I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17973

Assessing the Statistical Significance of Inequality Differences: The Problem of Heavy Tails

Nicolas Herault
Stephen Jenkins

JUNE 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17973

Assessing the Statistical Significance of Inequality Differences: The Problem of Heavy Tails

Nicolas Herault
University of Bordeaux and IZA

Stephen Jenkins
London School of Economics and IZA

JUNE 2025

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Assessing the Statistical Significance of Inequality Differences: The Problem of Heavy Tails*

Because finite sample inference for inequality indices based on asymptotic methods or the standard bootstrap does not perform well, Davidson and Flachaire (Journal of Econometrics, 2007) and Cowell and Flachaire (Journal of Econometrics, 2007) proposed inference based on semiparametric methods in which the upper tail of incomes is modelled by a Pareto distribution. Using simulations, they argue accurate inference is achievable with moderately large samples. We provide the first systematic application of these and other inferential approaches to real-world income data (high-quality UK household survey data covering 1977–2018), while also modifying them to deal with weighted data and a large portfolio of inequality indices. We find that the semiparametric asymptotic approach provides a greater number of statistically significant differences than the semiparametric bootstrap which in turn provides more than the conventional asymptotic approach and the 'Student-t' approach (Ibragimov et al., Econometric Reviews, 2025), especially for year-pair comparisons within the period from the late-1980s onwards.

JEL Classification: C14, C46, C81, D31

Keywords: income inequality, Pareto distribution, asymptotic approach, semiparametric bootstrap approach, semiparametric asymptotic approach, t-statistic approach

Corresponding author:
Stephen P. Jenkins
Department of Social Policy
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
United Kingdom
E-mail: s.jenkins@lse.ac.uk

* We thank Tod Wright, and audiences at WU Wien, Roma Tré, EUI, and the 18th Winter School on Inequality and Social Welfare (Alba di Canazei), for comments on early versions of this paper, and Emmanuel Flachaire for patiently responding to our queries about semiparametric methods. Thanks also to Ben Jann and Philippe Van Kerm for their Stata programs (dstat and paretofit_obre, respectively). The ONS's HIE team helped us make their ETB datasets available to other potential users. The paper was drafted when Jenkins was a Visiting Fellow at EUI's Department of Economics. Hérault acknowledges financial support from the French Ministry of Higher Education and Research, the French National Research Agency, the IdEx University of Bordeaux and the GPR HOPE.

1. Introduction

You might think that it is straightforward to assess whether a difference in two income inequality indices is statistically significant: simply derive estimates of the indices and their standard errors using standard methods and then do a two-sample t-test of the inequality difference – analogous to what you might do for assessing the statistical significance of differences in means of other scalar socioeconomic indicators. However, this strategy is flawed (see below). This problem has led researchers to propose improved inferential methods, but they have rarely been applied to real-world datasets. In this paper, we provide the first systematic application of several recently proposed approaches to assess the difference in inequality between a pair of years using high-quality UK household survey data, also showing how the methods can be adapted to account for survey weights and be applied to wider range of inequality indices than has been considered to date.

Standard asymptotic and bootstrap approaches to assessing the statistical significance of inequality differences have been criticized as having poor statistical performance even in large samples. See, e.g., Cowell and Flachaire (2007, 2015), Davidson (2012), Davidson and Flachaire (2007), Schluter (2012), and Schluter and van Garderen (2009). The problem is that distributions of studentized (differences in) inequality indices derived using these approaches, and which are the basis of t-statistics used for statistical tests, tend towards skewed distributions rather than standard normal distributions. (These skewed distributions are illustrated by, e.g., Schluter, 2012, Figure 1.) The source of the problem is that income distributions are typically heavy-tailed – skewed, with a long right-hand tail and Pareto-like shape at the top. Sparseness of observations in the top income range is an additional complication with household survey data.

Naturally, the question arises whether there are methods for inference about income inequality that have better statistical performance for heavy-tailed income distributions than the standard asymptotic and bootstrap methods. The answer according to the research cited above is affirmative, and several approaches have been proposed, as we discuss below. However, assessments of these approaches have been based almost entirely on simulated data, with few substantive applications using real-world survey datasets.

Our first contribution arises from our comparisons of the proposed inferential approaches in a real-world setting, using high-quality yearly UK household survey data for 1977–2018 covering periods when inequality changed a lot and when it changed relatively little. We document whether the methods provide similar or different conclusions about the

statistical significance of income inequality differences between pairs of years for multiple pairs of years.

Undertaking inference about inequality differences is important because income inequality statistics are key social indicators, with inequality levels and trends the subject of public attention and debate, with national and international statistical agencies routinely publishing statistics.

To illustrate the issues, Figure 1 shows inequality time series derived from the data the UK's Office for National Statistics has used to compile official inequality statistics (see, e.g., Office for National Statistics, 2019) and which we also use later in this paper. In addition to the Gini coefficient, the ONS's headline indicator, we show series for five other indices.

<Figure 1 near here>

According to all six indices, there was an increase in income inequality between the end of the 1970s and the start of the 1990s, about 10 percentage points according to the Gini coefficient, which is substantial and likely to be statistically significant (we verify this later). But what about the inequality changes thereafter? All indices fluctuate in value over the subsequent three decades, and it is of interest to know which year-pair inequality differences are statistically significant and which represent sampling variability. We expect answers to depend on which specific pair of years is compared and on which inferential approach is used, and we document this. We also report how answers depend on the inequality index.

Statistical agency bulletins typically do not report standard errors along with inequality index estimates nor undertake formal statistical tests of inequality differences, but instead use approaches such as reporting rounded estimates to minimize the chances of incorrectly interpreting differences as statistically significant. For example, the UK Department for Work and Pensions' *Households Below Average Income* reports show Gini coefficient estimates to two decimal places (Department for Work and Pensions, 2025b). Our research sheds light on whether assessing inequality differences using informal rounding rules is consistent with the results of applying appropriate inference.

Our second contribution is to extend the proposed inferential methods to increase their useability. Instead of employing method of moments estimators as Davidson and Flachaire (2007) and Cowell and Flachaire (2007) do, we use asymptotically equivalent estimators based directly on a survey's unit record data. This switch has two advantages. The first is that researchers can straightforwardly incorporate survey weights in derivations of estimates and their sampling variances. Research proposing new inferential approaches has ignored weights

and yet virtually every household survey contains weights that adjust for survey design features such as differential sampling probabilities, or help address issues such as non-response and, more generally, better represent the target population of interest. Not using the weights when deriving inequality indices (or other descriptive statistics) leads to biased estimates.

A further advantage of the unit record approach is that one can use a wide range of inequality indices. Almost all the research proposing improved inference has focused on the Gini index and the Theil index and sometimes other Generalized Entropy indices. With our approach, one can undertake inference for almost any inequality index (subject to caveats about their 'sensitivity'). We assess inequality differences using a portfolio of indices that are widely used in official statistics and by inequality researchers: the Gini coefficient, members of the Generalized Entropy (GE) family, the $p90/p10$ quantile ratio, and the shares of total income held by the richest 10% and by the richest 1% of income units. We use all these indices but focus on the Gini and Theil indices for empirical reasons we discuss later. (We do not consider Atkinson inequality indices because, for each member of that family, there is an ordinally equivalent member in the GE family.) The $p90/p10$ quantile ratio and top income share measures are also widely used.

We assess UK inequality differences using four approaches to inference about inequality change (reviewed in more detail later). Approach 1, and the reference point against which we compare more recently proposed methods, is the conventional *asymptotic* approach, with formulae derived using influence functions or linearization (also known as the delta method). Problems with this approach spurred proposals for improved inferential methods.

Inference approaches 2 and 3 are semiparametric because the upper tail of each pair of income distributions being compared is assumed to be well-described by a Pareto distribution. Approach 2 is the *semiparametric asymptotic* approach proposed by Cowell and Flachaire (2007) in which conventional asymptotic methods are used to compare inequality between a pair of distributions, where each is described by a mixture of a Pareto distribution for the upper tail and the observed data for the rest of the distribution. Approach 3 is that of David and Flachaire (2007) who proposed a *semiparametric percentile-t bootstrap* method in which inference is based on repeated sampling from semiparametric mixture distributions (defined as above). Our analysis adapts approaches 2 and 3 to use unit record data rather than method of moments estimators.

The fourth approach is the *Student-t* method evaluated by Ibragimov, Kattuman, and Skrobotov (2025) and Midões and de Crombrugghe (2023), applying results derived by Ibragimov and Müller (2010). The approach involves a random split of each distributional sample into a relatively small number of groups, deriving group-specific estimates, and then combining them to derive an overall estimate for the index and its sampling variance. Test statistics are easy to compute for this method by comparison with the semiparametric methods.

Three other approaches to inference for inequality differences are not considered here. Davidson and Flachaire (2007) considered an M out of N ('Moon') bootstrap method, but their analysis of statistical performance leads them to favour their semiparametric percentile-t bootstrap method, which is what we use. Dufour et al. (2019) develop permutation tests (and associated permutation bootstrap tests) for tests of inequality differences, investigating their performance using simulated data. We have not used permutation tests because they are not applicable when the survey data include (non-integer) weights, which is the ubiquitous real-world situation including in our application. Schluter and van Garderen (2009) and Schluter (2012) propose an approach based on variance stabilising transformations that do not have analytical solutions in general and are "more demanding in terms of moments of the underlying distribution" than other approaches (Midões and de Crombrugghe, 2023, p. 920). This and the complexity of the numerical solutions required for implementation militates against using this approach.

Our paper has some similarities to that by Alfons, Templ, and Filzmoser (2013). Common to their paper and ours is a unit record data approach and incorporation of survey weights, plus use of robust estimators of Pareto distribution shape parameters. However, Alfons et al. consider only semiparametric asymptotic estimators and not also semiparametric percentile-t bootstrap (or Student-t) estimators, they measure inequality using the Gini coefficient alone, their substantive application examines only two countries (Belgium and Austria) for two years (2005, 2006), and they do not formally test for inequality differences. We compare a larger portfolio of estimators and inequality indices and undertake formal tests of pairwise inequality differences on a relatively large scale (year-pairs drawn from 42 years of UK survey data up to 2018).

The various inferential methods can also be applied to inequality comparisons across regions or countries. In their empirical illustration, Ibragimov et al. (2025) test for inequality differences between Moscow and every other Russian region, comparing the findings of conventional asymptotic, permutation and permutation bootstrap, and Student-t approaches

for inference about pairwise differences in Gini coefficients. Midões and de Crombrugghe (2023) undertake inference about differences in Theil coefficients estimated from two different Russian surveys using the permutation and Student-t approaches. In both papers, estimates are based on unweighted survey data, and neither employs the semiparametric approaches in their Russian data applications.

The headline findings from our application to UK data are, first, that the conventional asymptotic and Student-t methods yield similar conclusions about the statistical significance of year-pair inequality comparisons. Second, compared to these approaches, we find that application of the semiparametric asymptotic approach increases the number of statistically significant inequality differences for pairs of years in the 30-year period following the late-1980s. The semiparametric bootstrap method yields conclusions similar to those for the semiparametric asymptotic approach for middle-sensitive inequality indices. In addition, we demonstrate that using naïve rules of thumb such as 'count differences of at least one or two percentage points in rounded Gini coefficients as statistically significant' are not reliable.

Although there remains a need for further research that compares the statistical performance of all four approaches in a single study, our findings suggest that users could use the semiparametric asymptotic approach for assessing the statistical significance of inequality differences rather than the other approaches. Our unit record data variant of this approach is straightforward to implement using available software, can incorporate survey weights, and provides improved inference.

The rest of the paper unfolds as follows. In Section 2 we describe the key elements of the four inference approaches cited above. We provide details of our UK household survey data and its income variables in Section 3. We report inequality estimates and tests of pairwise inequality differences in Section 4, comparing results across methods. Section 5 contains our conclusions. Supplementary materials cited in the main text are in Appendices A–D.

2. Approaches to assessing the statistical significance of income inequality differences

This section explains the four approaches to inference that we apply, providing an overview of their key features (but not repeating the detailed expositions available in the original articles), and explaining how we have adapted the semiparametric approaches to real-world data. Throughout, we assume that there are unit record data from at least two household

surveys with independent samples. Each survey contains suitably defined income variables and there are survey weights.

2.1. Conventional asymptotic approach

The asymptotic approach is the one most used. It takes no account of the heavy-tailed nature of income distributions.

A point estimate of inequality index *I* can be derived either by direct calculation from the unit record data or via the method of moments. For example, for the Theil index, *T*, the researcher calculates the estimate $\hat{T}$:

$$\hat{T} = \sum_{i=1}^{N} f_i \, (y_i/m)\log(y_i/m)$$

where y_i is the income of unit i, f_i is the survey weight for i normalized by the sum of the weights over the N sample units, and m is the weighted sample mean. Alternatively, but equivalently, the Theil index can be written in terms of its moments:

$$T = (v/\mu) - \log(\mu)$$

where $v = \mathrm{E}[y\log(y)]$ and μ is mean income, $\mathrm{E}[y]$. *T* can be estimated by replacing the moments by their (weighted) sample counterparts. For the weighted data case, Cowell (1989) provides moment formulae for all Generalized Entropy indices including the Theil index.

Sampling variances can be derived for the unit record approach using linearization or influence function methods – they lead to identical formulae (Cowell and Flachaire, 2015). Formulae for the weighted data case are provided by Biewen and Jenkins (2006) for Generalized Entropy and Atkinson indices and by Langel and Tillé (2013) for the Gini coefficient. Methods of moments variance estimators using weighted data are derived using the linearization (delta) method and the requisite formulae are functions of the covariance matrix for the moments: see Cowell (1989).

To test the hypothesis of equality for years *A* and *B*, the Studentized (t-type) test statistic W_d is given by:

$$W_d = \frac{\hat{I}_B - \hat{I}_A}{\left[\hat{V}(\hat{I}_A) + \hat{V}(\hat{I}_B)\right]^{0.5}} \qquad (1)$$

where the $\hat{I}$ and $\hat{V}(.)$ are the inequality index and variance estimates. See Davidson and Flachaire (2007, eqn. 16). The *p*-value for the null hypothesis of no difference in inequality according to index *I* is:

$$P^* = 2N(-|W_d|) \qquad (2)$$

where $N(.)$ is the standard normal cumulative distribution function.

2.2. Semiparametric asymptotic approach

The semiparametric asymptotic approach addresses heavy-tail and top-sparsity issues by replacing the observed survey data at the top of the distribution with a Pareto distribution that has been fitted to those same data. The parametric assumption not only fills in the distribution within the top income range in the observed survey data but also extrapolates it beyond the observed range. Each of the sample income distributions compared is a mixture of Pareto-distributed incomes for Rich top-income units (identified by indicator $R = 1$) and observed incomes for the remaining units ($R = 0$). For a given survey dataset, the asymptotic and semiparametric asymptotic point estimates of a specific inequality index are not necessarily equal (see Figure 2 below).

To derive point and variance estimates, Cowell and Flachaire (2007) use the method of moments. They exploit the assumption that the survey units are independently distributed and derive the moments for the overall mixture distribution as a population-share weighted sum of the moments for $R = 0$ units (as in the conventional asymptotic approach) and the moments implied by the (fitted) Pareto distribution for $R = 1$ units. Cowell and Flachaire consider five Generalized Entropy indices ($GE(\alpha)$, $\alpha \in [-1, 2]$) but ignore survey weights. Allowing for weights is difficult because, in the mixture distribution case, the already complex formulae shown in Cowell and Flachaire's (2007) equations 21–25 need to be generalized to incorporate bivariate moments (for incomes and weights), i.e., one needs to extend Cowell's (1989) formulae for the non-semiparametric case to the semiparametric case.

Our unit record data variant of Cowell and Flachaire's (2007) approach is as follows. For each survey dataset, we use the observed incomes $\boldsymbol{y}$ for $R = 0$ units but, for each $R = 1$ unit i, we replace its observed income y_i with a value y_i^P which is a random draw from the fitted Pareto distribution. To preserve the joint ordering of weights and incomes in the original data, we sort the distribution of observed incomes $\boldsymbol{y}$ in ascending order and also sort the distribution of imputed incomes $\boldsymbol{y}^P$ in ascending order. Then we allocate the survey weight of the r^{th} richest unit in $\boldsymbol{y}$ to the r^{th} richest unit in $\boldsymbol{y}^P$. Differently from Cowell and Flachaire (2007), we fit the Pareto distribution by a more robust method (OBRE rather than maximum likelihood, explained in §2.4 below).

Our unit record data procedure is similar to the Alfons et al. (2013) 'replacement of non-representative outlier' approach. However, differently from them (and motivated by

Blanchet, Flores, and Morgan, 2022), we improve coverage of the top income range by randomly drawing $M = 50$ imputed values for each $R = 1$ unit, proportionally reducing each of the cloned units' sampling weights at the same time. Improving coverage also implies closer correspondence to Cowell and Flachaire's (2007) moment-based approach which is akin to using $M = +\infty$. (We thank Emmanuel Flachaire for this insight.) We experimented with values of M ranging from 1 to 100 and found that $M = 50$ provided a good balance between improved coverage (assessed using comparisons of empirical densities and densities implied by the fitted Pareto distributions) and increasing computational burden. Further details are available on request.

