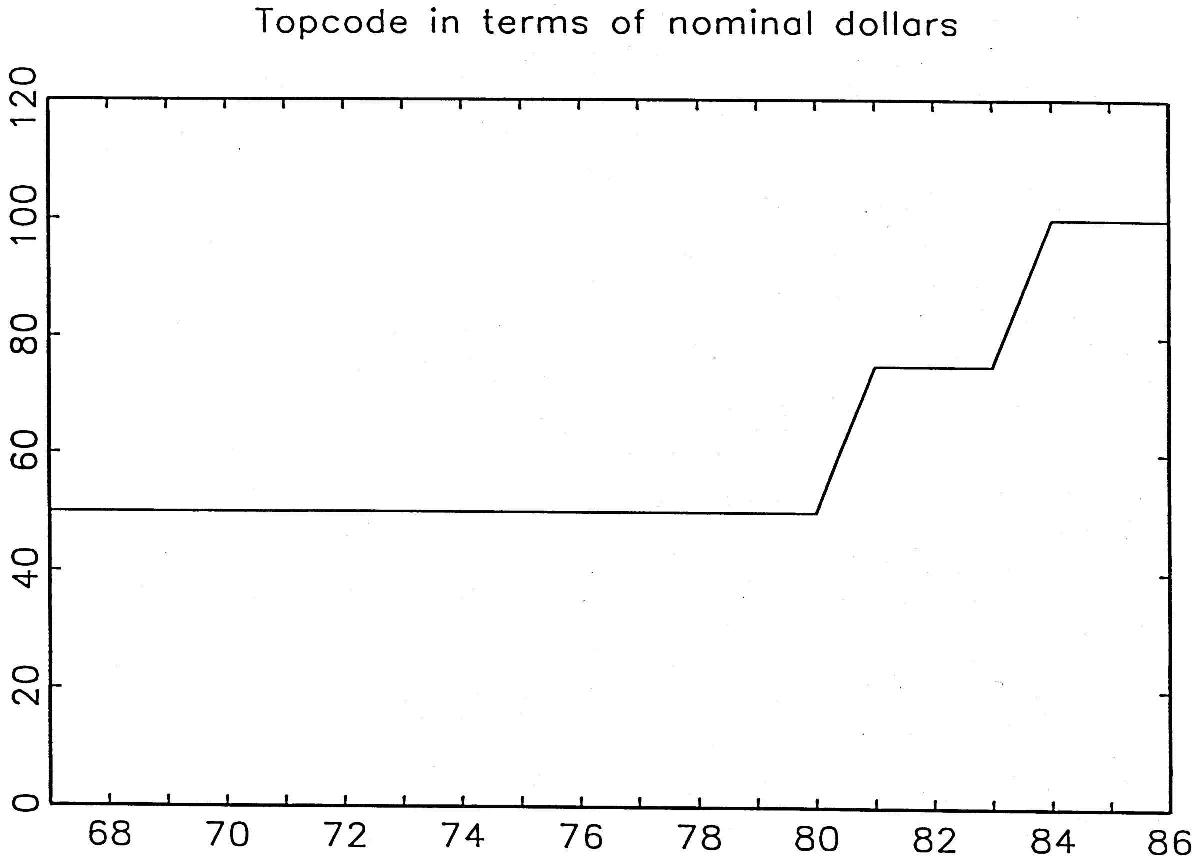
Note: These estimates were published in figure 1 of Tolbert and Lyson (1992). Each data point in the graph is a 3-year-moving average of a measure of inequality of earned income. Earned income in Tolbert and Lyson (1992) is defined as the sum of wage and salary income plus income from self-employment and farm income for residents of rural areas.
Source: Charles Tolbert, Lousiana State University and Thomas Lyson, Cornell University, based on estimates from the March Current Population Survey.

Figure 2. Year-to-year rate of inflation, 1968-87



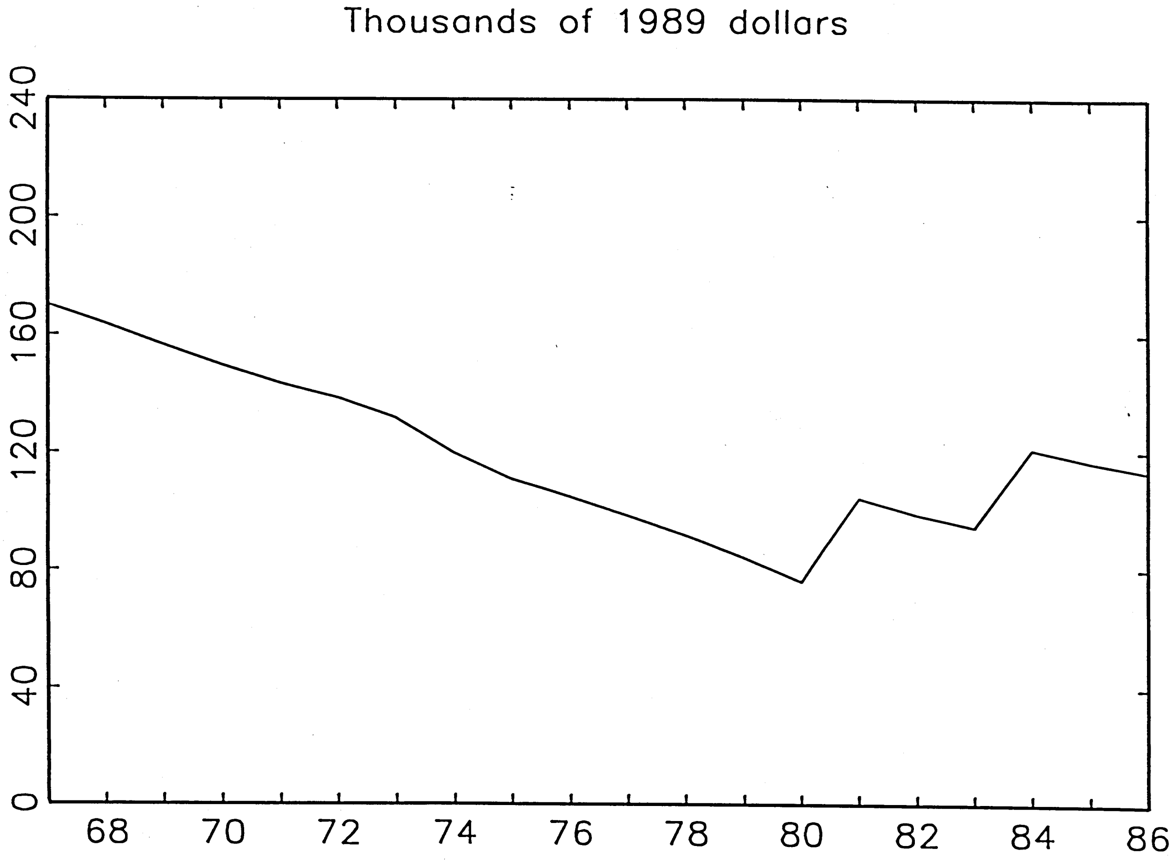
Source: Council of Economic Advisers, 1996. "Total personal consumption expenditure" index of Table B-3, "Chain-type price indices for gross domestic product, 1959-95."

