THE DISTRIBUTION OF HOUSEHOLD INCOME IN THE UNITED STATES, 1946-2015*

John Albert Schwendel Jr.† Hamid Mohtadi‡

September 18, 2020

Abstract

This paper estimates annual, state-level distributions of household income for the US from 1946-2015. In contrast to much of the existing literature, we focus on the main body of earners. We also propose a method to estimate continuous distributions from binned data. Simulations demonstrate this strategy eliminates a bias found in other methods. Growth of inequality is found to be concentrated in the 1940s-1950s and in the 1960s. Trends in real earnings support the well-known hollowing out effect and add previously unavailable detail to the mid-20th century. While intrastate inequality dominates national inequality, interstate inequality displays an inverted W-shaped pattern.

JEL Classification: C81, D31

Keywords: Household Income, Parametric Estimation, History of Inequality

*We would like to thank David Autor, Benjamin Bridgman, Scott Drewianka, Matthew Fisher-Post, John Heywood, and Thomas Piketty for their many constructive comments and insights. In addition, we greatly appreciate the input of all participants from the Econometric Society North American Summer meeting, the Midwest Economics Association Conference, the University of Wisconsin-Milwaukee Department of Economics Labor workshop, and the Wisconsin Economics Association Conference.

†Corresponding Author, Merton M. Sealts Jr. Visiting Professor, Department of Economics, The College of Wooster, 1189 Beall Ave, Wooster, OH 44691, jschwendel@wooster.edu

‡Professor, Department of Economics, University of Wisconsin-Milwaukee, mohtadi@uwm.edu, and Affiliated Professor, Department of Applied Economics, University of Minnesota, mohtadi@umn.edu
1 Introduction

Earnings inequality has emerged as one of the single most important issues of public concern and policy debate. Yet, many of the well-known studies of inequality focus either on top earners at the expense of the remainder of the distribution, on national trends at the expense of intraregional and interregional trends, or on the short or intermediate run at the expense of the long run.\(^1\) Focusing on US data, we address all three gaps. We do this by presenting a methodology that allows us to generate a new dataset, significantly extending the historical data on the distribution of household incomes for the main body of US earners at both the national and state-level. The results reveal distinct new patterns of US income inequality that have not been fully revealed in prior work and expand our understanding of historical income inequality trends in the US. They are likely to prove pivotal to future inequality research.

Our proposed method, which allows us to generate the new data, is based on recognizing and fixing a key shortcoming of many of the existing estimation techniques used for constructing continuous distributions from \textit{binned} income data. The shortcoming is rooted in the relationship between mean earnings across bins and the structure of the bins.\(^2\) We show that ignoring this relationship produces a systematic bias in measures of inequality, such as the Theil or Gini. The key insight is that income bins across states or time will exhibit a \textit{differential} pattern of crowding traceable to the differences in mean income across states or years. Since many interpolation methods are sensitive to this crowding, inequality measures based on these methods will be influenced by the level of crowding, producing a systemic measurement error which is unrelated to the real underlying level of inequality. Because this error creates spurious results in estimation both across time and between states, eliminating it is critically important to the historical and regional inequality trends that will be presented here. We formally demonstrate these issues by examining several existing candidate methodologies. We then demonstrate a generalized least squares method to be the most suitable for the type of binned data used

\(^1\)A recent paper by Aaberge, Atkinson and Modalsli (2020) shares this triple gap perspective with us.

\(^2\)While this problem has been acknowledged in some cases, its effect has never been formally estimated.
Applying our approach to the binned Adjusted Gross Income (AGI) data of the Internal Revenue Service’s (IRS) Statistics of Income (SOI) reports, we create a dataset of continuous earnings distributions for the main body of United States’ earners by state and year for the contiguous 48 states from 1946-2015, thereby significantly extending the existing inequality data spatially and temporally. From these distributions we develop state and national measures of inequality for the main body of earners, between the 5th and 95th percentile of the income distribution. This effort leads us to uncover distinct changes in the U.S. inequality profile not previously revealed. For example, we observe dramatic changes in the central segments of the household earnings distribution from the 1940s to the 1960s that entail large episodic increases in inequality and divergence in real earnings.

Apart from our methodological/diagnostic innovation, two features of the inequality data that we generate are worth mentioning. First, because we focus on household earnings, our new findings reveal interesting and important dynamics concerning the extensive margin of earnings of both individuals and married couples that cannot be demonstrated with individual wage data. For example, this level of observation allows for examination of issues such as non-wage sources of income and dual income households, which have significant implications for behavior, consumption, and welfare. This makes our focus distinct from existing studies on this topic. In addition, the dataset created here will extend the coverage of these studies along the spatial and temporal dimensions discussed above. A second important feature of our dataset is attention to the main body of earners which reveals trends involving labor income which has been recognized as the primary

3 Additionally, we adopt a process which allows us to select the correct distributional assumptions for this methodology. We do this by analyzing individual level data from the subsamples of those used to construct the binned income data. This is discussed later.

4 Our work differs markedly but complements the numerous and valuable contributions in the area of historical US income inequality, such as the studies by Goldin and Katz (1999, 2007) and Goldin and Margo (1992.) These and other similar analyses have had foci such as; various and specific segments of the population including full time workers and industry and gender specific trends, intensive measures of income including hourly, weekly, and individual level wages, and returns to human capital. In contrast, the dataset we will present covers all annual household income and from this more extensive margin of income, we create continuous distributions.

5 The prior work in this area often focused on a national level or individual state-level datasets and often had data at larger intervals than the annual results to be presented here. Our period of coverage offers a rich panel of observations as it covers the contiguous 48 states individually as well as the aggregate(national) level from 1946-2015.
driver of overall U.S. income inequality (Piketty, 2014). Because the dataset uniquely includes the late 1940s to early 1960s, preceding other such studies, it will provide an opportunity for researchers to analyze various contemporary issues such as minimum wage policies, unionization, and returns to skill by exploiting historical and geographic heterogeneity. Some of the largest variations in these labor market institutions can be traced back to educational shifts produced by the G.I. Bill in 1944, changes in the rate and type of unionization beginning in the 1940s, and legislative changes to the national minimum wage in 1961 and 1966. These specific historic events shed greater light on the evolution of the middle class and the popular notion of their rising affluence after World War II. The subsequent alleged decline of such affluence—the so-called hollowing out of the middle class—requires understanding the context of the full post WWII history, rather than just a recent slice of it.

The importance of the main body of the income distribution is most recently emphasized in a paper by Aaberge, Atkinson, and Modalsli (2020) that studies Norway’s historical income trends. While our study differs from theirs in both geographic focus and diagnostic methodology, we share the authors’ position that a better understanding of the evolution of the main body of earners is both welcome and necessary. Also, as our diagnostic innovation produces measures of inequality that are comparable across US states and over time, we share the stated potential of the work by AAM to create inter-regionally comparable measures of inequality.

