ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Biewen, Martin (Ed.); Flachaire, Emmanuel (Ed.)

Book — Published Version

Econometrics and income inequality

Provided in Cooperation with:
MDPI – Multidisciplinary Digital Publishing Institute, Basel

Suggested Citation: Biewen, Martin (Ed.); Flachaire, Emmanuel (Ed.) (2018) : Econometrics and income inequality, ISBN 978-3-03897-367-6, MDPI, Basel, https://doi.org/10.3390/books978-3-03897-367-6

This Version is available at:
https://hdl.handle.net/10419/193987

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.


https://creativecommons.org/licenses/by-nc-nd/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

econometrics

Econometrics and Income Inequality

Edited by
Martin Biewen and Emmanuel Flachaire

Printed Edition of the Special Issue Published in *Econometrics*

www.mdpi.com/journal/econometrics

MDPI

Econometrics and Income Inequality

Econometrics and Income Inequality

Special Issue Editors

Martin Biewen
Emmanuel Flachaire

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade

Special Issue Editors
Martin Biewen
Universität Tübingen
Germany

Emmanuel Flachaire
Aix-Marseille Université
France

Editorial Office
MDPI
St. Alban-Anlage 66
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Econometrics* (ISSN 2225-1146) from 2017 to 2018 (available at: https://www.mdpi.com/journal/econometrics/special_issues/inequality)

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. *Journal Name* **Year**, *Article Number*, Page Range.

ISBN 978-3-03897-366-9 (Pbk)
ISBN 978-3-03897-367-6 (PDF)

Articles in this volume are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles even for commercial purposes, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications. The book taken as a whole is © 2018 MDPI, Basel, Switzerland, distributed under the terms and conditions of the Creative Commons license CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents

About the Special Issue Editors

Martin Biewen, Prof. Dr., is a professor of Econometrics at the University of Tübingen. He received his doctoral degree in Economics from the University of Heidelberg and his habilitation from the University of Mannheim. He held prior academic positions at the University of Essex, the University of Frankfurt, and the University of Mainz. His current research interest include topics in labor economics, education economics, and the distribution of income. He has published articles in journals such as the *Journal of Econometrics*, the *Journal of Applied Econometrics, The Review of Economics and Statistics*, the *Oxford Bulletin of Economics and Statistics*, and the *Journal of Labor Economics*.

Emmanuel Flachaire, Professor of Economics at Aix-Marseille University. He received his doctoral degree in Economics from Aix-Marseille University. He held prior positions at CORE, Université Catholique de Louvain, the London School of Economics, and the University Paris 1 Panthéon-Sorbonne. He has also been a visiting professor at McGill University and HEC Montréal. His current research interest include topics in econometrics, inequality and mobility measurement, and the distribution of income. He has published articles in journals such as the *Journal of Econometrics*, the *Journal of Business and Economic Statistics, Quantitative Economics, Economica*, and in the Handbook of Income Distribution. He is the coauthor of Non-Parametric Econometrics (OUP, 2010).

econometrics

MDPI

Editorial

Econometrics and Income Inequality

Martin Biewen [1,*] and Emmanuel Flachaire [2,*]

[1] School of Business and Economics, University of Tübingen, Mohlstraße 36, 72074 Tübingen, Germany
[2] Aix-Marseille Université, AMSE, EHESS and CNRS, 5 bd Maurice Bourdet, CS 50498, 13205 Marseille CEDEX 01, France
* Correspondence: martin.biewen@uni-tuebingen.de (M.B.); emmanuel.flachaire@univ-amu.fr (E.F.)

Received: 11 October 2018; Accepted: 11 October 2018; Published: 15 October 2018

It is well-known that, after decades of non-interest in the theme, economics has experienced a proper surge in inequality research in recent years. In addition to numerous research articles in scientific journals, this has brought to us publications such as Thomas Piketty's *Capital in the 21st Century* and Tony Atkinson's *Inequality: What can be done?*, which are highly visible in the public domain. As in many other fields of the discipline, the analysis of inequality poses both interesting theoretical and statistical problems. The present Special Issue of *Econometrics* is a collection of 15 excellent papers that address some of these issues.

The articles range from purely methodological contributions on the measurement of inequality to questions of statistical inference and to substantive empirical contributions in various fields of empirical inequality research. Starting with contributions to the pure measurement of inequality, **"From the Classical Gini Index of Income Inequality to a New Zenga-Type Relative Measure of Risk: A Modeller's Perspective" by Francesca Greselin and Ričardas Zitikis** provides a fascinating theoretical treatment that unifies theoretical inequality indices and various measures of risk. Further contributions to the pure theory of inequality measurement come from **"On the Decomposition of the Esteban and Ray Index by Income Sources" by Elena Bárcena-Martín and Jacques Silber** and from **"Decomposing the Bonferroni Inequality Index by Subgroups: Shapley Value and Balance of Inequality" by Giovanni M. Giorgi and Alessio Guandalini**. Both contributions consider decomposability properties of inequality and polarization measurement procedures, an important topic with a long tradition in the literature. The paper **"Inequality and Poverty When Effort Matters" by Martin Ravaillion** tackles another long-standing question in the inequality literature, namely the incorporation of income differences due to differential effort. One of the many highlights of this Special Issue is the paper **"Decomposing Wage Distributions Using Recentered Influence Function Regressions" by Sergio P. Firpo, Nicole M. Fortin, and Thomas Lemieux**. This paper works out the so-called "hybrid" Re-centered Influence Function (RIF-) regression decomposition, which enjoys enormous popularity among applied researchers. The paper has already received a lot of citations as a working paper and we hope that it continues to do so as a part of this Special Issue.

An important part of the Special Issue deals with problems of statistical inference. As in other fields, statistical inference is a necessity when carrying out inequality analyses. Unfortunately, due to their usually complex and nonstandard nature, working out statistical inference procedures for methods of inequality measurement is often challenging. The article **"Statistical Inference on the Canadian Middle Class" by Russell Davidson** develops distribution-free methods for the measurement of middle-income shares and applies them in order to measure the size of the Canadian Middle Class. The paper **"Parametric Inference for Index Functionals" by Stéphane Guerrier, Samuel Orso, and Maria-Pia Victoria-Feser** proposes an inference procedure for inequality index functionals, based on a Generalized Method of Moment estimator for parametric data generating mechanisms, and evaluates its finite sample performance. In their article **"A Hybrid MCMC Sampler**

for Unconditional Quantile Based on Influence Function", El Moctar Laghlal and Abdoul Aziz Junior Ndoye develop a Bayesian estimation method for the unconditional RIF-regression that has a superior performance in the presence of heavy-tailed distributions. **"Using the GB2 Income Distribution" by Duangkamon Chotikapanich, William E. Griffiths, Gholamreza Hajargasht, Wasana Karunarathne, and D. S. Prasada Rao** is a highly useful survey article on the estimation and inference problems for the generalized beta distribution of the second kind (GB2). This distribution enjoys high levels of popularity among inequality researchers due to its flexibility and good fit in empirical applications.

The Special Issue contains two interesting papers on the popular topic of measuring top incomes. Such top incomes have been known to be on the rise in many advanced countries, so much so that adequate statistical estimation and inference procedures are of high interest. **"Top Incomes, Heavy Tails, and Rank-Size Regressions" by Christian Schluter** studies rank-size regressions of tail exponents. This method still represents the most popular estimation technique of this kind in applied studies in economics. The author shows, both theoretically and empirically, based on UK data, that the method may lead to size distortions that undermine statistical inference in practice. Another study focusing on the very top of the distribution is **"Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data" by Vladimir Hlasny and Paolo Verme**. Based on data for European countries, the paper provides an analysis of reweighting and replacing methods to correct inequality measure for top-income biases generated by data issues such as unit or item nonresponse. The authors show that income inequality may be substantially underestimated if no correction techniques are used.

Finally, this Special Issue contains a number of substantive empirical studies in a wide range of relevant settings. "**Polarization and Rising Wage Inequality: Comparing the U.S. and Germany" by Dirk Antonczyk, Thomas DeLeire, and Bernd Fitzenberger** provides an in-depth analysis of the differences in wage polarization trends in the US and Germany. The authors find that their evidence is consistent with a technology-driven polarization of the labor market, but that there are important country-specific factors, such as cohort effects. "**The Wall's Impact in the Occupied West Bank: A Bayesian Approach to Poverty Dynamics Using Repeated Cross-Sections" by Tareq Sadeq and Michel Lubrano** applies a sophisticated Bayesian modelling strategy to investigate the effect of the wall in occupied West Bank on poverty persistence for the affected population. The paper "**Income Inequality, Cohesiveness, and Commonality in the Euro Area: A Semi-Parametric Boundary-Free Analysis" by Gordon Anderson, Maria Grazia Pittau, Roberto Zelli, and Jasmin Thomas** studies the question of income cohesiveness in the Euro area using an approach based on mixture distributions. The authors conclude that the Eurozone is best described by a four-class, increasingly unequal polarizing structure with income growth in all four classes. Finally, in "**Foreign Workers and the Wage Distribution: What Does the Influence Function Reveal?", Chung Choe and Philippe Van Kerm** study the impact of Foreign Workers on Wage Distribution in Luxembourg. The case of Luxembourg is particularly interesting because of its extremely high share of foreign workers. The paper also makes a methodological contribution related to the RIF-methodology, thus nicely connecting to other papers in the Special Issue.

We are very grateful to all contributing authors, who have made considerable efforts to meet the standards of the journal. We believe that this Special Issue has been very successful in attracting topical and high-quality contributions, many from very well-known scholars in the field, proving that open access-publishing is a realistic option for our discipline. We also would like to thank the numerous reviewers who have greatly contributed to the quality of the published papers. Last but not least, we thank the editor-in-chief, Marc Paolella, and the team of assistant editors, Vera Zhu, Lu Liao, and Michele Cardani, for their excellent support.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

From the Classical Gini Index of Income Inequality to a New Zenga-Type Relative Measure of Risk: A Modeller's Perspective

Francesca Greselin [1] and Ričardas Zitikis [2,*]

[1] Dipartimento di Statistica e Metodi Quantitativi, Università di Milano–Bicocca, Milan 20126, Italy; francesca.greselin@unimib.it

[2] School of Mathematical and Statistical Sciences, Western University, London, ON N6A 5B7, Canada

* Correspondence: zitikis@stats.uwo.ca; Tel.: +1-519-432-7370

Received: 28 August 2017; Accepted: 22 January 2018; Published: 25 January 2018

Abstract: The underlying idea behind the construction of indices of economic inequality is based on measuring deviations of various portions of low incomes from certain references or benchmarks, which could be point measures like the population mean or median, or curves like the hypotenuse of the right triangle into which every Lorenz curve falls. In this paper, we argue that, by appropriately choosing population-based references (called societal references) and distributions of personal positions (called gambles, which are random), we can meaningfully unify classical and contemporary indices of economic inequality, and various measures of risk. To illustrate the herein proposed approach, we put forward and explore a risk measure that takes into account the relativity of large risks with respect to small ones.

Keywords: economic inequality; reference measure; personal gamble; inequality index; risk measure; relativity

JEL Classification: D63; D81; C46

1. Introduction

The Gini mean difference and its normalized version, known as the Gini index, have aided decision makers since their introduction by Corrado Gini more than a hundred years ago (Gini 1912, 1914, 1921); see also (Giorgi 1990, 1993, 1921); Ceriani and Verme 2012; and references therein). In particular, the Gini index has been widely used by economists and sociologists to measure economic inequality. Measures inspired by the index have been employed to assess the equality of opportunity (e.g., Weymark 2003; Kovacevic 2010; Roemer 2013) and estimate income mobility (e.g., Shorrocks 1978). Policymakers have used the Gini index in quantitative development policy analysis (e.g., Sadoulet and de Janvry 1995) and in particular for assessing the impact of carbon tax on income distribution (e.g., Oladosu and Rose 2007). The index has been employed for analysing inequality in the use of natural resources (e.g., Thompson 1976) and for developing informed policies for sustainable consumption and social justice (e.g., Druckman and Jackson 2008). Various extensions and generalizations of the index have been used to evaluate social welfare programs (e.g., Duclos 2000; Kenworthy and Pontusson 2005; Korpi and Palme 1998; Ostry et al. 2014) and to improve the knowledge of tax-base and tax-rate effects, as well as of temporal repercussions of distinct patterns of taxation and public finance on the society (e.g., Pfähler 1990; Slemrod 1992; Yitzhaki 1994; Van De Ven et al. 2001). Furthermore, Denneberg (1990) has advocated the use of the Gini mean difference as a safety loading for insurance premiums, with recent developments in the area by Furman and Zitikis (2017), and Furman et al. (2017).

