Web Appendix of "EARNINGS INEQUALITY AND MOBILITY IN THE UNITED STATES: EVIDENCE FROM SOCIAL SECURITY DATA SINCE 1937" by Wojciech Kopczuk, Emmanuel Saez, and Jae Song

A. Data Sample and Organization

• Covered Workers

Table 2.A1 of the *Annual Statistical Supplement* of SSA (2005) presents the evolution of covered employment and self-employment provisions from 1937 to date. At the start in 1937, only employees in commerce and industry were covered. There have been a number of expansions in coverage since 1937.

In 1951 most self-employed workers and all regularly employed farm and domestic employees became covered. The coverage has also been (in some cases electively or for new hires) extended to non-profit organizations and some state and local government employees. A further expansion to state and local employees covered under a state or local retirement system took place in 1954, followed by many smaller change expanding coverage to additional categories of state, local and federal government employees. For this reason, we eliminate from our main sample (referred to as "commerce and industry") workers that fall into categories that have not always been covered. Quantitatively, other than directly obvious categories of public administration, self-employed, farm workers and household employees, these expansions brought into the system a large number of workers in education and health care.

Self-employment and farm earnings are not reported on W-2 forms, instead SSA obtains this information from the IRS as reported on tax returns. As a result, self-employment earnings were effectively top-coded at the taxable maximum until 1993 (when the cap for Medicare tax was eliminated) and are never present in the data on a quarterly basis. All of it makes it impossible to pursue any reasonable imputation strategy above the top code in that group. Additionally, the presence of self-employment earnings may potentially interact with withholding and reporting of other types of income. Hence, we exclude individuals with other than occasional self-employment income, i.e. those who have self-employment income in two subsequent years (the number of observations affected is very small). Imputations above maximum taxable earnings from 1951 to 1977 (either our own imputations from 1951 to 1956 or the LEED imputations from 1957 to 1977) are also based solely on employment earnings excluding farm wages. Therefore, excluding self-employment earnings and farm employment earnings has no repercussions for imputations above the top code.

To exclude non-always covered industry categories, we rely on industry codes present in the LEED (starting with 1957). We exclude workers with main source of earnings in the following categories (using SIC classification): agriculture, forestry and fishing (01-09), hospitals (8060-8069), educational services (82), social service (83), religious organizations and non-classified membership organizations (8660-8699), private households (88), public administration (91-97). These categories were selected by looking at the fraction of individuals in each industry in 1957 who were present in the data in 1950, i.e. prior to expansions (when industry codes are not available). We selected categories with over 60 percent of newly covered workers (the average for the whole sample was 29 percent, with no large remaining categories exceeding 40 percent).

Between 1951 and 1956 no industry codes are present. Hence, we apply a heuristic to correct for the expansion of coverage during that period. We eliminate earnings in 1951-1956 for workers who worked in one of the excluded industries in 1957 or 1958 (we choose 1958 if there are no earnings in 1957) and who did not have any covered earnings in 1949-1950. We also eliminate 1951-1956 earnings for workers with no earnings in 1947-1950 and 1957-1960. For the remaining workers working in the excluded industries as of 1957 (who were by construction working in a covered occupation in 1949 or 1950), we randomly assign the date of joining that industry drawn from the uniform distribution on (1950,1957) and erase earnings in 1951-1956 preceding this imputed date. We verified that this procedure brings us close to matching the time pattern of employment dynamics in the 1950s.

- **Top Coding and Imputations Before 1978**

The general idea is to use earnings for quarters when they are observed to impute earnings in quarters that are not observed (because the annual taxable maximum has been reached) and to rely on a Pareto interpolations when the taxable maximum is reached in the first quarter. Pareto parameters are obtained from income tax statistics tabulations (published in U.S. Treasury Department: Internal Revenue Service [1916-2004] by size of wage income combined with the Piketty and Saez [2003] homogeneous series estimated based on the same tax statistics source. The important point to note is that we do a Pareto interpolation by brackets because the location of the top code (or 4 times the top code) changes overtime and the Pareto parameter is somewhat sensitive to the threshold of earnings defining the top tail. Each individual*year observation who reaches the annual taxable maximum is assigned a random iid uniformly distributed variable u_{it}. We describe our imputations from 1937 to 1977 by reverse chronological order as the complexity of the imputations is greater in the earlier years.

From 1957 to 1977, the 1% LEED file provides imputed earnings above the top code. This imputation was originally done using quarterly earnings information and Method II described below. The imputation was based on employment earnings (and excluding farm wages and self-employment earnings). Unfortunately, the quarterly earnings information has not been retained in the LEED file and hence we cannot replicate directly ourselves the imputation. The original Method II imputation for those above 4 times the top code was set equal to a given constant (which only varied by year and gender). From 1957 to 1977, we replace this LEED imputation for observations above 4 times the top code with a single Pareto interpolation:

$$z_{it} = (4 \cdot taxmax) \cdot u_{it}^{-1/a_t},$$

where a_t is the Pareto parameter estimated from the Piketty and Saez [2003] wage income series. a_t is estimated as $b/(b-1)$ where b is average earnings above the threshold $(4 \cdot taxmax)$ divided by the threshold. We pick as the threshold for the Pareto interpolation the percentile (P95, P99, P99.5 or P99.9) threshold from the Piketty and Saez [2003] series closest to the $4 \cdot taxmax$ threshold.

From 1951 to 1956, the 0.1% CWHS also reports the earnings by quarter (up to the point where the taxable maximum is reached). This information allows us to apply Method II [described further in Kestenbaum, 1976]. If the taxable maximum is reached in quarter 1, we do a Pareto interpolation as described above. If the taxable maximum is reached in quarter T $(T = 2, 3, 4)$, then earnings in quarters $T, .., 4$ are estimated as earnings in the most recent quarter with earnings exceeding earnings in quarter T or as earnings in quarter T if there is no earlier quarter with higher earnings.

From 1946 to 1950, the 0.1% CWHS reports the quarter in which the taxable maximum is reached (but does not report the amount of earnings in each quarter before the tax code is reached). This allows us to apply Method I to impute earnings. Method I is described in Kestenbaum [1976]. Method I assumes that earnings are evenly distributed over the year. Hence, if the taxable maximum X is reached in quarter 1, we assume that annual earnings are above $4 \cdot X$. If the taxable maximum is reached in quarter 2, we assume that annual earnings are between $2 \cdot X$ (when the taxable max is reached at the very end of quarter 2) and $4 \cdot X$ (when the taxable max is reach at the very beginning of quarter 2). Similarly, if the taxable maximum is reached in quarter 3, we assume that annual earnings are between $\frac{4}{3} \cdot X$ and $2 \cdot X$ and if the taxable maximum is reached in quarter 3, we assume that annual earnings are between X and $\frac{4}{3} \cdot X$. We assume that the distribution of earnings in each of those brackets follows a Pareto distribution estimated bracket by bracket from the wage income tax statistics. The formula for imputed earnings z_{it} in the bracket $[z_1, z_2)$ is:

$$z_{it} = z_1 \cdot \left(u_{it} + (1 - u_{it}) \cdot \frac{z_1}{z_2} \right)^{-\frac{1}{a}},$$

where a is the Pareto parameter which is specific to each year and bracket.[1] For the top bracket, the Pareto parameter is estimated as $b/(b-1)$ where b is average earnings above the threshold $(4 \cdot taxmax)$ divided by the threshold.

For each year b is obtained from the Piketty and Saez [2003] series. For brackets below the top, the Pareto parameter a is obtained from the tax statistics using the formula:

$$a = \frac{\log(p_2/p_1)}{\log(z_1/z_2)}, \tag{4}$$

where p_i is the fraction of earners above z_i and z_i are the cap thresholds X, $\frac{4}{3} \times X$, $2 \times X$, and $4 \times X$.

From 1937 to 1945, the 0.1% CWHS reports only earnings up to the top code with no additional information on quarterly earnings for those who reach the annual top code. Hence, the data are effectively top coded up to the social security taxable maximum of \$3,000 for those years. The number of top coded individuals in our main sample grows from about 3 percent in 1937-1939 to almost 20 percent in 1944 and 1945 [see Table A2 in Kopczuk et al.,

[1] The same formula applies for the top bracket where $z_2 = \infty$.

2007]. Because the relative location of the top code changes so much during these years, a single standard Pareto interpolation would not reproduce accurately the wage income distribution from the tax statistics.