We will establish that our methodology is the most suitable method for both the type of data used here and the measures derived from it. Specifically, our simulations show that the generalized least squares method eliminates the biases described earlier, without reducing

---

6 Autor, Manning, and Smith (2010) demonstrate that the real level of the minimum wage matters, not only for the left tail of earners, but may also have spillover effects into the main body of the earnings distribution.

7 Union membership is shown by Dinardo, Fortin, and Lemieux (1996) to drive inequality between the 50th and 90th percentiles of wage earners, and Card, Lemieux, and Riddell (2004) provide evidence that union membership also has differential effects on earnings based on skill level.

8 Autor (2014) stresses that returns to skill are a major driver in U.S. inequality and apply mostly to The Other 99%.

9 Wolfson (1994) demonstrates that polarization cannot be accurately described by a single index. The use of Theil index, a measure used for reasons discussed later, does not address Wolfson’s concern. But data on continuous distributions will.
performance in other ways. Because the bias is driven by mean income across states and time, the consistency of this method is important for comparability of the measures themselves but could also be especially helpful when assessing causal relationships via regressions.

Our state-level data provides an opportunity to investigate whether changes in national inequality are driven by distributional changes within states or regions, or by changes in the relative mean earnings between them. For the purpose of analyzing relative trends of inter- and intra-state inequality measures and their contribution to national inequality, the Theil Index is a natural choice due to its unique decomposition property. Additionally, this decomposition will allow us to independently analyze the historical trends in each measure. From this effort, a picture will emerge of a general increase in national inequality over the sample period with particularly acute increases in inequality from the mid-1940s to the mid-1950s, from the mid-1960s to early 1970s, and across the 1980s. We also observe a relatively steady growth in inequality in the 1970s, and from the 1990s forward. Within state inequality measures for individual states generally follow a similar pattern. Also, while between-state inequality is not a large portion of the national measure, this component of inequality is found to be considerably more volatile. The time trend of this component changes direction much more often and severely than the within state component. Many of these changes occur in time periods or segments of the earnings distribution that have not been explored in prior work, making it possible to not only describe the history of U.S. earnings inequality more fully and provide new insights into its evolution, but also to lay the basis for future analyses of its causes and its relevance to economic outcomes, thereby helping to inform policy decisions.

The rest of the paper will be organized as follows. Section two will discuss the features and advantages of the AGI data in the SOI over other income surveys, while section three will discuss the methodology necessitated by its particular features. Section four will discuss the results, and section five will conclude.
2 Data

Starting in 1916, the IRS has published annual data on individual earnings and taxes in the SOI. Of specific interest to this paper is the earnings data from 1946 onward, as prior to 1944 a majority of Americans were not required to file income taxes, and until 1946 the large number of earners deployed in World War II may drive the pattern of inequality. After a series of changes in the allowable deduction in the 1940s, the percentage of actual filers from the population of potential filers has consistently fluctuated between 85% and 95%, as demonstrated by Freeland and Hodge (2012). This makes the source from which the IRS draws its sample rather large.\footnote{Additionally, as noted by Frank (2014) the pattern of the number of filers follows that of population changes in the United States after World War II.}

While this population may exclude individuals with specific characteristics, Cilke (1998) shows that the majority of excluded individuals consists of dependents living at home and older individuals. These groups may be seen as outliers for the working population at large, and thus may represent an important, but distinct pattern from the ones drawn here.

The IRS SOI has several distinct advantages over the often-used Current Population Survey (CPS). First, the time period for which annual data are available offers two additional decades of information in the SOI as compared to many CPS surveys. Second, it limits misreporting by the population from which samples are drawn: while the CPS must rely on careful survey design and respondent training, information reported to the IRS carries with it a potential penalty for misreporting. As pointed out by Lemieux (2006), even if respondents honestly attempt to disclose earnings, some CPS measures are inherently subject to errors in estimation.\footnote{Lemieux shows that estimates of yearly earnings show greater variation than those of hourly earnings and asserts this is due to the high prevalence of hourly paid workers in the United States misestimating annual earnings.} IRS penalties should discourage inaccurate data whether due to purposeful misreporting of income or to misestimation on the part of tax payers. Finally, IRS data avoids top-coding and thus gives a much more precise measure of high earners and, in fact, tends to oversample these same individuals.\footnote{The sample drawn for the SOI represents a higher portion of the population for higher AGI brackets, the highest of which may include every individual in the actual population.} By contrast, the CPS top-codes above certain income thresholds. While top-coding is clearly problematic for observing the very highest earners,
Schmitt (2003) shows how this censoring extends to lower earners in the distribution.\textsuperscript{13}

With these advantages there are also specific challenges inherent to IRS data. First, there is no breakdown of the components of income at the state-level for all years in the sample period. To address this, we will rely on a measure of earnings that is consistently available: the pre-tax AGI.\textsuperscript{14} While this may mask the effect of taxation and redistribution on inequality, pre-tax data will be more useful for explaining structural changes in the economy that influence inequality. Additionally, while the use of the AGI may limit precision in discussing specific types of earnings, it offers a more complete picture of inequality. Many labor studies actively exclude classes of earnings such as self-employment or focus on hourly wages. The former leaves out an important stratum of the population, while the later ignores the extensive margin of earnings (i.e., hours worked) and other sources of income.

As in many other income surveys, AGI is presented in binned form. Binned income data do not allow for observing each individual surveyed, but instead present data on the count of the population and the total amount of income falling within a given range of income values. Because IRS SOI binned data report this population in terms of “tax-units,” which include both single and joint filers, our results can be interpreted as measures of household inequality, distinct from individual level data which are common in the literature. The difference between the two allows for a different set of questions to be asked. In the case of household level observations, the data allow for addressing issues such as the complementarity or substitutability of spouses’ labor supply choices. For example, if a spouse experiences a decrease in their hourly wage rate, it may induce their partner to seek additional hours at their employment, or if they had not been employed, to enter the workforce. Effects of this type that occur on the extensive margin would be missed in literature concentrating on hourly wages and individual workers. Yet, understanding such effects is critical for many policy questions, especially in cases where households’ consumption ability is of concern. In addition, household measures are arguably more directly related to certain inequality questions such as the strength or the hollowing of the middle class. Other studies dealing exclusively with individuals beg the question of what qualifies as middle class for households

\textsuperscript{13}Within the years studied, top-coding can be present for up to 10.3\% of the workforce.

\textsuperscript{14}State-level data on this measure were not published by the IRS for the years 1982-1988.
with two earners.