Figure 3. Maximum disclosable income in March CPS, 1967-86



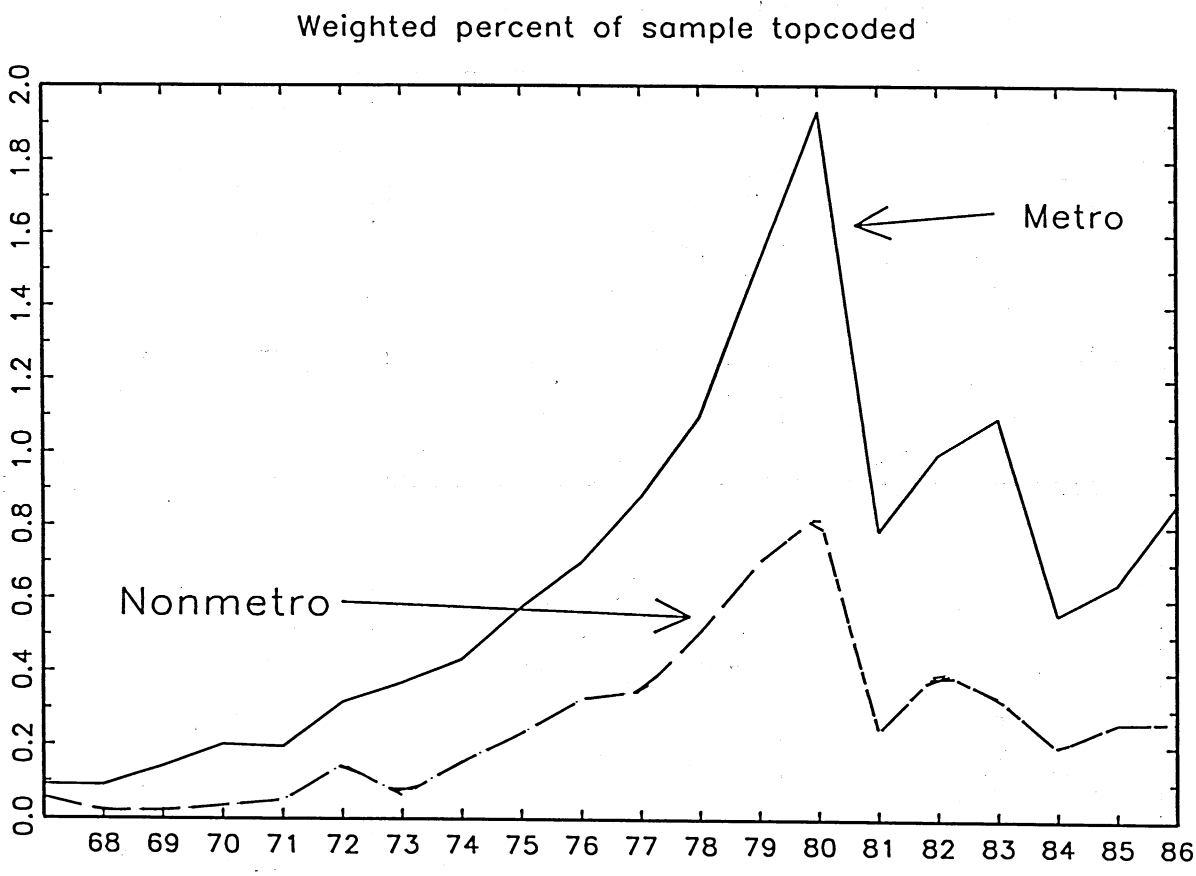
Note: Dollars are in terms of 1,000's of nominal (current) dollars.
Source: Maximum disclosable incomes are identified in the March Current Population Survey (CPS) documentation provided by Unicon, Inc., a data reseller (Unicon, 1997) and by direct inspection of March CPS files.

Figure 4. Maximum disclosable wage and salary income in March CPS, 1967-86



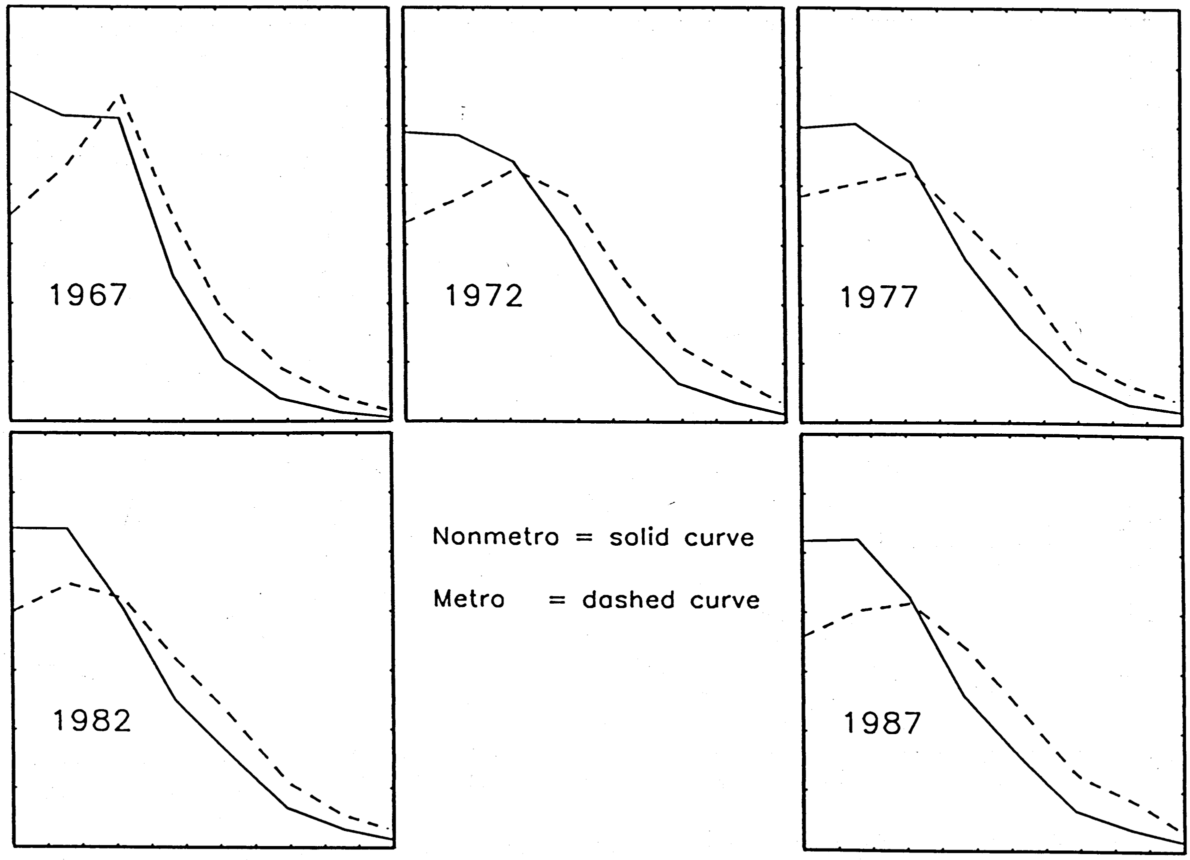
Source: Maximum disclosable incomes as identified in March Current Population Survey (CPS) documentation provided by Unicon, Inc., a data reseller (Unicon, 1997), and by direct inspection of the CPS files, are adjusted to 1989 dollar values by using the "total personal consumption expenditure" index of Table B-3, "Chain-type price indices for gross domestic product, 1959-95" (Council of Economic Advisers, 1996).