Therefore, for that period, we have imputed earnings above the top code using a Pareto interpolation by brackets calibrated on the top wage income shares from Piketty and Saez [2003]. More precisely, we replicate the Piketty and Saez [2003] wage income shares for P90-95, P95-99, and P99-100 up to a multiplicative factor (constant across years) to paste our series in 1951.

From 1937 to 1956, the 0.1% CWHS contains relatively few observations at the top, hence the Pareto imputation for the top bracket can sometimes generate extreme values which can have a large impact on top income shares. To remedy this noise issue in the imputation, we randomly order top-coded observations and space them equally in the corresponding c.d.f. underlying the Pareto imputation. This method guarantees that we match the top income share exactly without sampling noise.

Note that imputations in various years are independent and that imputations are independent of any earnings information in other years that we may know. In other words, we do not try to impute the mobility patterns for top-coded observations. This procedure is innocuous for the annual income shares of groups bigger than the top-coded group because by construction it matches those share exactly. It is important to note that it also provides an unbiased estimate of top income share based on averages over a number of years if all individuals with imputed income remain in the top income group. Because in 1951-1977 imputations apply to at most 1 percent of the sample and, empirically, the likelihood of an observation falling out from the top quintile for reason other than death or retirement is extremely low, this procedure is expected to provide a good approximation of the income share of the top quintile of distribution averaged over a number of years.

• **Data cleaning**

As pointed out by [Utendorf, 2001/2002], there are a number of errors in the uncapped earnings for year 1978 to 1980 that are due to errors in the coding of the data and which bias severely top income shares and mobility measures if not corrected for. There are also some erroneous observations in some years after 1978 (although much less common).

We first explain the nature of these problems, and then describe our procedure. We also describe the procedure that to use in our ongoing work that deals with the problems more precisely and explain why it is not applied in this paper. The problems are already present in the administrative database (Master Earnings File, MEF) from which CWHS and LEED are derived. Among other things, the MEF contains information on total compensation (starting in 1978) and Social Security covered earnings derived from W-2. Each W-2 corresponds to one or more records in the database. A single W-2 may correspond to multiple records, either to accommodate multiple boxes on W-2 or to split large numbers. A single employment relationship may correspond to multiple W-2s, for example when the W-2 was later amended. Subsequent corrections of errors are also recorded as additional records in the MEF. The research databases are obtained from the MEF by aggregating information to the employer level (LEED) or individual level (CWHS). Any problems in the underlying MEF records are then potentially confounded and hence hard to detect due to aggregating them with other information. The problems in the administrative data take a variety of forms: some records are duplicated, adjustments may be made to FICA earnings but not to total compensation, typos are present and so on. Problems in the MEF are common in 1978-1980, the dominant (but not the sole) one being omission of the decimal point in total compensation figure.[2] The documentation for the MEF indicates that the total compensation in 1978 and soon after may reflect the decimal point as being in the wrong position but does not provide a way to identify affected observations. These problems affect total compensation. The (top-coded) FICA earnings are of very high quality, presumably because they are the critical input in computing benefits.

Using the MEF, these problems are hard but not impossible to identify and address by comparing FICA and total compensation, searching for duplicates, checking for the lack of adjustments to total compensation when adjustments to FICA are present and so on. An ideal correction routine would work directly on the MEF. In our ongoing work, we follow this path

[2] Another important type of problem arises when corrections to W-2 were made: they are implemented by adding two new records — one showing the amended income and another with negative income equal to the old value so that it gets offset when aggregating. In practice, these negative numbers are correctly included in the FICA field but sometimes missing from the total compensation field making aggregation of total compensation less reliable.

and work directly with extracts from the MEF. However, estimates presented in this paper rely on our earlier and more heuristic data cleaning procedure that incorporates information on total compensation and FICA earnings present in 1% CWHS and LEED. The main reason for this approach is our desire to retain consistency of pre- and post-1978 data. CWHS and LEED are derived from the MEF after about a year and are not subsequently updated to reflect any future adjustments and undergo some additional processing. Starting with 1978, CWHS and LEED can be thought of as (processed) extracts from the MEF, however prior to 1978 these datasets contain some information that is not present in the modern MEF.[3] Since MEF does not contain detail information for years prior to 1978, data cleaning procedure relying on the MEF would require replicating the process of creating LEED and CWHS to retain consistency with pre-1978 data, we did not attempt to do so. However, we rely on the 1% MEF in 1978-2004 to address another deficiency of the data. In some years a substantial number of observations is missing from CWHS but present in the MEF.[4] We investigated carefully the patterns of entry/exit from the sample and did not find evidence that such problems were present prior to 1978. Not addressing this issue would result in discrete changes in the number of observations used driven by factors other than Social Security coverage.

We proceed as follows to construct earnings variables in 1978-2004. We construct corrected total compensation for everyone as described below. However, we use FICA-covered earnings for individuals with earnings below taxable maximum and use the corrected total compensation only for those with earnings above the taxable maximum.

Our objective is to obtain a dataset that preserves information for high-income individuals and does not distort mobility patterns. In designing the data cleaning procedure, we compared income distributions, mobility patterns and joint distributions of incomes from all available sources with those for years that are not affected by these issues and with earnings distribution based on income tax records. The procedure was designed to be as conservative as possible so

[3] Obviously, how earnings histories are recorded and stored by the SSA evolved over time and the CWHS has not always been a simple extract from the administrative database. In fact, the CWHS predates the computer technology: it started in 1940, with information originally recorded on punch cards [Perlman and Mandel, 1944].

[4] The worst case in that respect is 1981, when 50,000 out of 900,000 observations are missing. The extent of this last problem generally falls over time, by 1987 it applies to less than 2 percent of observations and by the end of our sample it falls below 1 percent.

that we do not correct observations that need not be adjusted.

Unless otherwise indicated, the procedure is applied to all years starting with 1978 (but in practice affects few observations after 1980). We first supplement CWHS earnings by earnings from the MEF (using the same definition as one used for earnings in the CWHS to maintain consistency) if CWHS is missing. Next, we verified that virtually all 1978-1979 observations that are missing in LEED but present in the CWHS and that have total earnings greater than \$100,000 have FICA earnings (when below taxable max) and earnings in adjacent years smaller by the factor of the order 100. In many cases, FICA earnings are exactly 1/100th of total earnings. Consequently, we divide CWHS earnings in such cases by 100. There are 2400 cases of this nature in 1978 and about 1400 in 1979. We are confident that over-correction here, if any, is limited to a handful of cases.

In other cases, we use CWHS total earnings if (1) LEED earnings are missing (2) CWHS earnings are greater than 50 and smaller than 5 times LEED earnings or (3) (in 1978-1979) when CWHS earnings exceed LEED earnings by a multiple of 100,000 with CWHS above taxable max and earnings in at least one of the three following years equal to at least a half of CWHS earnings.[5] If none of these is the case, we start with LEED earnings.

We compare Social Security earnings with total compensation and if the latter is 100 times greater than the former (plus or minus \$100), we use Social Security earnings. For other observations we proceed with a more heuristic algorithm. Candidates to be corrected are defined as follows: an observations must have FICA earnings higher than taxable max minus 10 or total earnings must exceed FICA earnings by a factor of at least 5, with FICA earnings positive. We make adjustments only to those observations among the ones identified above that have earnings in adjacent years that are very much out of line. We use income in the three following years (fewer years in 2002-2004) and income in two preceding years with the exception of 1978-1980 when we use instead income in 1977. Starting with the last year, we correct by dividing by 100 or reverting to LEED in cases where LEED and CWHS were different by a multiple of 100,000 if and only if the following three conditions hold: (1) income in any of the adjacent years as specified above is not zero, (2) income in all the adjacent years is less than 20 of income in the

[5] We verified that W2-level earnings data in 1978-1979 in LEED never exceed 100,000 and in fact include only the last five digits (and decimal part).

year considered and (3) if 1977 income is used, it is not at the taxable max. We repeat this step one more time for 1979 and 1980 so that some additional corrections take place based on already corrected observations.

In our final dataset, in 1978, 50,000 out of approximately 870,000 observations have their origin in LEED and in 1979 this is the case for 100,000 of approximately 900,000. In other years, earnings have their source only in CWHS or MEF.[6] Due to the multitude of tests that we apply before an observation gets corrected, the number of observations that are affected by our correction procedure is small (and the numbers below are overestimates because we construct the corrected earnings measure for all observations, including those with earnings below the taxable maximum for which we end up using FICA earnings anyway). Other than the accurate adjustment of observations missing from LEED mentioned above, we end up correcting about 6900 observations in 1978, 5600 in 1979 and 800 in 1980. Afterwards, this procedure usually affects 500 or fewer observations, with the exception of 1982, 1987, 2002, 2003 and 2004 when it affects approximately 1000 cases. Although the number of affected observations is very small relative to the sample size, their pre-corrected values were heavily concentrated at the top and both mobility and inequality patterns at the top were obviously and very significantly incorrect. These adjustments bring earnings shares in line with tax statistics and generate mobility patterns that do not exhibit significant discontinuities.