Implementation of these steps yields a revised unit record dataset containing incomes and accompanying weights to which one can apply the standard asymptotic methods described in the previous section, except that they are now semiparametric estimates because the Pareto distribution is used. If there were no survey weights, our approach would mimic that of Cowell and Flachaire (2007). However, because we work with unit record data, it is straightforward to undertake estimation and inference for almost any inequality index, not only the Generalized Entropy ones that were their focus.

Test statistic W_d and the p-value for the hypothesis of no inequality difference are calculated as in (1) and (2) except that inequality and variance estimates now refer to estimates derived from semiparametric distributions.

2.3. Semiparametric percentile-t bootstrap approach

Davidson and Flachaire's (2007, p. 158) algorithm is as follows. The approach starts by calculating the same W_d statistic as the conventional asymptotic approach (eqn. 1) but, instead of using the standard normal distribution for inference (as in eqn. 2), one uses a bootstrap approach which resamples from semiparametric mixture distributions.

That is, one fits a Pareto distribution to each of the survey datasets for years A and B, followed by construction of two semiparametric distributions and estimates of an inequality index for each of the two years, $\tilde{I}_A$ and $\tilde{I}_B$. The index was the Theil index in Davidson and Flachaire's (2007) case, but the algorithm also works for other indices. The next step is to construct a pair of bootstrap samples from the pair of semiparametric distributions. For each distribution, this is derived by taking a standard bootstrap sample of $R = 0$ units and, for $R = 1$ units, replacing each unit's observed income with a single random draw from the fitted Pareto

distribution. Then, from the two semiparametric distributions that result, calculate the pairs of inequality index estimates and sampling variances.

The Studentized test statistic for each bootstrap sample, b, is:

$$W_b^* = \frac{\left[\left(\hat{I}_{Bb} - \tilde{I}_B\right) - \left(\hat{I}_{Ab} - \tilde{I}_A\right)\right]}{\left[\hat{V}\left(\hat{I}_{Ab}\right) + \hat{V}\left(\hat{I}_{Bb}\right)\right]^{0.5}}. \tag{3}$$

As Davidson and Flachaire explain, "the numerator is recentred so that the statistic tests a hypothesis that is true for the bootstrap samples" (2007, p. 161). After repeating this step B times, one has a distribution of bootstrap statistics W_b^* for $b = 1, \ldots, B$.

The percentile-t bootstrap p-value for the test of equality, P^*, is the proportion of bootstrap samples for which the bootstrap statistic is more extreme than the statistic calculated from the original data, W_d:

$$P^* = \left(\frac{1}{B}\right)\sum_{b=1}^{B} \iota(|W_b^*| > |W_d|) \tag{4}$$

where $\iota(.)$ is the indicator function. See Cowell and Flachaire (2007, eqn. 18).

Our implementation of Davidson and Flachaire's (2007) approach uses the same value formula but modifies the algorithm. In our variant, we fit the Pareto distribution using OBRE rather than maximum likelihood, calculate inequality indices and sampling variances using the unit record approach described earlier (instead of the method of moments), allow for survey weights, and undertake calculations for more inequality indices (not only the Theil index). When calculating inequality indices $\tilde{I}_A$ and $\tilde{I}_B$ from the semiparametric distributions, we expand the data ($M = 50$) for $R = 1$ observations, as discussed in §2.2. However, to mimic Davidson and Flachaire's (2007) approach, we do not expand the data within the bootstrap replications. Were we to use $M > 1$ in this context, our variant of the bootstrap approach would artificially reduce variability. (We thank Emmanuel Flachaire for this insight.)

We also considered an additional bootstrap approach in preliminary analyses. Davidson and Flachaire's (2007) bootstrap approach does not take account of the uncertainty arising because the Pareto shape parameters are themselves estimates. To address this issue, we used a modified bootstrap approach in which the Pareto parameters were estimated within each bootstrap repetition. Cowell and Flachaire (2007) point out that this approach creates its own issues, and they rely on the semiparametric asymptotic method instead.

Calculating standard errors for Generalised Entropy indices using their method of moments formulae requires that all relevant moments are finite, specifically that $\hat{\theta} \geq \max\{0, 2\alpha\}$, where $\hat{\theta}$ is the estimated Pareto shape parameter and α is the sensitivity

parameter for the GE(α) inequality index (see their eqn. 26). For each bootstrap sample in the modified semiparametric bootstrap approach, "we need to estimate a new value $\hat{\theta}$: even if condition (26) is satisfied in the original sample, it can be violated many times in bootstrap samples" (Cowell and Flachaire, 2007, p. 1091). In our empirical analysis reported below, the $\hat{\theta}$ for each year at the initial step range between 4.5 and 2.5 and, consistent with Cowell and Flachaire's prediction, we found many violations of their condition 26 when fitting Pareto distributions within bootstrap replications. Using a unit record data approach like ours does not get rid of the problem. For example, estimates of GE(2) indices, and their sampling variances specifically, were erratic and unreliable. Consequently, like Cowell and Flachaire (2007), we rely on the semiparametric asymptotic method and do not report estimates from our additional semiparametric bootstrap analyses.

2.4. Researcher choices when applying semiparametric approaches

There are implementation choices, including how to fit the Pareto distributions.

First there is the issue of which units should be used to fit the Pareto distribution for each of the surveys available, i.e., how should one define the subgroup of $R = 1$ units? There is a trade-off between bias and variance. The Pareto assumption is likely a better description of top incomes the higher the top income range considered (Jenkins 2017, Charpentier and Flachaire 2022) but the smaller the number of observations, the greater the sampling variability and high-income outliers may have undesirable effects. There is a substantial literature about fitting of Pareto distributions and how to choose the k richest units for the estimation sample: see, e.g., Cowell and Flachaire (2007), Davidson and Flachaire (2007), and Jenkins (2017), who cite references to informal graphical methods and more formal statistical methods. The choice of k is often discussed in the context of fitting a Pareto model to a single sample distribution, whereas in our application we have 42 years of data, and we also have survey weights. So, rather than using a sample selection rule tailored to each year separately and framed in terms of the number of richest units, we adopted a rule referring to the (weighted) fraction of top-income observations ('ptail').

We set ptail = 5% in the results we report, in which case min(k) = 239 and max(k) = 339 households across our 42 years of survey data (see Section 3). The same fraction was used by Atkinson and Jenkins (2020) in their multi-year semiparametric inequality analysis (covering 1937–2010). Jenkins (2017) used ptail = 1%, 5%, and 10% when analysing yearly data for 1995–2010, and reports that semiparametric inequality estimates were little different

across variants. Other rules of thumb have been used: Davidson and Flachaire (2007, p. 157) set k equal to the square root of the sample size for each of their simulated distributions and Cowell and Flachaire (2007, p. 1059) and Midões and de Crombrugghe (2023) used a similar rule. The 'merge point' algorithm developed by Blanchet, Flores, and Morgan (2022) is inapplicable in this context because it requires external information from administrative record data.

To check the robustness of our results, we repeated all our inference analysis using ptail = 1% (selected results are reported in Appendix D). Pareto shape parameter estimates were a bit smaller (distributions were more heavy-tailed) than for the ptail = 5% variant, and there were more large swings in confidence intervals across years and larger two-sample test p-values. Notwithstanding these differences, the substantive conclusions about pairwise inequality differences that we draw assuming ptail = 5% are little affected.

A second choice concerns the method used to fit the Pareto model. Davidson and Flachaire (2007) and Cowell and Flachaire (2007), and many other top-income researchers, use the maximum likelihood (ML) estimator. The ML estimator of the Pareto shape parameter and its standard error is consistent, efficient, and asymptotically normal (Hill 1975; Quandt 1966). However, the ML estimator can be biased if there are a few high outlier incomes, whether genuine or reflecting error or data contamination in the sense of Cowell and Victoria-Feser (1996) and Cowell and Flachaire (2007): the influence function for the maximum likelihood estimator is unbounded in this situation. To address this potential problem, we use the ML 'Optimal b-robust estimator' (ML-OBRE) due to Victoria-Feser and Ronchetti (1994). The idea is to use the ML score function for most of the data, exploiting the efficiency of the ML estimator, but to place an upper limit c on it for high income values in the interests of robustness. Victoria-Feser and Ronchetti (1994) show that with 95% efficiency, the optimal value in the Pareto case is $c = 3$, and this is what we use. Brzezinski's (2016) Monte-Carlo study finds that ML-OBRE performs well compared to four other robust estimators, including the one favoured by Alfons et al. (2013). As it happens, our ML-OBRE and ML estimates of Pareto shape parameters were similar (further details available on request).

Another choice concerns the number of bootstrap replications used when implementing the semiparametric bootstrap approach. Davidson and Flachaire (2007), Cowell and Flachaire (2007), and Midões and de Crombrugghe (2023) all used $B = 199$. We use $B = 999$ to increase bootstrap precision while not increasing computational burden unduly.

2.5. The Student-t approach

The Student-t approach proposed by Ibragimov and Müller (2010) and implemented by Ibragimov et al. (2025) takes no specific account of the heavy-tailed nature of income distributions but produces symmetric test statistic distributions by construction. For each survey, first, randomly allocate the sample units to q groups, with $q \geq 2$. Second, calculate the inequality index I separately for each group. (Any inequality index can be used, but Ibragimov et al. 2025 focus on the Gini and Theil indices.) Third, derive the overall estimate of I for the survey as a simple average of the group estimates and its variance as the sample variance of the group estimates. Having repeated these steps for all the surveys available, undertake pairwise t-tests for inequality differences between surveys using expressions analogous to (1) and (2) for test statistics and p-values (hence the 'Student-t' label). Midões and de Crombrugghe (2023) also evaluate the Student-t approach extensively, remarking that "the method is simple, intuitive and computationally cheap, particularly in comparison to non-standard bootstrap methods" (2023, p. 908).

The developers of the Student-t approach have not considered survey weights. The validity of their tests is founded upon the independence and asymptotic normality of the various group estimators, and this is unaffected by the presence of weights. Hence, the desirable properties of the Student-t estimators carry over to the weighted data case.

We split each household survey sample into 8 equal-sized groups of households. In principle, the number of groups may differ across empirical distributions but our use of 8 equal-sized groups for all years is not only practical given our 42 years of data but also consistent with the recommendations of Ibragimov et al. (2025) and Midões and de Crombrugghe (2023), noting that our sample sizes are relatively large and do not differ markedly across years (Appendix Table A1).

2.6. Ranking the approaches in terms of statistical performance

In a 'beauty contest' between the four approaches, where beauty is assessed in terms of test size, there is one clear loser – the conventional asymptotic approach – but no clear winner. (Test size refers to the probability of rejection of the null hypothesis of index equality compared to the nominal benchmark, taken to be 5% in the literature we cite.) It is difficult to point to winners because the calculations of empirical test size for the various approaches have analyzed different types of tests (mostly one-sample tests, fewer two-sample tests), and used different inequality indices (but most commonly the Theil index). Moreover, although

all the assessments of test performance use a similar set of Singh-Maddala distributions, albeit with variations in tail-heaviness (and some also use other distributions such as the lognormal and Pareto), there is no assessment with performance comparisons of all four approaches. This is because the Student-t approach was developed relatively recently. Moreover, the focus of Ibragimov et al.'s (2025) and Midões and de Crombrugghe's (2023) performance comparisons is between the Student-t and permutation approaches (and the conventional asymptotic approach). Only Cowell and Flachaire (2007) analyze the performance of the semiparametric asymptotic method.

Davidson and Flachaire write regarding the Theil index that "accurate inference can be achieved with [their semiparametric bootstrap] method in moderately large samples" (2007, p. 141), by contrast with the conventional asymptotic, standard and Moon bootstrap methods. All tests are oversized (test size is larger than the nominal 5%; error rejection fractions are positive), but least so for the semiparametric method. Cowell and Flachaire conclude that "the semiparametric bootstrap outperforms the other methods and gives accurate inference in finite samples" and "in situations where semiparametric inequality measures can be used, they perform well in asymptotic tests and at least as well as semiparametric bootstrap methods" (2007, p. 1068). The conclusions are (mostly) based on one-sided one-sample tests. For details, see their Tables 4 (Theil index) and 5 (mean logarithmic deviation) comparing conventional asymptotic, standard bootstrap, semiparametric asymptotic, and semiparametric bootstrap methods. Performance also depends on the inequality index. Cowell and Flachaire's (2007) Figures 10 and 11 show that, for both semiparametric approaches, error rejection fractions are small and minimized if the inequality index is not too top- or bottom-sensitive. Davidson and Flachaire (2007, Figure 15) also include analysis of a two-sided two-sample test of equal Theil indices showing that error rejection fractions are smallest for the semiparametric bootstrap for sample sizes of at least 2,000 (but around 5% nonetheless), and largest for the conventional asymptotic approach (around 7.5%).

Ibragimov et al. (2025) and Midões and de Crombrugghe (2023) also consider two-sided two-sample tests. Ibragimov et al. (2025) report size calculations for the Theil and Gini indices for two distributions with the same number of observations ($N = 200$) in their Table 1, comparing conventional asymptotic, permutation and permutation bootstrap, and several variants of their Student-t approach. Their Table 2 shows similar calculations for two distributions with different numbers of observations. Overall, Student-t tests appear to have better size properties than conventional asymptotic tests (and at least as good as permutation

and permutation bootstrap tests). As long as the distributions compared are not too heavy-tailed, the Student-t tests tend to be slightly conservative (size just below 5%). Midões and de Crombrugghe's (2023) conclusions are similar. See, e.g., their Figure 6 for the Theil index showing the Student-t method "controls for size remarkably well" (p. 914) compared to the conventional asymptotic and semiparametric bootstrap approaches over sample sizes ranging from 500 to 10,000, though the differences between them diminish if the distributions are not heavy-tailed (p. 915).

All in all, although picking winners in the statistical performance beauty contest requires new research that compares all four methods on the same terms, it is already clear that performance is conditional on how heavy-tailed the distributions being compared are and the index used to summarize inequality. Even without the new research, it is of considerable interest to know whether the four methods provide similar or different inferential conclusions when applied to real-world survey data, and this paper delivers such information for the first time.

3. The ONS's ETB data and income concepts

Our analysis is based on a historical series of unit-record household survey data deposited by the Office for National Statistics (ONS) at the UK Data Service (ONS, 2022). The data were used by the ONS in their annual articles about the 'Effects of taxes and benefits on household incomes' (ETB) until recently: see, e.g., ONS (2019) for 2018. Since 2020 the ONS has revised its inequality data, incorporating a top-income adjustment based on income tax administrative data to improve survey coverage of the very top of the distribution, and supplemented the survey data (now labelled the Household Finances Survey): see, e.g., ONS (2020) and references therein. We restrict attention in this paper to the 42-year-long consistent series of survey data that finishes just prior to the onset of Covid-19. We return to discuss inference using ONS's most recent ETB data in Section 5.

The public-use ONS data we use derive from the Living Costs and Food Survey (LCFS, from 2008) and its predecessor, the Family Expenditure Survey (FES, to 2007). These are household surveys with a focus on household spending and income, each intended to be nationally representative of the UK private household population. The annual sample size is approximately 6,500 households per year on average (and ranges from around 4,900 to 7,500). Survey years refer to financial years (12-month periods starting 5 April each year)

from 1993/94 onwards and to calendar years before that. For brevity we label financial years by their first part: '2016' refers to financial year 2016/17, etc. ETB income data are not top-coded.

We use the ONS's definition of income, which is consistent across the 42 years. Income is net (disposable) household income, i.e., household income from employment and self-employment, plus income from capital and government cash transfers, minus personal income tax payments, employee national insurance contributions, and local taxes such as council tax. (Incomes are expressed in pounds per week.) This definition corresponds closely to that set out by the Canberra Group's (2011) guidelines and is also used by statistical offices in most high-income countries around the world, including Eurostat for official income statistics for EU member states.

Following the ONS, we adjust all household incomes and income components for differences in household size and composition using the modified-OECD equivalence scale. The ONS uses the same scale in its reports but our calculation of it differs slightly from theirs. This is because the modified-OECD scale defines children to be individuals aged 14 or under. In our public-use dataset, we only know whether an individual is a 'dependent child', i.e., aged 15 or less, or aged 16–19 and in full-time education. Thus, our equivalence scale calculations count slightly more children than the ONS do, but we expect the effects to be negligible.

The FES and LCFS include survey weights. Before 1996, a household's weight was simply the number of individuals in the household. From 1996 onwards, the weights also adjust for non-response and are calibrated to population totals. We use the survey weights in all calculations.