Figure 5. Percentage of wage and salary incomes topcoded, 1967-86



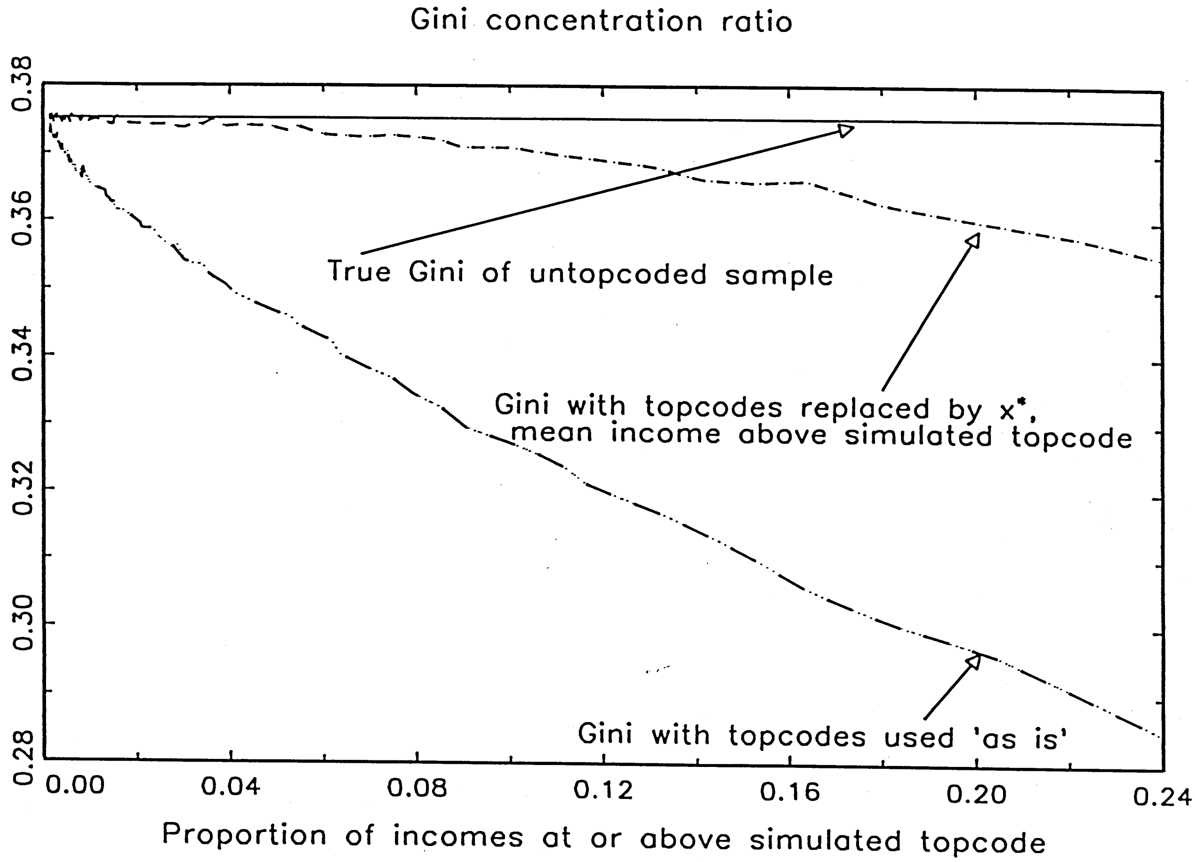
Note: Computed for people 25 to 65 years of age with at least $1 in wage and salary income.
Source: March Current Population Survey (CPS).

Figure 6. Relative frequency distributions of nonmetro and metro wage and salary income in 1989 dollars, selected years



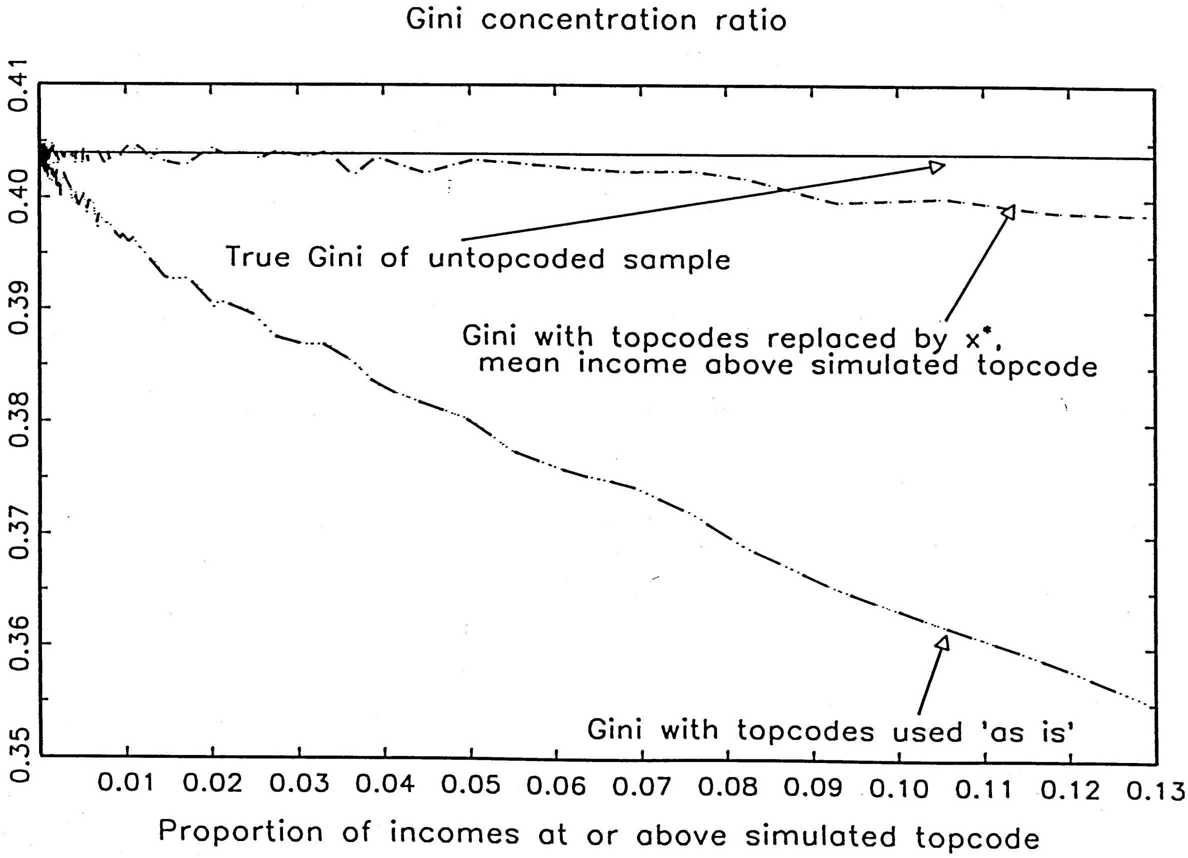
Note: Relative frequencies are proportions of sample falling into eight income bins (e.g., a range of incomes such as \$8,001 through \$16,000) from \$1 through \$64,000 in terms of 1989 dollars. Nominal dollar values were adjusted to constant 1989 dollars using the "total personal consumption expenditure" index of Table B-3, "Chain-type price indices for gross domestic product, 1959-95," (Council of Economic Advisers, 1996). X-axis is income from \$1 to \$64,000 in terms of 1989 dollars; y-axis is relative frequency from 0 to 0.35 of people aged 25 to 65 with at least \$1 of wage and salary income.
Source: March Current Population Surveys, as documented and recoded by Unicon, Inc., a data reseller (Unicon, 1997).