The usage of AGI as a measure of interest follows authors such as Frank (2009,2014) and Feenberg and Poterba (2000). All measures in these works report shares of income and measures of inequality that use AGI as the measure of group income and total AGI of taxpayers as the total income for the population of interest. This is a different treatment than those by Sommeiller and Price (2018) and Piketty and Saez (2003) who calculate total national and state income from Department of Commerce and Bureau of Economic Analysis data respectively. Both these studies make similar adjustments to AGI for their calculation of income shares. While these treatments differ slightly, the qualitative results in both cases remain extremely similar, suggesting that regardless of the precise definition of income the trends discussed would remain consistent.

3 Method

Because Adjusted Gross Income data are presented in “bins,” using AGI for measuring inequality presents specific challenges. The binned income data in the SOI only provides information on the number of filing units and the corresponding total AGI for each given bin. Additionally, the number of bins and their thresholds (i.e., the values of AGI at which each bin begins and ends) change throughout the sample period. This means constructing measures such as real earnings at the 10\textsuperscript{th}, 50\textsuperscript{th}, and 90\textsuperscript{th}, percentiles would be impossible from the raw data.\footnote{Unless population demographics allowed AGI brackets to somehow coincidentally fall exactly along these specific percentages, a very unlikely statistical fluke.}

Even more problematic is the fact that if the share of income and population within each bin were to be used for calculation of inequality indices, a systematic bias in measures such as the Theil or Gini would be introduced. This occurs because even when AGI brackets are identical across states or years, there will be variations in the share of income and population falling into these brackets due to several systematic factors. Specifically, crowding of populations inside lower and upper bins occurs respectively in wealthier or poorer states. Differential crowding also occurs in years with more coarsely defined bins, or even if bin
placement remains constant, in periods of inflation or deflation. Mean income and bin structure will influence the extent of this crowding. As greater numbers of individuals are binned into a given income category, calculations will be performed as if the population within that category had identical incomes, understating inequality to an increasing degree. Due to this mechanism, differences in mean income will create differences in the estimated Theil in a given year or state, even if the underlying level of inequality were to remain constant. This implies that as intra-bracket populations grow differently across states and time, the extent of crowding and the resulting bias of the indices will also vary, meaning that changes in the measure of inequality may be due to systematic differences in mean income instead of solely being reflective of underlying changes in inequality.

An alternative approach is to calculate continuous distributions from binned data. Unfortunately, as will be discussed, many of the methods available for this purpose are similarly sensitive to the issue of crowding within bins. This is because estimations derived from these methods are based on the data that are altered by crowding. To our knowledge, this important issue has been overlooked in the literature. What follows is a discussion of several popular methods for dealing with binned data, and a summary of why each method may be sensitive to such crowding. Along with the intuitive explanations of the mechanisms which influence these estimations, a series of simulations will be performed with their specific results presented in the methodology appendix. The goal of these simulations is to mirror the nature of the sampling and binning procedures that produce the data structure presented in the IRS SOI. To this end, an underlying distribution of household incomes will be assumed. This is meant to represent a population of tax units. From this continuous distribution, random samples are drawn to represent the way in which the IRS selects a sample of tax returns. The income from these samples are then binned in order to conform to the bins used by IRS in the SOI. As an underlying distribution is assumed, the underlying inequality measures for that distribution are known prior to the sampling and binning. It is then possible to compare the true measures calculated from the assumed distributions to those that can be estimated by performing various interpolation strategies on the constructed binned data.

To show how mean income differences across states or over time produce spurious
estimation in the measurement of inequality, we perform a scaling exercise. Specifically, we rescale the sample to effectively create several simulated populations. This way, the true Theil index is preserved and remains identical across our simulated populations as the relative income shares income per individual are unchanged. The next steps are crucial: First, the scaled data are binned according to the SOI structure. Second, from this binned data, continuous distributions are interpolated using various prominent methods that have been proposed or used in the literature (see below). Finally, the resulting Theil indices are calculated. This is meant to test whether crowding will create spurious results in measures of inequality. Ideally, if the population Theil index did not change, but the mean income levels across simulated populations did, a proper estimation technique performed on bin data would have produced the same Theil value for each simulated population. In this case, one would not have observed any pattern of misestimation caused by crowding induced in this way. Following this logic, if the estimated Theil is observed to differ systematically from the true value, the underlying source of discrepancy must be the combined effects of the changed presentation of the binned income data and the method used to create a continuous distribution from that data. Given this characterization, we will be able to establish the influence of bin crowding on various estimation methods. Specifically, for several of these methods, we will formally demonstrate the existence of a systematic bias in the measure of inequality that is correlated with the level of scaling and therefore the mean income of a given population.16

A commonly used and generally simple method for calculating inequality is the Mid-point approximation. Heitjan (1989) shows that this method yields reasonable estimates of the population mean and with some adjustment, the population standard deviation, although these results depend on strong assumptions on the nature of binning.17 While Heitjan focuses on the estimation of the parameters discussed using his method, the use of this method for calculation of inequality indices will involve a crowding issue very similar to that discussed above. This is because this method assumes all individuals within a bin

---

16For all the methods which follow, a more detailed discussion of how the methods are implemented as well as a discussion of particulars of the various simulations and their results can be found in the Method Appendix.

17Bins are assumed to be of equal width and bounded on both ends. For the tax data used here, the widths of bins are never equal, and the top bin is always unbounded.
have identical incomes, usually set equal to the mid-point of the bin involved. While this method does provide the ability to use constant percentages of the population in order to calculate inequality indices, it will still have the issue of assuming perfect equality among increasing portions of the population as the crowding becomes more severe. Thus, with inflation or deflation and high or low income states or years the Theil will be systematically underestimated.\footnote{For our purposes bin means will be used instead of mid-points. This method offers greater precision and does not require assumptions on the top income bin, which is unbounded in IRS data, and therefore has no known mid-point. Bin means were not available in Hetijan’s dataset, but this method will display a similar pattern of bias to his mid-point method.}

Cowell and Mehta (1982) offer several alternatives for estimations of inequality based on binned data. Among their proposed methods is the Split-Histogram method which has been extensively used in the inequality literature. This assumes two uniform distributions within a bin which are split at an arbitrary value within the bin, customarily the mean. Then the share of individuals assigned to either side of the split is determined by the relative distance between either end point of the bin and the chosen point within the bin. Following the most common practice in the literature, we will use the mean as our split point within the bin. Crowding can influence this estimation technique through two channels. First, crowding will change the placement of the mean within a bin, and therefore change where the split occurs. Secondly, and as recognized by Cowell and Mehta, the uniform distribution effectively creates a linear distribution function for income, and this assumption will create systematic misestimations dependent on the true shape of the distribution within a given bin. As further crowding occurs, and more of the underlying distribution falls into a given bin, this issue will be exacerbated. As demonstrated in the methodology appendix, these two effects will produce a downward bias in the measure of inequality.