The estimates we report are based on data that we trimmed lightly. Specifically, we drop a small number of incomes that are (i) zero or negative, (ii) between zero and one, or (iii) top incomes that are more than twice the next highest income. These selections remove on average 14, 1, and 0.2 incomes per year, respectively. (Over the 42 years, only 9 top incomes are dropped, no more than one per year.) In total the number of households dropped is 0.23% per year on average and is never more than 0.48% in any year. For more details of the sample sizes and numbers dropped by trimming, see Appendix A.

Our trimming rule is consistent with standard practice but is intentionally conservative to check how well the inference approaches deal with top-income issues. Many inequality indices are undefined for incomes less than or equal to zero and these values also raise questions about survey measurement error, hence rejection rule (i). Rules (ii) and (iii)

reduce the possibility that bottom- and top-sensitive inequality indices produce estimates unduly affected by influential outliers. On these, see the discussion by Cowell and Flachaire (2007) and Cowell and Victoria-Feser (1996). We have rerun all our analysis using data without using rule (iii) and found, as expected, that comparisons based on the top-sensitive GE(2) index were even less robust than the ones we report, as predicted by Cowell and Flachaire's (2007) simulation analyses. (Results are available on request.)

ONS ETB articles do not report standard errors (SEs) or confidence intervals (CIs) or statistical tests of inequality differences. The technical documentation accompanying the LCFS (see, e.g., Office for National Statistics, 2023) provides general discussion about sampling errors but reports no SEs or CIs for inequality indices. Hence, our inferential analysis is the first of its kind for ETB data that we are aware of.

4. Assessing the statistical significance of inequality differences: results

The running order for our presentation of results is as follows. First, we show for multiple indices how the approaches differ in terms of their (point) estimates and precision. Second, we summarize the tests of inequality differences between pairs of years. Summaries are essential because with 42 years of data there are 861 possible two-sample tests to report, for each approach and index. Hence, we begin by showing for each year A, the number of two-sample tests for which the p-value is less than 5% for the test of no inequality difference between year A and every other year B (maximum number = 41), by approach and index. (5% is the most used critical value for tests of inequality differences.)

We then focus on four specific years, $A \in \{1977, 1990, 2006, 2018\}$ and report the p-values for tests of no inequality difference for each of these four years and every other year, B, by approach and index. We chose the four specific years deliberately. They are approximately evenly spaced over the 42-year period. Moreover, 1977 is the year with lowest inequality, 1990 is when the 1980s inequality increase levelled off (Figure 1) and 2006 is a year when inequality was little different from the 1990 level according to most indices. In addition, 1990 and 2006 were years prior to recessions. All in all, we provide detailed examination of differences over the initial period when we expect statistically significant changes, but also within the later period when statistical significance is likely to be more at issue.

For brevity, we focus on results for the Gini and Theil indices because the former is the most used index among practitioners and the latter has been the focus of most simulation-based assessments of the semiparametric approaches to inference. For completeness, we also derive results for additional indices (GE(α), for $\alpha = -1, 0, 2$; top 10% share; top 1% share; $p90/p10$), reporting them in Appendix C. For a comprehensive survey of inequality indices providing formulae and discussing properties and interpretations, see inter alia Cowell (2000). The Gini coefficient is the most reported index. It is more sensitive to income differences around the middle of the distribution than to differences at the top or at the bottom ('middle-sensitive'). GE indices range from bottom-sensitive to top-sensitive, with sensitivity depending on parameter α, with researchers most commonly using values $\alpha = -1$ (bottom-sensitive), 0 (mean logarithmic deviation, MLD; middle-sensitive), 1 (Theil index; slightly top-sensitive), and 2 (half the squared coefficient of variation; distinctly top-sensitive). Top income shares and the $p90/p10$ quantile ratio are widely used inequality indices. We expect inference about $p90/p10$ differences to be similar across the four approaches we apply because the index takes no account of the income distribution above the 90th percentile. Following Cowell and Flachaire (2007), we also expect inference about MLD differences and Gini differences to be similar as they are both middle-sensitive.

4.1. Index estimates and indicative precision, by approach

The conventional asymptotic and semiparametric bootstrap approaches provide identical estimates of each inequality index – the series shown in Figure 1 – and the corresponding Student-t series of estimates are virtually identical to these series as well. This is also true for the GE(–1) and GE(2) indices. (See Appendix Figures B1–B3.) The semiparametric asymptotic series differs slightly from the others, with the differences more apparent, the more top-sensitive is the inequality index, which of course is where the Pareto imputations are most relevant. The differences are not a problem for inference because the semiparametric bootstrap approach computes the relevant test statistics accounting for initial-step inequality differences (see eq. 3).

Table 1 provides indicative information about the relative precision of the estimates, by index and approach. We summarize precision by the length of the one-sample 95% confidence interval (CI), expressed as a percentage of the corresponding estimate, to facilitate comparisons across indices and approaches. Larger values mean lower precision. Bootstrap CIs are calculated using Davidson and Flachaire, 2007, eqn. 17. (Note, however, the authors'

caveats about assessing CI accuracy in this context; hence our reference to 'indicative' precision.) For brevity, the table shows 42-year averages for each index and approach.

<Table 1 near here>

Table 1 shows that the two semiparametric approaches provide more precise estimates than the conventional asymptotic approach, as expected, for all indices (except GE(–1) to which we return). There is a gain in precision from using the semiparametric asymptotic approach rather than the semiparametric bootstrap (cf. cols. 3 and 2). Comparing indices, observe that, for all approaches, precision decreases the more top-sensitive is the inequality index (the larger is α for index GE(α), and the top 1% share versus the top 10% share), and especially so for GE(2). (The exception is the semiparametric asymptotic approach, for which GE(2) is relatively precisely estimated.) The precision of $p90/p10$ is much the same whatever the approach, as expected.

By comparison with the three other approaches, the Student-t approach stands out for its greater imprecision for all inequality indices, greater even than the conventional asymptotic approach (cf. cols. 5 and 1).

This finding requires careful interpretation. Arguably, if the three other approaches are oversized by comparison with the Student-t approach, as suggested by the discussion in §2.6, then one would expect those approaches to have narrower (standardized) CIs, arising from smaller standard errors, which is what is seen in Table 1. However, this argument is not fully persuasive. First, the (standardized) CIs for the conventional asymptotic and Student-t series are not too different, which is not what the research on test performance cited earlier would lead us to expect. Second, the same research shows that the conventional asymptotic method delivers the most over-sized tests, so one would expect it to deliver smaller, not larger, standardized CIs by comparison with the semiparametric methods. An alternative interpretation of the patterns shown in Table 1, and the one we favour, is that with these real-world data the semiparametric approaches are delivering test statistics closer to nominal size, and that the Pareto assumption for the top tail is 'buying' greater precision.

Providing more detail about indicative precision, Figure 2 shows standardized CIs for the Gini and Theil index estimates, year by year. The figure underlines the points made regarding Table 1 in terms of the relative precision of the different approaches but adds more. The semiparametric asymptotic series is remarkably stable over time by contrast with the others, for both indices. Two other features of the charts stand out.

First, the semiparametric bootstrap series is more like the conventional asymptotic series when considering the Theil index rather than the Gini. Second, there are large year-to-year fluctuations in the Student-t series, for both indices. (This feature reflects year-to-year variations in standard errors rather than estimates: see Figure 2.) We do not have a definitive explanation for this but conjecture it is due to 'grouping variability'. This is the issue arising because the Student-t method uses a single random allocation of survey units to groups. The problem is that a different random allocation to groups leads to different estimates and sampling variances and so, arguably, random chance underpins the observed variability. See Hérault and Jenkins (2025) for more discussion.

<Figure 2 near here>

Our principal interest is in two-sample tests, not indicative precision or its fluctuations. We summarize these tests next.

4.2. p-values for tests of no inequality difference between a pair of years

Figure 3 summarizes counts of *p*-values less than 5% for tests of no inequality difference between each year *A* between 1977 and 2018 and every other year *B*, for the Gini and Theil indices. The maximum count is 41. We use 5% as the threshold for statistical significance in accordance with common practice.

<Figure 3 near here>

There are high counts for every year from 1977 through 1986, for both indices, and all approaches. These high counts are followed by a fall in 1987, and fluctuating counts thereafter, as one might expect given the ups and downs in the inequality estimates shown in Figure 1. However, there are distinct differences by approach for this period.

As Table 1 and Figure 2 lead us to expect, the semiparametric asymptotic approach yields higher counts than the three other approaches. For example, for the Gini index, the counts for the former fluctuate around 30 (nearly three-quarters of the maximum count). The semiparametric bootstrap approach yields slightly fewer (typically around 5 per year). In contrast, the counts for the conventional asymptotic and Student-t approaches fluctuate around 20 (only around half the maximum count). For the Theil index, the patterns are like those for the Gini except that the counts for semiparametric asymptotic approach are generally slightly higher over the 1986–2017 period and those for the conventional asymptotic and Student-t approaches are lower. Also, for the Theil index comparisons, the *p*-value counts per year for semiparametric bootstrap series are more clearly in between the series for the other approaches.

There are similarities and differences in temporal patterns of counts when inequality indices other than these two are considered. For example, according to the counterparts to Figure 3 for the MLD (Appendix Figure C1), top 10% share (Figure C3), and top 1% share (Figure C5), and GE(2) (Figure C11), the semiparametric approach continues to deliver the highest counts across all years, and the conventional asymptotic and Student-t approaches the lowest. The semiparametric bootstrap series is more like the semiparametric asymptotic series for the middle-sensitive MLD index and the top 10% share. For the $p90/p10$ index (Figure C7), counts per year are similar for all four approaches, albeit a bit lower for the Student-t approach. For bottom-sensitive GE(–1) (Figure C9), counts per year are also similar for all approaches, albeit a bit lower for the semiparametric bootstrap approach.

We now turn to more detailed analysis of inequality differences, summarizing comparisons between each of four specific years (1977, 1990, 2006, and 2018) and every other year. In Figures 4 and 5, we show *p*-values for the two-sided two-sample tests for the Gini and Theil indices respectively, with dotted lines showing the 5% critical value for reference. It is useful to be able to see the *p*-values themselves, and thereby assess whether they are well below any specific threshold or close to it. It turns out that, if a *p*-value is less than 5%, it is typically close to zero.

<Figures 4 and 5 near here>

Figures 4 and 5 show there is a distinct difference in test *p*-values for the semiparametric asymptotic approach on the one hand and the conventional asymptotic and Student-t approaches on the other hand. The first yields more statistically significant *p*-values (using the 5% benchmark) for both the Gini and Theil indices for comparisons with years after 1986. The semiparametric bootstrap approach produces *p*-values that generally lie in between those for the other three approaches, closer to the semiparametric asymptotic values for the Gini coefficient and closer to the other two approaches' values for the Theil index.

What are the substantive findings if we focus on the test results for the semiparametric asymptotic approach? For both the Gini and Theil index, inequality in 1977 – a low inequality year – differed from inequality in every year from 1980 onwards. Inequality in 1990, a high inequality year, differs significantly from inequality in most other years except for the relatively high inequality years at the end of the 1990s and around 2010 (recall the estimates shown in Figure 1). Inequality in 2006 is slightly lower than in 1990 but differences are statistically significant for much the same set of years. In 2018, inequality is slightly lower again, and many of the significant differences for the Gini index are the same as for the comparisons between 1977 and 1990, except that (for example) comparisons for 1987 and

1988 are not statistically significant and there are only a few statistically significant differences from the Gini for years in the early 2000s and later. The story for the Theil index is similar except that there are more statistically significant year-pair differences for the years after the early 2000s. This period is also when the temporal changes in the Theil index were larger than for the Gini coefficient (Figure 1).

Continuing with the results of the tests for the semiparametric asymptotic approach, we now report the outcomes for additional inequality indices. The set of year-pair differences that are statistically significant is much the same for middle-sensitive MLD index and top 10% share (Appendix Figures C2, C4) as for the Gini coefficient. The same is true for $p90/p10$ tests except that there are almost no statistically significant differences between 2018 and each year after 2000 (Figure C8). The pattern of test results for the top-sensitive indices – the top 1% share (Figure C6) and GE(2) (Figure C12) – are broadly similar to those for the Theil index, except that there are more statistically significant GE(2) differences between 2018 and years after 1986. For the bottom-sensitive GE(–1) (Figure C10), there are few statistically significant comparisons except between the low inequality year 1977 and year after 1985, or between higher-inequality years (1990, 2006 or 2018) and each year before 1987.

4.3. Are there lessons for statistical agencies and practitioners?

National statistical agencies typically derive SEs and CIs for published statistics, including inequality estimates, using the conventional asymptotic approach or similar, albeit accounting for additional survey design features such as clustering and stratification that we have not considered here. How the sampling uncertainty is communicated differs, but no agency accounts for heavy-tail issues when calculating SEs and CIs.

Consider two leading examples, the UK Department for Work and Pensions's *Household Below Average Incomes* publication ('HBAI', annual) that reports estimates of Gini coefficients using the same income definitions as the ONS ETB articles (see, e.g., Department for Work and Pensions, 2025b) and the US Census Bureau's annual P-60 report on *Income in United States* (see, e.g., Guzman and Kollar, 2023). UK *HBAI* reports do not publish SEs or CIs but, instead, report Gini coefficients to 2 decimal places, i.e., the nearest percentage point (a practice to which we return). The associated technical documentation (e.g., Department for Work and Pensions, 2025a) provides general discussion of statistical uncertainty and detailed explanations of their derivation of SEs accounting for survey design features, but no SEs or CIs are reported, nor are design effect sizes (DEFFs). In contrast, US

P-60 reports do provide information about SEs and the statistical significance of Gini index and quintile group share changes in their main tables (see, e.g., Guzman and Kollar, 2023, Figure 3). The associated technical documentation (US Census Bureau, 2023, especially Appendix H) explains how SEs are derived using a replicate method that accounts for sample design features.

The research reviewed in §2.6 shows the semiparametric asymptotic approach provides improved inference over the conventional asymptotic approach and we have shown that it is relatively straightforward to implement. Statistical agencies and other practitioners could therefore account for the heavy-tailed nature of the income data in their household surveys relatively easily. Moreover, observe that the methods we have used can be generalized to address survey design features such as clustering and stratification, with limited additional computational burden, by employing suitable bootstrap methods. Alfons et al. (2013) demonstrate this point.

On the other hand, arguably the Department for Work and Pensions's rounding of their Gini estimates to two decimal places provides a straightforward and transparent way to guard against unwarranted claims of statistically significant inequality differences. That is, consider a naïve inference rule that is 'count a one percentage point difference in rounded Ginis as significant'. If the rule worked, we would not find differences in rounded Ginis of one percentage point that are statistically insignificant. How well does the rule work in practice?

In our UK data, there are many one percentage point differences in rounded Ginis that are not statistically significant. (There are also many such differences that are statistically significant.) To take one example, there are 4 years for which the Gini estimate is one percentage point more than the 1989 Gini, which rounds to 32%, and all differences are non-significant (relative to the 5% benchmark) according to the semiparametric asymptotic (and other approaches). Similarly, for the Theil index (not reported by the ONS ETB report), there are 7 years for which the index is one percentage point smaller than the 1992 estimate (19% when rounded), but the difference is not statistically significant.

What if the naïve inference rule were changed to 'count a two percentage point difference in rounded Ginis as significant'? According to the semiparametric asymptotic method, there are no non-significant differences for Gini coefficients and one non-significant difference for the Theil index (2010 and 2012). However, according to the Student-t approach, there are many more such cases. If the naïve inference rule is relaxed even further, to refer to differences of at least three percentage points in rounded Ginis, there remain many

differences that are assessed by the Student-t approach as non-significant relative to the 5% benchmark. There are even more such differences in rounded Theil indices according to both the Student-t and conventional asymptotic approaches.

We therefore conclude that assessing the statistical significance of inequality differences using naïve rounding rules is unreliable. Of course, one percentage point differences in the Gini that are statistically significant are also likely not substantively significant. Atkinson, for example, has argued that "a three percentage point reduction in the Gini coefficient does not seem unreasonable as a criterion of salience" (2015, p. 54). However, salience defined thus does not necessarily imply statistical significance, as we have shown. Our view is that one should not discuss substantive significance without first considering statistical significance.

5. Conclusions

Using real-world data, we have adapted and road-tested recent proposals for improved inference about inequality differences. With 42 years of UK household survey data, we have confirmed the conclusions of Cowell and Flachaire (2007) that semiparametric methods provide more precise estimates than conventional asymptotic estimates. Using simulation analysis, they found that the semiparametric asymptotic approach provides similar inference conclusions to Davidson and Flachaire's (2007) semiparametric percentile-t bootstrap approach. Using UK survey data, we find that the former provides more statistically significant test outcomes than the latter. Conclusions are most similar across these two approaches for middle-sensitive inequality indices like the Gini.