B. Sensitivity Analysis

Figure A.1 reports average and median earnings (in 2004 dollars) and the total number of covered workers (in the full population) from 1937 to 2004 in the core sample. As is well known, both the median and average earnings increased quickly from 1937 to 1973. After 1973, median earnings stagnated.[7]

We perform sensitivity analysis along three key dimensions: (1) commerce and industry restriction, (2) choice of the minimum earnings threshold, (3) imputations above the top code.

[6] In 1978-1980, few observations from MEF need to be used.

[7] They are almost identical in 2004 and 1973, even using the revised CPI-U-RS price deflator which incorporates 7-8 percent less cumulative price inflation (and hence 7-8 percent more real growth) than the official CPI from 1978 to 1992. After 1992, the official CPI includes the new methods of the CPI-U-RS. Before 1978, there are no CPI-U-RS series available.

We therefore construct three alternative samples to the core sample.

(1) In the “all industries” sample, we expand the core sample to include all workers with covered earnings from any industry (instead of restricting earnings to “commerce and industry” sectors) above the minimum threshold. In that case, earnings are defined as all covered earnings (instead of “commerce and industry” earnings). Note that we continue to exclude self-employment income and farm income. As described above, before 1951, the “all industries” and core samples coincides because only “commerce and industry” earnings are covered. In recent decades, the “all industries” sample includes about 95 percent of US employees as very few sectors remain uncovered. The primary goal of the “all industries” sample is to check (for recent decades) whether mobility in and out the commerce and industry sample affects substantially measures of mobility and long-term inequality.

(2) In the “4*minimum threshold” sample, we restrict the core sample to all workers with “commerce and industry” earnings above a minimum threshold of $10,300 in 2004 (and indexed using average wage for earlier years). This alternative threshold is four times as high as the core sample threshold of $2,575. The higher threshold corresponds to a full time and full year minimum wage annual earnings ($= 40 * 50 * \$5.15$). The goal of this alternative sample is to assess whether our results are sensitive to the arbitrary choice of the minimum threshold.

(3) In the “Pareto imputation fixed effect” sample, the sample remains the core sample but we estimate earnings above the top code using individual fixed effects random draws u_i instead of iid random draws u_{it} as in the core sample. This alternative method assesses the sensitivity of our *mobility* and multi-year inequality estimates with respect to top code imputation. The core sample method Pareto imputation is based on draws from a uniform distribution that are independent across individuals but also time periods. As there is persistence in ranking even at the top of the distribution, this method generates an upper bound on mobility within top coded individuals. In the alternative method, the uniform distribution draw is independent across individuals but fixed over time for a given individual. As there is some mobility in rankings at the top of the distribution, this method generates a lower bound on mobility.

Figure A.2 depicts average earnings and number of workers in the core sample, the all industries sample, and the 4*minimum threshold sample. Unsurprisingly, average earnings are higher in the 4*minimum sample and the number of workers is higher in the all industries sample and lower in the 4*minimum threshold sample.

Figure A.3 compares estimates of the Gini coefficient for our commerce-industry core sample and three alternative samples. Figure A.3 displays the Gini coefficient for the "all industries" sample. The overall evolution over time is the same. The Gini including all industries is lower today than the commerce and industry sample while the Gini for the two samples was almost identical in 1970. This is consistent with Katz and Krueger [1991] who show that inequality within the public sector has increased much less than in the private sector in the 1980s. We cannot document changes in inequality outside the commerce and industry sector during the Great Compression. However, Margo and Finegan [2002], using census data, showed that a similar compression took place within the public sector as well. This suggests that the overall U-shape evolution over time for the Gini should be robust to including all sectors.

Figure A.3 also displays the Gini coefficient when increasing the minimum threshold by a factor 4 (so that it is equal to a full-time full-year minimum wage $10,300 in 2004). Unsurprisingly, the Gini is lower for that sample. However, the overall U-shape over time and the key inflection points remain identical. Figure A.3 also displays the Gini coefficient when excluding the top percentile earners. The figure shows that the increase in the Gini in the 1980s and especially the 1990s is noticeably smaller when excluding the top 1 percent. This is not surprising that the top 1 percent share has increased dramatically and the share going to the top affects significantly the Gini (this can be easily seen by drawing the Lorenz curve). This shows that Gini estimates based on top coded data such as the CPS are likely to be severely biased relative to administrative data with no top code and good coverage at the top.

Figure A.4 compares the log-percentile ratios P80/P50 and P50/P20 in the core sample and in two alternative samples: the "all industries" sample and the "4*minimum threshold" sample. The time patterns are very similar. Note that the 4*minimum threshold displays more inequality at the bottom but less at the top (although the time patterns of the series are very close to those in the core sample). The "all industries" series display slightly less inequality increase over recent decades, especially in the upper part of the distribution, consistent the Gini sensitivity analysis above.

Figure A.5 compares the Gini coefficients based on 5-year earnings averages in the core sample and in the "all industries" sample, the "4*minimum threshold" sample, and the "Pareto imputation fixed effect" sample.[8] The figure shows that Pareto imputations have virtually no effect showing that very little bias comes from the Pareto imputations. The overall time pattern of the series is also very close for the "all industries" and "4*minimum threshold" sample (with the usual finding that the 4*minimum threshold displays a lower level of inequality and that the "all industries" series display less inequality increase in recent decades).

Figure A.6 compares the year-to-year rank correlation in the core sample and in the "all industries" sample, and the "Pareto imputation fixed effect" sample. The figure shows that rank correlation is virtually the same in those alternative samples.

Figure A.7 compares the transitory variance series in the core sample and in the "all industries" sample, the "4*minimum threshold" sample, and the "Pareto imputation fixed effect" sample. The overall time pattern of the series is also very close for all four series (we also note that the 4*minimum threshold displays a lower level of transitory variance which is not surprising). There is a small effect of the Pareto imputation strategy for the early years when top-coding is substantial and no effect thereafter.

Figure A.8 compares the Gini coefficients based on 11-year earnings averages in the core sample and in the "all industries" sample, the "4*minimum threshold" sample, and the "Pareto imputation fixed effect" sample. The figure shows again that Pareto imputations have virtually no effect showing that very little bias comes from the Pareto imputations. The overall time pattern of the series is also very close for the "all industries" and "4*minimum threshold" sample (with the usual finding that the 4*minimum threshold displays a lower level of inequality and that the "all industries" series display less inequality increase in recent decades).

Figure A.9 compares the long-term rank correlation in the core sample and in the "all industries" sample, the "4*minimum threshold" sample, and the "Pareto imputation fixed effect" sample. The figure shows again that Pareto imputations have virtually no effect showing that very little bias comes from the Pareto imputations. The overall time pattern of the series is also very close for the "all industries" and "4*minimum threshold" sample (we note that the 4*minimum threshold displays less correlation in levels than the other series).

Finally, Figure A.10 shows that our upward mobility findings by gender (Figure XI) are

[8] We did not display the Pareto imputation fixed effect series in Figures A.2, A.3, A.4 because Pareto imputations matter only when looking at longitudinal earnings.

robust to conditioning on birth cohort. The figure displays the probability of moving from P0-40 in early career to P80-100 in late career by year of birth. Such upward mobility measures increase in the full sample but decomposition by gender shows that this is entirely driven by women which experience a large increase in upward mobility while upward mobility for men stays stable over the period.