More recently, Blanchet, Fournier, and Piketty (2017) have proposed a non-parametric method. This method fits polynomial splines within bins in order to create a smooth, continuous income distribution from binned data. In order to fit such a distribution, their algorithm requires data on the beginning and end points of each bin, as well as the average income within that bin. It is clear that crowding will change the latter, but this method is also sensitive to such effects for another reason. It requires parametric fitting in the top
and bottom bins. As crowding becomes more severe, and more income falls into these bins, it is likely that the distributional assumptions for these bins become less accurate, thus increasing the misspecification error described by the authors. In fact, they even discuss the ideal concentration of population within bins, and the standards they set are often violated by the nature of the data found in the SOI and other income surveys. As detailed in the Appendix this method also produces biases similar to those of the Split-Histogram method. While the precise nature of the mechanism which creates the bias in this method is somewhat complex, due to the influences discussed, the fact remains that crowding produces systematic biases in measures of inequality.

These shortcomings motivate the need for a strategy which would allow estimation of population and income share data such that inequality estimates will be consistent and free of the type of biases outlined above and demonstrated in the Method Appendix. Our proposed strategy, which meets these requirements, is a least squares method. We will demonstrate that this technique is free of all these biases, enabling it to make meaningful comparisons over time and across states, an advantage that has been wanting thus far. While this technique has been used in the literature, it is less common than other strategies for dealing with binned data, and has not been used for the IRS data, or the inequality measures that are constructed from the estimation procedure adopted here.\(^{19}\)

The data reported in the SOI effectively provides measures of empirical quantiles for the distribution of AGI according to the brackets provided.\(^{20}\) For the least squares technique, using an assumed distribution, the parameters are allowed to vary in order to minimize the sum of squared distances between the percentiles calculated from a theoretical distribution and the empirically observed population percentiles that fall at or below a given level of AGI. The parameters which minimize this measure are then selected to describe a continuous distribution. If the assumption on distributional form is accurate and the measurements of the population within classes are unbiased, this strategy will result in unbiased estimates.

---

\(^{19}\)This technique has been used by Chotikapanich (2007) to test the appropriateness of using a mixture of gamma densities to estimate Canadian income data. Pinkovskiy and Sala-i-Martin (2009) use the same technique to estimate lognormal distributions of world incomes to study poverty rates.

\(^{20}\)AGI classes have upper and lower bounds and total numbers of filers within that class. By dividing the cumulative sum of all groups by the total number of tax units, it is straightforward to obtain the percent of individuals falling below a critical level of income.
of parameters of the underlying distribution of earnings. As a result, this method will also yield unbiased estimates of inequality because it is not influenced by scale driven differential crowding complications which affect the performance of the previously described methods. This point is demonstrated via the same simulations performed on the alternate methods. Finally, as noted by Pinkovskiy and Sala-i-Martin (2009), this type of parametric estimation helps to correct for misreporting at the tails of the distribution. This is an advantage over other strategies which tend to overfit to potentially inaccurate values at the tails.

The above strategy requires the assumption of an underlying distribution of earnings in the first place. In order to establish a method for making such an assumption, we will first discuss several candidate distributions whose use has prevailed in the literature. We will then present a method that permits us to select the distribution that has the best fit. This will also allow for the flexibility to assume a best fit distribution that can vary across time and states.

An extremely popular assumption for estimation of income has been the lognormal distribution. Gibrat (1931) offered the theoretical basis for why a lognormal distribution approximates earnings for a majority of the population\textsuperscript{21} and Atkinson (1975) and López and Servén (2006) demonstrated that the hypothesis of a lognormal distribution cannot be rejected. This distribution is also known to do well for the main body of the earnings distribution\textsuperscript{22} which is the focus of this paper.

However, despite the popularity of the lognormal distribution and its wide use, several other distributions have been found to be suitable, especially for households and for specific time periods. For example, Salem and Mount (1974) find that the gamma distribution provides a superior fit to lognormal for U.S. household income in the 1960s. Bandourian, McDonald, and Turley (2003), who also concentrate on household distributions, show that the Weibull distribution offers the best fit of a variety of two parameter distributions for several European countries as well as the U.S. for select years between 1967 and 1997.

The availability of public use micro-files from the IRS from 1962 onwards allows us to

\textsuperscript{21}Gibrat creates a theoretical model that assumes that current earnings result from past earnings multiplied by a random component. This results in a lognormal distribution.

\textsuperscript{22}Estimates of percentage cutoffs for best performance vary slightly, but suggest that, when excluding the extreme tails, this is the most appropriate distributional form.
select a method which choses the best distribution among the discussed candidates and to check the validity of our general strategy.\textsuperscript{23} The details of the method are explained in the methodology appendix. As an overview, we are able to demonstrate that the best performing distribution, i.e. the distribution that provides the closest estimate of an inequality index as compared to one calculated for the underlying population, is also the distribution which demonstrates the lowest Jensen-Shannon divergence. Divergence measures are especially appropriate in this case, as they allow for a comparison of the best fitting distribution when there are several candidates to select from. In addition, other goodness of fit tests such as the Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Chi-square (CS) have been shown by Jäntschi and Bolboacă (2017) to reject the null hypothesis of zero divergence between two distributions when outliers or heavy tailed distributions are present. This could be especially problematic for our purposes as the Weibull and Lognormal, two of the candidate distributions, are themselves heavy tailed. The divergence here is calculated between the empirical quantiles from the binned data, and the theoretical quantiles calculated from the parameters of each estimated distribution. When this criteria is used to select distributions, the estimates of inequality obtained from the least squares method performed on binned sample data are extremely similar to the true measures for the underlying sample. As such, the distribution for each state and year will be selected by this criterion.

Using this strategy, distributional parameters will be recovered by year and by state that fully describe their respective earnings distributions. From these parameters, percentages of the population and of total earnings can be calculated in order to measure annual state-level Theil indices in a manner which makes the calculated values comparable across states and time. The Theil index is of specific interest due to its decomposability property. It is possible to separate the measure of an entire population into mutually exclusive components which can be summed to the total value.\textsuperscript{24} Here, this will allow a comparison of how much of the national level of the Theil is due to inequality inside of states, and how much is due to the

\textsuperscript{23}While these micro files provide an excellent check on methodology, they themselves are not appropriate for estimating the measures of interest here. They represent a shorter time period, are not weighted for state-level estimation, and are a subsample of the data in the SOI.