Figure 7. Effect of topcoding on estimates of metro wage and salary incomes



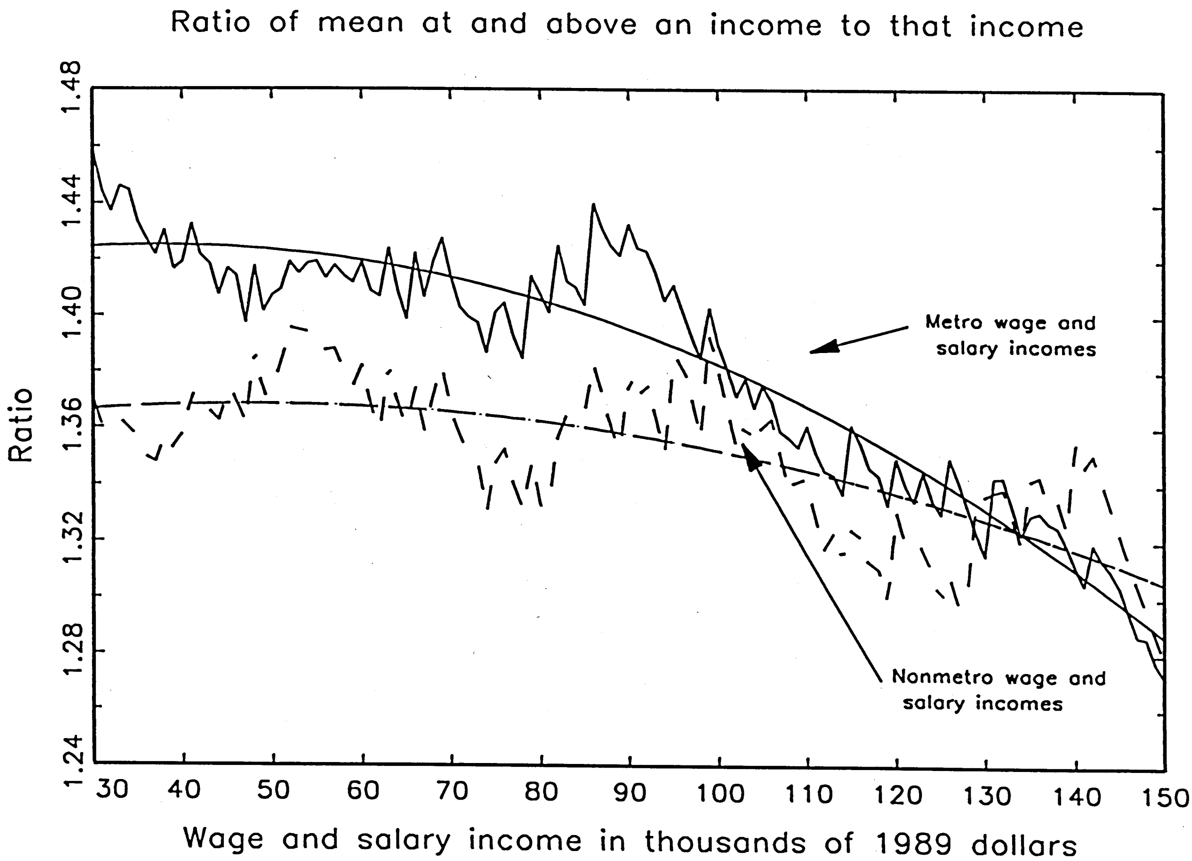
Note: Gini concentration ratios are estimated by simulating the maximum disclosable income falling from $150,000 to $30,000 in 1989 dollars.
Source: March Current Population Surveys from 1968 through 1971, surveys with a high maximum disclosable income in terms of purchasing power.

Figure 8. Effect of topcoding on estimates of nonmetro wage and salary incomes



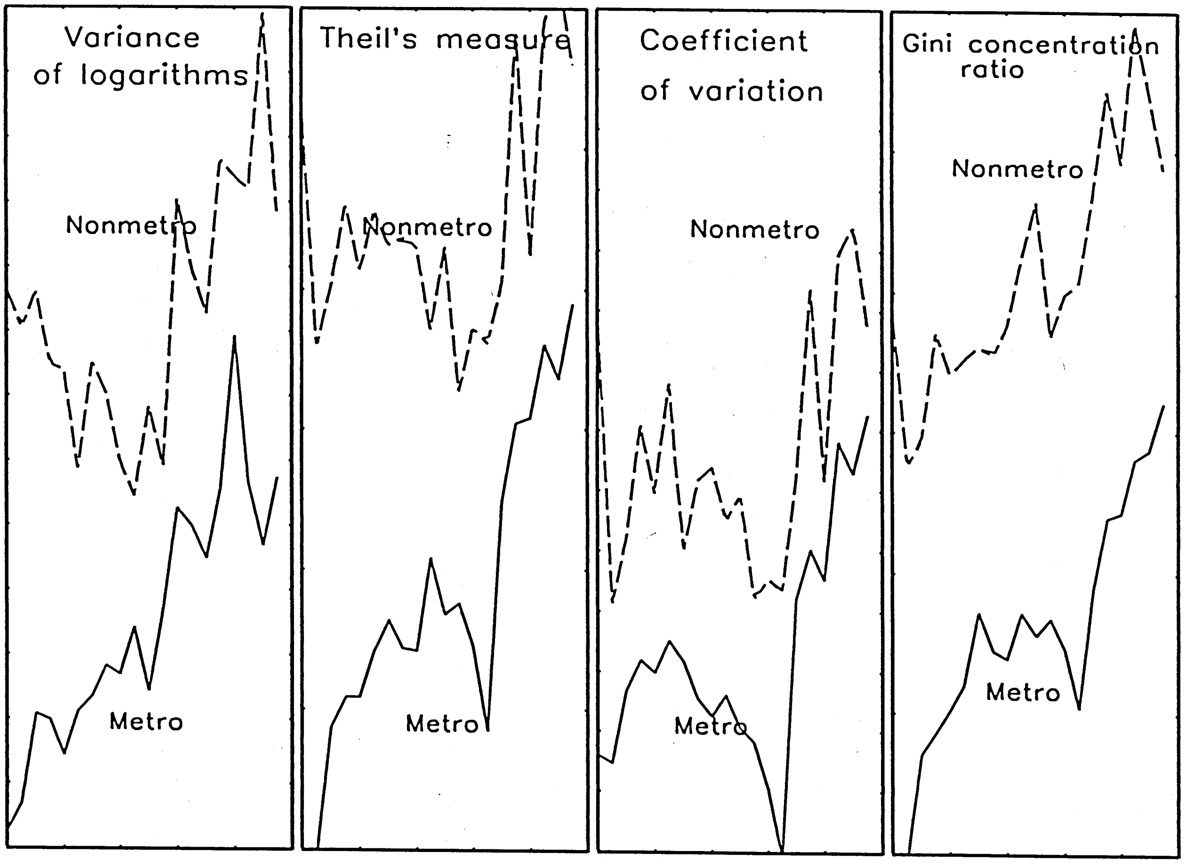
Note: Gini concentration ratios are estimated by simulating the maximum disclosable income falling from $150,000 to $30,000 in 1989 dollars.
Source: March Current Population Surveys from 1968 through 1971, surveys with a high maximum disclosable income in terms of purchasing power.