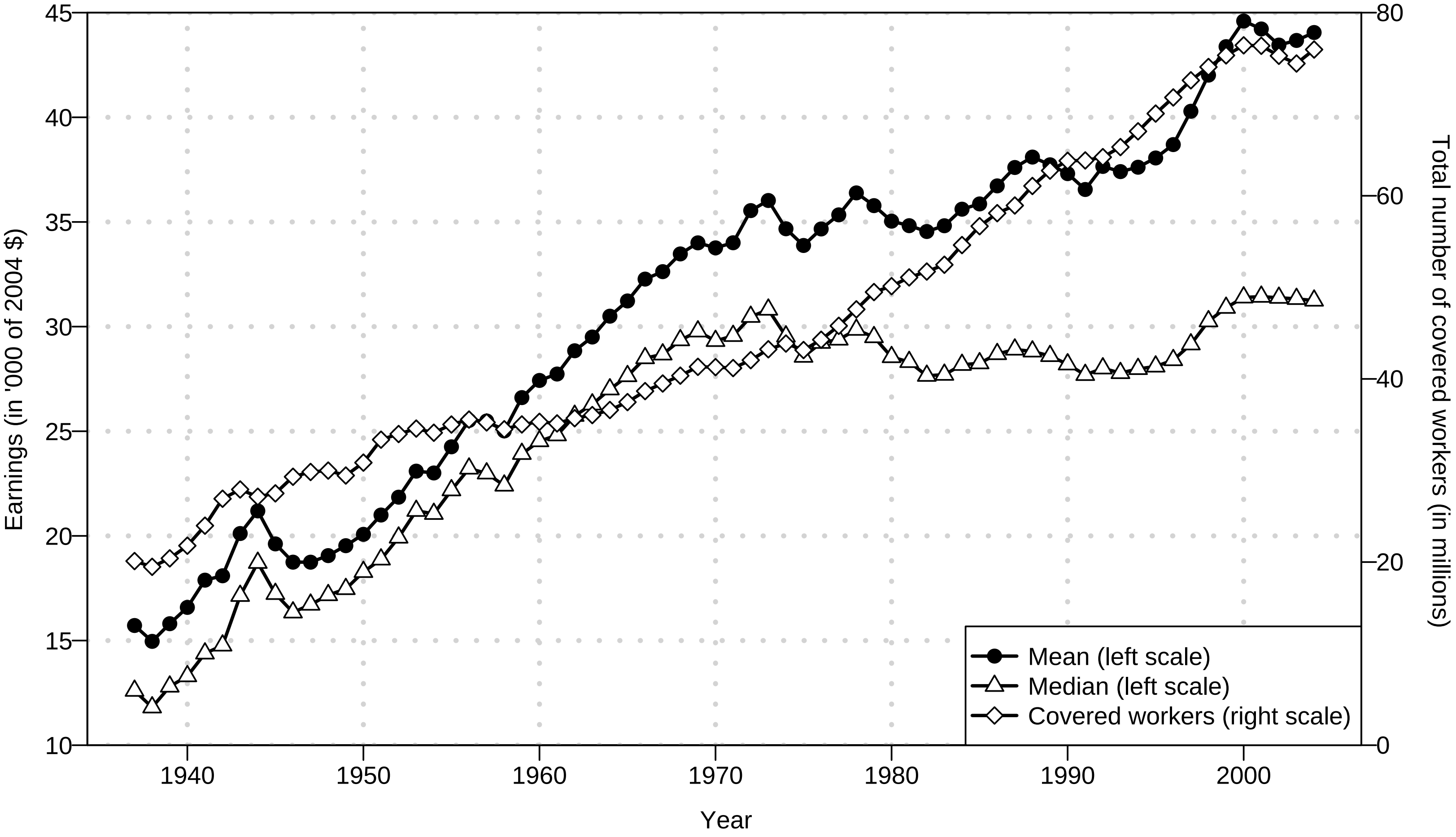

Figure A.1
Aggregate SSA Earnings and Workers in Commerce and Industry Core Sample

Sample is the core sample defined as all employees in Commerce and Industry with earnings above minimum threshold ($2,575 in 2004 and indexed using average wage for earlier years) and aged 25 to 60 (by January 1st of a given year t). Commerce and Industry is defined as all industrial sectors excluding government employees, agriculture, hospitals, educational services, social services, religious and membership organizations, and private households. Only commerce and industry earnings are included. Self-employment earnings are fully excluded. Average Earnings are reported in 2004 dollars (using the CPI and the CPI-U-RS after 1978).

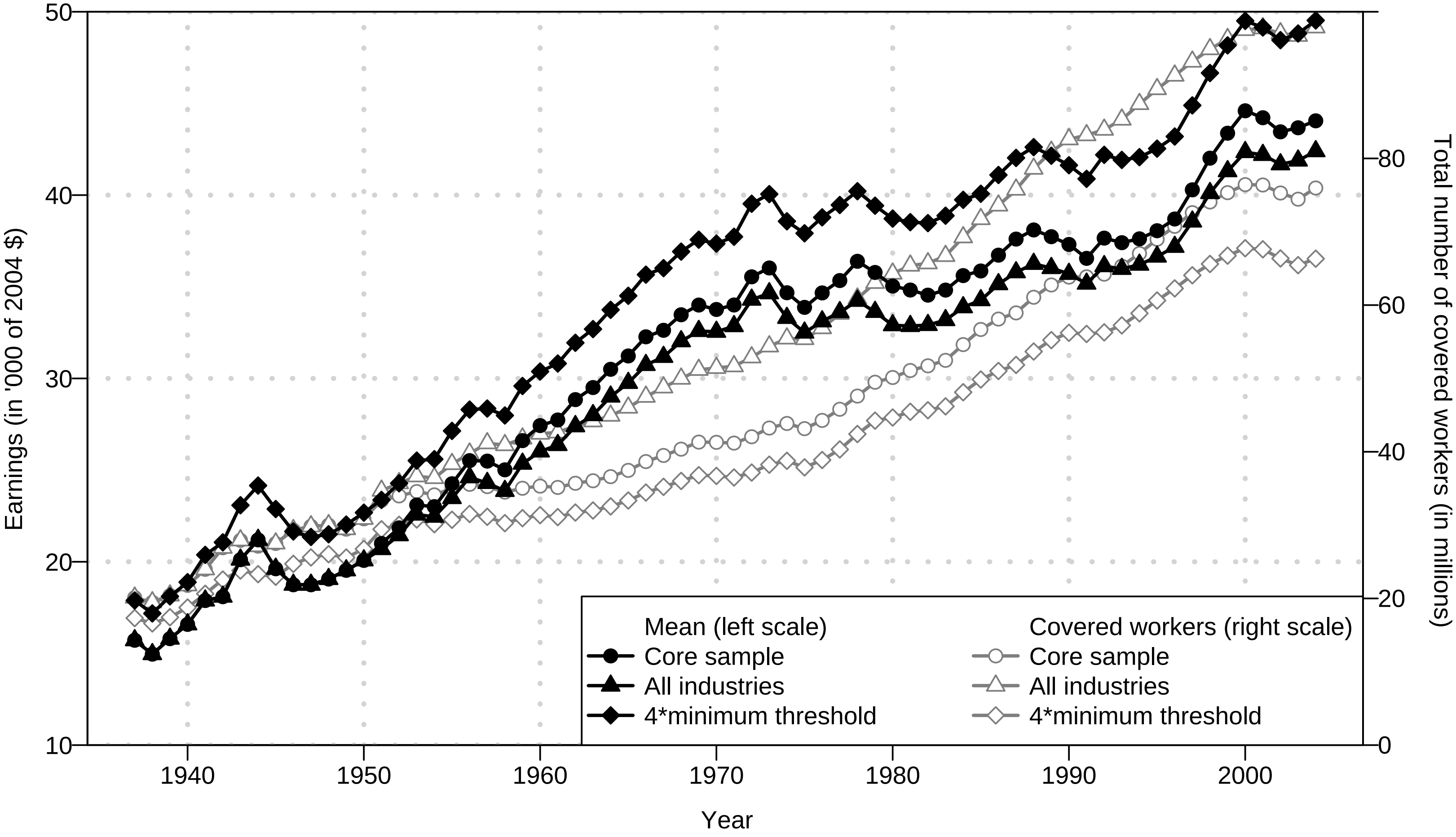

Figure A.2
Average Earnings and Number of Workers in Alternative Samples

The figure displays average earnings and number of workers in three different samples. The first sample (core sample) is the core sample: employees aged 25 to 60 with Commerce and Industry earnings above a minimum threshold of $2,575 in 2004 (and indexed using average wage for earlier years). The second sample (4*minimum threshold) restricts to core sample to employees with Commerce and Industry earnings above a higher minimum threshold equal to $10,300 (=4*$2,575) in 2004 (and indexed using average wage for earlier years). The third sample (all industries) extends the core sample to include all employees with covered earnings (in any industry, not only “Commerce and Industry”) above $2,575 in 2004 (and indexed using average wage for earlier years). In this all industry sample, earnings include earnings from all industries. In all three samples, average earnings are reported in 2004 dollars (using the CPI and the CPI-U-RS after 1978).