We also find that the recently-proposed Student-t approach of Ibragimov et al. (2025) yields similar inferential outcomes to the conventional asymptotic approach when applied to UK data. Consistent with this, we note that in Ibragimov et al.'s application to pairwise differences in Gini coefficients between Moscow and 83 other Russian regions, "the conclusions of all the approaches – the asymptotic, [standard] bootstrap, permutation, and the *t*-statistic based robust tests – to testing equality of the Gini coefficients *GM* and *GR* agree among themselves" (2025, p. 401).

In addition, we have shown how switching to a unit record data variant of the semiparametric approaches enables researchers to consider inference for a larger portfolio of inequality indices and to incorporate survey design features such as weights. It would be

straightforward to add additional inequality indices to the portfolio, e.g., S80/S20 (the ratio of the income share of the richest fifth to the income share of poorest fifth) which is reported by Eurostat, and the Palma ratio (ratio of the income share of the richest tenth to the income share of the poorest 40%).

More research is needed about the statistical performance of the various approaches for two-sample tests. Most assessments of the statistical performance of semiparametric approaches have focused on one-sample tests, and yet two-sample tests are the most relevant for applied researchers. Student-t proponents have undertaken more two-sample tests but the performance comparisons between the Student-t and other methods have rarely included semiparametric methods, and none consider the semiparametric asymptotic approach. Another fruitful direction for future research is development of tests for multiple comparisons. As in all previous research on inequality differences, we have restricted ourselves to inference for pairwise comparisons and yet there is much interest in assessing hypotheses about upward or downward trends over multiple years.

There also remain issues concerning how income data from household surveys should be trimmed to remove egregious influential outliers at the bottom and especially top of the distribution. More radical trimming than we have used runs the risk of dropping valid values and thence introducing bias. Trimming the top 1% and bottom 1% may make estimates robust to high-leverage outliers but inequality trends for the middle 98% of the distribution are unlikely to describe trends for the population as a whole: what is going on at the very top and the very bottom is important too. Relatedly, we have illustrated with our empirical application how Cowell and Flachaire's (2007) remarks about the interconnectedness of inference and high-income outlier issues. For example, it is harder to obtain reliable inference for differences in top-sensitive indices like the GE(2) than for middle-sensitive indices like the Gini coefficient.

New inference methods are also required given recent developments to address non-sampling error, specifically the under-coverage of the top of the income range by household surveys. Unless addressed, this form of systematic error introduces under-estimation of inequality indices. To counter this problem, the UK Department for Work and Pensions has used a top-income adjustment – the 'SPI' adjustment – since 1992. The UK ONS recently introduced its own top-income adjustment building on the work of Burkhauser et al. (2018*a*, 2018*b*) and Jenkins (2017), and it underpins recent ETB reports. Blanchet, Flores, and Morgan (2022) propose a general methodology for making top-income adjustments. For a recent review of methods and practice, see Jenkins (2022).

The distinctive feature of these top-income adjustments is that they use external information about top incomes taken from income tax administrative record data. By contrast, the current paper and the 'improved inference' literature we build on uses no external information: the Pareto distributions for top incomes are estimated from the household survey data to hand. To develop inference for inequality indices calculated from top-income adjusted survey data might involve a modification of the semiparametric approaches considered to date. It might entail fitting Pareto distributions to the administrative data (as in Jenkins 2017) rather than the survey data, and perhaps also using a bootstrap approach that resamples from the top incomes in the administrative data.

Developments such as these are only possible if suitable administrative record data on incomes are available, and that is not the case for most countries in the world. They must continue to rely on the household survey data that are available, and the findings of this paper remain relevant for this common situation.

6. References

Alfons, A., Templ, M., and Filzmoser, P. (2011). Robust estimation of economic indicators from survey samples based on Pareto tail modelling. *Journal of the Royal Statistical Society (C): Applied Statistics*, 62 (2), 271–286. https://doi.org/10.1111/j.1467-9876.2012.01063.x

Atkinson, A. B. (2015). *Inequality. What Can Be Done?* London: Harvard University Press. https://www.jstor.org/stable/j.ctvjghxqh

Atkinson, A. B. and Jenkins, S. P. (2020). A different perspective on the evolution of UK income inequality. *Review of Income and Wealth*, 66 (2), 253–266. https://doi.org/10.1111/roiw.12412

Biewen, M. and Jenkins S. P. (2006). Variance estimation for Generalized Entropy and Atkinson inequality indices: the complex survey data case. *Oxford Bulletin of Economics and Statistics*, 68 (3), 371–383. https://doi.org/10.1111/j.1468-0084.2006.00166.x

Blanchet, T., Flores, I., and Morgan, M. (2022). The weight of the rich: improving surveys using tax data, *Journal of Economic Inequality,* 20 (1), 119–150. https://doi.org/10.1007/s10888-021-09509-3

Brzezinski, M. (2016). Robust estimation of the Pareto tail index: a Monte Carlo analysis, *Empirical Economics*, 51 (1), 1–30. https://doi.org/10.1007/s00181-015-0989-9

Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2018*a*). Top incomes and inequality in the UK: Reconciling estimates from household survey and tax return data. *Oxford Economic Papers*, 70 (2), 301–326. https://doi.org/10.1093/oep/gpx041

Burkhauser, R. V., Hérault, N., Jenkins, S. P., and Wilkins, R. (2018*b*). Survey under-coverage of top incomes and estimation of inequality: What is the role of the UK's SPI adjustment? *Fiscal Studies*, 39 (2), 213–240. https://doi.org/10.1111/1475-5890.12158

Canberra Group (2011). *Handbook on Household Income Statistics*, second edition. Geneva: United Nations Economic Commission for Europe. https://digitallibrary.un.org/record/719970?ln=en&v=pdf

Charpentier, A. and Flachaire, E. (2022). Pareto models for top incomes and wealth. *Journal of Economic Inequality*, 20 (1), 1–25. https://doi.org/10.1007/s10888-021-09514-6

Cowell, F. A. (1989). Sampling variance and decomposable inequality measures. *Journal of Econometrics* 42 (1), 27–41. https://doi.org/10.1016/0304-4076(89)90073-0

Cowell, F. A. (2000). Measurement of inequality. In: A. B. Atkinson and F. Bourguignon (eds), *Handbook of Income Distribution, Volume 1*. Amsterdam: Elsevier, 87–166. https://doi.org/10.1016/S1574-0056(00)80005-6

Cowell, F. A. and Flachaire, E. (2007). Income distribution and inequality measurement: the problem of extreme values. *Journal of Econometrics*, 141 (2), 1044–1072. https://doi.org/10.1016/j.jeconom.2007.01.001

Cowell, F.A. and Flachaire, E. (2015). Statistical methods for distributional analysis. In: A. B. Atkinson and F. Bouguignon (eds), *Handbook of Income Distribution, Volume 2*. Amsterdam: Elsevier, 359–465. http://dx.doi.org/10.1016/B978-0-444-59428-0.00007-2

Cowell, F. A. and Victoria-Feser, M.-P. (1996). Robustness properties of inequality measures. *Econometrica*, 64 (1), 77–101. https://doi.org/10.2307/2171925

Davidson, R. (2012). Statistical inference in the presence of heavy tails. *Econometrics Journal*, 15 (1), C31–C53. https://doi.org/10.1111/j.1368-423X.2010.00340.x

Davidson, R. and Flachaire, E. (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics*, 141 (1), 141–166. https://doi.org/10.1016/j.jeconom.2007.01.009

Department for Work and Pensions (2025*a*). *Households Below Average Income series: quality and methodology information report FYE 2023*, https://www.gov.uk/government/statistics/households-below-average-income-for-financial-years-ending-1995-to-2023/households-below-average-income-series-quality-and-methodology-information-report-fye-2023.

Department for Work and Pensions (2025b). *Households Below Average Income: an analysis of the UK income distribution: FYE 1995 to FYE 2023*, https://www.gov.uk/government/statistics/households-below-average-income-for-financial-years-ending-1995-to-2023/households-below-average-income-an-analysis-of-the-uk-income-distribution-fye-1995-to-fye-2023.

Dufour, J.-M., Flachaire, E., and Khalaf, L. (2019). Permutation tests for comparing inequality measures, *Journal of Business and Economic Statistics*, 37 (3), 457–470. https://doi.org/10.1080/07350015.2017.1371027

Guzman, G. and Kollar, M. (2023). *Income in the United States: 2022*, U.S. Census Bureau, Current Population Reports, P60-279. Washington, DC: U.S. Government Publishing Office. https://www.census.gov/content/dam/Census/library/publications/2023/demo/p60-279.pdf

Hérault, N. and Jenkins, S. P. (2025). The t-statistic approach to inference for inequality indices: the issue of 'grouping variability'. IZA Discussion Paper, forthcoming.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3 (5), 1163–1174.

Ibragimov, R., Kattuman, P., and Skrobotov, A. (2025). Robust inference on income inequality: t-statistic based approach. *Econometric Reviews*, 44 (4), 384–415. https://doi.org/10.1080/07474938.2024.2432362

Ibragimov, R. and Müller, U. K. (2010). *t*-statistic based correlation and heterogeneity robust inference. *Journal of Business and Economic Statistics*, 28 (4), 453–468. https://doi.org/10.1198/jbes.2009.08046

Jenkins, S. P., (2017). Pareto models, top incomes, and recent trends in UK income inequality. *Economica*, 84 (334), 261–289. https://doi.org/10.1111/ecca.12217

Jenkins, S. P., (2022). Top-income adjustments and official statistics on income distribution: the case of the UK. *Journal of Economic Inequality*, 20 (1), 151–168. https://doi.org/10.1007/s10888-022-09532-y

Langel, M. and Tillé, Y. (2013). Variance estimation of the Gini index: revisiting a result several times published. *Journal of the Royal Statistical Society, Series A*, 176 (2), 521–540. https://doi.org/10.1111/j.1467-985X.2012.01048.x

Midões, C., and de Crombrugghe, D. (2023). Assumption-light and computationally cheap inference on inequality measures by sample splitting: the Student *t* approach. *Journal of Economic Inequality*, 21 (4), 899–924. https://doi.org/10.1007/s10888-023-09574-w

ONS (Office for National Statistics) (2019). Effects of taxes and benefits on UK household income: financial year ending 2018. London: Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/theeffectsoftaxesandbenefitsonhouseholdincome/financialyearending2018

ONS (Office for National Statistics) (2020). Household income inequality, UK: Financial year ending 2019. London: Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/incomeandwealth/bulletins/householdincomeinequalityfinancial/financialyearending2019

ONS (Office for National Statistics) (2020). *Effects of Taxes and Benefits on Household Income Time Series, 1977-2017*. [data collection]. Office for National Statistics, [original data producer(s)]. Office for National Statistics. UK Data Service Study Number: 8683, DOI: http://doi.org/10.5255/UKDA-SN-8683-1

ONS (Office for National Statistics) (2023). *Living Costs and Food Survey technical report: financial year ending March 2023*, https://www.ons.gov.uk/peoplepopulationandcommunity/personalandhouseholdfinances/expenditure/articles/livingcostsandfoodsurveytechnicalreport/financialyearendingmarch2023

Quandt, R. E. (1966). Old and new methods of estimation and the Pareto distribution. *Metrika: International Journal for Theoretical and Applied Statistics*, 10 (1), 55–82. https://doi.org/10.1007/BF02613419

Schluter, C. (2012). On the problem of inference for inequality measures for heavy-tailed distributions. *Econometrics Journal*, 15 (1), 125–153. https://doi.org/10.1111/j.1368-423X.2011.00356.x

Schluter, C. and van Garderen, K. J. (2009). Edgeworth expansions and normalizing transforms for inequality measures. *Journal of Econometrics*, 150 (1), 16–29. https://doi.org/10.1016/j.jeconom.2008.12.022

US Census Bureau (2023). Technical documentation accompanying the 2023 Current Population Survey, https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar23.pdf.

Victoria-Feser, M.-P. and Ronchetti, E. (1994). Robust methods for personal-income distribution models. *Canadian Journal of Statistics*, 22 (2), 247–258. https://doi.org/10.2307/3315587

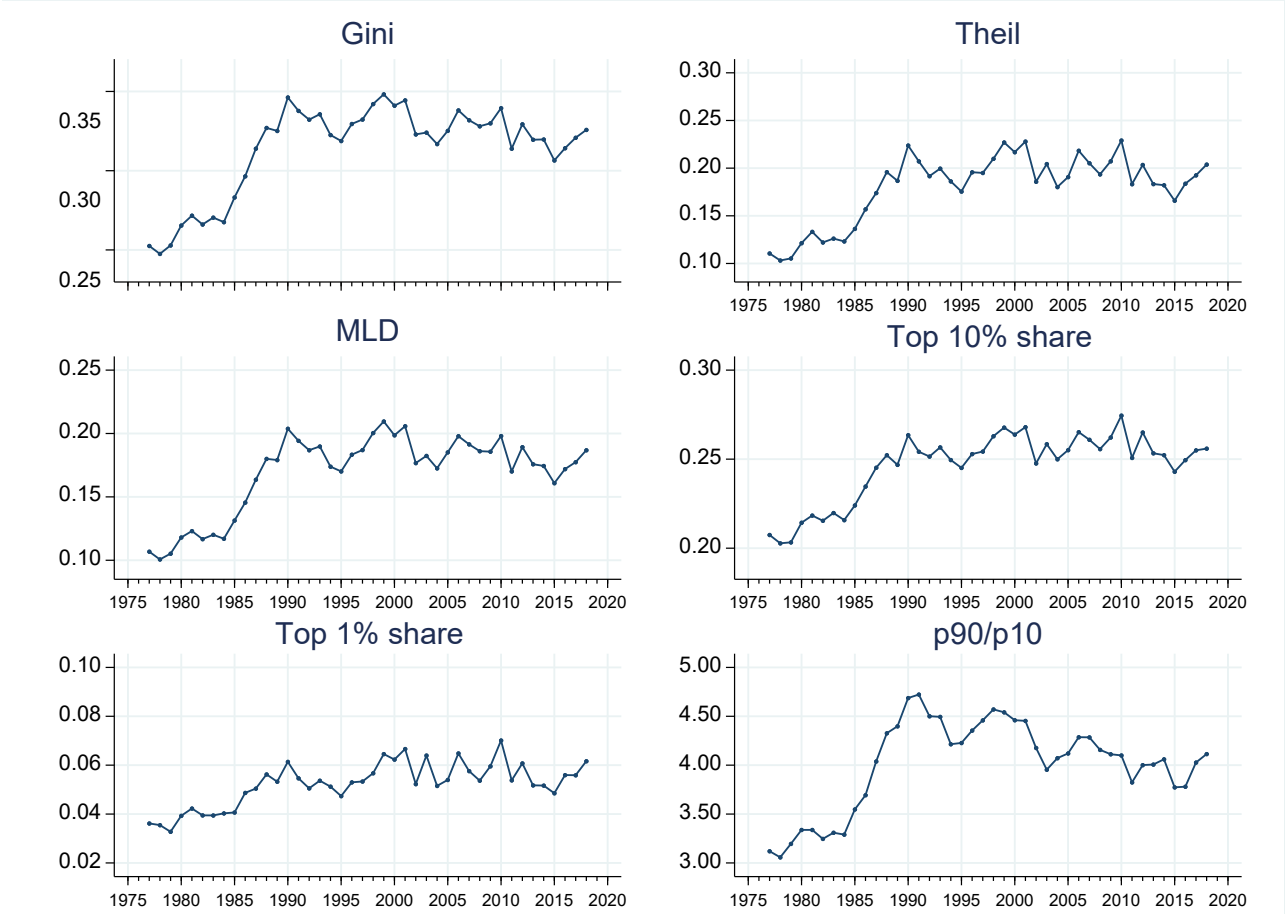
Data Availability Statement

Our research uses unit record data from the UK Office for National Statistics (ONS), 'Effects of Taxes and Benefits on Household Incomes' (ETB) datasets. The yearly datasets, covering 1977–2018, are the versions of the ETB datasets that do not include the top-income adjustments discussed in the Conclusions section of the main text. (The newer ONS ETB datasets incorporating top-income adjustments are catalogued at the UK Data Service (UKDS): see https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8856.)

The ETB datasets we use were previously made available to registered users of the UK Data Service, with the following Study Numbers: 8683 (for years 2001–2017), 8660 (2018), 7065 to 7081 (1977–1993), and 3657, 3780, 3948, 4070, 4398, 4401, and 4577 (1994–2000). These datasets are currently 'decatalogued' but are available on request via the UKDS. First, you need to be a registered user of the UKDS: see https://ukdataservice.ac.uk/help/registration/registration-login-faqs/. Once registered, apply to the UKDS to access the datasets, citing the persistent identifier in the UKDS catalogue (https://beta.ukdataservice.ac.uk/datacatalogue/doi/?id=8683#!#1) and the Study Numbers for the years required. Requests will be passed on to the ONS to authorise. Once authorisation is granted, the UKDS will allow access to the datasets requested.