Figure 9. Metro and nonmetro $x^{*}/\hat{x}$ ratio



Note: Each smooth curve is regression of $x^{*}/\hat{x}$ ratio on $\hat{x}$ and $\hat{x}^2$. x^{*} is mean income above a particular income.
Source: Data are from the March Current Population Surveys of 1968 through 1971. These are years when the maximum disclosable income of $50,000 was high in terms of purchasing power and few incomes were topcoded. The few incomes topcoded are estimated by the mean of the fitted Pareto pdf above the maximum disclosable income. The Henson method was used to estimate the parameter of the fitting Pareto pdf. In terms of 1989 dollars the nominal topcode of $50,000 was worth $170,000, $163,704, $156,738, and $149,831 in 1967, 1968, 1969, and 1970, respectively.

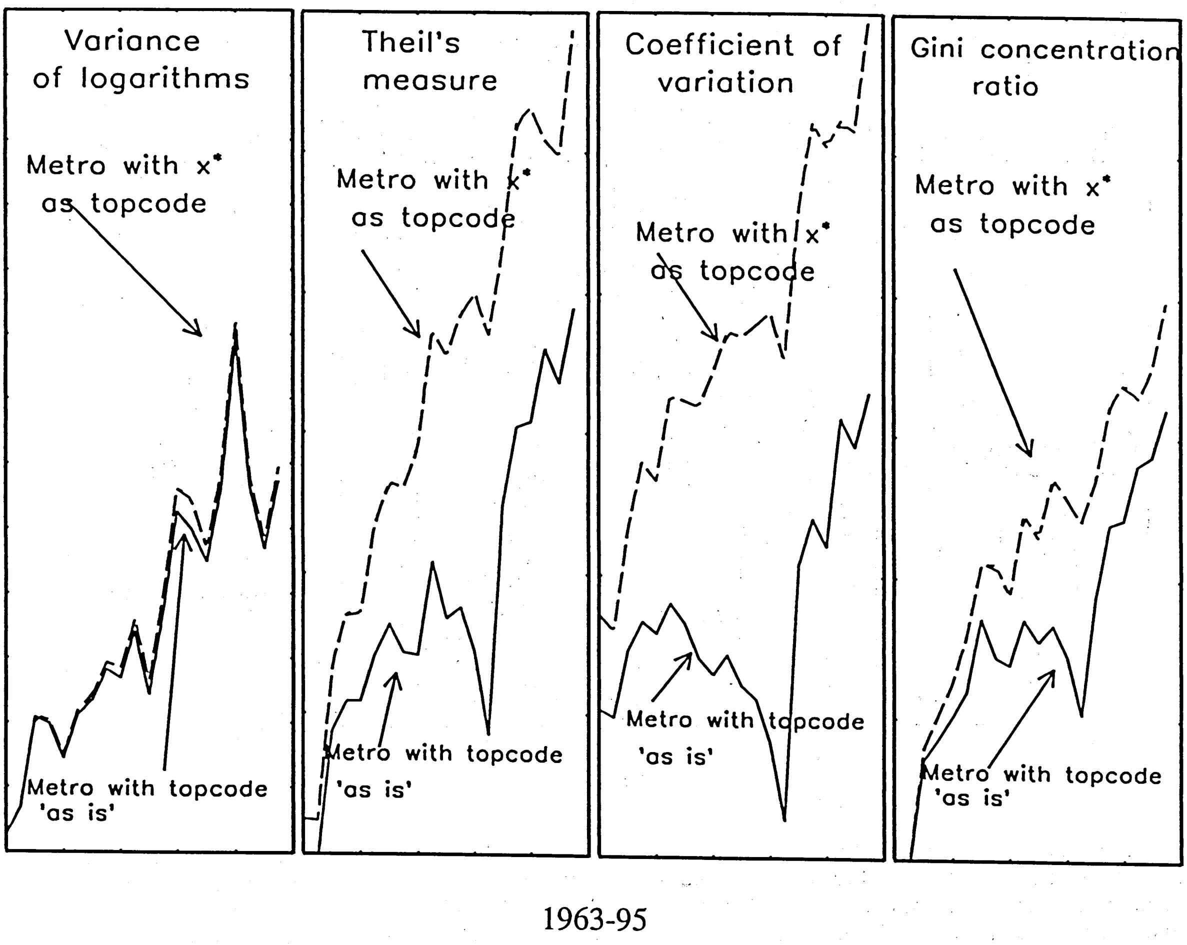
Figure 10. Four measures of inequality of wage and salary income estimated with topcodes used "as is," 1963-95