Naturally, a multitude of interpretations, mathematical expressions, and generalizations of the index have manifested in the literature. As noted by Ceriani and Verme (2012), Corrado Gini

himself proposed no less than thirteen formulations of his original index. Yitzhaki (1998, 2003), and Yitzhaki and Schechtman (2013) have discussed a great variety of interpretations of the Gini index. Many monographs and handbooks have been written on measuring economic inequality, where the Gini index and its various extensions and generalizations have played prominent roles: Amiel and Cowell (1999), Atkinson and Bourguignon (2000, 2015), Atkinson and Piketty (2007), Banerjee and Duflo (2011), Champernowne and Cowell (1998), Cowell (2011), Kakwani (1980a), Lambert (2001), Nygård and Sandström (1981), Ostry et al. (2014), Piketty (2014), Sen (1997), Silber (1999), Yitzhaki and Schechtman (2013), to name a few.

Given the diversity, one naturally wonders if there is one underlying thread that unifies all these indices. The population Lorenz function, as well as its various distances from the hypotenuse of the right triangle into which every Lorenz function falls, have traditionally provided such a thread. However, recent developments in the area of measuring economic inequality (e.g., Palma 2006; Zenga 2007; Greselin 2014; Gastwirth 2014; Kośny and Yalonetzky 2015) have highlighted the need for departure from the population mean, which is inherent in the definition of the Lorenz function as the benchmark, or reference point, for measuring economic inequality. The newly developed indices have deviated from the aforementioned unifying thread and thus initiated a fresh rethinking of the problem of measuring inequality.

Bennett and Zitikis (2015) ventured in this direction by suggesting a way to bridge the Harsanyi (1953) and Rawls (1971) conceptual frameworks via a spectrum of random societal positions. In this paper, we make a further step by developing a mathematically rigorous approach for unifying and interpreting numerous classical and contemporary indices of economic inequality, as well as those of risk. Briefly, the approach we have developed is based on appropriately chosen

1. societal references such as the population mean, median, or some population distribution-tail based measures, and
2. distributions of random personal positions, or gambles, that determine person's position on a certain population-based function.

Certainly, the literature is permeated by discussions related to points 1 and 2. Relativity issues have been explored in virtually every work, empirical and theoretical, due to the simple reason that they are a fact of life (e.g., Amiel and Cowell 1997, 1999). Naturally, fundamental measures of inequality, such as the Lorenz function, are also relative quantities, e.g., with respect to the population mean income. For discussions of various choices of reference measures and inherent relativity issues, we refer to, e.g., Sen (1983, 1998); Amiel and Cowell (1997, 1999); Zoli (1999, 2012); Duclos (2000); and references therein. To illustrate the point, which will become pivotal in our following deliberations, we recall a remark by Claudio Zoli, who wrote:

> In particular, Amiel and Cowell (1997, 1999) find evidence that "the appropriate inequality equivalence concept depends on the income levels at which inequality comparisons are made." Moreover, they show that, as income increases, the equivalence concept moves from the relative attitude to the absolute one, a pattern consistent with our intuition (Zoli 2012, p. 4).

This remark leads us towards the use of what we call relative-value functions, which, as we shall see later in this paper, offer a flexible way for coupling fundamental measures of economic inequality, or risk, with appropriate reference points, such as the mean (e.g., Equation (7) below). This is very much in the spirit of Definition 3 by Cowell (2003). We shall come back to the latter work in the second half of Section 4.

Finally, we note that the construction of distributions that govern personal random positions on population-based functions have been explored within the dual or rank-dependent utility theory (Quiggin 1982, 1993; Schmeidler 1986, 1989; Yaari 1987), other non-expected utility theories (e.g., Puppe 1991; Machina 1987, 2008; and references therein), distortion

risk measures (Wang 1995, 1998), and weighted insurance premium calculation principles (Furman and Zitikis 2008, 2009).

The rest of the paper is organized as follows. In Section 2, we revisit the classical Gini index and, in particular, express it in two ways—absolute and relative—within the framework of expected utility theory using appropriately chosen gambles and societal functions (i.e., Lorenz and Bonferroni). In Section 3, we step aside from the Lorenz and Bonferroni functions and, crucially for this paper, suggest using a (financial) average value at risk as the underlying societal function on which various personal gambles are played; however, the reference measure remains the mean income μ_F. In Section 4, we depart from the latter reference and introduce a general index that accommodates any population-based reference measure. In Sections 5 and 6, we show how the Donaldson-Weymark-Kakwani index and the Wang (or distortion) risk measure, as well as their generalizations, fall into the expected utility framework with collective mean-income references and appropriately chosen personal gambles. In Section 7, we argue for the need for incorporating personal preferences into reference measures, and, in Section 8, we demonstrate how this yields a new measure of risk that takes into account the relativity of large risks with respect to smaller ones. Section 9 finishes the paper with a general index of inequality and risk.

2. The Classical Gini Index Revisited

Naturally, we begin our arguments with the classical index of Gini (1914). Let X be a random variable (think of 'income') with non-negatively supported cdf $F(x)$ and finite mean $\mu_F = \mathbf{E}[X]$. The Gini index, which we denote by G_F, is usually interpreted as twice the area between the actual population Lorenz function (Lorenz 1905; Pietra 1915; Gastwirth 1971)

$$\mathrm{L}_F(p) = \frac{1}{\mu_F}\int_0^p F^{-1}(t)dt$$

and the egalitarian Lorenz function $\mathrm{L}_E(p) = p$, $0 \le p \le 1$, which is the hypotenuse of the right triangle that we have alluded to in the abstract. For parametric expressions of $\mathrm{L}_F(p)$, we refer to Gastwirth (1971), Kakwani and Podder (1973), as well as to more recent works of Sarabia (2008), Sarabia et al. (2010), and references therein. Hence, the Gini index is

$$\begin{aligned}\mathrm{G}_F &= 2\int_0^1 \big(\mathrm{L}_E(p) - \mathrm{L}_F(p)\big)\,dp \\ &= 2\mathbf{E}\big[\mathrm{L}_E(\pi) - \mathrm{L}_F(\pi)\big],\end{aligned} \tag{1}$$

where the gamble π follows the uniform density on the unit interval $[0,1]$, that is, $f(p) = 1$ for all $p \in [0,1]$. Intuitively, π governs person's position in terms of income percentiles, and we thus call it *personal* gamble. In other words, barring the normalizing constant 2, the Gini index G_F is the expected *absolute*-deviation of person's position π on the actual Lorenz function $\mathrm{L}_F(p)$ from his/her position on the reference (egalitarian) Lorenz function $\mathrm{L}_E(p)$. Naturally, the position π is random, and we have already seen in the case of the Gini index that it follows the uniform on $[0,1]$ distribution. This means that the person has an equal chance of receiving any income among all the available incomes which are, in terms of percentiles, identified with the unit interval $[0,1]$.

In general, the personal gamble π can follow various distributions on $[0,1]$, and we shall see a variety of examples throughout this paper. The choice of distribution of π carries information about person's probable positions and is thus inevitably subjective, but many of the examples that we have encountered in the literature follow the beta distribution

$$f_{\mathrm{Beta}}(p \mid \alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} \quad \text{for} \quad 0 < p < 1,$$

which we have visualized in Figure 1. We succinctly write $\pi \sim \text{Beta}(\alpha, \beta)$ and so, for example, the Gini index (cf. Equation (1)) is based on $\pi \sim \text{Beta}(1, 1)$. For illuminating statistical and historical notes on the beta and other related distributions in the context of measuring economic inequality, we refer to Kleiber and Kotz (2003). For very general yet remarkably tractable beta-generated families of distributions for greater modeling flexibility, we refer to Alexander et al. (2012), and references therein.

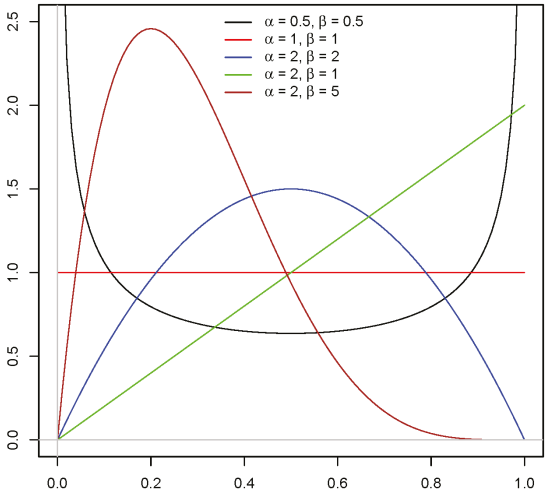


Figure 1. Beta densities of gambles π for various values of α and β.

Importantly for our following discussion, the Gini index G_F can also be viewed as the expected *relative*-deviation of person's position π on the actual Lorenz function $\text{L}_F(p)$ from his/her position on the reference Lorenz function $\text{L}_E(p)$, as seen from the equations:

$$\begin{aligned} \text{G}_F &= \int_0^1 \left(1 - \frac{\text{L}_F(p)}{\text{L}_E(p)}\right) 2p\, dp \\ &= \mathbf{E}\left[1 - \frac{\text{L}_F(\pi)}{\text{L}_E(\pi)}\right], \end{aligned} \tag{2}$$

where $\pi \sim \text{Beta}(2, 1)$, which is a considerable change from $\pi \sim \text{Beta}(1, 1)$ used in the absolute-deviation based representation (1) of the Gini index. Note that the right-hand side of Equation (2) can be succinctly written as $\mathbf{E}[\text{B}_F(\pi)]$, where

$$\text{B}_F(p) = 1 - \frac{\text{L}_F(p)}{\text{L}_E(p)} = 1 - \frac{\text{L}_F(p)}{p} \tag{3}$$

is the Bonferroni function of inequality (cf. Bonferroni 1930), which is also known in the literature as the Gini function of inequality because it appeared in Gini (1914). For details on the Bonferroni function and the corresponding Bonferroni index, we refer to Tarsitano (1990) and references therein.

In addition to its role when studying income and poverty, the Bonferroni function $\text{B}_F(p)$ has also found many uses in other fields such as reliability, demography, insurance, and medicine (e.g., Giorgi and Crescenzi 2001; and references wherein). For detailed historical notes and references with explicit expressions of the Lorenz and Bonferroni functions, as well as of the Gini and Bonferroni indices, for many parametric distributions, we refer to Giorgi and Nadarajah (2010). The role of

the Bonferroni function within the framework of L-functions for measuring economic inequality and actuarial risks can be found in Tarsitano (2004), and Greselin et al. (2009).

3. From Egalitarian Lorenz to the Mean Reference

Not only the classical Gini index but also a multitude of other indices of economic inequality can be viewed as deviation measures (e.g., functional distances) between the actual and egalitarian Lorenz functions (cf., e.g., Zitikis 2002). Note, however, that the actual Lorenz function $\mathrm{L}_F(p)$ itself is a relative measure that compares $p \times 100\%$ lowest incomes with the population mean income μ_F. This two-stage relativity—first with respect to the egalitarian Lorenz function and then with the mean income—warrants a rethinking of the inequality measurement.