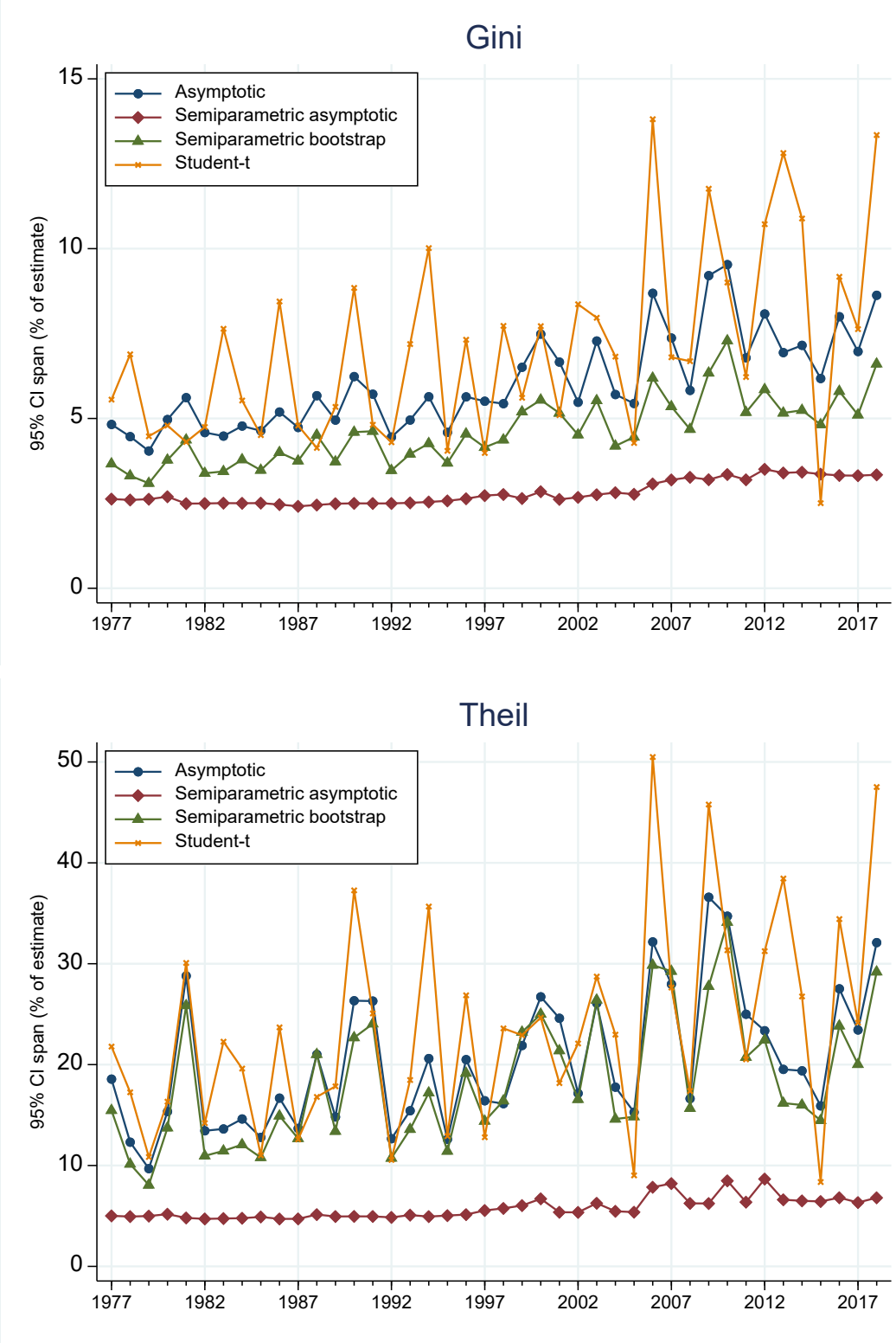
Stata do-file code used to produce the material reported in this paper is available on request from the authors.

Figure 1. UK income inequality, 1977–2018, by inequality index



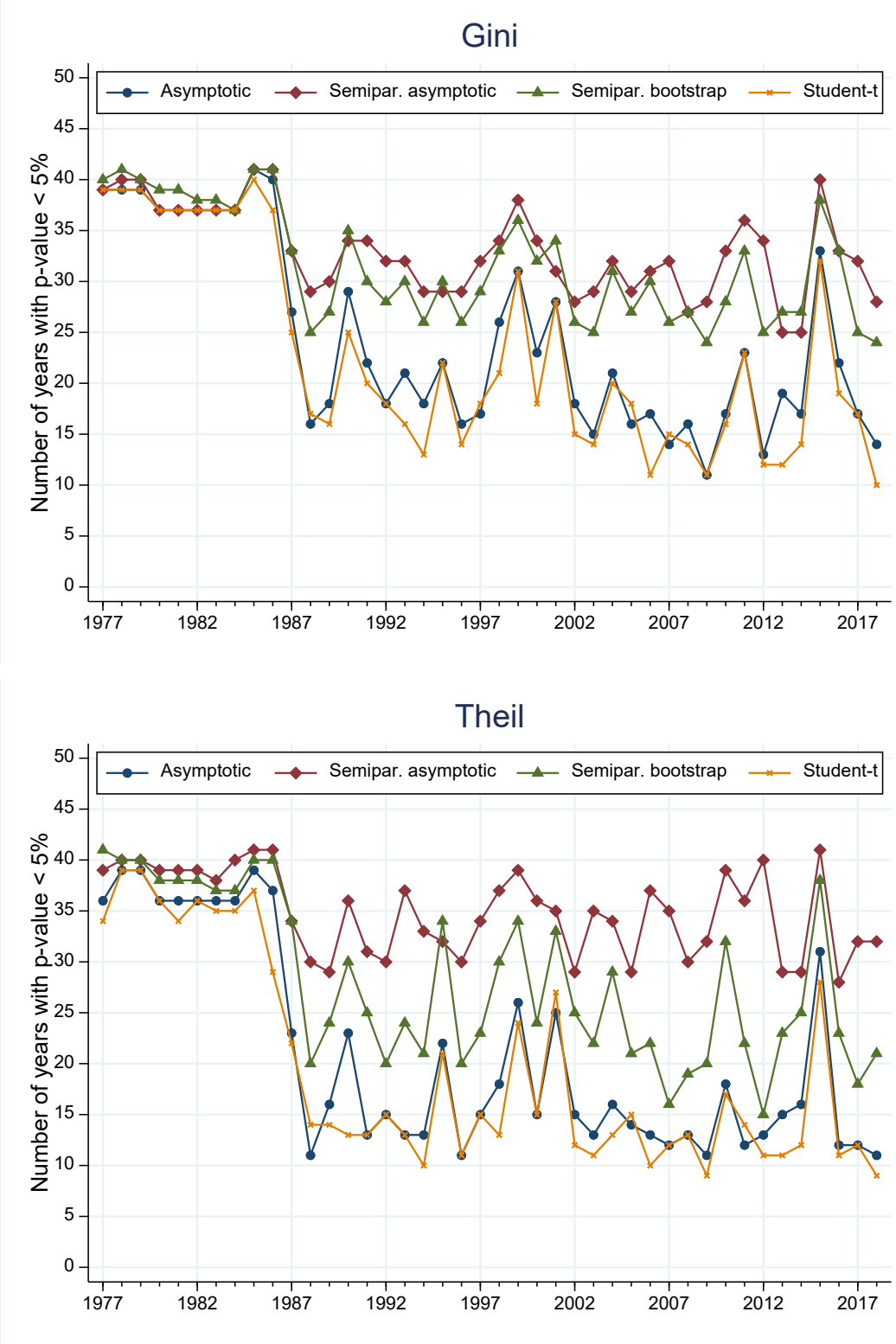
Notes. Authors' calculations based on ONS ETB data using the conventional asymptotic approach. The data are described in more detail in Section 3.

Figure 2. 95% confidence intervals (as % of estimate) by inequality index and approach



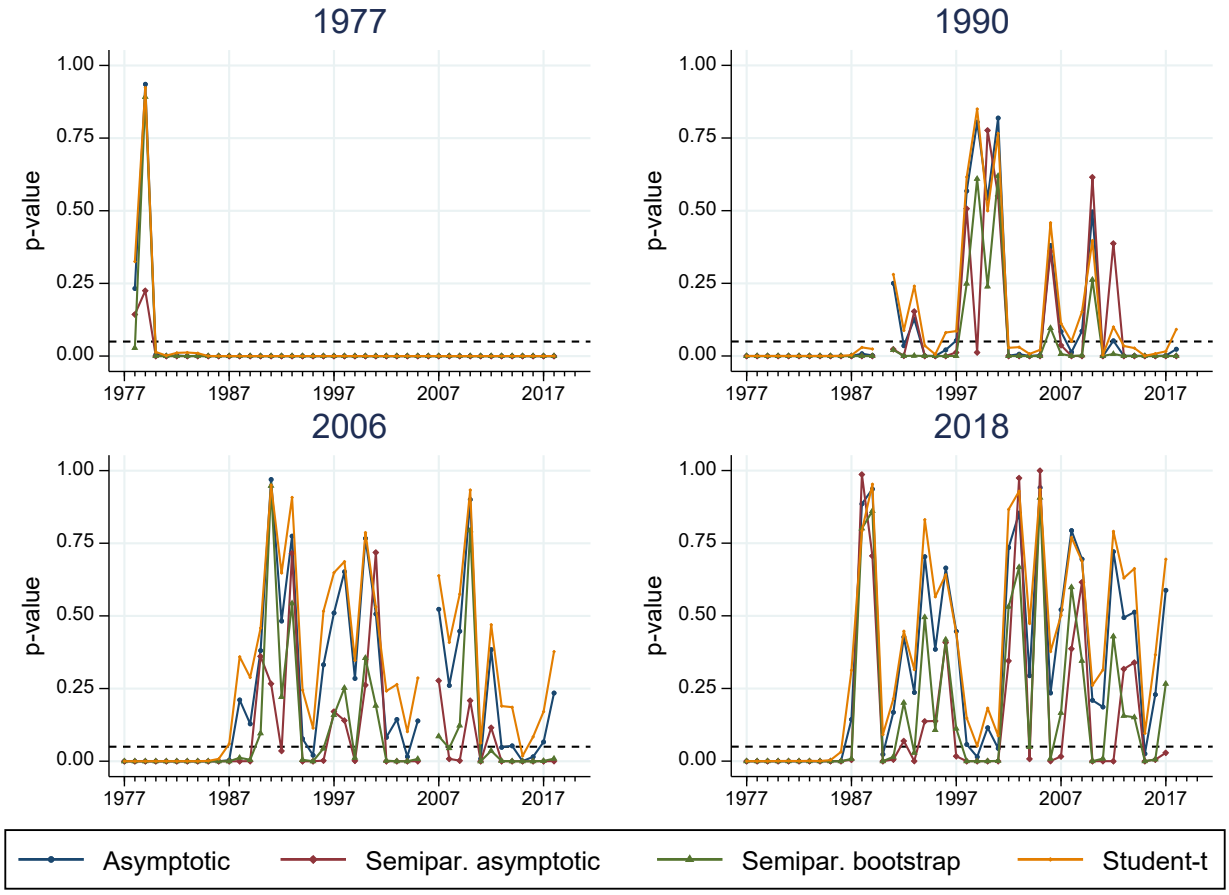
Notes. As for Figure 1.

Figure 3. Count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by inequality index and approach



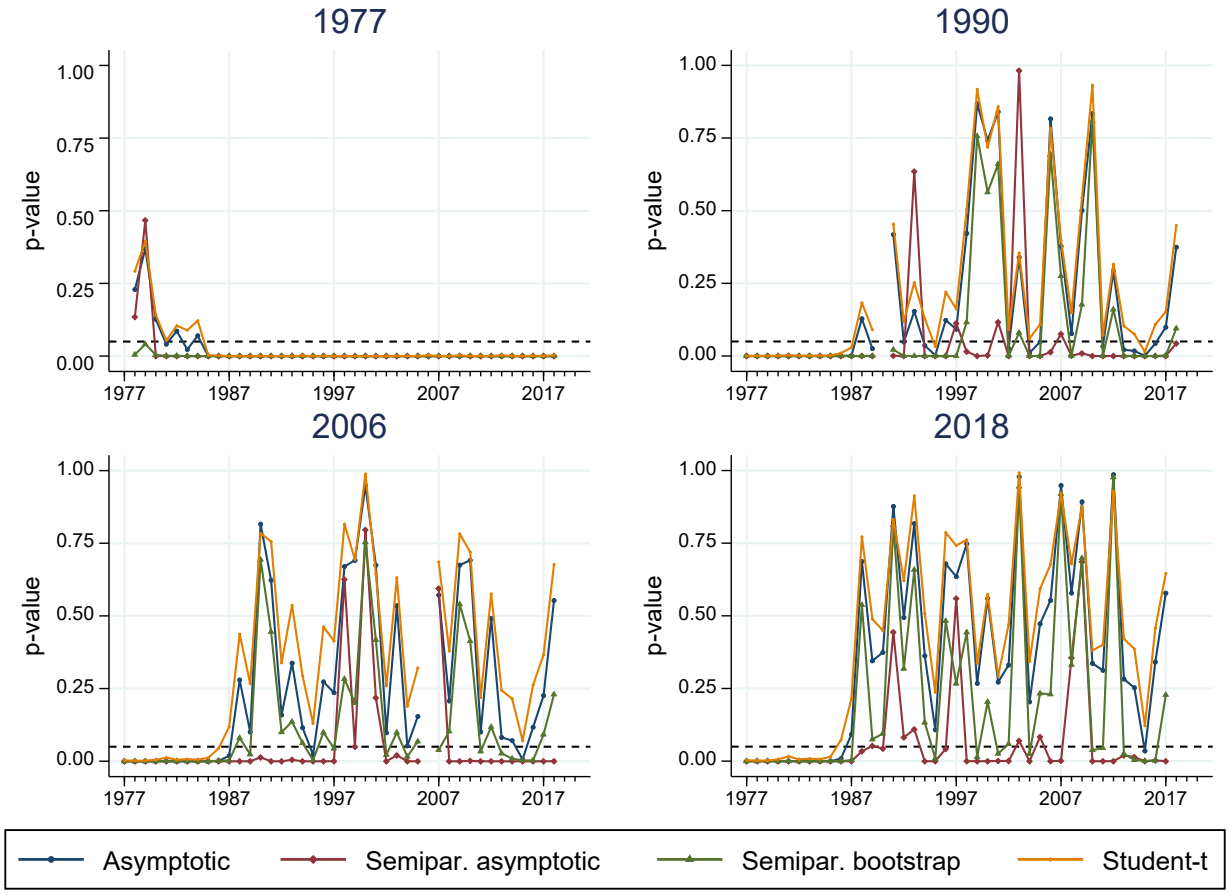
Notes. As for Figure 1. The maximum count is 41.

Figure 4. Gini index: *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach



Notes. As for Figure 1.

Figure 5. Theil index: *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach



Notes. As for Figure 1.

Table 1. 95% confidence intervals (as % of estimate), 42-year averages, by inequality index and approach

	Asymptotic	Semiparametric asymptotic	Semiparametric bootstrap	Student-t (8 groups)
Gini	6.1	2.8	4.6	7.1
MLD = GE(0)	13.1	6.6	10.8	15.2
Theil = GE(1)	20.4	5.8	18.4	23.6
Top 10% share	7.6	2.4	5.8	8.9
Top 1% share	30.0	4.6	30.5	34.4
GE(–1)	28.9	27.2	45.9	34.8
GE(2)	46.6	10.5	60.5	54.6
$p90/p10$	6.9	6.0	6.0	7.9

Notes. As for Figure 1. Bootstrap confidence intervals calculated as per Davidson and Flachaire (2007, eqn. 17).