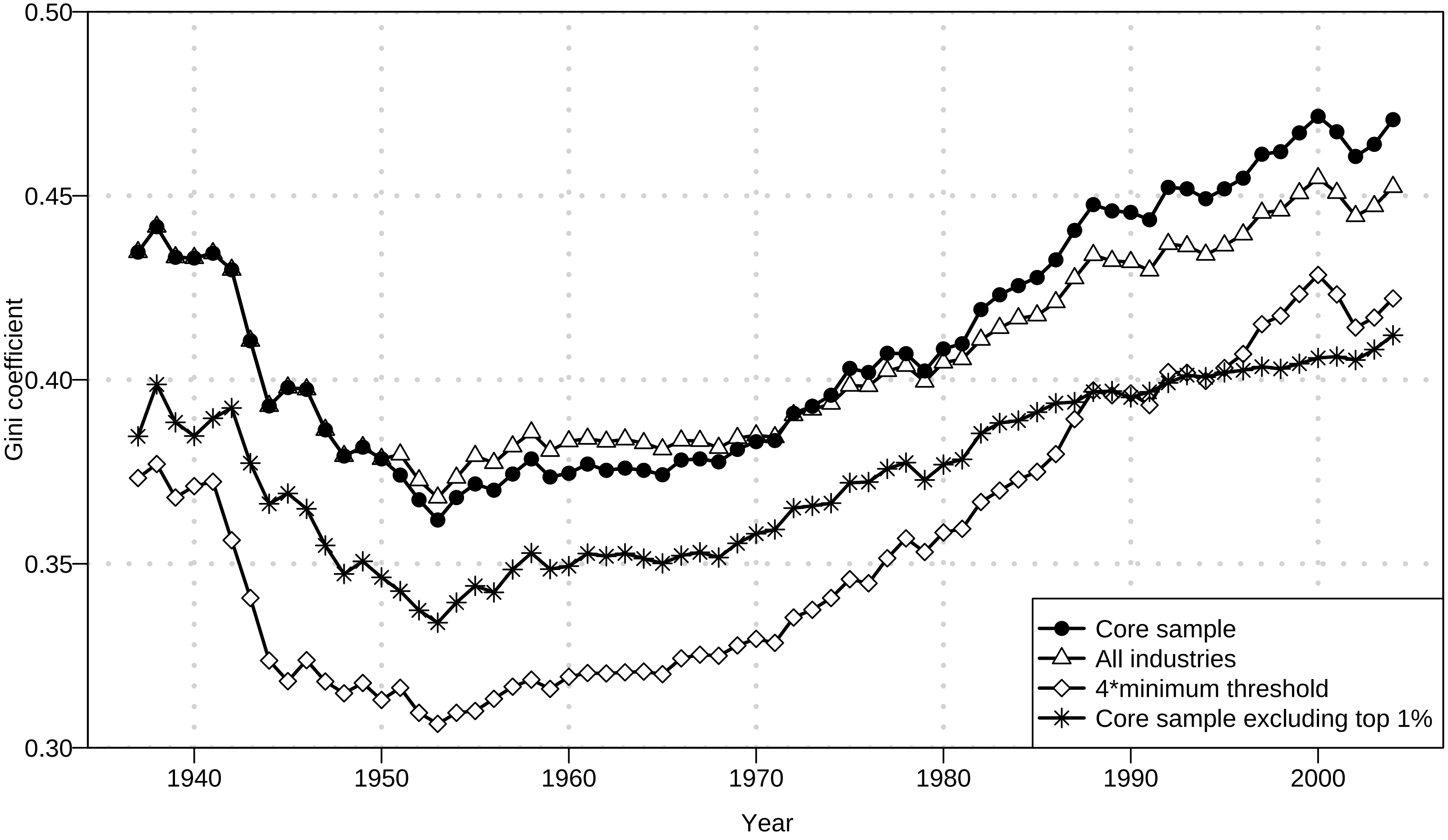

Figure A.3
Annual Earnings Gini Coefficients Sensitivity

The figure reports the annual earnings Gini coefficients in various samples: (a) in the core sample (as in Figure II in the text, series all workers), (b) in the sample including workers in all covered industries (instead of only commerce and industry), (c) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text).

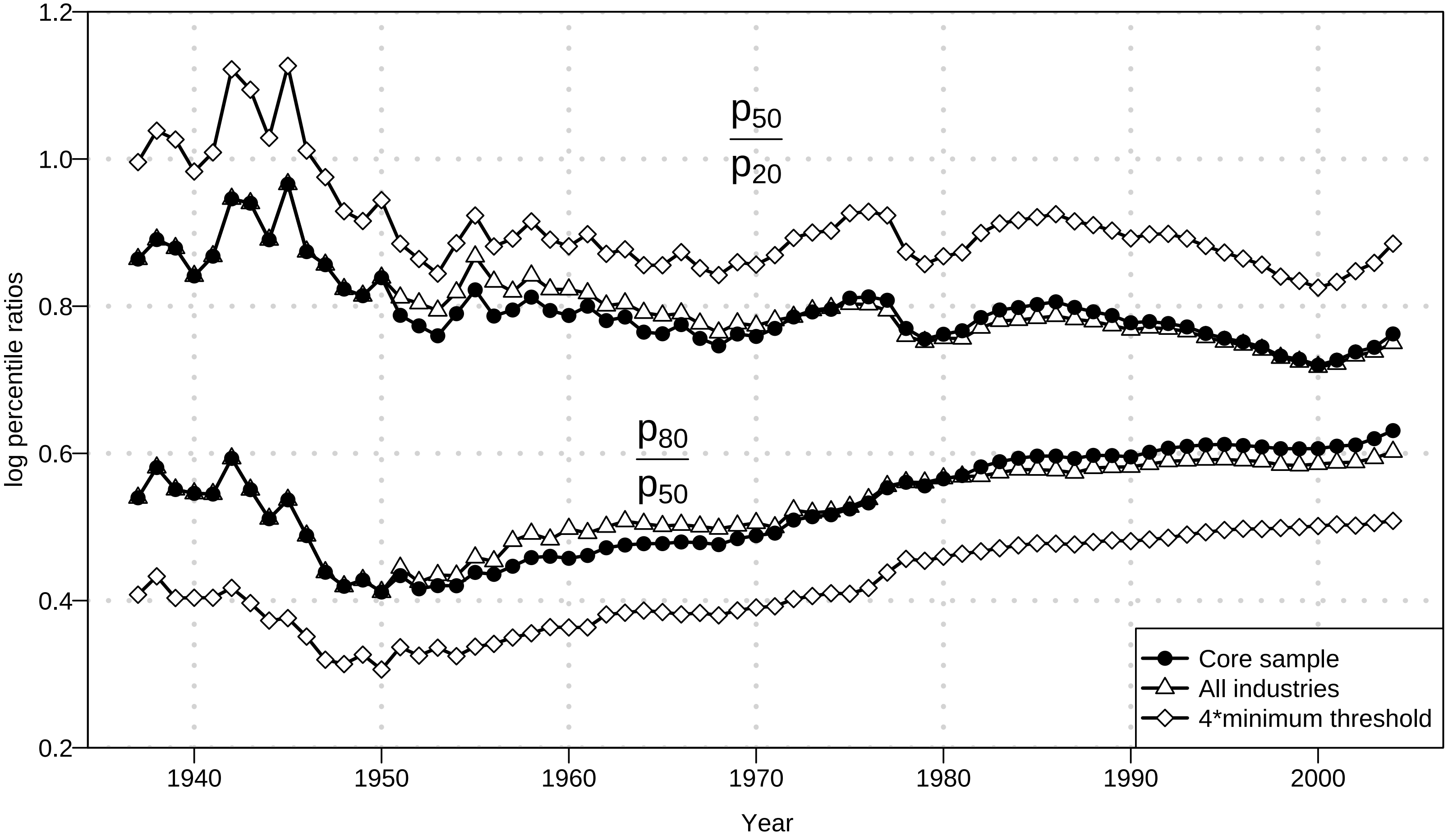

Figure A.4
Percentile Ratios Sensitivity

The figure reports the percentile Ratios Log(P80/P50) and Log(P50/P20) (a) in the core sample (as in Figure 2 in text, series all workers), (b) in the sample including workers in all covered industries (instead of only commerce and industry), (c) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text).