Toward this end, we next rephrase the definition of the Gini index G_F by first rewriting the Bonferroni function $\mathrm{B}_F(p)$ as follows:

$$\mathrm{B}_F(p) = 1 - \frac{\mathrm{AV@R}_F(p)}{\mu_F}, \tag{4}$$

where

$$\mathrm{AV@R}_F(p) = \frac{1}{p}\int_0^p F^{-1}(t)dt$$

is the (financial) average value at risk of X. Indeed, with a little mathematical caveat, $\mathrm{AV@R}_F(p)$ is the conditional expectation $\mathbf{E}[X \mid X \le F^{-1}(p)]$, which is the mean income of those who are below the 'poverty line' $F^{-1}(p)$. In summary, Equation (2) becomes

$$\mathrm{G}_F = \mathbf{E}\left[1 - \frac{\mathrm{AV@R}_F(\pi)}{\mu_F}\right] \tag{5}$$

with the gamble $\pi \sim \mathrm{Beta}(2,1)$. If, instead of the latter gamble, we use $\pi \sim \mathrm{Beta}(1,1)$ on the right-hand side of Equation (5), then the expectation turns into the Bonferroni index

$$\mathrm{B}_F = \int_0^1 \left(1 - \frac{\mathrm{AV@R}_F(p)}{\mu_F}\right) dp. \tag{6}$$

For details on the Bonferroni index, we refer to Tarsitano (1990) and references therein. For a comparison of the two weighting schemes, that is, of the gambles π employed in the Gini and Bonferroni cases, we refer to De Vergottini (1940). Implications of using the Bonferroni index on welfare measurement have been studied by, e.g., (Benedetti 1986; Aaberge 2000; Chakravarty 2007). Nygård and Sandström (1981) give a wide-ranging discussion of the use of Bonferroni-type concepts in the measurement of economic inequality. Giorgi and Crescenzi (2001), and Chakravarty and Muliere (2004) propose poverty measures based on the fact that the Bonferroni index exhibits greater sensitivity on lower levels of the income distribution than the Gini index. A general class of inequality measures inspired by the Bonferroni index has been explored by Imedio-Olmedo et al. (2011). Giorgi (1998) provides a list of Bonferroni's publications.

Equations (5) and (6) suggest that the Gini and Bonferroni indices are members of the following general class of indices

$$\mathcal{A}_F = \mathbf{E}[v(\mathrm{AV@R}_F(\pi), \mu_F)], \tag{7}$$

where $v(x,y)$ can be any function for which the expectation is well-defined and finite. In the case of the Gini and Bonferroni indices (e.g., Greselin 2014), we have $v(x,y) = 1 - x/y$, which is the relative value of x with respect to y. We call any function $v(x,y)$ used in expressions like (7) a relative-value function throughout this paper. Hence, we can view the index $\mathcal{A}_F$ as the expected utility of being in the society whose income distribution is depicted by the function $\mathrm{AV@R}_F(p)$ and compared with the reference mean income μ_F using an appropriately chosen relative-value function $v(x,y)$. We should

note at this point that even though the class of relative-value functions $v(x,y)$ may look large, it is nevertheless prudent to restrict our attention to those that are of the form

$$v(x,y) = \ell(x/y) \tag{8}$$

for some function $\ell(t)$. Indeed, under the natural assumption of positive homogeneity, which means that the equation $v(\lambda x, \lambda y) = v(x,y)$ holds for all $\lambda > 0$, Euler's classical theorem says that we must have Equation (8) for some function $\ell(t)$. The Gini and Bonferroni indices give rise to $\ell(t) = 1 - t$.

Another example of the function $\ell(t)$ arises from the E-Gini index of Chakravarty (1988):

$$\begin{aligned} \mathrm{C}_{F,\alpha} &= 2\bigg(\int_0^1 (t - \mathrm{L}_F(t))^\alpha dt \bigg)^{1/\alpha} \\ &= 2\bigg(\int_0^1 \bigg(1 - \frac{\mathrm{AV@R}_F(\pi)}{\mu_F}\bigg)^\alpha t^\alpha dt \bigg)^{1/\alpha} \\ &= \frac{2}{(\alpha+1)^{1/\alpha}} \Big(\mathbf{E}[v(\mathrm{AV@R}_F(\pi), \mu_F)] \Big)^{1/\alpha}, \end{aligned} \tag{9}$$

where the reference-value function is $v(x,y) = (1 - x/y)^\alpha$, that is, $\ell(t) = (1-t)^\alpha$, and the gamble $\pi \sim \mathrm{Beta}(\alpha+1, 1)$. Zitikis (2002) suggests using $(\alpha+1)^{1/\alpha}$ instead of 2 in the definition of the E-Gini index (see also Zitikis (2003) for additional notes) in which case the right-hand side of Equation (9) turns into the index

$$\widetilde{\mathrm{C}}_{F,\alpha} = \Big(\mathbf{E}[v(\mathrm{AV@R}_F(\pi), \mu_F)] \Big)^{1/\alpha}.$$

In either case, note from the expressions of $\mathrm{C}_{F,\alpha}$ and $\widetilde{\mathrm{C}}_{F,\alpha}$ that it is sometimes useful to transform the index $\mathcal{A}_F$ by some function $w(x)$. We shall elaborate on this point in the next section.

Coming now back to the index $\mathcal{A}_F$, we note that, with the generic relative-value function $v(x,y) = \ell(x/y)$, the index can be rewritten as $\mathbf{E}[\overline{\ell}(B_F(\pi))]$, where $\overline{\ell}(t) = \ell(1-t)$. Hence, we are dealing with the distorted Bonferroni function $\overline{\ell}(B_F(p))$, $0 < p < 1$, which is analogous to the distorted Lorenz function upon which Sordo et al. (2014) have built their research (see Aaberge (2000) for earlier results on the topic). We do not pursue this research venue in the present paper because the Bonferroni function, just like that of Lorenz, incorporates a pre-specified reference measure, which is the mean income μ_F. In what follows, we argue in favour of more flexibility when choosing reference measures, which may even include personal preferences in addition to those of the entire population.

4. From the Mean to Generic Societal References

We now extend the index $\mathcal{A}_F$ to arbitrary references, which we denote by θ_F. Namely, let

$$\mathcal{B}_F = w\Big(\mathbf{E}[v(\mathrm{AV@R}_F(\pi), \theta_F)] \Big),$$

where $w(x)$ is a normalizing function whose main role is to fit the index into the unit interval $[0,1]$, with the value 0 meaning perfect equality (i.e., everybody has the same amount) and 1 meaning extreme inequality (i.e., only one person has something, and thus everything, with the others having nothing). Having the flexibility to manipulate references is important due to a variety of reasons. For example, the use of the mean μ_F can become questionable when population skewness increases, and this has already been noted by, e.g., Gastwirth (2014) who, in his research on the changing income inequality in the U.S. and Sweden, has suggested replacing the mean μ_F by the median $m_F = F^{-1}(0.5)$.

Another example of θ_F that differs from μ_F is provided by the Palma index; we refer to Cobham and Sumner (2013a, 2013b, 2014) for details. Namely, let θ_F be the average of the top 10% of the population incomes, that is, $\theta_F = \frac{1}{0.1}\int_{0.9}^1 F^{-1}(t)dt$. Furthermore, let the normalizing function be

$w(x) = x$, the relative-value function $v(x, y) = y/x$, and the (deterministic) gamble $\pi = 0.4$. Under these specifications, the index $\mathcal{B}_F$ becomes the Palma index of economic inequality:

$$\mathrm{P}_F^{40,90} = \frac{\frac{1}{0.1}\int_{0.9}^{1} F^{-1}(t)dt}{\frac{1}{0.4}\int_{0}^{0.4} F^{-1}(t)dt}.$$

Instead of the underlying random variable (e.g., income) X, the researcher might be primarily interested in its transformation (e.g., utility of income) $u(X)$. To tackle this situation, we first incorporate the transformed incomes into our framework by extending the definition of the (financial) average value at risk as follows:

$$\mathrm{AV@R}_{F,u}(p) = \frac{1}{p}\int_0^p u(F^{-1}(t))dt.$$

Note that $\mathrm{AV@R}_{F,u}(1) = \mathbf{E}[u(X)]$, which we can view as the expected utility of X. We have arrived at the extension

$$\mathcal{C}_F = w\Big(\mathbf{E}[v(\mathrm{AV@R}_{F,u}(\pi), \theta_F)]\Big) \tag{10}$$

of the index $\mathcal{B}_F$.

The index $\mathcal{C}_F$ appears to be a minor generalization of the extended intermediate index of Cowell (2003) (see Equation (12) therein), which has been shown to include a large number of well-known indices (in particular, the Generalized Entropy class of indices) and far-reaching new ones. Namely, $\mathcal{C}_F$ reduces to the index of Cowell (2003), which for referencing purposes we denote by $\mathcal{C}_{F,k}$, by choosing $w(x) = A_k(x-1)$ for a certain constant A_k, $u(x) = \phi_k(x)$ for a certain function $\phi_k(x)$, the reference $\theta_F = u(\mu_F)$, the relative-value function $v(x) = x/y$, and the (deterministic) gamble $\pi = 1$; here are the aforementioned quantities that we have not yet specified:

$$A_k = \frac{1+k^2}{\alpha_k^2 - \alpha_k}, \quad \alpha_k = \gamma + \beta k, \quad \phi_k(x) = \frac{1}{\alpha_k}(x+k)^{\alpha_k},$$

where $\gamma \in (-\infty, \infty)$, $\beta \geq 0$, and $k \geq 0$ are parameters. Hence, even though the reason for our use of the letter $\mathcal{C}$ for index (10) is alphabetical, it would only be natural to call $\mathcal{C}_F$ the Cowell general intermediate index, whose special case, called extended intermediate index, appears in Cowell (2003).

The Atkinson (1970) index, which we denote by $\mathrm{A}_{F,\gamma}$, is a special case of $\mathcal{C}_F$. (For many other special cases, we refer to Cowell (2003).) Namely, let the utility function be $u(x) = x^\gamma$ for some $\gamma \in (0, 1)$. Furthermore, let the (deterministic) gamble be $\pi = 1$, the reference $\theta_F = u(\mu_F)$, and the relative-value function $v(x, y) = 1 - x/y$. Under these specifications, the index $\mathcal{C}_F$ turns into $1 - \mathbf{E}[X^\gamma]/\mu_F^\gamma$, which after the transformation with the function $w(x) = 1 - (1-x)^{1/\gamma}$ becomes the Atkinson index

$$\mathrm{A}_{F,\gamma} = 1 - \frac{(\mathbf{E}[X^\gamma])^{1/\gamma}}{\mu_F}.$$

This index has been highly influential in measuring economic inequality (e.g., Cowell (2011), and references therein) and inspired a variety of extensions and generalization of the Gini index. In addition, Mimoto and Zitikis (2008) have found the Atkinson index useful for developing a statistical inference theory for testing exponentiality, which has been a prominent problem in life-time analysis and, particularly, in reliability engineering.

5. The Donaldson-Weymark-Kakwani Index Revisited and Extended

The Donaldson-Weymark-Kakwani index (Donaldson and Weymark 1980, 1983; Kakwani 1980a, 1980b; Weymark 1981)

$$\mathrm{DWK}_{F,\alpha} = \alpha(\alpha-1)\int_0^1 (1-p)^{\alpha-2}(p - \mathrm{L}_F(p))dp,$$

which is also known as the *S*-Gini index, has arisen following Atkinson (1970) observation that the Gini index G_F does not take into account social preferences. Via the parameter $\alpha > 1$, the index $\mathrm{DWK}_{F,\alpha}$ can reflect different social preferences, with the classical Gini index arising by setting $\alpha = 2$. We note in this regard that a justification for a family of indices to be based on the theory of relative deprivation has been provided by Yitzhaki (1979, 1982).