Appendix A. Sample numbers

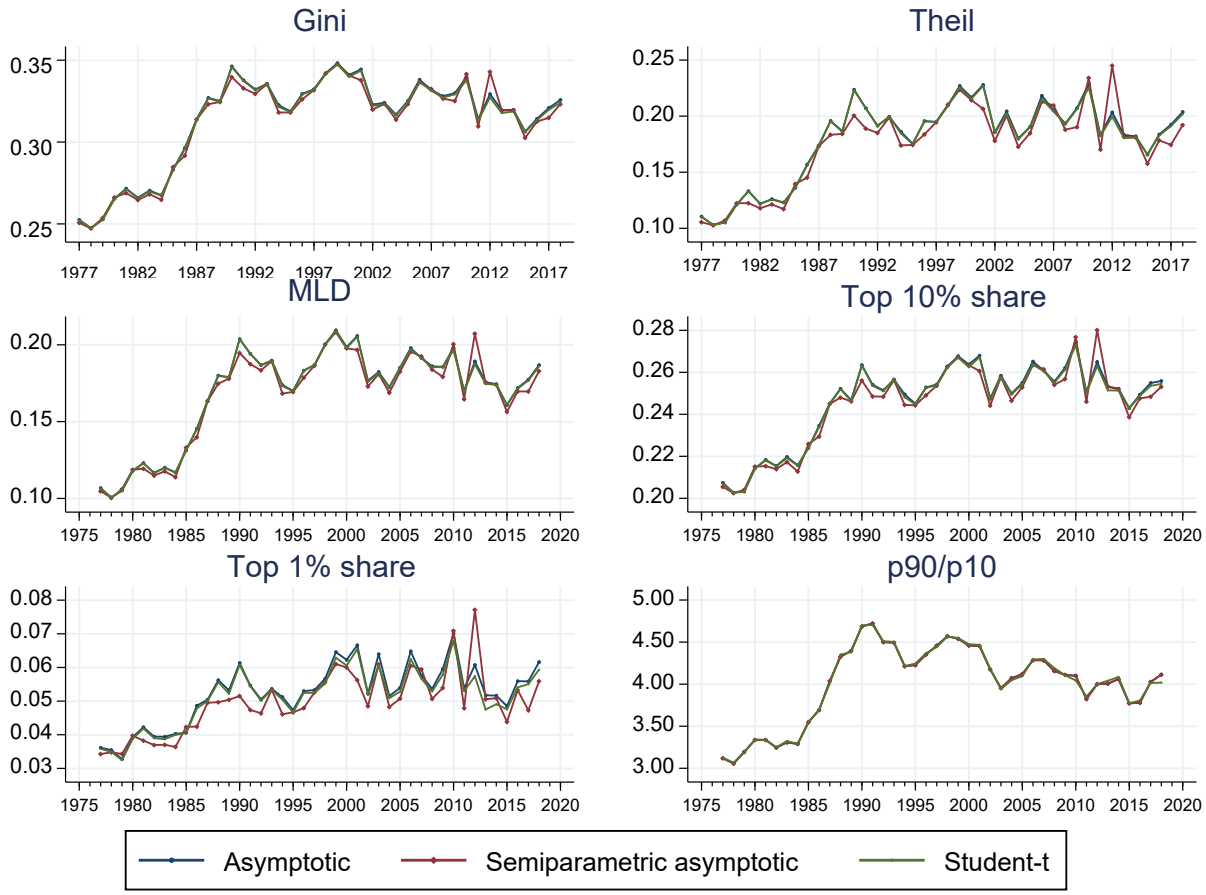
Table A1. Sample numbers before and after dropping extreme outliers, by year

Year	Number of households in LCFS	Negative or zero income	Income between 0 and 1	Top-income outlier	(2) + (3) + (4) as % of (1)	Number of households in analysis sample	Weighted number of individuals
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1977	7,193	11	3	0	0.19	7,179	19,797
1978	6,996	7	2	0	0.13	6,987	18,939
1979	6,768	8	0	0	0.12	6,760	18,202
1980	6,942	9	3	0	0.17	6,930	18,799
1981	7,520	8	1	0	0.12	7,511	20,462
1982	7,420	7	2	1	0.13	7,410	19,900
1983	6,964	9	0	0	0.13	6,955	18,412
1984	7,077	9	0	0	0.13	7,068	18,498
1985	7,007	6	0	0	0.09	7,001	18,140
1986	7,175	7	0	1	0.11	7,167	18,285
1987	7,395	1	0	0	0.01	7,394	18,723
1988	7,264	7	1	1	0.12	7,255	18,244
1989	7,406	6	1	0	0.09	7,399	18,526
1990	7,038	12	0	0	0.17	7,026	17,333
1991	7,054	6	0	0	0.09	7,048	17,063
1992	7,417	10	2	0	0.16	7,405	18,144
1993	6,975	7	2	0	0.13	6,966	17,236
1994	6,849	16	0	0	0.23	6,833	16,550
1995	6,794	11	0	0	0.16	6,783	16,532
1996	6,413	9	0	0	0.14	6,404	57,712
1997	6,409	6	0	0	0.09	6,403	58,089
1998	6,629	14	1	0	0.23	6,614	58,215
1999	7,096	26	2	0	0.39	7,068	58,449
2000	6,634	20	0	0	0.30	6,614	58,691
2001	7,466	31	1	0	0.43	7,434	58,790
2002	6,926	31	1	1	0.48	6,893	57,700
2003	7,047	22	2	1	0.35	7,022	57,945
2004	6,794	15	1	0	0.24	6,778	58,059
2005	6,778	11	1	0	0.18	6,766	58,014
2006	6,387	18	2	1	0.33	6,366	58,457
2007	6,108	15	3	1	0.31	6,089	59,300
2008	5,764	14	5	1	0.35	5,744	60,307
2009	5,575	18	0	0	0.32	5,557	60,550
2010	5,253	10	1	0	0.21	5,242	61,408
2011	5,672	13	0	0	0.23	5,659	61,335
2012	5,456	12	1	1	0.26	5,442	62,805
2013	5,089	22	0	0	0.43	5,067	63,155
2014	5,095	19	0	0	0.37	5,076	63,514
2015	4,912	20	0	0	0.41	4,892	63,704
2016	5,041	21	0	0	0.42	5,020	64,243
2017	5,407	23	0	0	0.43	5,384	64,350
2018	5,473	24	2	0	0.48	5,447	64,801
Average	6,540	13.60	0.95	0.21	0.23	6,525	41,366

Notes. Income is equivalized household disposable income (£ per week) among individuals derived from yearly UK Living Costs and Food Survey (LCFS) data. Top income outliers are those for which the highest income is at least twice as large as the next highest income. Before 1996, the LCFS weights referred only to the number of persons per household. From 1996 onwards, the weights also include 'grossing up' factors to adjust to the UK private household population (weighted totals refer to millions of individuals). The analysis reported in the main text is based on the trimmed samples with the yearly numbers shown in columns (6) and (7).

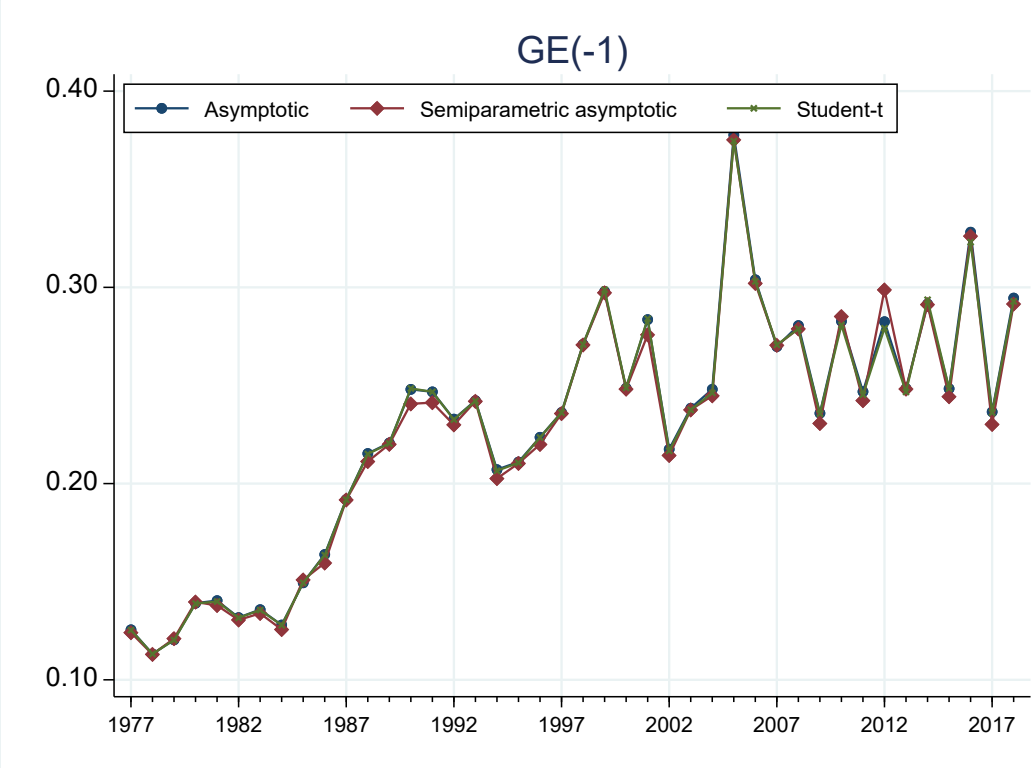
Appendix B. UK income inequality, 1977–2018, by index and approach

Figure B1. UK income inequality, 1977–2018, by index and approach



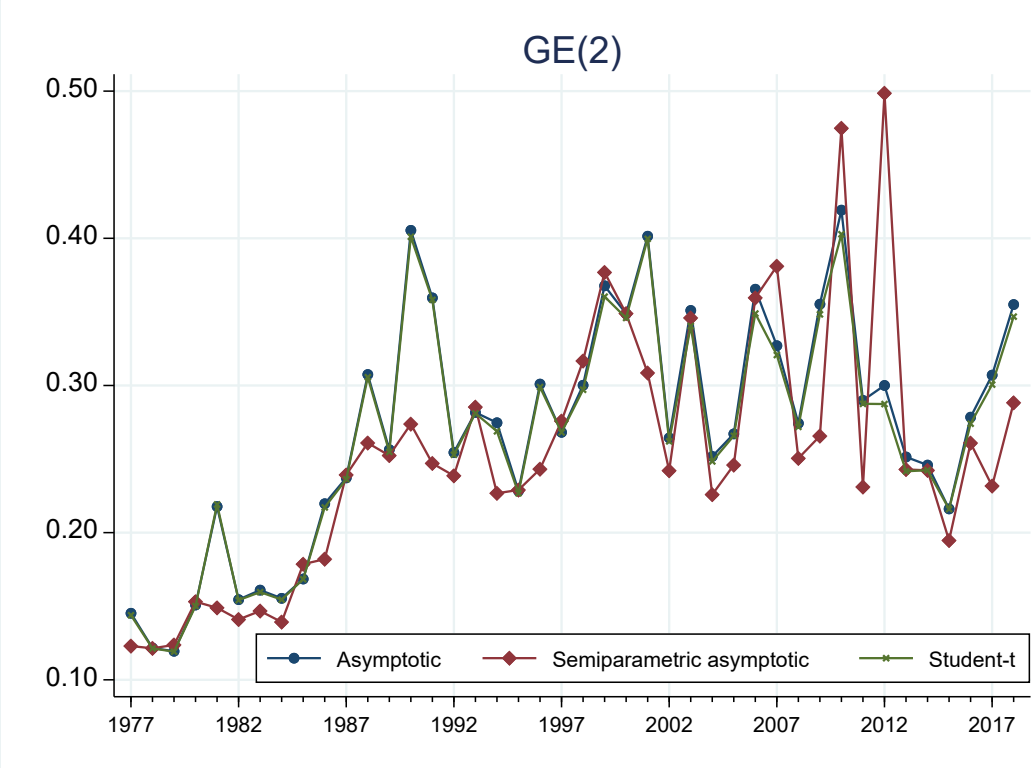
Notes. As for Figure 1. The (conventional) asymptotic series is the same as shown in Figure 1. The series for the semiparametric bootstrap is identical to that series (by construction – see main text) and thus not shown. The derivations of the semiparametric asymptotic and Student-t series (8 groups) are explained in Section 2.

Figure B2. GE(–1): estimates by year and approach



Notes. As for Figure B1.

Figure B3. GE(2): estimates by year and approach



Notes. As for Figure B1.

Appendix C. Test *p*-values for additional indices

Figure C1. MLD = GE(0): Count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by approach

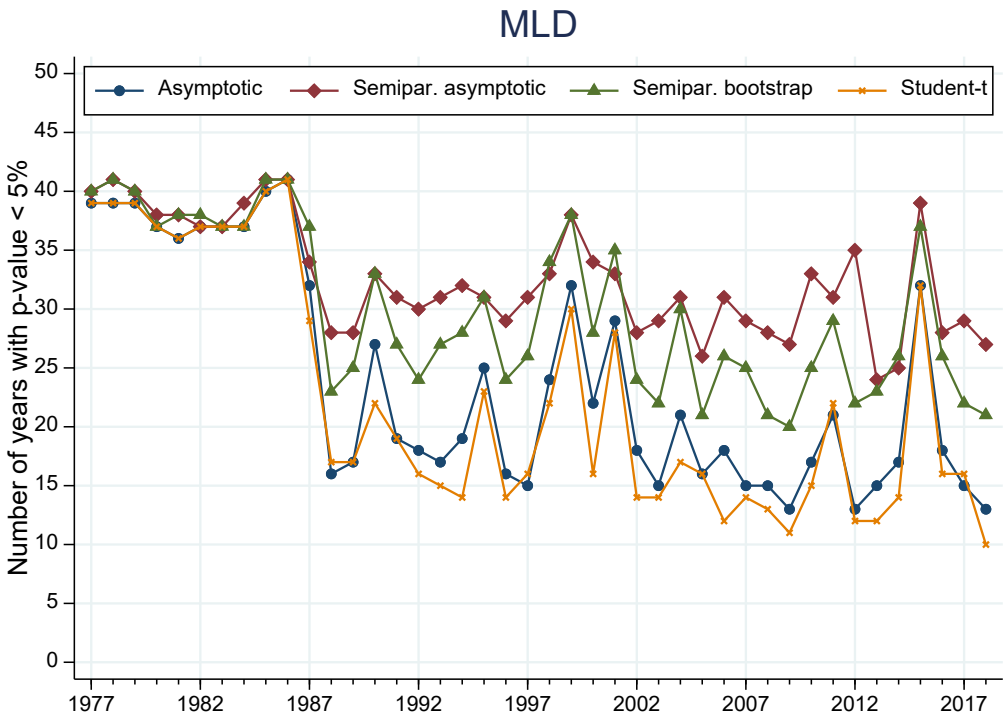


Figure C2. MLD = GE(0): *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach

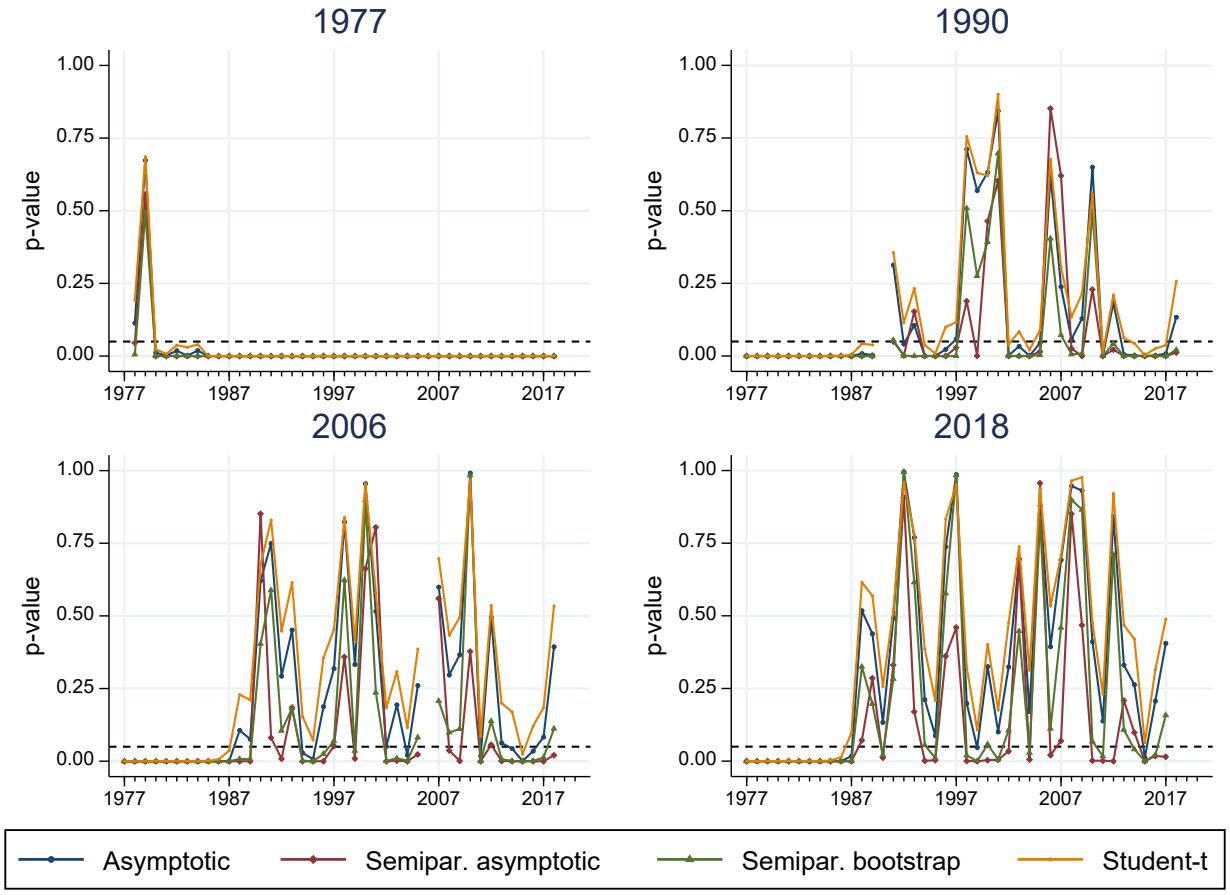


Figure C3. Share of top 10%: count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by approach