1963-95

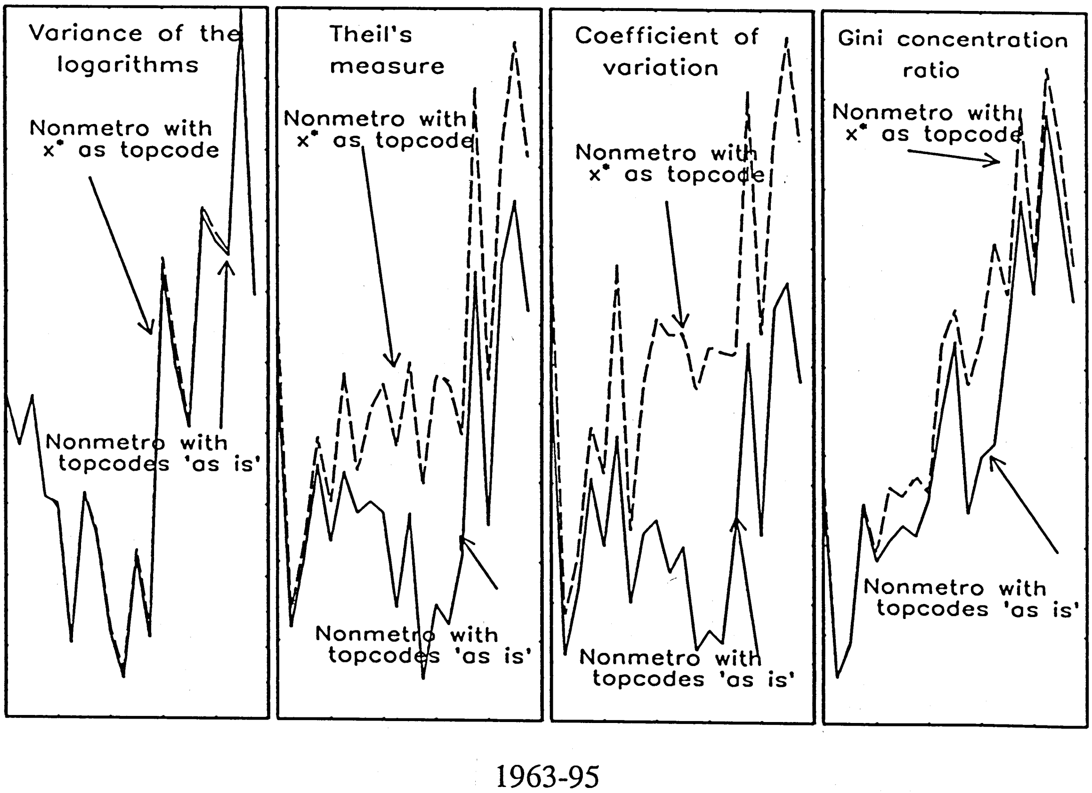
Note: Estimated on data from people aged 25 to 65 with at least $1 in wage and salary income. Income topcodes assigned by the U.S. Bureau of the Census used as if they were valid income observations, i.e., "as is." X-axis is year from 1967 through 1987. Y-axis is value of each of the four measures. Range of values is 1.05 to 1.70 for variance of the logarithms, 0.24 to 0.30 for Theil's measure, 0.70 to 0.85 for coefficient of variation, and 0.37 to 0.43 for the Gini concentration ratio.
Source: March Current Population Survey.

Figure 11. Four measures of metro inequality, 1963-95



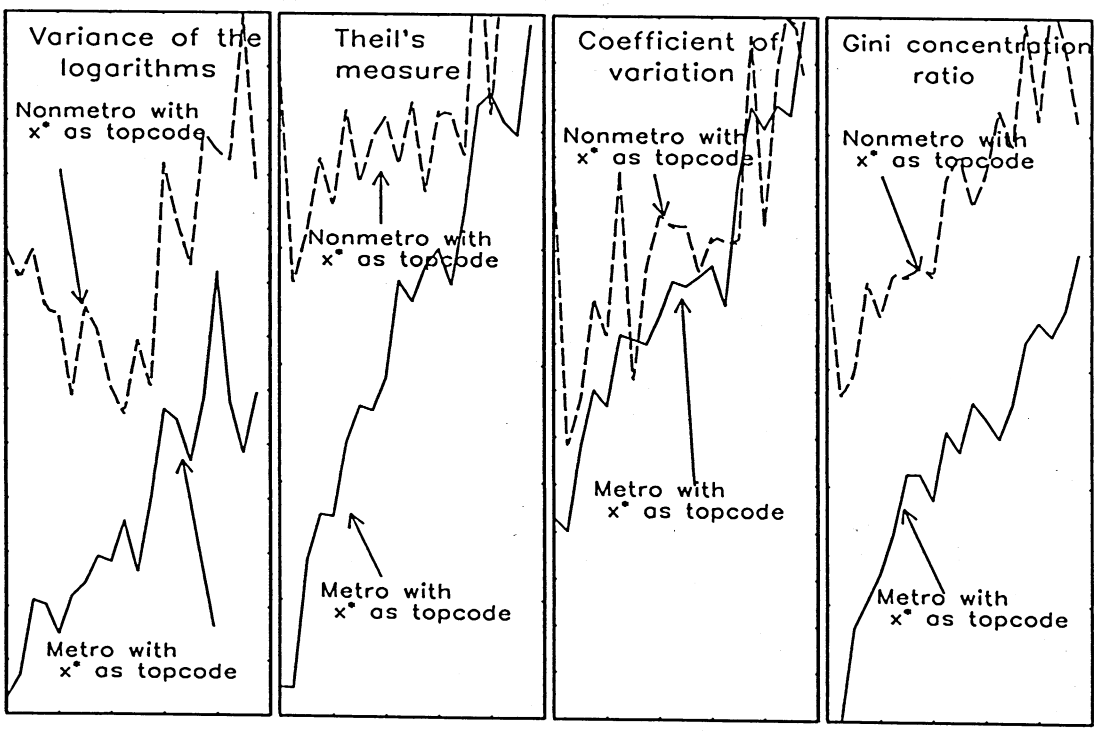
Note: Estimated on data on people aged 25 to 65 with at least $1 of wage and salary income. One set of statistics is estimated with Bureau of the Census topcodes; the other set is estimated with incomes equal to the maximum disclosable income replaced with an estimate of the mean of incomes equal to and greater than the maximum disclosable income. X-axis is year from 1967 through 1987. Y-axis is value of each of the four measures. Range of values is 1.05 to 1.70 for variance of the logarithms, 0.24 to 0.30 for Theil's measure, 0.70 to 0.85 for coefficient of variation, and 0.37 to 0.43 for the Gini concentration ratio.
Source: March Current Population Survey.

Figure 12. Four measures of nonmetro inequality, 1963-95



Note: Estimated on data on people aged 25 to 65 with at least $1 in wage and salary income. One set of inequality measures is estimated with Bureau of the Census topcodes; the other set is estimated with incomes equal to the maximum disclosable income replaced with estimate of mean of incomes equal to or in excess of the maximum disclosable income. X-axis is year from 1967 through 1987. Y-axis is value of each of the four measures. Range of values of 1.05 to 1.70 for variance of logarithms, 0.24 to 0.30 for Theil's measure, 0.70 to 0.85 for coefficient of variation, and 0.37 to 0.43 for the Gini concentration ratio. Source: March Current Population Survey.