\textsuperscript{24}The Theil is defined as: $T_t = \sum_{i=1}^{n} y_i \ln \left( \frac{y_i}{x_i} \right)$, with $y_i \& x_i$ as the percentage of income and population comprised by individual $i$ respectively. This can be decomposed into a within and between group component: $T_t = \sum_{k=1}^{n} y_k T_k + \sum_{k=1}^{m} y_k \ln \left( \frac{y_k}{x_k} \right)$, with $k$ groups or regions.
divergence in the average income across states. In addition, the individual patterns of each of these measures can be observed for the sample period. The estimated parameters will also allow discussion on changes in the structure of real earnings.

4 Results

Using the above method, Figure 1 presents estimates of the Theil index at the national level, as well as its inter- and intra-state components. The first striking feature is a general increase in inequality, starting in the immediate aftermath of World War II. Even with this general trend, one observes episodic and relatively more severe increases in inequality for the periods 1947-1956, 1963-1969, and to a lesser extent, across the 1980s. This is especially notable as the first two periods pre-date (completely or partially) the coverage that has been available to most previous studies concerning the group of interest here. This result challenges the popular belief that the rise in inequality is a more recent phenomenon, at least

![Figure 1: National Theil, between-state, and within-state components, calculated using state-level distributions, estimated with the least squares method, using Jensen-Shannon divergence to select distribution for each state-year observation.](image)

---

\(^{25}\)Many data points are missing for the 1980s as the IRS did not publish the necessary tables for these years. But because data are available for the beginning and the end of this decade, a general trend over this time period can still be discussed.
among households. In fact in most other periods, the growth of inequality is surprisingly steady, with more extreme upward trends limited to the discussed time frames. Another result emerges from the decomposition of the National Theil index in Figure 1. Here it is shown that the largest component and primary driver of national inequality is within-state inequality. While the importance of this component is at its lowest in the 1950s, it never represents less than 90% of the national measure. This result is not surprising, given that the between-state component is effectively calculated using the average income in each state. The relative size of these measures reflects that there is much greater disparity of income within states than there is between states’ average income.

Nonetheless, the between-state component of the Theil is still informative for analyzing inequality across regions. Figure 2 summarizes changes in this measure. Divergence between state average incomes rises sharply for nearly a decade after the start of the sample period. It then declines until the late 1970s, albeit with greater variability. This is then followed by an increase in the inter-state divergence over the 1980s. Inter-state divergence then declines and is stable for the 1990s but oscillates after that point. To recap, the most important lesson from this aspect of the paper, afforded by our new data, is that inter-state divergence

---

**Figure 2:** Between-state component of National Theil calculated using state level distributions, estimated with the least squares method, using Jensen-Shannon divergence to select distribution for each state-year observation.
increased dramatically in the 1950s and was followed by an equally dramatic decline in mid 1960s to mid 1970s. Subsequent increases over the next five decades never reach the 1950s level of inter-state divergence.

One potential issue with using the Theil or any other single-valued measure of inequality is that a given change in an index could be the result of different underlying changes in the distribution of earnings. To address this issue, and to uncover the underlying changes in our case, the above estimation technique used for each state is repeated for the country as a whole. AGI groups were aggregated for the United States, and distributional parameters were estimated at the national level. The Theil calculated from this aggregate estimation is presented in Figure 3, along with the Theil calculated by state-level estimates of distributions, which are subsequently used to estimate the within and between state components of a National Theil. As can be seen, the National Theil calculated by estimating a distribution for the nation as a whole shows an extremely similar pattern to the prior treatment, both qualitatively and quantitatively, implying that the prior results are not driven by varied state-level distributional assumptions.

Estimating the Theil index at the aggregate (national) level, one can then explore the

![Figure 3: Comparison of National Theil calculated by state-level distributions versus national distribution estimated with least squares method, using Jensen-Shannon divergence to select distribution.](image)

16
underlying dynamics that influence the shifts in the overall index. To do this, we compare the share of income held by the top and bottom of the population over the entire period. Figure 4 shows the relative share of total national earnings of the top and bottom 10% of the group of interest here (the 5th-14th percentile and the 86th-95th percentile of the total population). By comparing this trend to that in Figure 3, we note that nearly all of the “dips” in the Theil index correspond to similar dips in the percent of earnings of the top 10%. Of course, the resulting decrease in inequality is only temporary as it bounces back along with the share of the top income group. On the whole, the share held by the relatively wealthy has risen over the entire post-WWII period. Moreover, the trend of the income share held by the bottom is largely downward sloping. This decline is seen as early as the mid-1940s to 1950s. These early post-war observations are documented in this paper for the first time.

But this does not tell the full story. So far, we have exclusively studied trends in relative shares, not in the levels of real earnings. This distinction is important to differentiate between possible causes for the same change in the discussed measures. For example, small “dips” in inequality, or decreases in the percentage of income held by top earners, could be caused

![Figure 4: Percent of total national income held by the top and bottom 10% of earners of interest (5-95% of total population), calculated using Jensen-Shannon divergence to select distribution.](image-url)
by; decreased real income at the top with no changes at the bottom, income gains solely for lower earners, or simultaneous changes in both groups. The reverse could hold for an increase in inequality. To understand such changes better, the real earnings of the 10th, 50th, and 90th percentile of earners are shown in Figure 5. As can be seen, the lowest earning households show stagnant levels of real earnings for the entire sample period. The middle of the distribution shows steady, if modest, growth from the mid 1940s until the mid 1970s with a somewhat flat and even slightly declining profile thereafter. The highest earning households however, experience a rapid growth of earnings until 1968, followed by a smaller upward trend over the 1980s, and from the late 2000s until the end of the sample, interrupted by small short-run dips. The figure demonstrates decisively that the gap between the top and middle earners grew over the entire period. The dramatic jump in the earnings of the 90th percentile begins in 2011, likely associated with the sharp increase in the equity market boom post Great Recession, accompanied by the absence of any rise in real income of the mid-income group, dramatically increasing the extent of the real earnings gap.

It is again apparent that changes among high earners are the primary driver to changes in the Theil. As can be seen by comparing Figures 3 and 5, dips in inequality represented by

![Figure 5: Real earnings at the 10th, 50th, and 90th percentile of the national population in 1982-1984 adjusted dollars and using Jensen-Shannon divergence to select distribution.](image-url)
the Theil are principally linked with drops in real earnings in this top income group and less by gains among others. The fact that the episodic declines in inequality, when they occur, are due to the reduced real earnings of the top earners, rather than increased earnings of the bottom, is somewhat at odds with the popular notion of the source of increased equality.

Finally, one of the purposes of this work is to establish a dataset for future inequality studies. The process which provided the national results above required first estimating state-level, annual distributions of earnings. While a full analysis of inequality across 48 states and 70 years is beyond the scope of this paper, these estimations lay the groundwork for future inequality studies, especially those involving regional effects. As both an application of these estimations and an illustration of the income trends they suggest, state-level measures of inequality are estimated and presented in the Results Appendix.