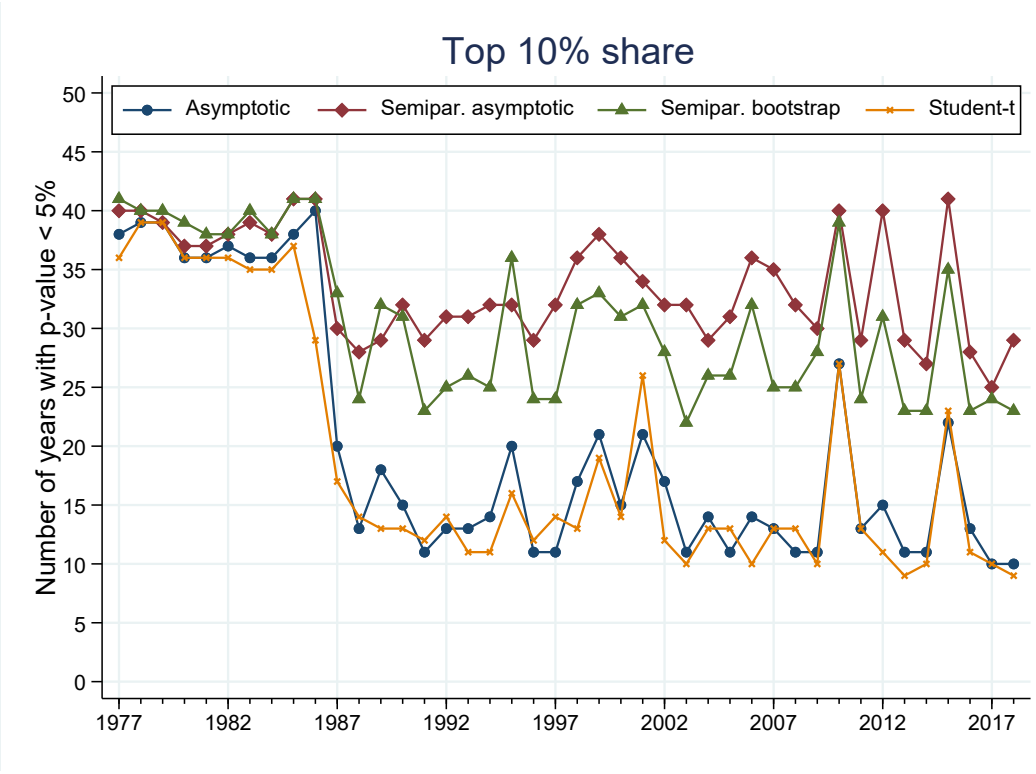


Figure C4. Share of top 10%: *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach

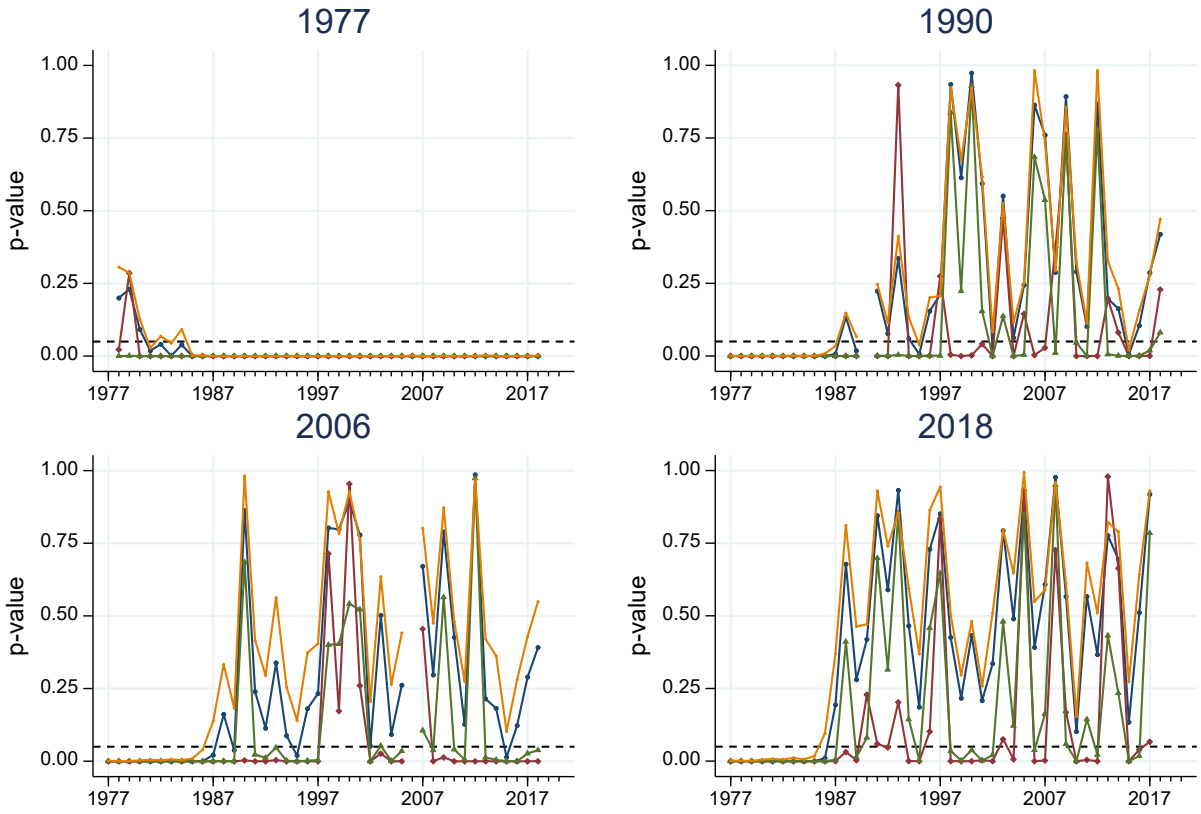


Figure C5. Share of top 1%: count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by approach

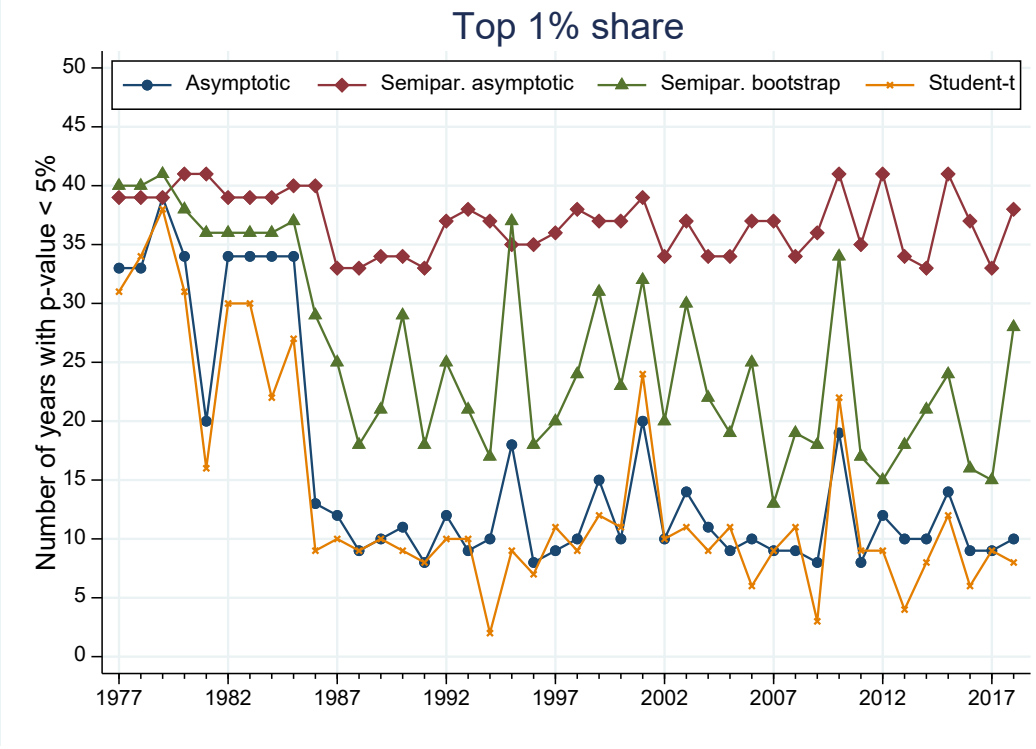


Figure C6. Share of top 1%: *p*-values for test of no inequality difference between year *A* ∈ {1977, 1990, 2006, 2018} and every other year (*B*), by approach

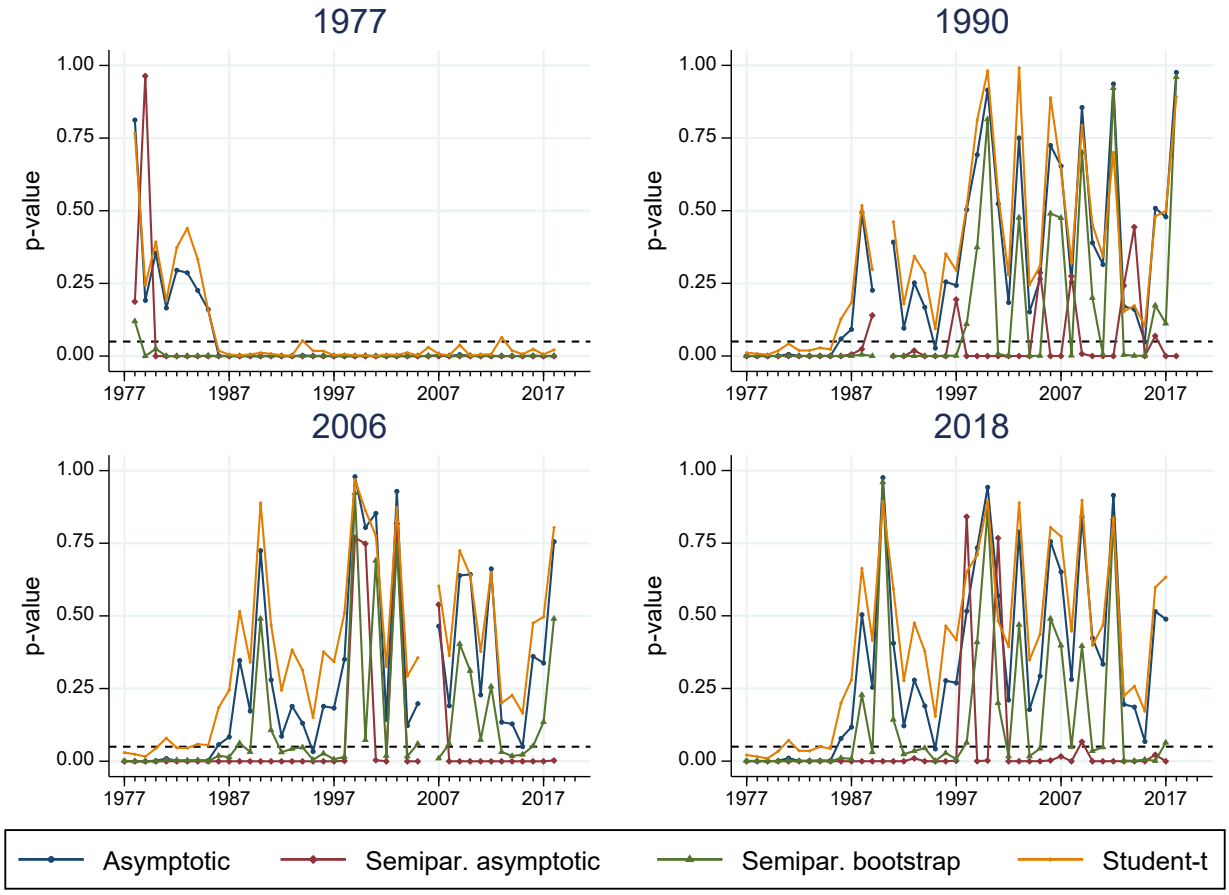


Figure C7. *p*90/*p*10: count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by approach

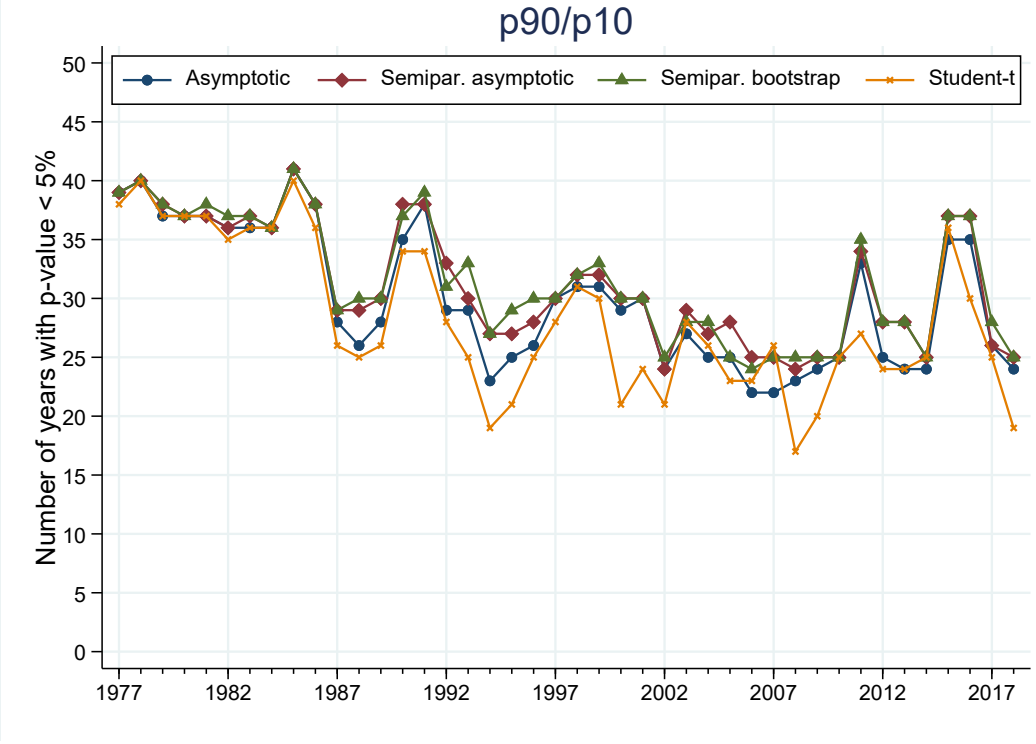


Figure C8. *p*90/*p*10: *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach

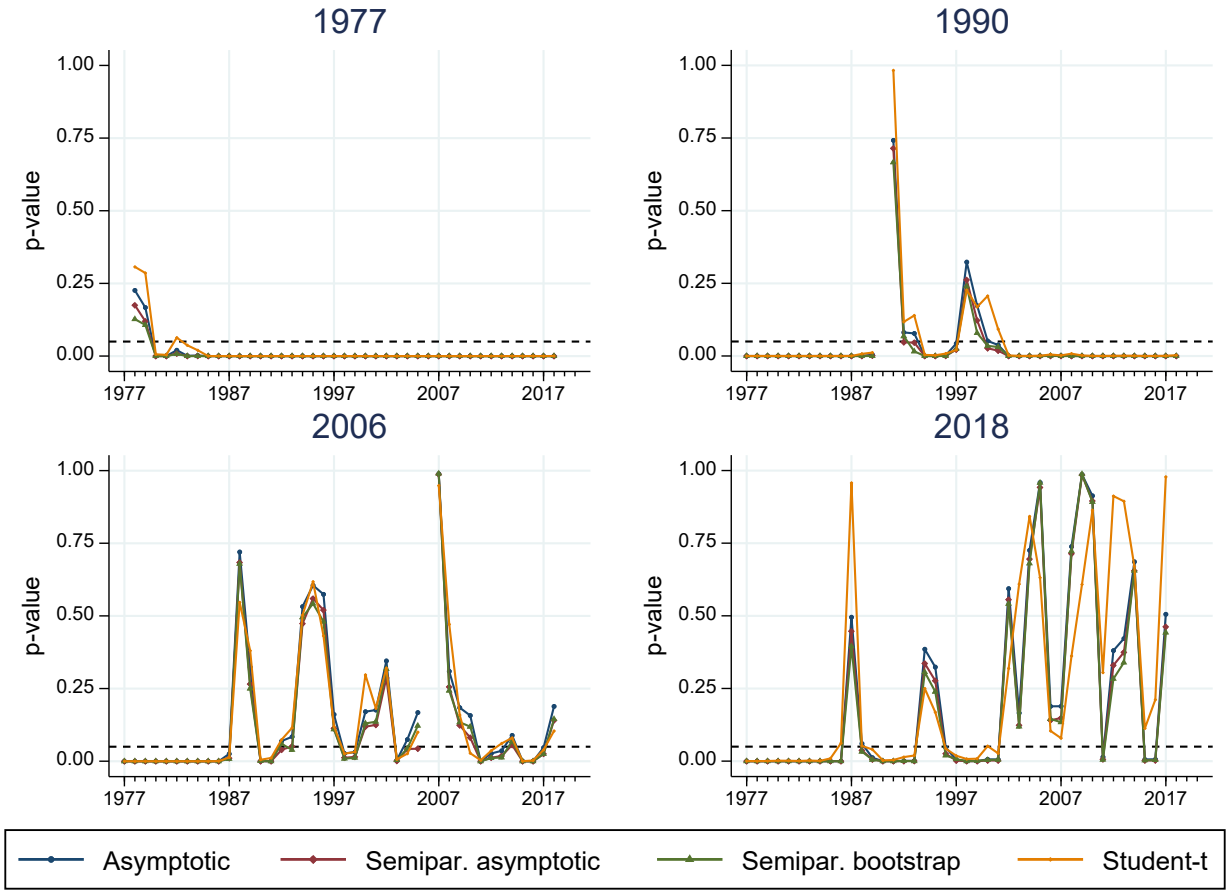


Figure C9. GE(–1): count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by approach