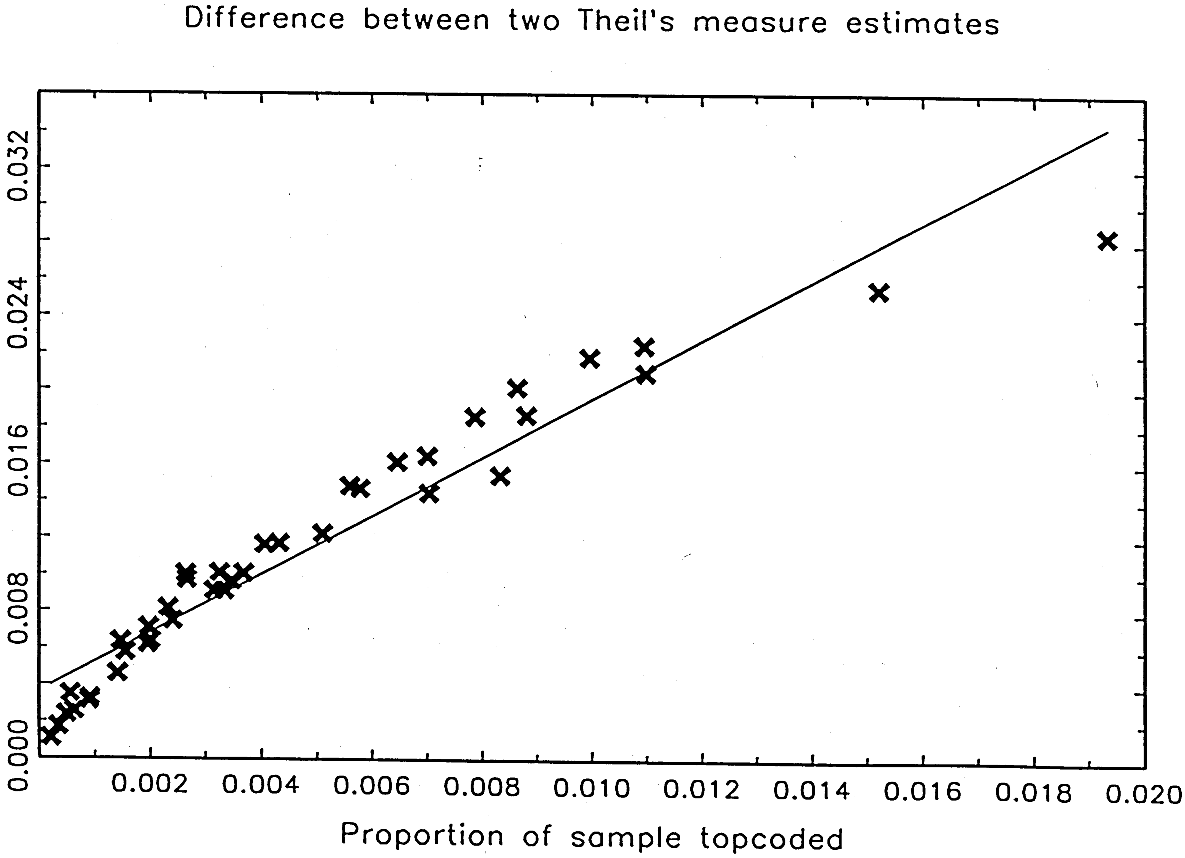
Figure 13. Four measures of metro and nonmetro wage and salary income inequality estimated with x^* instead of maximum disclosable income, $\hat{x}$, as topcode, 1963-95



1963-95

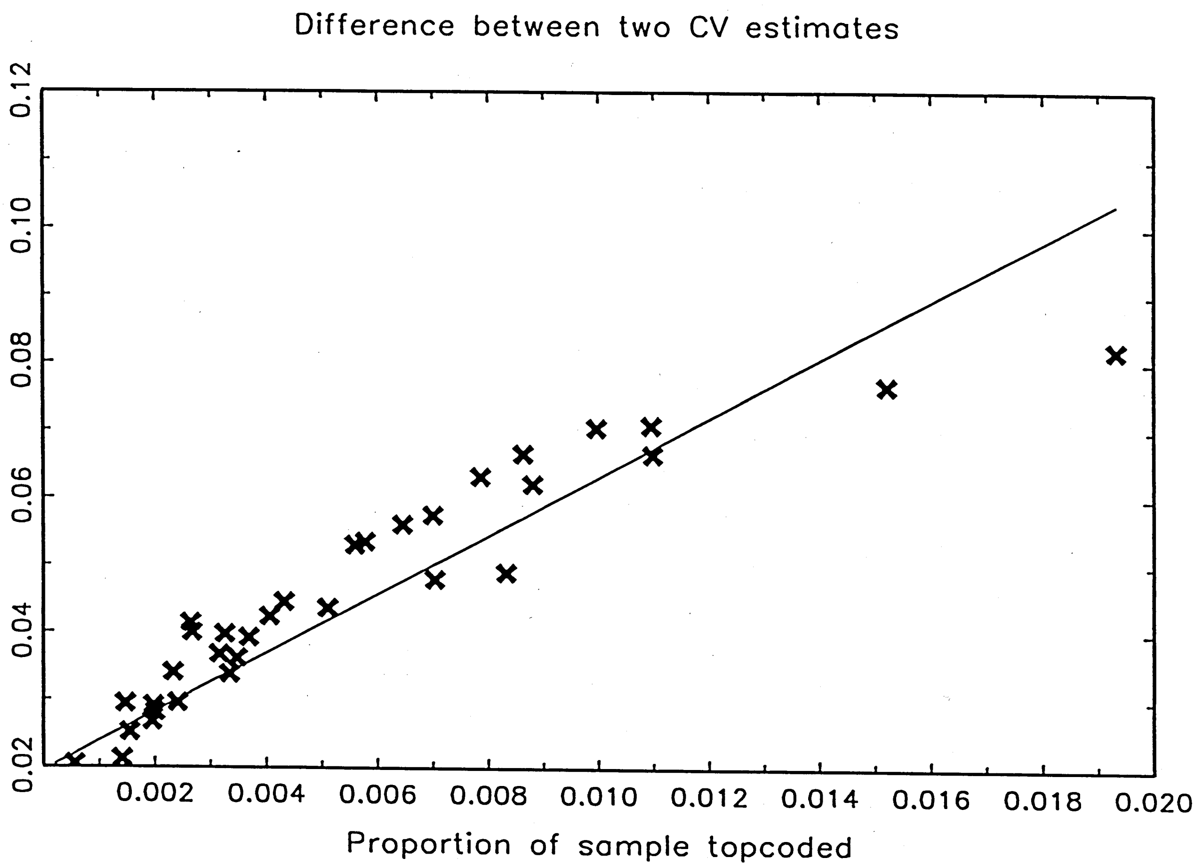
Note: Estimated on data on people aged 25 to 65 with at least $1 of wage and salary income. Inequality measures estimated with incomes equal to the Bureau of the Census maximum disclosable income replaced with an estimate of the mean of incomes equal to and in excess of the maximum disclosable income. X-axis is year from 1967 through 1986. Y-axis is value of each of the four measures. Range of values is 1.05 to 1.70 for variance of the logarithms, 0.24 to 0.30 for Theil's measure, 0.70 to 0.85 for coefficient of variation, and 0.37 to 0.43 for the Gini concentration ratio.
Source: March Current Population Survey.