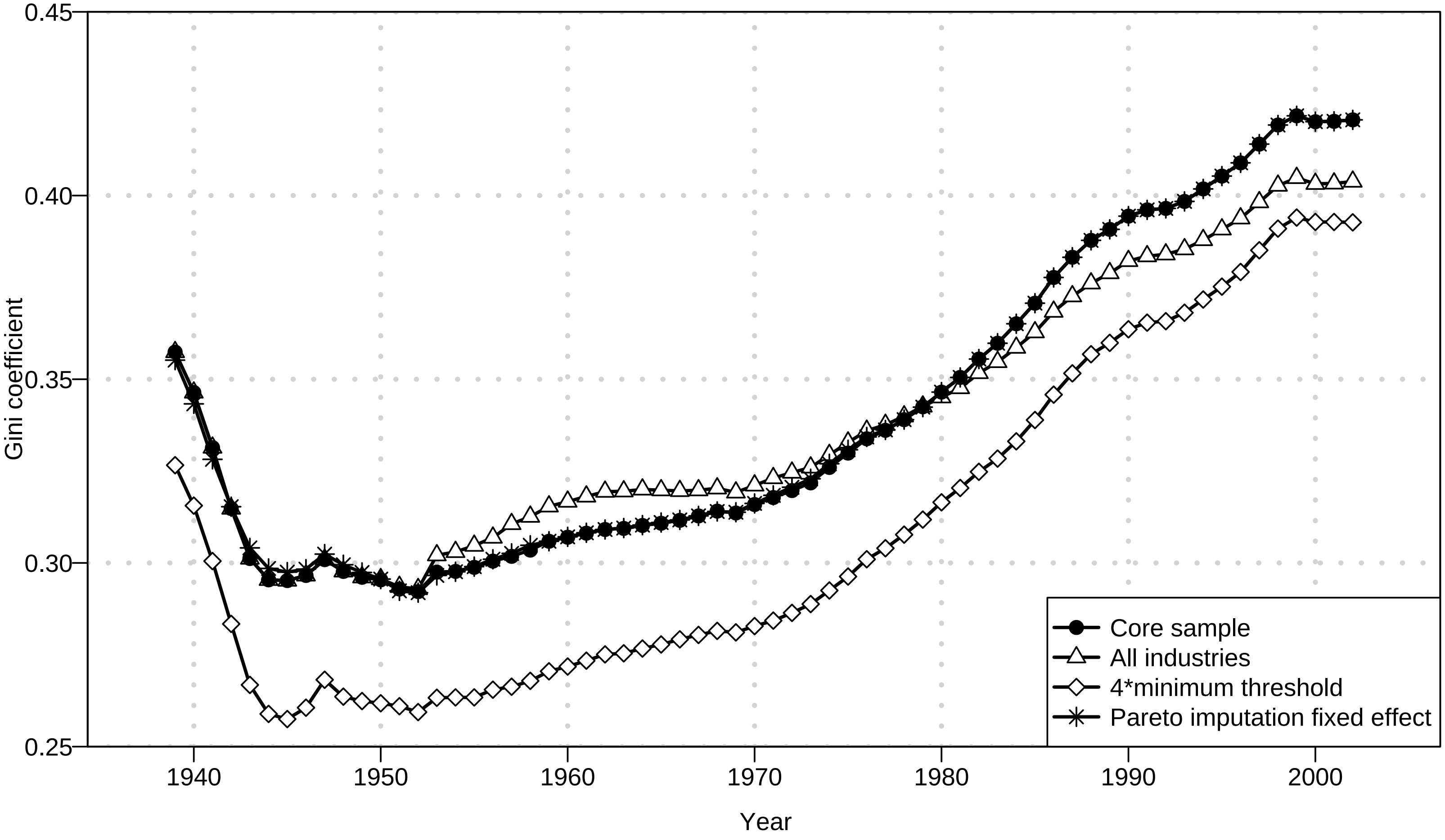

Figure A.5
5-Year Earnings Average Gini Coefficients Sensitivity

The figure reports the Gini coefficients for 5-year earnings averages in various samples: (a) in the Figure 3 sample in the text (series 5 year earnings, all workers), (b) in the sample including workers in all covered industries (instead of only commerce and industry), (c) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text), (d) in the sample where imputations above the top code are based on random draws that are constant over time for each individual (instead of being iid as in the text).

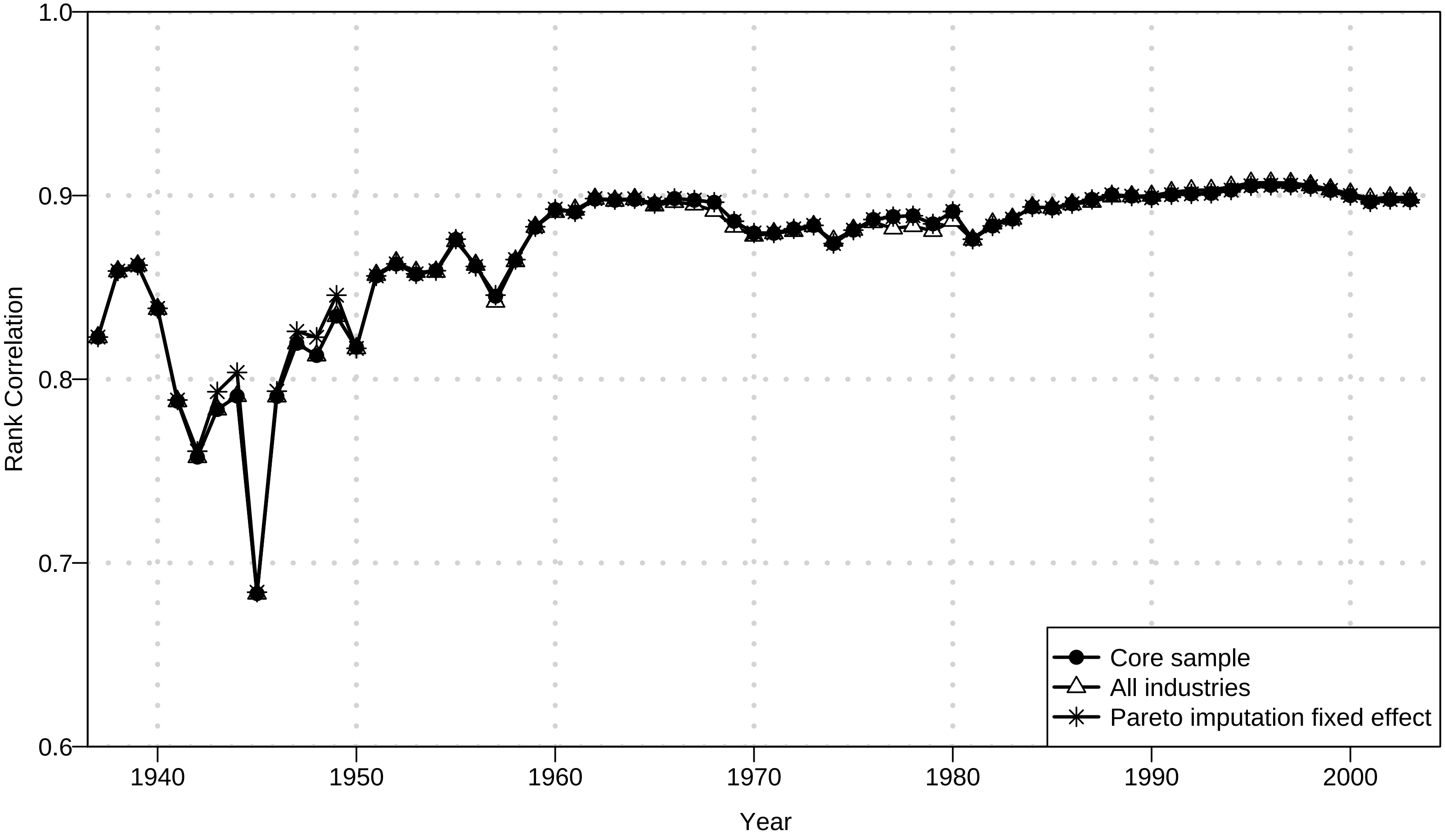

Figure A.6
Year to Year Rank Correlation Sensitivity

The figure reports the year to year rank correlation in various samples: (a) in the Figure 4 sample in the text (series rank correlation, all workers), (b) in the sample including workers in all covered industries (instead of only commerce and industry), (c) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text), (d) in the sample where imputations above the top code are based on random draws that are constant over time for each individual (instead of being iid as in the text).