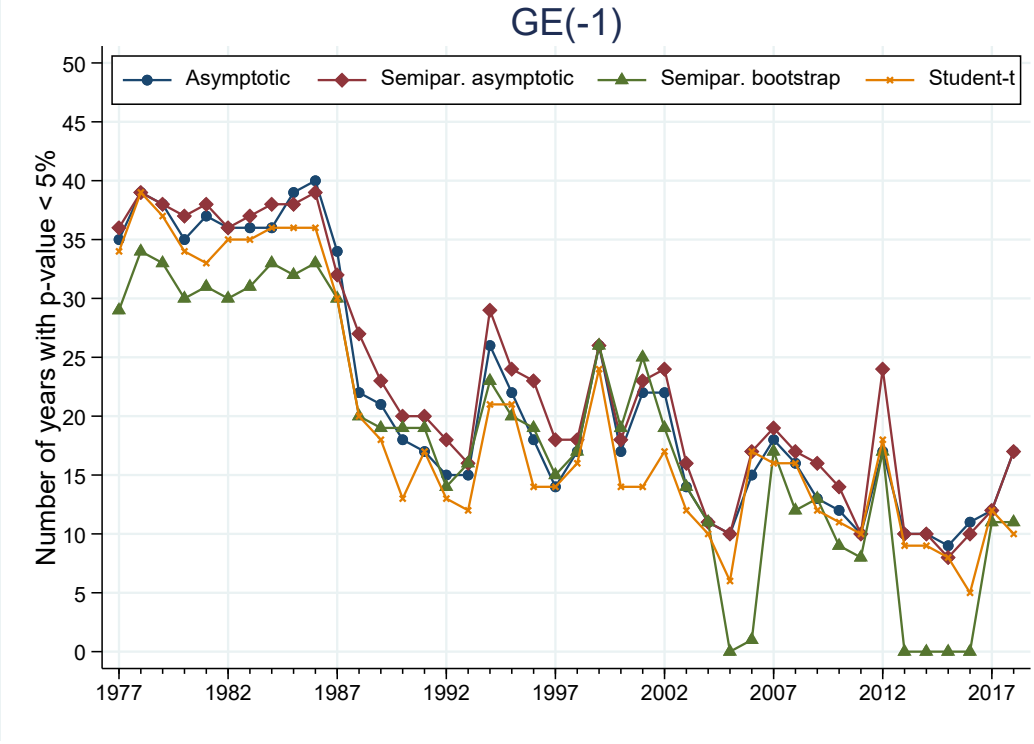


Figure C10. GE(–1): *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach

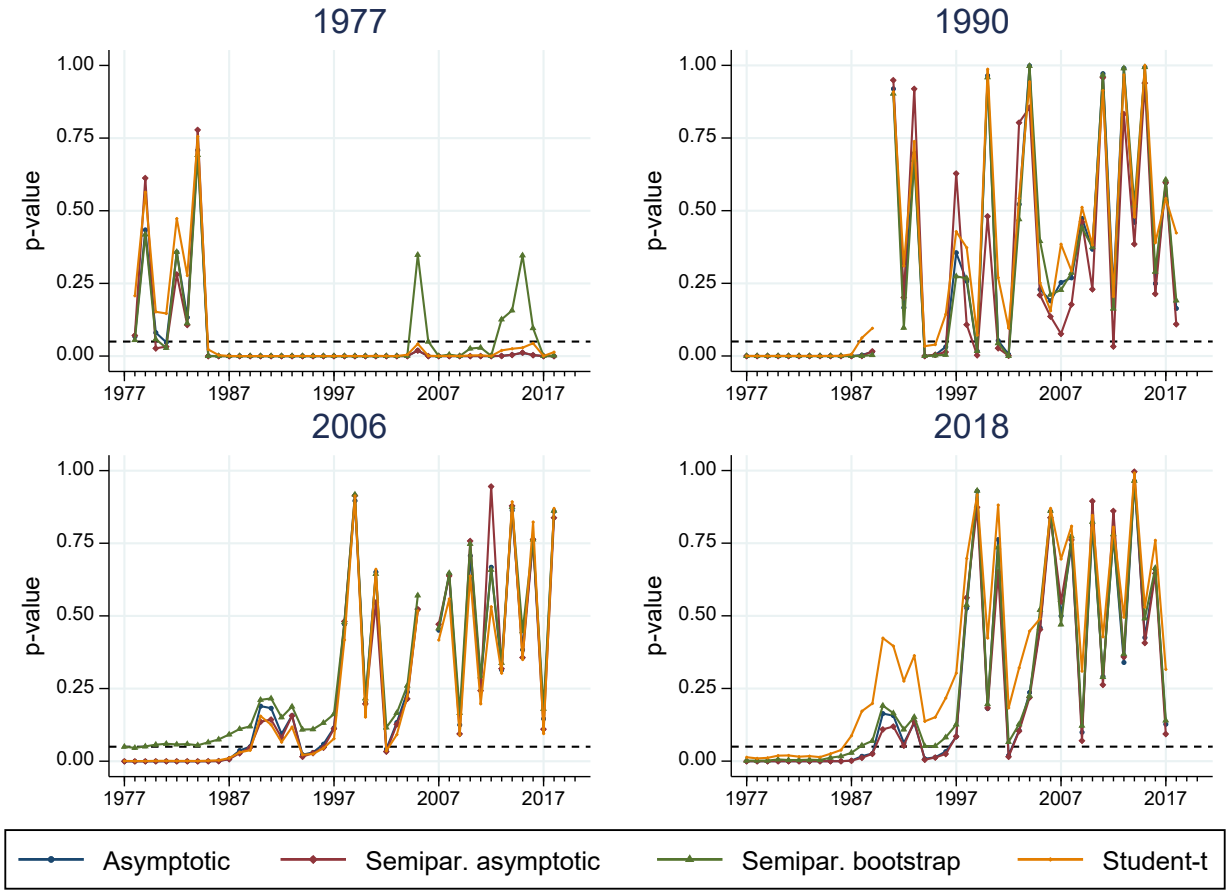


Figure C11. GE(2): count of *p*-values under 5% for test of no inequality difference between every pair of years (1977–2018), by approach

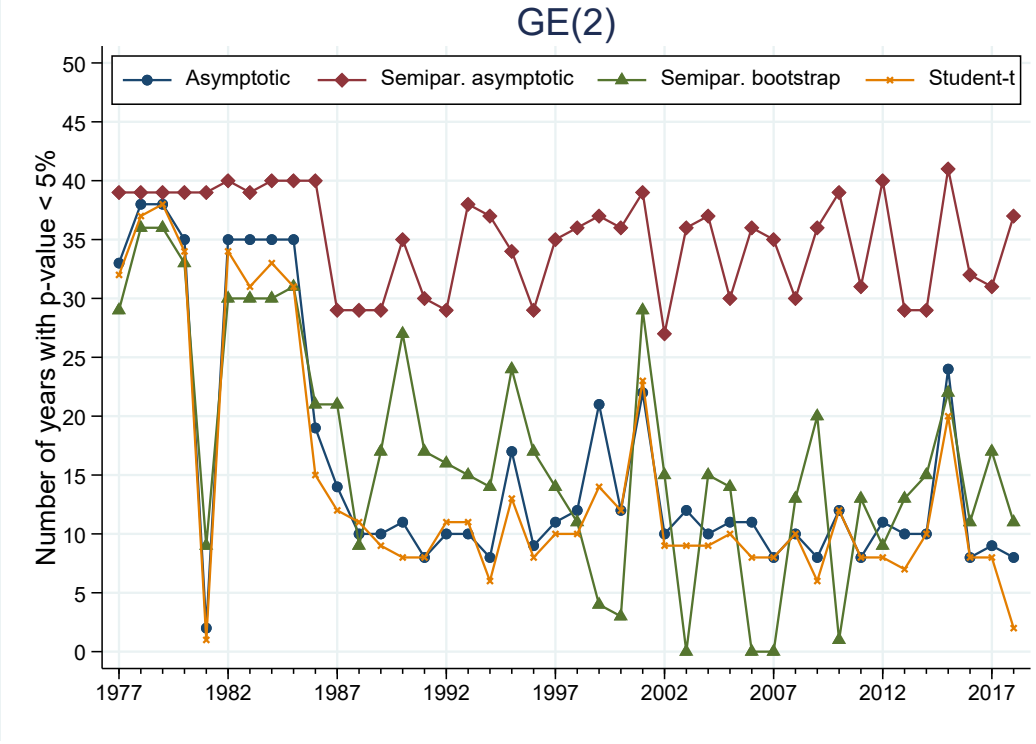
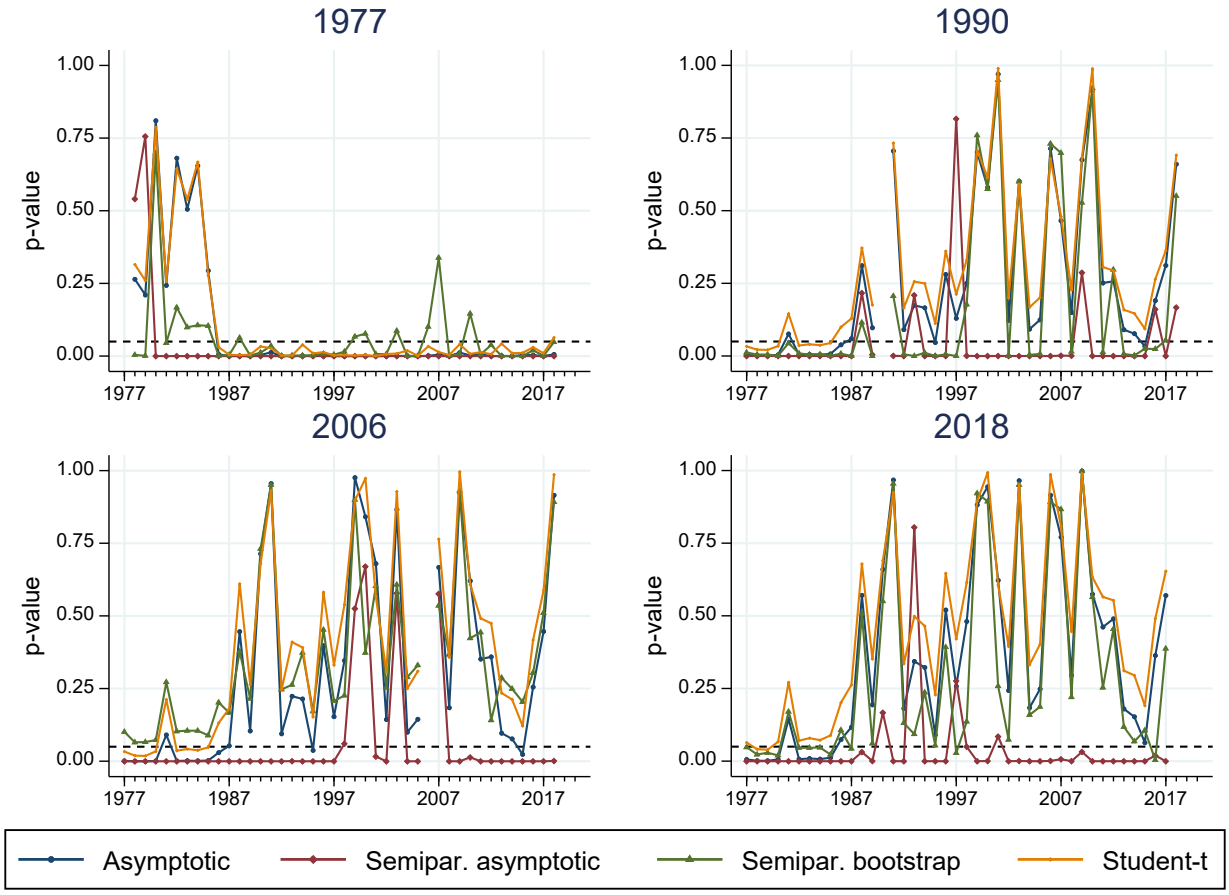


Figure C12. GE(2): *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B), by approach



Appendix D. Test *p*-values for Gini and Theil index comparisons: ptail = 1%

Figure D1. Gini index: *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B)

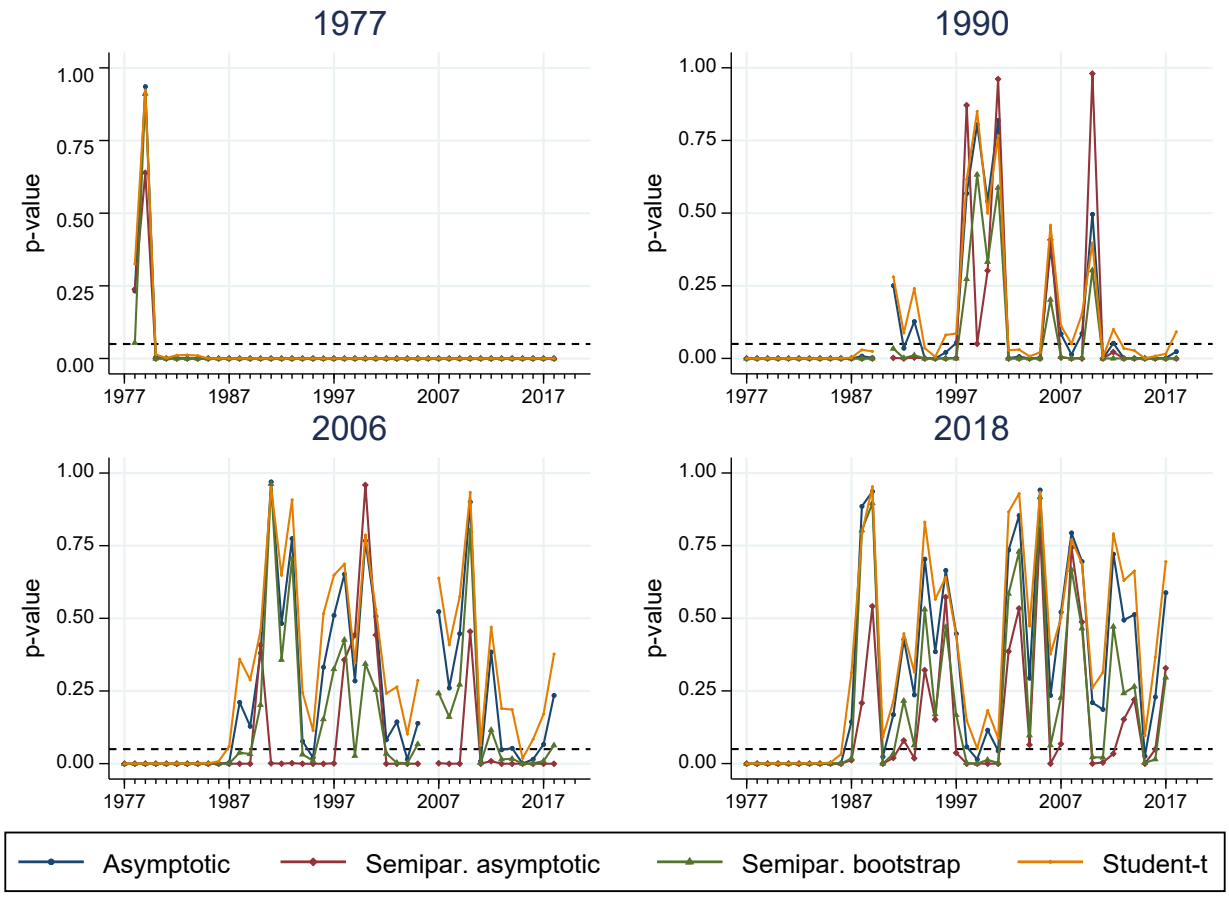


Figure D2. Theil index: *p*-values for test of no inequality difference between year $A \in \{1977, 1990, 2006, 2018\}$ and every other year (B)