Figure 14. Theil's measure estimated with x* minus Theil's measure estimated with Bureau of the Census topcodes "as is"



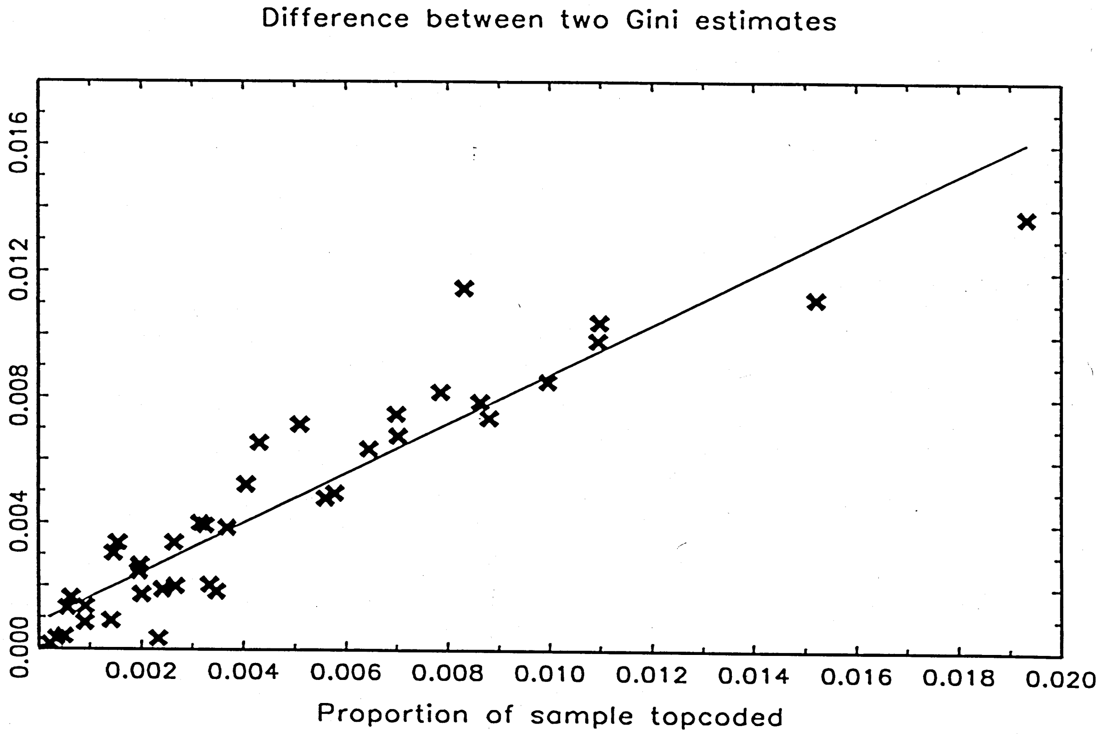
Note: Each data point plotted is one of 40 differences. Each difference is between Theil's measures estimated two different ways in a year (20 years) in either metro or nonmetro areas. The range of years is from the March 1968 through March 1987 CPS. Estimates are based on data on people aged 25 to 65 with at least $1 of wage and salary income. Theil's measure is estimated with incomes equal to the maximum disclosable income replaced with an estimate of the mean of incomes equal to and in excess of the maximum disclosable income. Then Theil's measure is estimated with Bureau of the Census topcodes "as is." The difference between the two is graphed.
Source: March Current Population Survey.

Figure 15. CV estimated with x* minus CV estimated with Bureau of the Census topcodes "as is"



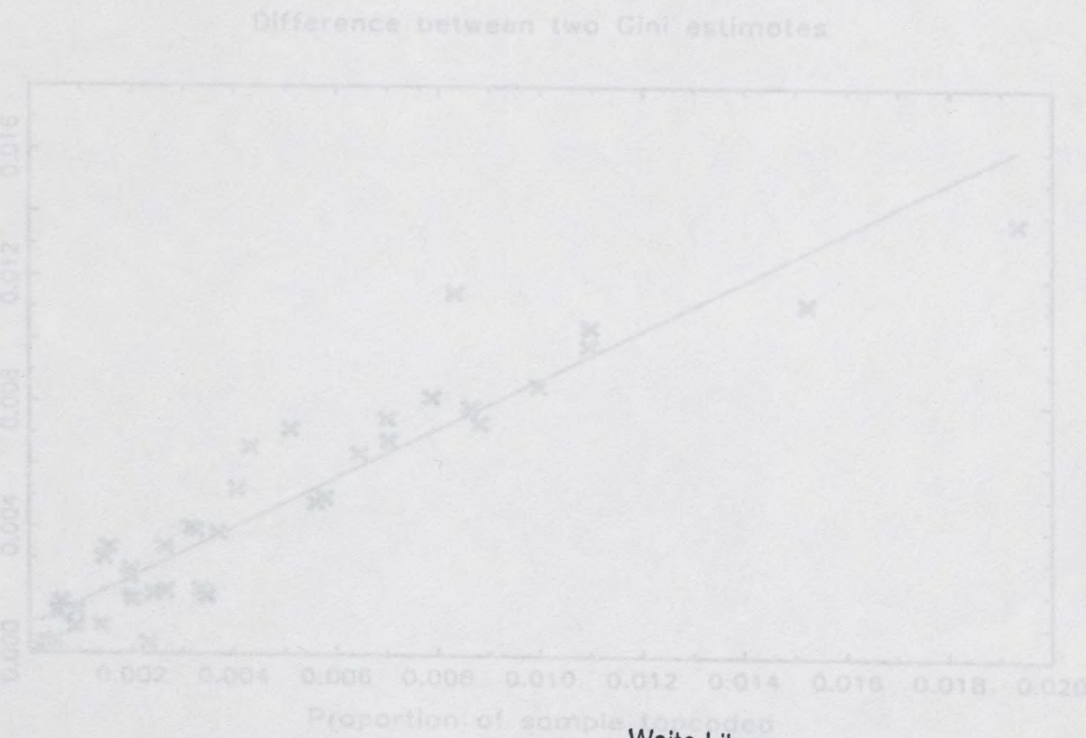
Note: Each data point plotted is one of 40 differences. Each difference is between CV's (coefficients of variation) estimated two different ways in a year (20 years) and in either metro or nonmetro areas. The range of years is from the March 1968 through March 1987 CPS. Estimates are based on people aged 25 to 65 with at least $1 of wage and salary income. The CV is estimated with incomes equal to the maximum disclosable income replaced with an estimate of the mean of incomes equal to and in excess of the maximum disclosable income. Then the CV is estimated with Census Bureau topcodes "as is." The difference between the two is graphed.
Source: March Current Population Survey.

Figure 16. Gini estimated with x* minus Gini estimated with Bureau of the Census topcodes "as is"



Note. Each data point plotted is one of 40 differences. Each difference is between Ginis estimated two different ways in a year (20 years) and in either metro or nonmetro areas. The range of years is from March 1968 through March 1987 CPS. Estimates are based on data on people aged 25 to 65 with at least $1 of wage and salary income. The Gini is estimated with incomes equal to the maximum disclosable income replaced with an estimate of the mean of incomes equal to and in excess of the maximum disclosable income. Then the Gini is estimated with Bureau of the Census topcodes "as is." The difference between the two is graphed.
Source: March Current Population Survey.

Figure 16. Gini estimated with x' minus Gini estimated with Bureau of the Census topcodes "as is"



Note. Each data point plotted is one of 40 differences. Each [illegible] estimated two different ways in a year (20 years) and in either metro or non[illegible] years is from March 1968 through March 1987 CPS. Estimates are based on data [illegible] at least $1 of wage and salary income. The Gini is estimated with incomes [illegible] income replaced with an estimate of the mean of incomes equal to and in excess of the maximum disclosable income. Then the Gini is estimated with Bureau of the Census topcodes "as is." The difference between the two is graphed.
Source: March Current Population Survey.