Just like the Gini index G_F, the index $\mathrm{DWK}_{F,\alpha}$ can also be placed within the framework of expected relative value. Indeed, using Equations (3) and (4), we have

$$\begin{aligned}\mathrm{DWK}_{F,\alpha} &= \int_0^1 \left(1 - \frac{\mathrm{L}_F(p)}{p}\right) f_{\mathrm{Beta}}(p \mid 2, \alpha-1)dp \\ &= \int_0^1 \left(1 - \frac{\mathrm{AV@R}_F(p)}{\mu_F}\right) f_{\mathrm{Beta}}(p \mid 2, \alpha-1)dp \\ &= \mathbf{E}[v(\mathrm{AV@R}_F(\pi_\alpha), \mu_F)]\end{aligned} \tag{11}$$

with the relative-value function $v(x,y) = 1 - x/y$ and the gamble $\pi_\alpha \sim \mathrm{Beta}(2, \alpha-1)$, whose density is visualized in Figure 2.

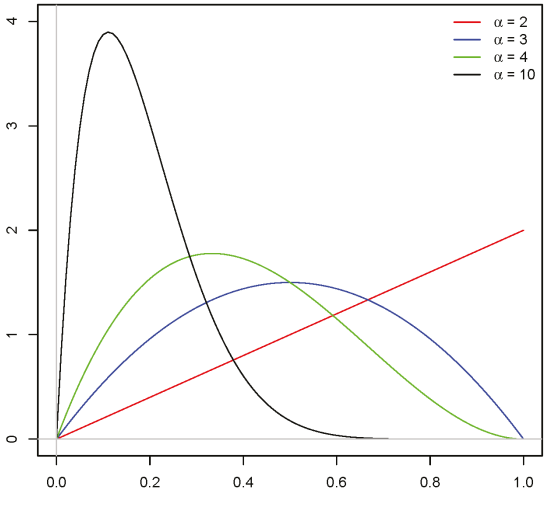


Figure 2. The density of π_α for various values of α.

We next introduce a more flexible index than $\mathrm{DWK}_{F,\alpha}$ that allows us to employ more general gambles than π_α. For this, we first introduce a class of generating functions:

(H) Let $h : [0,1] \to [0,1]$ be any twice differentiable and convex function (i.e., $h''(p) \geq 0$ for all $p \in (0,1)$) that satisfies the boundary conditions $h(0) = 0$ and $h(1) = 1$, and such that $h'(0) \neq 1$.

Let π_h denote the gamble whose density $f(p)$ is given by the formula

$$f(p) = \frac{p\,h''(1-p)}{1 - h'(0)} \tag{12}$$

for all $p \in (0,1)$, and $f(p) = 0$ elsewhere. With the relative-value function $v(x,y) = 1 - x/y$, we have (details in Appendix A)

$$\begin{aligned} \mathrm{DWK}_{F,h} &:= \mathbf{E}[v(\mathrm{AV@R}_F(\pi_h), \mu_F)] \\ &= \frac{1}{1-h'(0)}\left(1 - \frac{1}{\mu_F}\int_0^1 F^{-1}(p)h'(1-p)\,dp\right) \\ &= \frac{1}{1-h'(0)}\left(1 - \frac{1}{\mu_F}\int_0^\infty h(1-F(x))\,dx\right). \end{aligned} \tag{13}$$

To illustrate, we choose the function $h(p) = p^\alpha$ with any $\alpha > 1$, in which case the gamble π_h follows the density $\alpha(\alpha-1)p(1-p)^{\alpha-2}$; that is, $\pi_h \sim \mathrm{Beta}(2, \alpha-1)$, which means that π_h has the same distribution as the earlier noted gamble π_α. Consequently, $\mathrm{DWK}_{F,h}$ reduces to $\mathrm{DWK}_{F,\alpha}$, and thus Equation (13) reduce to the following expressions of the Donaldson-Weymark-Kakwani index:

$$\begin{aligned} \mathrm{DWK}_{F,\alpha} &= 1 - \frac{\alpha}{\mu_F}\int_0^1 F^{-1}(p)(1-p)^{\alpha-1}dp \\ &= 1 - \frac{1}{\mu_F}\int_0^\infty (1-F(x))^\alpha dx \end{aligned} \tag{14}$$

(cf. Donaldson and Weymark (1980, 1983); Yitzhaki (1983); Muliere and Scarsini (1989)).

6. The Wang Risk Measure Revisited and Extended

The index $\mathrm{DWK}_{F,h}$ is based on gambles generated by *convex* functions h. A similar index but based on *concave* generating functions g is called the Wang (or distortion) risk measure, which has been used in actuarial science and financial mathematics for measuring risks. In detail, the risk measure is defined by the formula

$$\mathrm{W}_{F,g} = \int_0^\infty g(1-F(x))\,dx,$$

where $g : [0,1] \to [0,1]$ is a distortion function, meaning that it is non-decreasing and satisfies the boundary conditions $g(0) = 0$ and $g(1) = 1$.

Hence, unlike in the previous section, we now work with concave distortion functions, denoted by g, under which the risk measure $\mathrm{W}_{F,g}$ is coherent (Wang et al. 1997; Wang and Young 1998; Wirch and Hardy 1999; see Artzner et al. (1999) for a general discussion). A classical example of such a distortion function is $g(p) = p^\alpha$ for any $\alpha \in (0,1)$, in which case the Wang risk measure $\mathrm{W}_{F,g}$ reduces to the proportional-hazards-transform risk measure (Wang 1995)

$$\mathrm{PHT}_{F,\alpha} = \int_0^\infty (1-F(x))^\alpha\,dx.$$

For more information on concave versus convex distortion functions in the context of measuring risks, their variability and orderings, we refer to Sordo and Suárez-Llorens (2011), Giovagnoli and Wynn (2012), and references therein.

We next show that the Wang risk measure $\mathrm{W}_{F,g}$ can be placed within the framework of expected relative value. When compared with the index $\mathrm{DWK}_{F,\alpha}$, there are two major changes: First, the function of interest is now the (insurance) average value at risk:

$$\mathrm{AVaR}_F(p) = \frac{1}{1-p}\int_p^1 F^{-1}(t)dt.$$

(Note that when $p = 0$, then $\mathrm{AVaR}_F(p)$ is equal to the mean μ_F.) Second, the function g that generates the distribution of the random position is concave. Specifically, we introduce the following class of generating functions:

(G) Let $g : [0,1] \to [0,1]$ be twice differentiable and concave function (i.e., $g''(p) \leq 0$ for all $p \in (0,1)$) that satisfies the boundary conditions $g(0) = 0$ and $g(1) = 1$, and such that $g'(1) \neq 1$.

Any such function g generates the density $f(p)$ of the gamble π_g given by the formula

$$f(p) = \frac{-(1-p)g''(1-p)}{1-g'(1)} \tag{15}$$

for all $p \in (0,1)$, and $f(p) = 0$ elsewhere. With the relative-value function $v(x,y) = y/x - 1$, we have (details in Appendix)

$$\begin{aligned} \mathbf{E}[v(\mu_F, \mathrm{AVaR}_F(\pi_g))] &= \frac{1}{1-g'(1)}\left(\frac{1}{\mu_F}\int_0^1 F^{-1}(p)g'(1-p)\,dp - 1\right) \\ &= \frac{1}{1-g'(1)}\left(\frac{1}{\mu_F}\int_0^\infty g(1-F(x))\,dx - 1\right). \end{aligned} \tag{16}$$

Consequently, the Wang risk measure $\mathrm{W}_{F,g}$ can be expressed in terms of the expected relative value $\mathbf{E}[v(\mu_F, \mathrm{AVaR}_F(\pi_g))]$ as follows:

$$\mathrm{W}_{F,g} = \mu_F\Big(\mathbf{E}[v(\mu_F, \mathrm{AVaR}_F(\pi_g))](1-g'(1)) + 1\Big). \tag{17}$$

When the generating function is $g(t) = t^\alpha$ for any $\alpha \in (0,1)$, then the gamble π_g follows $\mathrm{Beta}(1,\alpha)$ whose density function $\alpha(1-p)^{\alpha-1}$ is depicted in Figure 3.

From Equation (16), we have

$$\mathbf{E}[v(\mu_F, \mathrm{AVaR}_F(\pi_\alpha))] = \frac{1}{1-\alpha}\left(\frac{1}{\mu_F}\int_0^\infty (1-F(x))^\alpha\,dx - 1\right). \tag{18}$$

Finally, we note the following expression for the proportional-hazards-transform risk measure:

$$\mathrm{PHT}_{F,\alpha} = \mu_F\Big(\mathbf{E}[v(\mu_F, \mathrm{AVaR}_F(\pi_g))](1-\alpha) + 1\Big).$$

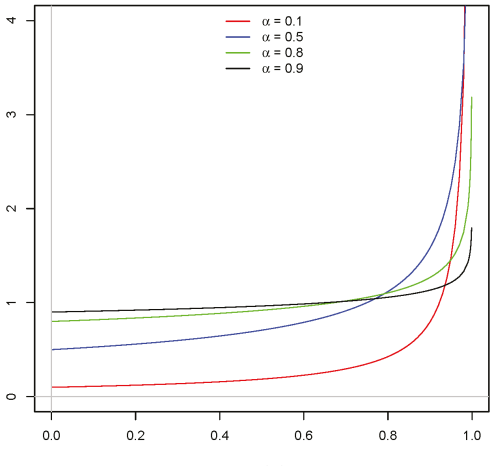


Figure 3. The density of π_g when $g(p) = p^\alpha$ for various values of α.

7. From Collective to Individual References

So far, we have worked with *collective* references. They do not depend on the outcomes of personal gambles and thus apply to all members of the society. Such references may not, however, be always desirable or justifiable. For example, given the outcome 0.4 of the gamble π, meaning that the person is considered to be among the 40% lowest income earners, the person may wish to compare the current position with the hypothetical one of being among the 60% highest income earners. In such situations, we are dealing with *individual* references: their values may depend on outcomes of the personal gamble π.

Hence, for example, the mean μ_F and the median $m_F = F^{-1}(0.5)$ are collective references, but $\theta_F = F^{-1}(\pi)$ is an individual reference because its value depends on the outcome of π. Would the quantile $F^{-1}(\pi)$ be a good reference? There are at least two major reasons against the use of the quantile, which is known in the risk literature as the value-at-risk:

1. The quantile $F^{-1}(p)$ is not robust with respect to realized values p of the random gamble π, in the sense that the quantile may change drastically even for very small changes of p.
2. For a realized value p of π, the quantile $F^{-1}(p)$ is not informative about the values of $F^{-1}(q)$ for $q > p$. Indeed, we may have the same value of $F^{-1}(p)$ irrespective of whether the cdf F is heavy- or light-tailed.

These are serious issues when constructing sound measures of economic inequality and risk. In the risk literature (cf., e.g., McNeil et al. (2005); Meucci (2007); Pflug and Römisch (2007); Cruz (2009); Sandström (2010); Cannata and Quagliariello (2011); and references therein), the problem with quantiles has been overcome by using $\mathrm{AVaR}_F(p)$ whose definition was given in the previous section. For example, adopting $\mathrm{AVaR}_F(p)$ as our (individual) reference θ_F and using the normalizing function $w(x) = x$, the earlier introduced index $\mathcal{B}_F$ turns into the Zenga (2007) index

$$\begin{aligned} \mathrm{Z}_F &= \int_0^1 \left(1 - \frac{\mathrm{AV@R}_F(p)}{\mathrm{AVaR}_F(p)}\right) dp \\ &= \mathbf{E}[v(\mathrm{AV@R}_F(\pi), \mathrm{AVaR}_F(\pi))] \end{aligned} \tag{19}$$

with the relative-value function $v(x, y) = 1 - x/y$ and the gamble $\pi \sim \mathrm{Beta}(1, 1)$. Hence, the Zenga index Z_F is the average with respect to all percentiles $p \in (0, 1)$ of the relative deviations of the mean income of the poor (i.e., those whose incomes are below the poverty line $F^{-1}(p)$) from the corresponding mean income of the rich, that is, of those whose incomes are above the poverty line $F^{-1}(p)$. We refer to Greselin et al. (2013) for a more detailed discussion of the relative nature of the Gini and Zenga indices, and their comparison.