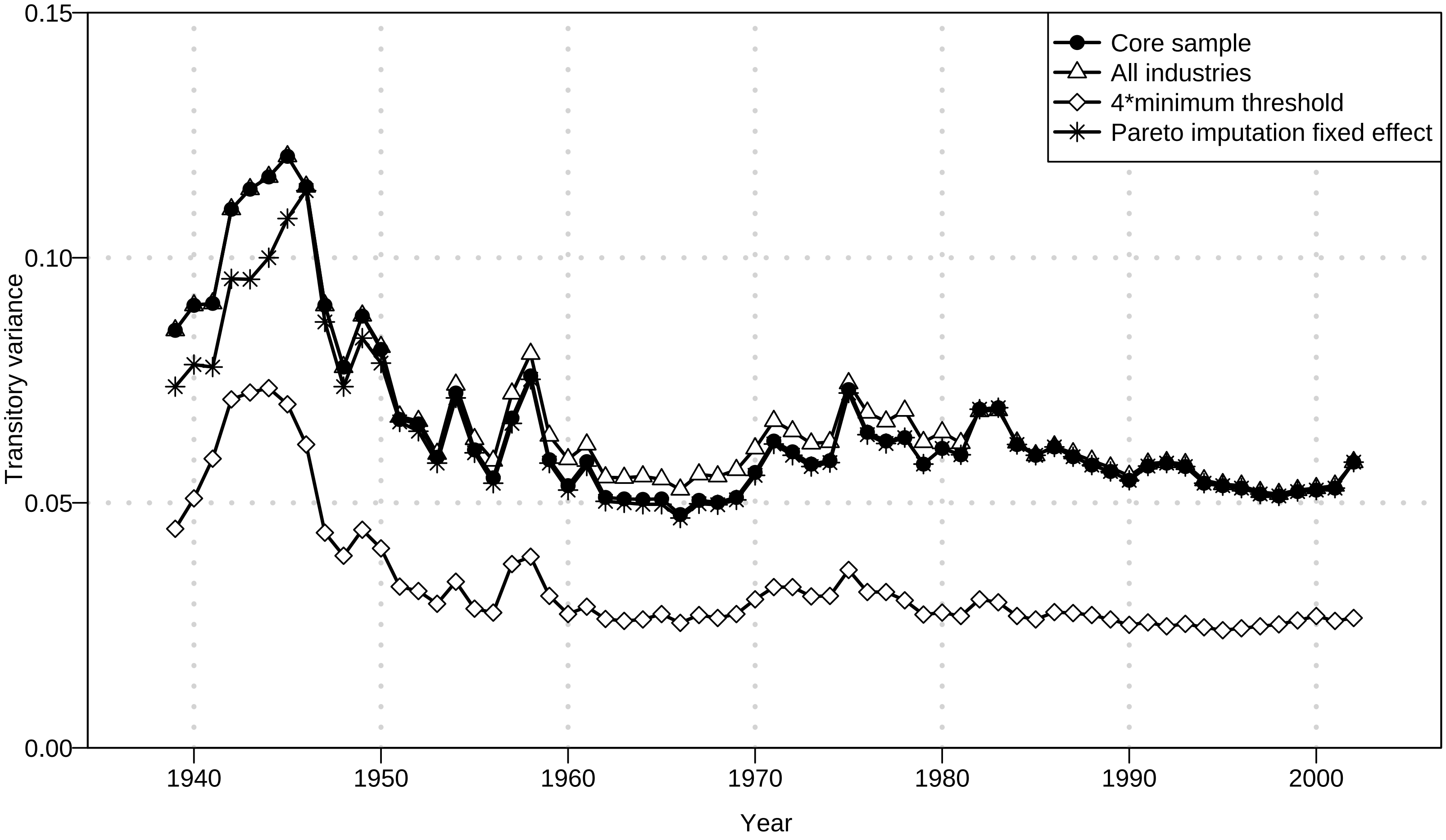

Figure A.7
Transitory Variance Sensitivity

The figure reports the transitory variance of log-earnings (defined as deviations of annual log-earnings from 5-year average log-earnings) in various samples: (a) in the Figure 5 sample in the text (series transitory earnings, all workers), (b) in the sample including workers in all covered industries (instead of only commerce and industry), (c) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text), (d) in the sample where imputations above the top code are based on random draws that are constant over time for each individual (instead of being iid as in the text).

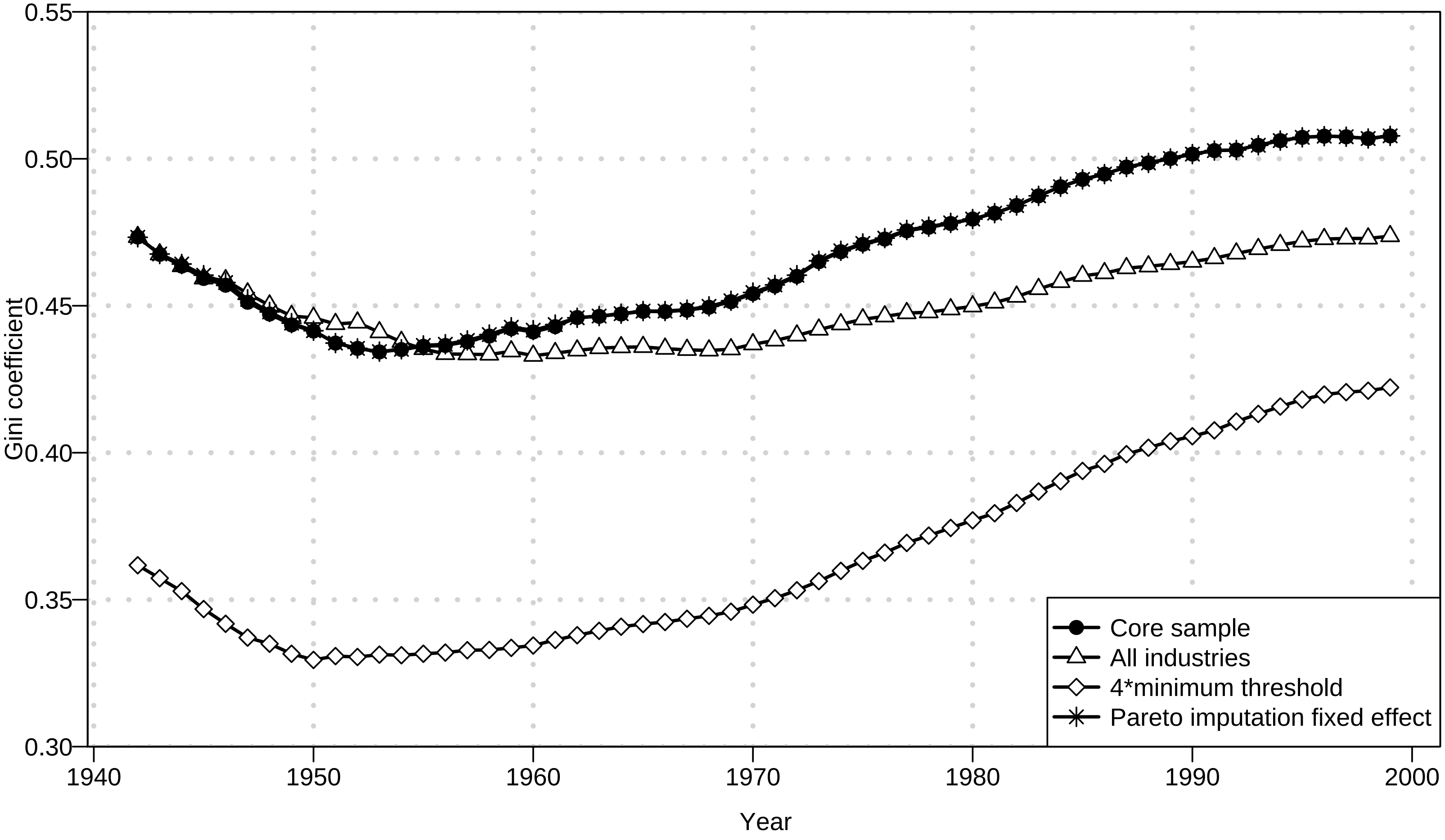

Figure A.8
11-Year Earnings Average Gini Coefficients Sensitivity

The figure reports the Gini coefficients for 11-year earnings averages in various samples: (a) in the Figure 7 sample in the text (series transitory all workers), (b) in the sample including workers in all covered industries (instead of only commerce and industry), (c) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text), (d) in the sample where imputations above the top code are based on random draws that are constant over time for each individual (instead of being iid as in the text).