5 Conclusion

The purpose of this paper has been to establish a dataset of state-level earnings for the main body of earners (5th-95th percentile of the population), across a longer time period (extending back to WWII), and with a level of geographic precision (the 48 contiguous states) that had not been previously achieved. This is done for two reasons, to provide a more complete picture of the history of U.S. earnings inequality, and to help with future analysis of the causes and outcomes of inequality. These contributions are especially important given the intense debate among scholars and policy makers over the drivers and the consequences of inequality in the United States. The omission of the main body of earners in the early post-WWII period has deprived researchers from subjecting various hypotheses to the data from this formative and critical period. This work is now possible with our contribution.

The mapping from IRS binned data to a continuous distribution for these purposes has called for selecting a method that outperforms alternatives in a non-trivial manner. The results presented in the method appendix show that the least squares estimation here to not only be consistent, but also superior to others for dealing with binned data for this our purposes here and similar applications.

The main findings here are that while inequality has generally grown in the U.S. over
this period, there are noticeable spikes, some of which occur earlier than prior literature has demonstrated. Additionally, when decreases in the measure of inequality are observed, this is generally due to income losses by high earners, rather than gains of low earners. A clearer picture of the history of the middle class also emerges. While the gap between the middle and top has consistently increased across the sample period, the middle class did exhibit gains in real earnings early on. However, this trend has flattened in the recent past. These results are especially important as much of the current literature seeks to explain changes that occur in the later part of the sample period here. Those explanations, while sufficient to explain changes in inequality in their respective time periods, are often specific to that time and therefore may not explain what factors influence inequality in the U.S. historically. Due to this, they do not establish the only avenue for changes in inequality, and do not provide a complete roadmap for avoiding or promoting certain changes in the income structure of the nation. Also, by following a strand of the literature which estimates inequality at the household level, the dataset created here is more suitable for discussing specific trends in inequality and policy implications. These qualities of the data make it especially well-suited for answering a host of economic questions, including those regarding labor market institutions, skill premia, and the middle class.
References


Cilke, J., “A Profile of Non-filers,” U.S. Department of the Treasury. OTA paper 78.


6 Data Appendix

Slight variations in the way that AGI data are structured from year to year require some effort to make sure distributional estimates are based on as similar data as possible. Before 1955 individuals with AGI summing to zero or negative values were entirely excluded from the publications. For all the years after this date the IRS either includes these returns in the lowest AGI bracket or includes them as a separate category. It is impossible to perfectly disentangle the sub-categories within these bins, which include both positive and negative AGI. As such, for those years for which the number of tax units with negative AGI is available as a separate measure, those units will be included in the lowest AGI bracket. Because of this, the first ten years may not be entirely comparable to the rest of the sample, as they are the only years lacking data on negative AGI completely, but most of the years will be of similar construction for use in estimation. A separate issue is the presentation of taxable and non-taxable income. For the years 1959 until 1962, taxable and non-taxable returns are presented separately. For these years all AGI brackets of taxable and non-taxable income are combined where possible, excluding any category of non-taxable AGI which does not have an upper bound. Those tax units most likely constitute several taxable categories and require strong assumptions for combining them into overall totals. Due to these factors, income distribution measures for the specific years discussed may also not be fully comparable. In years when negative and positive, or taxable and non-taxable AGI are presented separately, estimations of inequality are made both including and excluding negative and non-taxable AGI for the purpose of comparison. This exercise shows that while some quantitative results slightly change, the qualitative trends in inequality are robust to both treatments.

Additionally, several years are absent from the sample completely, from 1982-1988 the necessary data is not made publicly available in any form.26 Also, in 1952 the counts for Washington and Alaska are combined, in 1961 Maryland and the District of Columbia are combined, and in 1962 Delaware is not included in the reported IRS SOI data.

---

26 Additionally, the years 1975, 1977, and 1981 are currently excluded due to low quality scans of the documents from which the data originates, meaning the underlying values cannot be currently verified. 2002 is also excluded, as it is not clear in what manner it deals with negative AGI.
7 Methodology Appendix

A particular quality of the Theil index, or similar indices such as the Gini, is their scale independence, due to the fact that these indices are based on relative population and income shares. An estimate of the Theil should then exhibit the same ranking of inequality regardless of the scale of income. This poses a problem for any source of data which bins by values of income and not by percentiles of the population. While prior literature has alluded to this problem by discussing the importance of the structure of bins to the calculation of various inequality measures, it has not directly addressed this issue, related to average incomes, or the difficulties that it causes for measures of inequality. The simulations which follow will build upon this prior work by showing that not only the placement of bins, but the relative crowding within bins is extremely important to these patterns. Specifically, by increasing or decreasing mean earnings in a given population, concentration per bin is altered, just as it would be if the underlying structure of survey reporting was changed. In particular, as nominal income increases or decreases crowding will occur in high and low income bins, respectively. As crowding becomes extreme, increasing portions of total population and of total income will be crowded into those bins, causing a variety of technical challenges for many estimation techniques. What follows is a demonstration of the biases created in the Theil estimated by various techniques, as well as a lack of such bias in the technique used here.

The simulations at hand demonstrate conceptually what would happen to the estimates of the Theil, if different states had the same segments of the population receiving the same percentages of earnings, but differed in average earnings. The critical relevance of this exercise rests on the fact that individual incomes within each state are binned in a uniform manner, not by percentages of the population, but by the AGI brackets set in the SOI. Thus any differences in subsequent estimates would be due solely to the differences in average earnings across states. It should be also noted that the mechanism for these differences is

\[27\text{Early work by Seiver (1979) demonstrates the size and number of bins can affect inequality measures, but only uses two alternative binning schemes. More recent work by Blanchet, Fournier, and Piketty (2017) demonstrates that the number and placement of thresholds for bins can affect the error on estimates for top earners.}\]

\[28\text{The index will be calculated by several different methods which will be described below.}\]
that differences in inter-state average earnings influence the number of individuals in the bins (e.g., more crowding at the top (bottom) bins for higher (lower) income states) and this in turn influences the performance of various estimation techniques.

For this purpose, income bins of the most recently available year (2015) of the IRS SOI are used. This year represents some of the coarsest bins in the sample period. It thus demonstrates the problem in its most severe form, making these bins ideal for illustrative purposes. Other years provide income data using more closely spaced bins at both the high and low end of the distribution. While performing this analysis on these years will result in extremely similar patterns to those to be outlined below, the bias present for such years would be less severe. For 2015 the lognormal distribution is found to be the most appropriate distribution, both nationally and across states, based on various goodness of fit tests. It should be pointed out however, that when other distributions prevail in other years of our sample, again based on goodness of fit tests, the pattern of bias for each method remains similar and the resulting estimates from the least squares methodology remain consistent.