8. Relative Measure of Risk

Many risk measures that we find in the literature are designed to measure *absolute* heaviness of the right-hand tail of the underlying loss distribution. Suppose now that we wish to measure the severity of large (e.g., insurance) losses *relative* to small ones. Note that this problem is very similar to that tackled by Zenga (2007) in the context of economic inequality. Hence, following the same path but now using the relative-value function $v(x, y) = y/x - 1$ and generic gamble π, we arrive at the relative measure of risk

$$\mathrm{R}_F = \mathbf{E}[v(\mathrm{AV@R}_F(\pi), \mathrm{AVaR}_F(\pi))], \tag{20}$$

which, in the spirit of expected utility, can be rewritten as

$$\mathrm{R}_F = \mathbf{E}[\mathrm{R}_F(\pi)], \tag{21}$$

where the role of utility function is played by the risk function

$$\mathrm{R}_F(p) = \frac{\mathrm{AVaR}_F(p)}{\mathrm{AV@R}_F(p)} - 1.$$

In what follows, we explore properties of this risk measure, using the notation R_X instead of R_F to simplify the presentation.

Proposition 1. *We have the following statements:*

1. *If the risk X is constant, that is, $X = d$ for some constant $d > 0$, then $\mathrm{R}_X = 0$.*
2. *Multiplying X by any constant $d > 0$ does not change the relative measure of risk, that is, $\mathrm{R}_{dX} = \mathrm{R}_X$.*
3. *Adding any constant $d > 0$ to the risk X decreases the relative measure of risk, that is, $\mathrm{R}_{X+d} \leq \mathrm{R}_X$.*

We have relegated the proof of Proposition 1 to Appendix. We next comment on the meaning of the three properties spelled out in the proposition. First, given that we are dealing with a *relative* measure of risk, properties 1 and 2 are self-explanatory. As to property 3, it says that lifting up the risk by any positive constant decreases its riskiness. This is natural because lifting up diminishes the relative variability of the risk. This, in turn, suggests that ordering of the relative risk measures should be done, for example, in terms of the Lorenz ordering, which is one of the most used tools for comparing the variability of economic-size distributions. This leads to the following property:

Proposition 2. *If risks X and Y follow the Lorenz ordering $X \leq_{\mathrm{L}} Y$, then $\mathrm{R}_X \leq \mathrm{R}_Y$.*

The proof of Property 2 is provided in Appendix, where the basic definition of Lorenz ordering can also be found. It is related to the notion of ordering based on the generalized, also called absolute, Lorenz curve (e.g., Ramos et al. 2000; Sriboonchita et al. 2010; and references therein). This leads us directly to a closely related property called the Pigou-Dalton principle of transfers. In the context of economic inequality, the principle says that progressive (i.e., from rich to poor) rank-order and mean-preserving transfers should decrease the value of inequality measures. Hence, in the context of risk, the transfers should be risk decreasing. Formally (cf., Vergnaud 1997), X is less risk-unequal than Y in the Pigou-Dalton sense, denoted by $X \leq_{\mathrm{PD}} Y$, if and only if $\mu_X = \mu_Y$ and $X \leq_{\mathrm{L}} Y$. Hence, $X \leq_{\mathrm{PD}} Y$ is sometimes denoted by $X \leq_{\mathrm{L},=} Y$ (cf. Denuit et al. 2005). The following property is now obvious.

Proposition 3. *If a Pigou-Dalton risk-increasing transfer turns risk X into Y so that $X \leq_{\mathrm{PD}} Y$, then $\mathrm{R}_X \leq \mathrm{R}_Y$.*

To have an idea of how the Pigou-Dalton transfers act, we recall (e.g., Shaked and Shanthikumar 2007; Sriboonchita et al. 2010) that given X and Y with densities f_X and f_Y, respectively, and assuming that their means are equal, if the sign of the difference $f_X - f_Y$ changes twice according to the pattern $(+,-,+)$, then $X \leq_{\mathrm{L}} Y$. Examples of parametric distributions with such pdf's can be found in, e.g., Kleiber and Kotz (2003); see also references therein.

In what follows, we discuss an example based on the Zenga (2010) distribution that has shown remarkably good performance in terms of goodness-of-fit on a number of real income data sets. It is a very flexible three-parameter distribution with Pareto-type right-hand tail and whose density is

$$f_{\mathrm{Zenga}}(x \mid \mu, \alpha, \theta) = \begin{cases} \dfrac{1}{2\mu\,\mathrm{Beta}(\alpha,\theta)} \left(\dfrac{x}{\mu}\right)^{-1.5} \displaystyle\int_0^{x/\mu} t^{\alpha+0.5-1}(1-t)^{\theta-2}dt, & x < \mu, \\ \dfrac{1}{2\mu\,\mathrm{Beta}(\alpha,\theta)} \left(\dfrac{\mu}{x}\right)^{1.5} \displaystyle\int_0^{\mu/x} t^{\alpha+0.5-1}(1-t)^{\theta-2}dt, & x \geq \mu, \end{cases}$$

where μ is the scale parameter, which also happens to be the mean of the distribution, and θ and α are two shape parameters that affect, respectively, the center and the tails of the distribution. We have depicted the Zenga density in Figure 4. For further details on this distribution and its uses, we refer to Zenga (2010), Zenga et al. (2011), Zenga et al. (2012), and Arcagni and Zenga (2013).

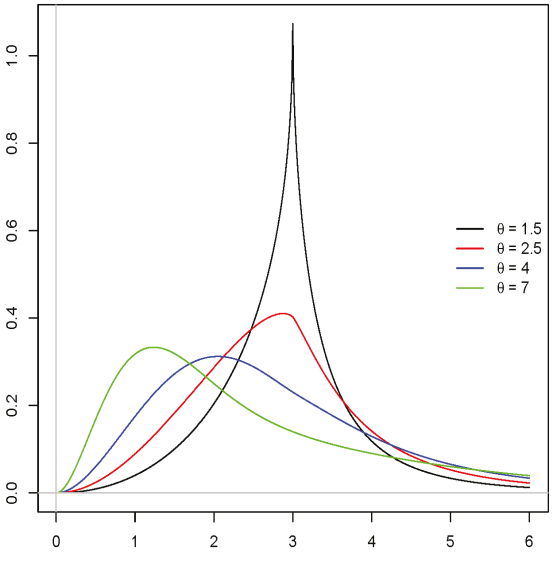


Figure 4. Zenga$(2, 2, \theta)$ density for various values of θ.

To see the effects of the Pigou-Dalton transfers in the case of the Zenga distribution, the following theorem is particularly useful.

Theorem 1 (Arcagni and Porro 2013)**.** *Assume* $X \sim \text{Zenga}(\mu_X, \alpha_X, \theta_X)$ *and* $Y \sim \text{Zenga}(\mu_Y, \alpha_Y, \theta_Y)$*, where all the parameters are positive. When* $\alpha_X \geq \alpha_Y$ *and* $\theta_X \leq \theta_Y$*, then* $X \leq_{\text{PD}} Y$*.*

9. Conclusions: A General Index of Inequality and Risk

The right-hand sides of Equations (19) and (20), which are identical, barring their different relative-value functions $v(x, y)$, give rise to a very general measure of inequality:

$$\mathbf{E}[v(\text{AV@R}_F(\pi), \text{AVaR}_F(\pi^*))],$$

where π and π^* are two gambles, which could be dependent or independent, degenerate or not. Obviously, when $\pi = \pi^*$, then we have either the Zenga index of economic inequality or the relative measure of risk, depending on the choice of the relative-value function. Furthermore, if $\pi^* = 0$, then we have $\text{AVaR}_F(\pi^*) = \mu_F$ and thus $\mathbf{E}[v(\text{AV@R}_F(\pi), \mu_F)]$, which is the Bonferroni index B_F. By appropriately choosing relative-value functions and personal gambles, we can reproduce a number of other measures of economic inequality and risk, but the Chakravarty and Atkinson indices require some little extension:

$$\mathcal{E}_F = w\Big(\mathbf{E}\big[v\big(\text{AV@R}_{F,u}(\pi), \text{AVaR}_{F,u^*}(\pi^*)\big)\big]\Big), \tag{22}$$

where u and u^* are two utility functions, and

$$\text{AVaR}_{F,u^*}(p) = \frac{1}{1-p}\int_p^1 u^*(F^{-1}(t))dt.$$

Note that $\text{AVaR}_{F,u^*}(0) = \mathbf{E}[u^*(X)]$. All the examples that we have mentioned in this paper, and also many other ones that appear in the literature, are special cases of the just introduced index $\mathcal{E}_F$. Table 1 provides a summary.

Table 1. Special cases of index (22) with $u^*(x) = x$ in all the rows.

	π	π^*	$w(x)$	$v(x,y)$	$u(x)$
Atkinson $A_{F,\alpha}$	1	0	$1-(1-x)^{1/\gamma}$	$1-x/y$	x^γ
Bonferroni B_F	Beta$(1,1)$	0	x	$1-x/y$	x
Chakravarty $C_{F,\alpha}$	Beta$(\alpha+1,1)$	0	$2(\alpha+1)^{-1/\alpha}x^{1/\alpha}$	$(1-x/y)^\alpha$	x
Inequality index $\widetilde{C}_{F,\alpha}$	Beta$(\alpha+1,1)$	0	$x^{1/\alpha}$	$(1-x/y)^\alpha$	x
Cowell $\mathcal{C}_{F,k}$	1	0	$A_k(x-1)$	x/y	$\phi_k(x)$
Cowell's Generalized Entropy class	1	0	linear	x/y	$\phi(x)$
Donaldson-Weymark-Kakwani $DWK_{F,\alpha}$	Beta$(2,\alpha-1)$	0	x	$1-x/y$	x
Inequality index $DWK_{F,h}$	π_h	0	x	$1-x/y$	x
Gini G_F	Beta$(2,1)$	0	x	$1-x/y$	x
Palma $P_F^{40,90}$	0.4	0.9	x	y/x	x
Risk measure R_F	Any	$\pi^*=\pi$	x	$y/x-1$	x
Wang $W_{F,g}$	1	π_g	$\mu_F(x(1-g'(1))+1)$	$y/x-1$	x
Proportional hazards transform $PHT_{F,\alpha}$	1	Beta$(1,\alpha)$	$\mu_F(x(1-\alpha)+1)$	$y/x-1$	x
Zenga Z_F	Beta$(1,1)$	$\pi^*=\pi$	x	$1-x/y$	x

We conclude with the note that, in the examples throughout this paper, the gambles π and π^* have been such that either they are identical (i.e., $\pi=\pi^*$) or one of them is degenerate (e.g., $\pi=1$ or $\pi^*=0$). There is no reason why this should always be the case: the two gambles can be dependent but not necessarily identical or degenerate. This suggests that, in general, modeling probability distributions of the pair (π,π^*) can be conveniently achieved by, for example, specifying marginal distributions of the gambles π and π^*, as well as dependence structures between them using, e.g., appropriately chosen copulas. For methodological and applications-driven developments related to copulas, we refer to the monographs of Nelsen (2006), Jaworski et al. (2010), Jaworski et al. (2013), and references therein.

Acknowledgments: We are indebted to Academic Editors and anonymous reviewers for suggestions, insightful comments, and constructive criticism that guided our work on the revision. The second author is grateful to the University of Milano-Bicocca for making his most inspiring scientific visit at the university possible. The research has been supported by the grant "From Data to Integrated Risk Management and Smart Living: Mathematical Modelling, Statistical Inference, and Decision Making" awarded by the Natural Sciences and Engineering Research Council of Canada to the second author.