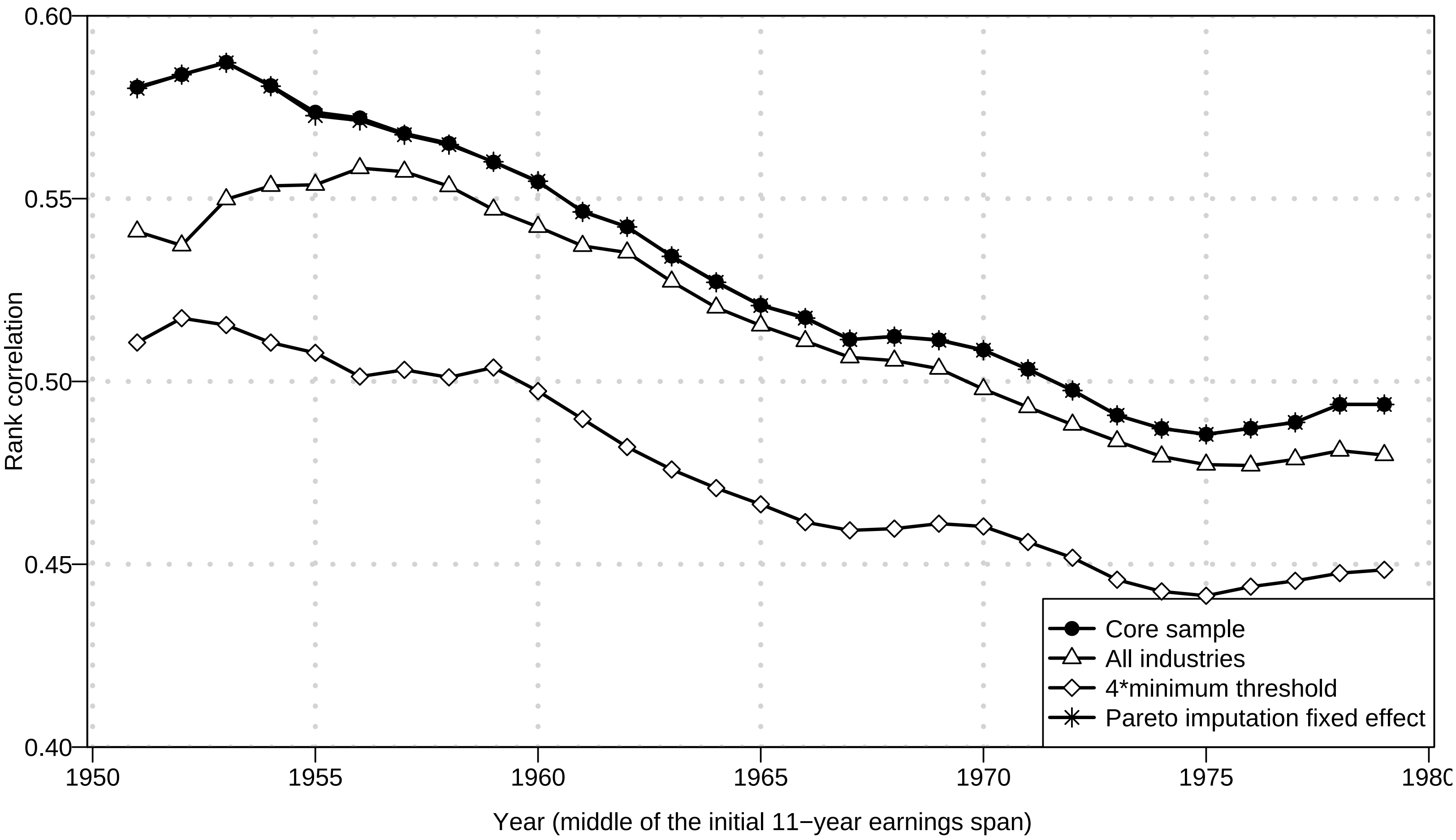

Figure A.9
Long-Term Rank Correlation Sensitivity

The figure reports the rank correlations of 11-earnings averages after 20 years in various samples: (a) in the Figure VIII sample in the text (series all workers and men), (b) in the sample with a higher minimum threshold equal to $10,300 in 2004 dollars (instead of $2,575 as in the text), (c) in the sample where imputations above the top code are based on random draws that are constant over time for each individual (instead of being iid as in the text).

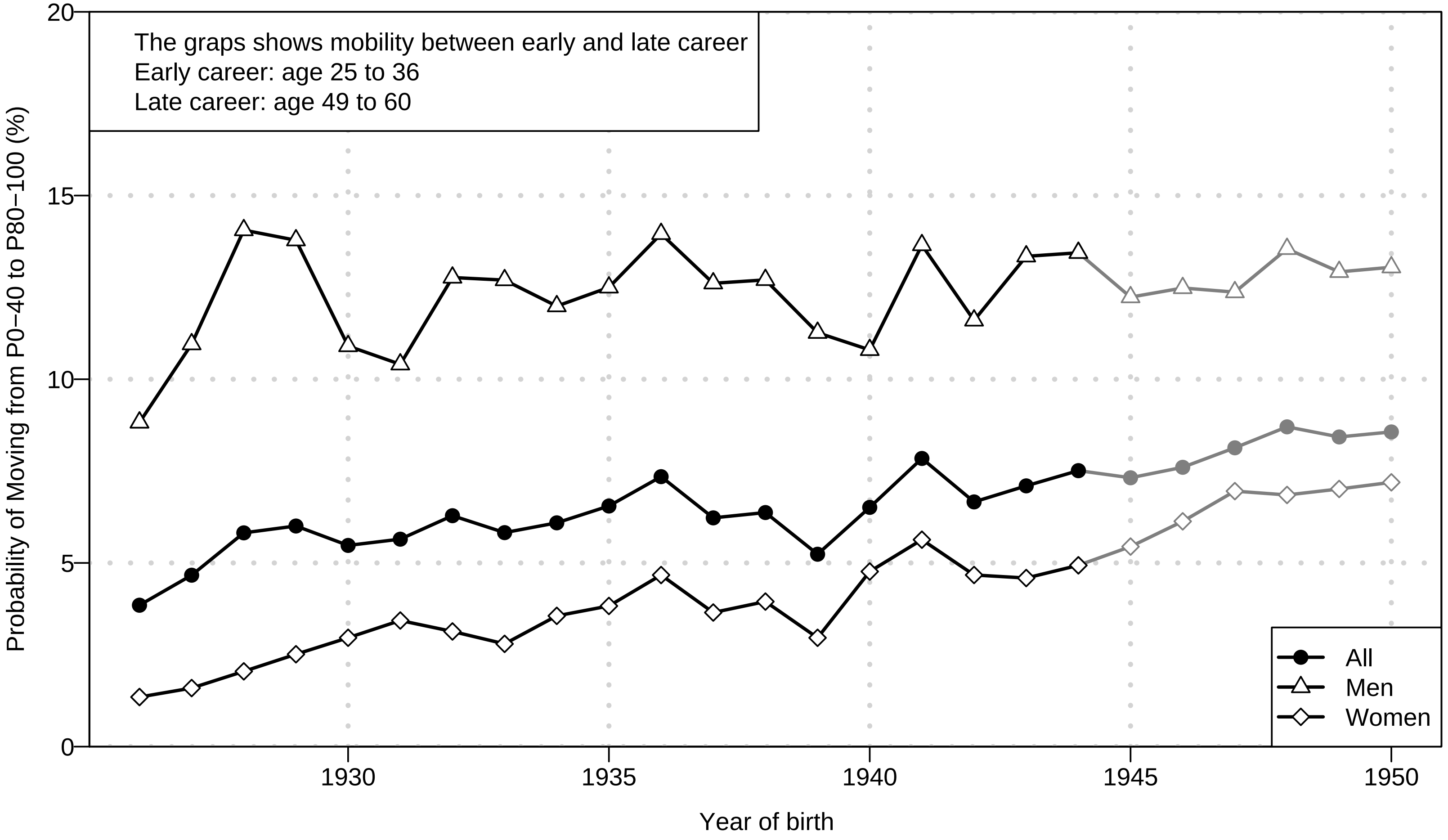

Figure A.10
Long-Term Upward Mobility and Gender: Cohort-Based Estimates

The figure displays, by birth cohort, the probability of moving to the top quintile group (P80-100) for late career earnings (age 48 to 59) conditional on having early career earnings (age 26 to 37) in the bottom two quintile groups (P0-40). The series are reported for all workers, men only, and women only. In all three cases, quintile groups are defined based on the sample of all workers. Estimates in lighter grey are imputed based on less than 12 year of earnings (as the career stage is right-censored in 2004), see Kopczuk et al. [2007] for details.