Given the lognormal distributional assumption, parameters are then chosen to allow bins to be populated in a manner similar to those observed across states in the actual IRS sample for 2015. From this distribution 1000 random draws are taken representing a sample of 1000 incomes. Once drawn, each observation of individual incomes in that particular sample is then scaled by a factor of 0.5 to 5 in 0.1 increments. To illustrate, each of the 1000 observed incomes comprising a given sample is first multiplied by 0.5 to create one theoretical population, next the original sample is multiplied by 0.6 to create another population and so on. This effectively creates 46 theoretical states in which each of the 1000 individual incomes preserve their percentage of total income while average incomes differ. Of course, this also implies that for any given portion of the population chosen, the percentage of income held by that portion of the population will be the same in any of the constructed “states.” As such, each theoretical state will only exhibit different measures of inequality

29 Using the bounds 0.5 to 5 allows for a range of scaled incomes even greater than may normally be seen across states or time. In addition, these bounds were selected as they allow for consistent comparison of all techniques which follow. Each technique requires some minimum number of bins to be populated in order for any estimation to be performed.

30 It should be noted that this exercise is not meant to create “states” representative of the contiguous U.S. states. Any number of theoretical states could be produced at a given scale parameter, strictly for the purpose of observing bias created by a given estimation method. In order to create the type of spread seen
due to the influence that crowding has on a given estimation technique.

Each of these theoretical states then have their 1000 observed incomes binned according to the AGI classes discussed. This binned data can then be used to estimate inequality by several methods proposed in the literature for such a purpose. This approach addresses inter-state and inter-temporal comparability of relative inequality measures created using these methods, with differences in estimated measures across the theoretical states arising from state-level average income differences, rather than from actual distributional differences.

Before we consider the results, we address possible sampling errors by repeating the above simulation 1000 times, creating 1000 samples of 1000 draws which each in turn are transformed into the 46 theoretical states as explained. If an estimation technique is consistent, as more samples are drawn, again barring any effect from binning, the median Theil calculated at each multiple (each individual theoretical state) should approach the value of the Theil calculated for the underlying population which was assumed. This value can be calculated directly from the density of the distribution from which the described samples are drawn.

The entire process described will then be repeated, allowing the scale parameter of the underlying population to vary from 1 to 1.3 in .02 increments. These will allow observed bins to vary in a manner that is similar to that observed across states in data presented in the IRS SOI. It is important to note that the step of multiplying random draws of income is equivalent to allowing the location parameter, assumed to be 3.5 here, to vary. That is to say, changing the location parameter of an underlying lognormal distribution should not change the value of the Theil for that distribution, as long the scale parameter remains constant, again barring any influence the binning procedure may have on the estimation technique. So, theoretical states could also be created by changing the location parameter, but some intuition about the underlying relative levels of average income would be lost.

---

across actual states, the scale parameter is allowed to vary, as discussed later.
7.1 Direct calculation from empirical population and income percentages

The first, and perhaps most obvious strategy arises from the fact that indices such as the Theil and the Gini can be calculated using only portions of the population and their corresponding portion of total income which fall into given groups. The SOI provides these measures directly, with the groups being defined by AGI brackets. Given this, it may seem attractive to calculate these measures directly from such data. The results of such a strategy are presented in Figure A.1. The horizontal line in this figure is the Theil that would be calculated using the underlying continuous distribution segmented at the critical values from the AGI buckets described. As can be seen, the median value of the samples for a given income multiplier shows a downward bias for richer states. This occurs because no inequality is observed within bins. The result for both relatively high income states is to understate actual inequality even if underlying income and population percentages match across states. Additionally, bands in which 90% of the samples (the middle 900 Theil indices for each sample generated by scaling incomes) fall demonstrate that this range often does not cover the Theil calculated using the underlying distribution and percentiles implied by the 2015 AGI brackets. It is immediately clear that some other strategy is necessary if this type of data is to be used to construct Theil indices. To this end and in search of a preferred strategy, each of the following strategies will be examined using the exact same simulated data.

There are interpolation methods available to deal with this data, and these strategies can be used to estimate a Theil which can be compared directly to a “true” underlying Theil. The following methods allow for separating the population into consistent portions, namely 10% sections of the population, instead of varying population percentages represented by AGI bins. As such the “true” Theil here will be calculated by using the underlying distribution which is assumed, and calculating the total income that falls into each decile of the total distribution. By performing the same segmenting with each strategy that follows, it can be seen which does the best job at estimating inequality of the underlying population.
Figure A.1: Comparison of the Theil found for a theoretical state with unmultiplied income, to the Theil of “states” with varying average incomes. Theil calculated directly from proportions of income and population in binned data. Distributions are assumed to be lognormal with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Bands represent the central 90% of estimates from random sampling from the distribution.
7.2 “Mid-Point” Approximation and the “Mean” method

A very similar strategy, in both implementation and results, to the first strategy described above, is the so called “Mid-Point” approximation. This method is performed by assuming all incomes within a bin have an income equal to the average of the upper and lower thresholds for that bin. The method we will refer to as the “Mean” method is very similar, but assumes all individuals within a bin have income equal to the mean income in that bin. The only advantage of the “Mid-Point” method is that it does not require information on group means. Both methods make an assumption on the income of every individual in the population, and as such the population can be segmented into any group structure desired, and from the income and population share of those groups, the Theil or Gini can be calculated. The “Mean” method offers slightly greater precision in measurement and, as such, is the method that is demonstrated here. However, both methods will result in extremely similar results. The results for the “Mean” method can be seen in Figure A.2. The major differences between the results here and those from the first strategy are that while the Theil index is generally underestimated in both, and while the bias tends to exhibit a similar curve in both, the curve is more irregular in Mean method. The shape of this curve is likely due to how the crowding in bins interacts with the subsequent segmenting into equal population portions.