Author Contributions: Both authors, with the consultation of each other, carried out this work and drafted the manuscript together. Both authors read and approved the final manuscript.

Conflicts of Interest: The authors declare that they have no competing interests.

Appendix A. Technicalities

Proof of Equation (13). Since the relative-value function is $v(x,y)=1-x/y$, we have

$$\mathrm{DWK}_{F,h} = 1 - \frac{1}{\mu_F}\int_0^1 \mathrm{AV@R}_F(p) f(p)\,dp, \tag{A1}$$

where $f(p)$ is the density function of the gamble π_h defined by Equation (12). The following are straightforward calculations:

$$\begin{aligned}
\int_0^1 \mathrm{AV@R}_F(p) f(p)\,dp &= \int_0^1 \frac{1}{p}\left(\int_0^p F^{-1}(t)\,dt\right) f(p)\,dp \\
&= \int_0^1 F^{-1}(t)\left(\int_t^1 \frac{1}{p} f(p)\,dp\right) dt \\
&= \frac{1}{1-h'(0)}\int_0^1 F^{-1}(t)\big(h'(1-t)-h'(0)\big)\,dt \\
&= \frac{1}{1-h'(0)}\left(\int_0^1 F^{-1}(t)h'(1-t)\,dt - h'(0)\mu_F\right).
\end{aligned}$$

Combining this result with Equation (A1), we obtain the first equation of (13). Since

$$\begin{aligned}\int_0^1 F^{-1}(t)h'(1-t)\,dt &= \int_0^\infty \left(\int_0^1 \mathbf{1}\{F^{-1}(t) > x\}h'(1-t)\,dt\right)dx \\ &= \int_0^\infty \left(\int_0^1 \mathbf{1}\{t > F(x)\}h'(1-t)\,dt\right)dx \\ &= \int_0^\infty \left(\int_{F(x)}^1 h'(1-t)\,dt\right)dx \\ &= \int_0^\infty h(1-F(x))\,dx,\end{aligned} \tag{A2}$$

we have the second equation of (13). □

Proof of Equation (16). Since the relative-value function is $v(x,y) = y/x - 1$, we have

$$\mathbf{E}[v(\mu_F, \mathrm{AVaR}_F(\pi_g))] = \frac{1}{\mu_F}\int_0^1 \mathrm{AVaR}_F(p)f(p)\,dp - 1, \tag{A3}$$

where $f(p)$ is the density function of the gamble π_g defined by Equation (15). The following are straightforward calculations:

$$\begin{aligned}\int_0^1 \mathrm{AVaR}_F(p)f(p)\,dp &= \int_0^1 \frac{1}{1-p}\left(\int_p^1 F^{-1}(t)\,dt\right)f(p)\,dp \\ &= \int_0^1 F^{-1}(t)\left(\int_0^t \frac{f(p)}{1-p}\,dp\right)\,dt.\end{aligned}$$

Applying definition (15) of the density function $f(p)$, we obtain

$$\begin{aligned}\int_0^1 \mathrm{AVaR}(p)f(p)\,dp &= \frac{1}{1-g'(1)}\int_0^1 F^{-1}(t)\big(g'(1-t) - g'(1)\big)\,dt \\ &= \frac{1}{1-g'(1)}\left(\int_0^1 F^{-1}(t)g'(1-t)\,dt - g'(1)\mu_F\right).\end{aligned} \tag{A4}$$

Combining Equations (A3) and (A4), we obtain the first equation of (16). Using Equation (A2) with g instead of h, we arrive at the second equation of (16). □

Remark A1. From the mathematical point of view, Equation (A4) is elementary, but it was a pivotal observation that allowed Jones and Zitikis (2003) to initiate the development of statistical inference for the Wang (or distortion) risk measure. Since then, numerous statistical results have appeared on risk measures: parametric and non-parametric, light- and heavy-tailed cases have been explored in great detail by many authors. To illustrate the challenges that arise in the heavy-tailed context, we refer to Necir and Meraghni (2009), and Necir et al. (2007) for the proportional hazards transform; Necir et al. (2010), and Rassoul (2013) for the tail conditional expectation; and Brahimi et al. (2012) for general distortion risk measures.

Proof of Proposition 1. Property 1 follows from the fact that, if $X = d$ for any constant $d > 0$, then $F_X^{-1}(p) = d$ and so $\mathrm{AVaR}_X(p) = \mathrm{AV@R}_X(p)$ for every $p \in (0,1)$. Property 2 follows from the fact that if $d > 0$, then $F_{dX}^{-1}(p) = dF_X^{-1}(p)$ and so $\mathrm{AVaR}_{dX}(p)/\mathrm{AV@R}_{dX}(p) = \mathrm{AVaR}_X(p)/\mathrm{AV@R}_X(p)$ for every $p \in (0,1)$. Property 3 follows from the fact that $F_{X+d}^{-1}(p) = F_X^{-1}(p) + d$ for every d, and so the bound $\mathrm{AV@R}_X(p) \le \mathrm{AVaR}_X(p)$ together with the assumed positivity of d imply

$$\frac{\mathrm{AVaR}_{X+d}(p)}{\mathrm{AV@R}_{X+d}(p)} = \frac{\mathrm{AVaR}_X(p)+d}{\mathrm{AV@R}_X(p)+d} \le \frac{\mathrm{AVaR}_X(p)}{\mathrm{AV@R}_X(p)}.$$

The latter bound is equivalent to $\mathrm{R}_{X+d}(p) \leq \mathrm{R}_X(p)$ for every $p \in (0,1)$, which establishes the bound $\mathrm{R}_{X+d} \leq \mathrm{R}_X$. □

Proof of Proposition 2. We first recall (Arnold 1987; Aaberge 2000) that the Lorenz ordering $X \leq_{\mathrm{L}} Y$ means the bound $\mathrm{L}_X(p) \geq \mathrm{L}_Y(p)$ for all $p \in [0,1]$. Since

$$\begin{aligned} \mathrm{R}_X(p) &= \frac{1-\mathrm{L}_X(p)}{\mathrm{L}_X(p)} \, \frac{p}{1-p} - 1 \\ &= \frac{p}{(1-p)\mathrm{L}_X(p)} - \frac{p}{1-p} - 1, \end{aligned}$$

the Lorenz ordering $X \leq_{\mathrm{L}} Y$ is equivalent to the R-ordering $X \leq_{\mathrm{R}} Y$, which means $\mathrm{R}_X(p) \leq \mathrm{R}_Y(p)$ for all $p \in (0,1)$. The latter bound and Equation (21) conclude the verification of Proposition 2. □

Remark A2. With the above introduced notion of R-ordering, we can rephrase Proposition 2 as follows: if $X \leq_{\mathrm{R}} Y$, then $\mathrm{R}_X \leq \mathrm{R}_Y$. For detailed treatments of various notions of stochastic orders, we refer to Shaked and Shanthikumar (2007); Li and Li (2013); and Sriboonchita et al. (2010).

References

Aaberge, Rolf. 2000. Characterizations of Lorenz curves and income distributions. *Social Choice and Welfare* 17: 639–53.

Alexander, Carol, Gauss M. Cordeiro, Edwin M. M. Ortega, and José María Sarabia. 2012. Generalized beta-generated distributions. *Computational Statistics and Data Analysis* 56: 1880–97.

Amiel, Yoram, and Frank A. Cowell. 1997. *Income Transformation and Income Inequality*. Discussion Paper DARP 24; London, UK: London School of Economics.

Amiel, Yoram, and Frank A. Cowell. 1999. *Thinking About Inequality*. Cambridge: Cambridge University Press.

Arcagni, Alberto, and Francesco Porro. 2013. On the parameters of Zenga distribution. *Statistical Methods and Applications* 22: 285–303.

Arcagni, Alberto, and Michele Zenga. 2013. Application of Zenga's distribution to a panel survey on household incomes of European Member States. *Statistica and Applicazioni* 11: 79–102.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath. 1999. Coherent measures of risk. *Mathematical Finance 9*: 203–28.

Arnold, Barry C. 1987. *Majorization and the Lorenz Order: A Brief Introduction*. New York: Springer.

Atkinson, Anthony B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–63.

Atkinson, Anthony B., and Francois Bourguignon. 2000. *Handbook of Income Distribution*. Amsterdam: Elsevier, vol. 1.

Atkinson, Anthony B., and Francois Bourguignon. 2015. *Handbook of Income Distribution*. Amsterdam: Elsevier, vol. 2.

Atkinson, Anthony B., and Thomas Piketty. 2007. *Top Incomes Over the Twentieth Century: A Contrast between Continental European and English-Speaking Countries*. Oxford: Oxford University Press.

Banerjee, A.V., and Esther Duflo. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: Public Affairs.

Bennett, Christopher J., and Ricardas Zitikis. 2015. Ignorance, lotteries, and measures of economic inequality. *Journal of Economic Inequality* 13: 309–16.

Benedetti, C. 1986. Sulla interpretazione benesseriale di noti indici di concentrazione e di altri. *Metron* 44: 421–29.

Bonferroni, C. E. 1930. *Elementi di Statistica Generale*. Firenze: Libreria Seeber.

Brahimi, B., F. Meddi, and A. Necir. 2012. Bias-corrected estimation in distortion risk premiums for heavy-tailed losses. *Afrika Statistika* 7: 474–90.

Cannata, F., and M. Quagliariello. 2011. *Basel III and Beyond*. London: Risk Books.

Ceriani, Lidia, and Paolo Verme. 2012. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *Journal of Economic Inequality* 10: 421–43.

Chakravarty, Satya R. 1988. Extended Gini indices of inequality. *International Economic Review* 29: 147–56.

Chakravarty, Satya R. 2007. A deprivation-based axiomatic characterization of the absolute Bonferroni index of inequality. *Journal of Economic Inequality* 5: 339–51.

Chakravarty, Satya R., and Piero Muliere. 2004. Welfare indicators: A review and new perspectives. 2. Measurement of poverty. *Metron* 62: 247–81.

Champernowne, D. G., and F. A. Cowell. 1998. *Economic Inequality and Income Distribution*. Cambridge: Cambridge University Press.

Cobham, Alex, and Andy Sumner. 2013a. *Putting the Gini Back in the Bottle? 'The Palma' As a Policy-Relevant Measure of Inequality*. Working Paper 2013-5; London, UK: King's College.

Cobham, Alex, and Andy Sumner. 2013b. *Is It All about the Tails? The Palma Measure of Income Inequality*. Working Paper 343; Washington, DC, USA: Center for Global Development.

Cobham, Alex, and Andy Sumner. 2014. Is inequality all about the tails?: The Palma measure of income inequality. *Significance* 11: 10–13.

Cowell, Frank A. 2003. *Theil, Inequality and the Structure of Income Distribution*. Discussion Paper DARP 67; London, UK: London School of Economics.

Cowell, Frank A. 2011. *Measuring Inequality*, 3rd ed. Oxford: Oxford University Press.

Cruz, M. 2009. *The Solvency II Handbook*. London: Risk Books.

Denneberg, Dieter. 1990. Premium calculation: Why standard deviation should be replaced by absolute deviation. *ASTIN Bulletin* 20: 181–90.

Denuit, Michel, Jan Dhaene, Marc Goovaerts, and Rob Kaas. 2005. *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. Chichester: Wiley.

De Vergottini, Mario. 1940. Sul significato di alcuni indici di concentrazione. *Giornale degli Economisti e Annali di Economia* 11: 317–47.

Donaldson, David, and John A. Weymark. 1980. A single-parameter generalization of the Gini indices of inequality. *Journal of Economic Theory* 22: 67–86.

Donaldson, David, and John A. Weymark. 1983. Ethically flexible Gini indices for income distributions in the continuum. *Journal of Economic Theory* 29: 353–58.

Druckman, A., and T. Jackson. 2008. Measuring resource inequalities: The concepts and methodology for an area-based Gini coefficient. *Ecological Economics* 65: 242–52.

Duclos, Jean-Yves. 2000. Gini indices and the redistribution of income. *International Tax and Public Finance* 7: 141–62.

Furman, Edward, Ruodu Wang, and Ricardas Zitikis. 2017. Gini-type measures of risk and variability: Gini shortfall, capital allocations, and heavy-tailed risks. *Journal of Business and Finance* 83: 70–84.

Furman, Edward, and Ricardas Zitikis. 2008. Weighted premium calculation principles. *Insurance: Mathematics and Economics* 42: 459–65.

Furman, Edward, and Ricardas Zitikis. 2009. Weighted pricing functionals with applications to insurance: An overview. *North American Actuarial Journal* 13: 483–96.

Furman, Edward, and Ricardas Zitikis. 2017. Beyond the Pearson correlation: Heavy-tailed risks, weighted Gini correlations, and a Gini-type weighted insurance pricing model. *ASTIN Bulletin* 47: 919–42.

Gastwirth, Joseph L. 1971. A general definition of the Lorenz curve. *Econometrica* 39: 1037–39.

Gastwirth, Joseph L. 2014. Median-based measures of inequality: Reassessing the increase in income inequality in the U.S. and Sweden. *Journal of the IAOS* 30: 311–20.

Gini, Corrado. 1912. *Variabilità e Mutabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche*. Bologna: Tipografia di Paolo Cuppini.

Gini, Corrado. 1914. On the measurement of concentration and variability of characters (English translation from Italian by Fulvio de Santis). *Metron* 63: 3–38.

Gini, Corrado. 1921. Measurement of inequality of incomes. *Economic Journal* 31: 124–26.

Giorgi, Giovanni M. 1990. Bibliographic portrait of the Gini concentration ratio. *Metron* 48: 183–221.

Giorgi, Giovanni M. 1993. A fresh look at the topical interest of the Gini concentration ratio. *Metron* 51: 83–98.

Giorgi, Giovanni M. 1998. Concentration index, Bonferroni. In *Encyclopedia of Statistical Sciences*. Edited by S. Kotz, D. L. Banks and C. B. Read. New York: Wiley, vol. 2, pp. 141–46.

Giorgi, Giovanni Maria, and M. Crescenzi. 2001. A look at the Bonferroni inequality measure in a reliability framework. *Statistica* 91: 571–83.

Giorgi, Giovanni Maria, and Saralees Nadarajah. 2010. Bonferroni and Gini indices for various parametric families of distributions. *Metron* 68: 23–46.

Giovagnoli, Alessandra, and Henry P. Wynn. 2012. (U, V)-Ordering and a Duality Theorem for Risk Aversion and Lorenz-Type Orderings. LSE Philosophy Papers; London, UK: London School of Economics and Political Science.

Greselin, Francesca. 2014. More equal and poorer, or richer but more unequal? *Economic Quality Control* 29: 99–117.

Greselin, Francesca, Leo Pasquazzi, and Ricardas Zitikis. 2013. Contrasting the Gini and Zenga indices of economic inequality. *Journal of Applied Statistics* 40: 282–97.

Greselin, Francesca, Madan L. Puri, and Ricardas Zitikis. 2009. *L*-functions, processes, and statistics in measuring economic inequality and actuarial risks. *Statistics and Its Interface* 2: 227–45.

Harsanyi, John C. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* 61: 434–35.

Imedio-Olmedo, Luis J., Elena Bárcena-Martín, and Encarnacion M. Parrado-Gallardo. 2011. A class of Bonferroni inequality indices. *Journal of Public Economic Theory* 13: 97–124.

Jaworski, Piotr, Fabrizio Durante, and Wolfgang Karl Härdle. 2013. *Copulae in Mathematical and Quantitative Finance*. Berlin: Springer.

Jaworski, Piotr, Fabrizio Durante, Wolfgang Härdle, and Tomasz Rychlik. 2010. *Copula Theory and Its Applications*. Berlin: Springer.

Jones, Bruce L., and Ricardas Zitikis. 2003. Empirical estimation of risk measures and related quantities. *North American Actuarial Journal* 7: 44–54.

Kakwani, Nanak C. 1980. *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. New York: Oxford University Press.

Kakwani, Nanak. 1980. On a class of poverty measures. *Econometrica* 48: 437–46.

Kakwani, N.C., and N. Podder. 1973. On the estimation of Lorenz curves from grouped observations. *International Economic Review* 14: 278–92.

Kenworthy, Lane, and Jonas Pontusson. 2005. Rising inequality and the politics of redistribution in affluent countries. *Perspectives on Politics* 3: 449–71.

Kleiber, Christian, and Samuel Kotz. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken: Wiley.

Korpi, Walter, and Joakim Palme. 1998. The paradox of redistribution and strategies of equality: Welfare state institutions, inequality, and poverty in the Western countries. *American Sociological Review* 63: 661–87.

Kośny, Marek, and Gaston Yalonetzky. 2015. Relative income change and pro-poor growth. *Economia Politica* 32: 311–27.

Kovacevic, Milorad. 2010. *Measurement of Inequality in Human Development—A Review*. Human Development Research Paper 2010/35; New York, NY, USA: United Nations Development Programme.

Lambert, Peter J. 2001. *The Distribution and Redistribution of Income*, 3rd ed. Manchester: Manchester University Press.

Li, Haijun, and Xiaohu Li. 2013. *Stochastic Orders in Reliability and Risk: In Honor of Professor Moshe Shaked*. New York: Springer.

Lorenz, M. O. 1905. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 9: 209–19.

Machina, Mark J. 1987. Choice under uncertainty: Problems solved and unsolved. *Economic Perspectives* 1: 121–54.

Machina, Mark J. 2008. Non-expected utility theory. In *The New Palgrave Dictionary of Economics*, 2nd ed. Edited by S. N. Durlauf and L. E. Blume. New York: Palgrave Macmillan, pp. 74–84.

McNeil, Alexander J., Rüdiger Frey, and Paul Embrechts. 2005. *Quantitative Risk Management*. Princeton: Princeton University Press.

Mimoto, Nao, and Ricardas Zitikis. 2008. The Atkinson index, the Moran statistic, and testing exponentiality. *Journal of the Japan Statistical Society* 38: 187–205.

Meucci, Attilio. 2007. *Risk and Asset Allocation*. Berlin: Springer.

Muliere, Pietro, and Marco Scarsini. 1989. A note on stochastic dominance and inequality measures. *Journal of Economic Theory* 49: 314–23.

Necir, Abdelhakim, and Djamel Meraghni. 2009. Empirical estimation of the proportional hazard premium for heavy-tailed claim amounts. *Insurance: Mathematics and Economics* 45: 49–58.

Necir, Abdelhakim, Djamel Meraghni, and Fatima Meddi. 2007. Statistical estimate of the proportional hazard premium of loss. *Scandinavian Actuarial Journal* 2007: 147–61.

Necir, Abdelhakim, Abdelaziz Rassoul, and Ricardas Zitikis. 2010. Estimating the conditional tail expectation in the case of heavy-tailed losses. *Journal of Probability and Statistics* 2010: 596839.

Nelsen, Roger B. 2006. *An Introduction to Copulas*, 2nd ed. New York: Springer.

Nygård, Fredrik, and Arne Sandström. 1981. *Measuring Income Inequality.* Stockholm: Almqvist and Wiksell.
Oladosu, Gbadebo, and Adam Rose. 2007. Income distribution impacts of climate change mitigation policy in the Susquehanna River Basin Economy. *Energy Economics* 29: 520–44.
Ostry, Jonathan D., Andrew Berg, and Charalambos G. Tsangarides. 2014. Redistribution, Inequality, and Growth. In *IMF Staff Discussion Note SDN/14/02.* Washington: International Monetary Fund.
Palma, José Gabriel. 2006. *Globalizing inEquality: 'centrifugal' and 'centripetal' Forces at Work*. DESA Working Paper No 35; Washington, DC, USA: United Nations Department of Economics and Social Affairs.
Pfähler, Wilhelm. 1990. Redistributive effect of income taxation: decomposing tax base and tax rates effects. *Bulletin of Economic Research* 42: 121–29.
Pflug, Georg Ch., and Werner Römisch. 2007. *Modeling, Measuring and Managing Risk*. Singapore: World Scientific.
Pietra, Gaetano. 1915. On the relationship between variability indices (Note I). (English translation from Italian by P. Brutti and S. Gubbiotti). *Metron* 72: 5–16.
Piketty, Thomas. 2014. *Capital in the Twenty-First Century.* Cambridge: Harvard University Press.
Puppe, Clemens. 1991. *Distorted Probabilities and Choice under Risk.* Berlin: Springer.
Quiggin, John. 1982. A theory of anticipated utility. *Journal of Economic Behavior and Organization* 3: 323–43.
Quiggin, John. 1993. *Generalized Expected Utility Theory: The Rank-Dependent Model.* Dordrecht: Kluwer.
Ramos, Hector M., Jorge Ollero, and Miguel A. Sordo. 2000. A sufficient condition for generalized Lorenz order. *Journal of Economic Theory* 90: 286–92.
Rassoul, Abdelaziz. 2013. Kernel-type estimator of the conditional tail expectation for a heavy-tailed distribution. *Insurance: Mathematics and Economics* 53: 698–703.
Rawls, John. 1971. *A Theory of Justice.* Cambridge: Harvard University Press.
Roemer, John E. 2013. Economic development as opportunity equalization. *World Bank Economic Review* 28: 189–209.
Sadoulet, Elisabeth, and Alain de Janvry. 1995. *Quantitative Development Policy Analysis.* Baltimore: John Hopkins University Press.
Sandström, Arne. 2010. *Handbook of Solvency for Actuaries and Risk Managers: Theory and Practice.* Boca Raton: Chapman and Hall.
Sarabia, José María. 2008. Parametric Lorenz curves: Models and applications. In *Modeling Income Distributions and Lorenz Curves*. Edited by D. Chotikapanich. Berlin: Springer, pp. 167–90.
Sarabia, José María, Faustino Prieto, and María Sarabia. 2010. Revisiting a functional form for the Lorenz curve. *Economics Letters* 107: 249–52.
Schmeidler, David. 1986. Integral representation without additivity. *Proceedings of the American Mathematical Society* 97: 255–61.
Schmeidler, David. 1989. Subjective probability and expected utility without additivity. *Econometrica* 57: 571–87.
Sen, Amartya. 1983. Poor, relatively speaking. *Oxford Economic Papers* 35: 153–69.
Sen, Amartya. 1997. *On Economic Inequality (expanded Edition With a Substantial Annexe by J. E. Foster and A. Sen)*. Oxford: Clarendon Press.
Sen, Amartya. 1998. *Choice, Welfare and Measurement*. Cambridge: Harvard University Press.
Shaked, Moshe, and J. George Shanthikumar. 2007. *Stochastic Orders*. New York: Springer.
Shorrocks, Anthony. 1978. Income inequality and income mobility. *Journal of Economic Theory* 19: 376–93.
Silber, Jacques. 1999. *Handbook on Income Inequality Measurement.* Boston: Kluwer.
Slemrod, Joel. 1992. Taxation and inequality: A time-exposure perspective. In *Tax Policy and the Economy*. Edited by J. M. Poterba. Chicago: University of Chicago Press, vol. 6, pp. 105–27.
Sordo, Miguel A., and Alfonso Suárez-Llorens. 2011. Stochastic comparisons of distorted variability measures. *Insurance: Mathematics and Economics* 49: 11–17.
Sordo, Miguel A., Jorge Navarro, and José María Sarabia. 2014. Distorted Lorenz curves: Models and comparisons. *Social Choice and Welfare* 42: 761–80.
Sriboonchita, Songsak, Wing-Keung Wong, Sompong Dhompongsa, and Hung T. Nguyen. 2010. *Stochastic Dominance and Applications to Finance, Risk and Economics*. Boca Raton: Chapman and Hall/CRC.
Tarsitano, Agostino. 1990. The Bonferroni index of income inequality. In *Income and Wealth Distribution, Inequality and Poverty*. Edited by C. Dagum and M. Zenga. New York: Springer, pp. 228–42.
Tarsitano, Agostino. 2004. A new class of inequality measures based on a ratio of *L*-statistics. *Metron* 62: 137–60.
Thompson, W. A., Jr. 1976. Fisherman's luck. *Biometrics* 32: 265–71.