7.3 Split-Histogram

Another common strategy for binned income data that is used pervasively is the split-histogram method. This strategy requires data on the endpoints $[a_{Θ}, a_{Θ+1}]$ for each bin, an arbitrary point in each bin $b_{Θ}$, the mean for each bin $μ_{Θ}$, the population for each bin $n_{Θ}$, and the total population $n$, all of which are available from the SOI AGI data. From this data the distribution within each bin can be found using the function:
Figure A.2: Comparison of the true population Theil to an estimate found using the “Mean” method on binned data when assuming a lognormal distribution with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Bands represent the central 90% of estimates from random sampling from the distribution.
\[ f(y) = \frac{n_\Theta}{n} \frac{\alpha_{\Theta+1} - 2\mu_\Theta + b_\Theta}{\left[\alpha_{\Theta+1} - \alpha_\Theta\right]\left[b_\Theta - \alpha_\Theta\right]}, y \in [\alpha_\Theta, b_\Theta) \]

or

\[ f(y) = \frac{n_\Theta}{n} \frac{2\mu_\Theta - b_\Theta - \alpha_\Theta}{\left[\alpha_{\Theta+1} - \alpha_\Theta\right]\left[\alpha_{\Theta+1} - b_\Theta\right]}, y \in [b_\Theta, \alpha_{\Theta+1}) \]

Performing interpolation using this method, one can again estimate the income of every individual in a population. Using consistent segments of this population and calculating the Theil also leads to a pattern of bias correlated with average income. This pattern, while distinct from the methods discussed thus far, is still problematic. As can be seen in Figure A.3 this strategy will result in underestimation of the Theil to an increasing degree as the “states” become wealthier. While this bias is smaller than that created by the “Mean” method, it is still systematic.

7.4 Generalized Pareto

In their work, BFP establish a technique for estimation of continuous distributions from binned data. Referred to as the “Generalized Pareto” method this strategy is non-parametric in nature. For estimation, the portion of the population within given bins, the thresholds for the bin, and the average income within the bin is needed to fit polynomial splines of degree five within each bin. The top bin is fit to a generalized Pareto distribution, and the bottom bin is fit using methods not fully outlined by the authors, but described as being parametric in nature. Using the simulated data described, the consistency of this method can be checked in the same manner as the other methods. The results of this exercise can be seen in Figure A.4 and are strikingly similar to those of the split histogram method. Here, much of the bias observed at higher incomes may be due to the fitting of a Pareto distribution to an increasing portion of the population which falls into the top bin. As much of the literature suggests this to only be a proper fit for top earners, it is clear that increasing the population in the top bin is problematic for this method.
Figure A.3: Comparison of the true population Theil to an estimate found using the “Split-Histogram” technique on binned data when assuming a lognormal distribution with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Bands represent the central 90% of estimates from random sampling from the distribution.
Figure A.4: Comparison of the true population Theil to an estimate found using the “Generalized Pareto” technique on binned data when assuming a lognormal distribution with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Bands represent the central 90% of estimates from random sampling from the distribution.
All the biases described so far are related to the problem of “misspecification error” arising from a difference between an underlying distribution and the distributional form used to estimate it. This issue is described by BFP and is distinct from measurement error. While BFP establish where the biases described arise from, they do not formally demonstrate the systematic nature of the biases described here.

7.5 Least Squares Method

We will now demonstrate that, under a correct distributional assumption, the application of the least squares methodology will produce unbiased results. As discussed in the methodology section, while least squares methodology is far from novel, its application to the estimation of inequality in the present context is. To demonstrate consistency the same binned data of theoretical states described earlier will now be used. The method consists of estimating the underlying parameters of the distribution of the theoretical population from which the samples are drawn, using the least squares estimation technique. A Theil index is then calculated for each of the samples, using consistent 10% shares of the population. The median, as well as a band composed of the central 90% of the estimated values from this sampling and estimation procedure are shown in Figure A.5. As can be seen, the line representing the median sample estimate and the true Theil for the population overlap almost exactly. Further, given the similarities in the spread of the 90% band between the split-histogram, generalized Pareto, and least squares results, it would appear that the loss of this bias does not come at any obvious expense in other measures of performance.

This exercise is not only important for cross state analysis in a given year, but for comparison of the Theil between years. If AGI bins remain the same, but states or the nation experience inflation or deflation or economic growth or recession, this would create a bias in the same manner discussed for relative income levels. These biases and the consistency of the estimation technique demonstrate a similar pattern for any given year’s survey structure and income data and therefore the values of the Theil presented in the paper should show no such bias either between years or between states. Further, the work here suggests that even the best survey reporting design is subject to producing these types of biases, if bins
Figure A.5: Comparison of the true population Theil to an estimate found using the Generalized Least Squares technique on binned data when assuming a lognormal distribution with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Bands represent the central 90% of estimates from random sampling from the distribution.
are chosen by the value of a given variable as opposed to consistent population shares.

As noted in the main body of the paper, the least squares methodology does require a strategy for selecting the assumed distribution. This strategy can be developed and tested using available public use micro files provided for select years by the IRS from 1962 onward. These files provide information on a subsample of taxpayers from the sample that the IRS uses to construct data in the SOI. This data has individual level observations with both AGI and relative weighting for that tax unit. As such it is possible to calculate inequality measures of interest for the sample population. It is also possible to bin the individual level data based on the structure present in the SOI for that year. Then that binned data can be used to estimate parameters for the three distributional candidates. These parameters are used to estimate the same measures of inequality that were calculated for the sample population. Then it is possible to see which distribution provides the best estimate of these measures. What is found by this exercise is that an entropic measure of divergence, namely the Jensen-Shannon divergence, can be used to determine the distribution that will provide the closest estimate of the Theil and other measures. The divergence is measured between population percentiles falling into the empirical bins created by SOI AGI grouping, and those which can be constructed from the distributional estimates. For every year analyzed, the distribution which provides the lowest divergence is also the distribution which provides the closest estimate of the inequality measures of interest. It is possible to calculate these divergence scores for all states and years in the sample, and thus pick a distribution which not only creates the closest estimate, but which strongly mirrors the pattern of inequality that would likely be seen if individual level data were available.

8 Results Appendix

Below we present measures of inequality which are derived from the state-level distributions discussed in the Results section. Figure A.6 presents time trends in individual state Theil indices while Figures A.7 and A.8 are the state-level equivalent of Figures 4 and 5 respectively. While the development of state-level measures of inequality will be especially significant for future studies, some general trends can be observed here. First,
the national patterns discussed above are largely mirrored within states, with a general rise in inequality and stagnant earnings on the low part of the distribution. Yet, there is also substantial heterogeneity among states. Given that many of the drivers of inequality discussed above, such as the minimum wage, unionization and returns to skills, exhibit state-level heterogeneity, our dataset will be useful for statistical inference to assess the contribution of these and other factors to the earnings distribution.
Figure A.6: Theil index for the contiguous 48 states, calculated using least squares method, using Jensen-Shannon divergence to select distribution by state and year.
Figure A.7: Percent of total income held by the top and bottom 10% of earners of interest (5-95% of total population) for the contiguous 48 states, calculated using least squares method, using Jensen-Shannon divergence to select distribution by state and year.
Figure A.8: Real earnings at the 10th, 50th, 90th percentile of the population in 1982-1984 adjusted dollars, for the contiguous 48 states, calculated using the least squares method, using Jensen-Shannon divergence to select distribution by state and year.