Van De Ven, Justin, John Creedy, and Peter J. Lambert. 2001. Close equals and calculation of the vertical, horizontal and reranking effects of taxation. *Oxford Bulletin of Economics and Statistics* 63: 381–94.

Vergnaud, J. C. 1997. Analysis of risk in a non expected utility framework and application to the optimality of the deductible. *Revue Finance* 18: 155–67.

Wang, Shaun. 1995. Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics* 17: 43–54.

Wang, Shaun. 1998. An actuarial index of the right-tail risk. *North American Actuarial Journal* 2: 88–101.

Wang, Shaun S., and Virginia R. Young. 1998. Ordering risks: Expected utility theory versus Yaari's dual theory of risk. *Insurance: Mathematics and Economics* 22: 145–61.

Wang, Shaun S., Virginia R. Young, and Harry H. Panjer. 1997. Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics* 21: 173–83.

Weymark, John A. 1981. Generalized Gini inequality indices. *Mathematical Social Sciences* 1: 409–30.

Weymark, John. A. 2003. Generalized Gini indices of equality of opportunity. *Journal of Economic Inequality* 1: 5–24.

Wirch, Julia Lynn, and Mary R. Hardy. 1999. A synthesis of risk measures for capital adequacy. *Insurance: Mathematics and Economics* 25: 337–47.

Yaari, Menahem E. 1987. The dual theory of choice under risk. *Econometrica* 55: 95–115.

Yitzhaki, Shlomo. 1979. Relative deprivation and the Gini coefficient. *Quarterly Journal of Economics* 93: 321–24.

Yitzhaki, Shlomo. 1982. Stochastic dominance, mean variance, and Gini's mean difference. *American Economic Review* 72: 178–85.

Yitzhaki, Shlomo. 1983. On an extension of the Gini inequality index. *International Economic Review* 24: 617–28.

Yitzhaki, Shlomo. 1994. On the progressivity of commodity taxation. In *Models and Measurement of Welfare and Inequality*. Edited by W. Eichhorn. Berlin: Springer, pp. 448–66.

Yitzhaki, Shlomo. 1998. More than a dozen alternative ways of spelling Gini. *Research on Economic Inequality* 8: 13–30.

Yitzhaki, Shlomo. 2003. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron* 51: 285–16.

Yitzhaki, Shlomo, and Edna Schechtman. 2013. *The Gini Methodology: A Primer on a Statistical Methodology.* New York: Springer.

Zenga, Michele. 2007. Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica and Applicazioni* 5: 3–27.

Zenga, Michele. 2010. Mixture of Polisicchio's truncated Pareto distributions with beta weights. *Statistica and Applicazioni* 8: 3–25.

Zenga, Michele, Leo Pasquazzi, M. Polisicchio, and Mariangela Zenga. 2011. More on M. M. Zenga's new three-parameter distribution for non-negative variables. *Statistica and Applicazioni* 9: 5–33.

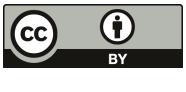
Zenga, Michele, Leo Pasquazzi, and Mariangela Zenga. 2012. First applications of a new three-parameter distribution for non-negative variables. *Statistica and Applicazioni* 10: 131–47.

Zitikis, Ricardas. 2002. Analysis of indices of economic inequality from a mathematical point of view. *Matematika* 8: 772–82.

Zitikis, Ricardas. 2003. Asymptotic estimation of the *E*-Gini index. *Econometric Theory* 19: 587–601.

Zoli, Claudio. 1999. A generalized version of the inequality equivalence criterion: A surplus sharing characterization, complete and partial orderings. In *Logic, Game Theory and Social Choice*. Edited by H. C. M. de Swart. Tilburg: Tilburg University Press, pp. 427–41.

Zoli, Claudio. 2012. *Characterizing Inequality Equivalence Criteria*. Working Paper 32; Verona, Italy: Department of Economics, University of Verona.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

On the Decomposition of the Esteban and Ray Index by Income Sources

Elena Bárcena-Martín [1] and Jacques Silber [2,3,4,*]

[1] Facultad de Ciencias Económicas y Empresariales, Universidad de Málaga, Calle Ejido 6, Málaga 29071, Spain; barcenae@uma.es
[2] Department of Economics, Bar-Ilan University, 52900 Ramat-Gan, Israel
[3] LISER, L-4366 Esch-sur-Alzette, Luxembourg
[4] Centro Camilo Dagum, Tuscan Interuniversity Centre, Advanced Statistics for Equitable and Sustainable Development, Università di Pisa, Dipartimento di Economia e Management, Pisa 56124, Italy
* Correspondence: jsilber_2000@yahoo.com; Tel.: +972-54-9327554

Received: 18 January 2018; Accepted: 11 March 2018; Published: 26 March 2018

Abstract: This paper proposes a simple algorithm based on a matrix formulation to compute the Esteban and Ray (*ER*) polarization index. It then shows how the algorithm introduced leads to quite a simple decomposition of polarization by income sources. Such a breakdown was not available hitherto. The decomposition we propose will thus allow one to determine the sign, as well as the magnitude, of the impact of the various income sources on the *ER* polarization index. A simple empirical illustration based on EU data is provided.

Keywords: decomposition; Esteban and Ray polarization index; income sources; polarization

JEL Classification: D31; D63; J31

1. Introduction

During the past 25 years, many studies attempted to measure the extent of the middle class and stressed the link between the concept of bipolarization and the importance of the middle class. Another strand of the economic literature emphasized the concept of polarization (or multi-polarization). The basic contribution here is that of Esteban and Ray (1994) who linked the concept of polarization to the notions of identification, alienation, and potential social conflict. Identification refers to the idea that an individual feels some degree of identification with those who are 'close' to him/her. Identification is thus an increasing function of the number of individuals who are in the same income class as that individual. The alienation function on the contrary characterizes the antagonism caused by income differences so that an individual will feel alienated from those who are 'far away' from him/her. While Esteban and Ray (1994), as well as Esteban et al. (2007), assumed that the number of groups was determined ex ante, Duclos et al. (2004) extended the analysis of polarization to the continuous case, letting the data determine the number of relevant groups and poles.

The focus of most empirical studies of bi-polarization and polarization was on the distribution of total income. There have however been a few attempts to decompose bipolarization and polarization indices by income sources (e.g., Araar 2008; Deutsch and Silber 2010) but the procedures are not very simple. More recently, Bárcena-Martín et al. (2017) proposed a simple matrix formulation to decompose the Foster and Wolfson bi-polarization index by income sources.

The main contribution of the present paper is to introduce a simple algorithm to compute the Esteban and Ray (1994) polarization index. We derive this algorithm from the simple matrix formulation suggested by Silber (1989) to compute the Gini index. We then show that, with such an

approach, it is easy to derive the contribution of various income sources (or explanatory variables in the case of an earnings function) to the degree of polarization of the distribution of total income.

Section 2 describes the algorithm allowing the simple computation of the ER index polarization index while Section 3 shows how such a formulation simplifies the decomposition of this index by income sources. Section 4 presents a simple empirical illustration and Section 5 concludes the paper.

2. Matrix Representation of the Esteban and Ray (1994) ER Polarization Index

The Esteban and Ray (1994) polarization index ER is expressed as

$$ER = \sum_{i=1}^{n} \sum_{j=1}^{n} v_i^{\beta} v_j |\mu_i - \mu_j| \tag{1}$$

where v_k is the relative population frequency of population subgroup k, μ_k the mean income[1] of group k and β a parameter which varies between 2 and 2.6 (see, Esteban and Ray 1994).

We can also write expression (1) as

$$ER = \sum_{i=1}^{n} \mu_i v_i \left(\sum_{j=1}^{i-1} v_j^{\beta} - \sum_{j=i+1}^{n} v_j^{\beta} \right) + \sum_{i=1}^{n} \mu_i v_i^{\beta} \left(\sum_{j=1}^{i-1} v_j - \sum_{j=i+1}^{n} v_j \right) \tag{2}$$

where the mean incomes μ_i are ranked by increasing values.

More generally, assuming n population subgroups, expression (2) becomes

$$ER = t'Gs + v'Gr = ER^A + ER^B \tag{3}$$

In (3), ER^A and ER^B are the two components of the ER index, t' is a (1 by n) row vector, written as $t' = \left[v_1^{\beta} \, v_2^{\beta} \ldots v_n^{\beta} \right]$, s is a (n by 1) column vector which, as row vector, would be written as $s' = [\mu_1 v_1 \; \mu_2 v_2 \ldots \; \mu_n v_n]$, v' is a (1 by n) row vector written as $v' = [v_1 \, v_2 \ldots v_n]$ and r is a (n by 1) column vector which, as a row vector would be expressed as $r' = \left[\left(\mu_1 v_1^{\beta} \right) \left(\mu_2 v_2^{\beta} \right) \ldots \left(\mu_n v_n^{\beta} \right) \right]$. G is a square n by n matrix, called G-matrix, whose typical element g_{ij} is equal to 0 if $i = j$, to -1 if $j > i$ and to +1 if $i > j$ (see, Silber 1989, for more details on this G-matrix[2]). It is important to stress that the elements $\mu_i v_i$ in vector s' and the elements $\left(\mu_i v_i^{\beta} \right)$ in vector r' have both to be ranked by decreasing values of the mean incomes μ_i.

Let τ' be a (1 by n) row vector, written as $\tau' = \left[\left(\frac{v_1^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}} \right) \ldots \left(\frac{v_i^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}} \right) \ldots \left(\frac{v_n^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}} \right) \right]$. Let also θ be a (n by 1) column vector of the income shares $\left(\frac{\mu_i v_i}{(\sum_{i=1}^{n} \mu_i v_i)} \right)$. In other words, if we call $\left(\frac{v_i^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}} \right)$ the 'identification modified population share' of population subgroup i, the expression $\tau' G \theta$ is a kind of Gini index comparing a priori shares which are the 'identification modified population shares' with a posteriori shares which are the actual income shares of the various population subgroups, the comparison being made via the linear operator G, the G-matrix.

Similarly, let η' be a (n by 1) row vector whose typical element η_i is written as $\eta_i = \left(\frac{\mu_i v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}} \right)$. η_i will be labeled the 'identification modified income share' of population subgroup i. The expression $v' G \eta$ is then a kind of Gini index, comparing a priori shares, the actual population shares, with a posteriori shares, the 'identification modified income shares' of the various population subgroups. This comparison is made again via the linear operator G, the G-matrix.

---

[1] Esteban and Ray (1994) refer to the natural logarithm of income rather than to income. We will make a somehow similar assumption by stating that the mean income of a given group refers in fact to its mean income relative to the mean income in the whole population. To simplify the notations, we do not introduce the population mean income in the formulations.

[2] As stressed already in Silber (1989), the first matrix formulation of the Gini index was proposed by Pyatt (1976).

Expression (3) is then rewritten as

$$ER = \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left[\tau' G\theta\right] + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left[v' G\eta\right] \quad (4)$$

In other words, the polarization index is equal to the corrected sum of two Gini-related indices. The first one compares the 'identification modified population shares' with the actual income shares of the different population subgroups. The second one compares the actual population shares with the 'identification modified income shares' of the different population subgroups. The first correction factor is equal to the product $\left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)$ while the second correction factor is equal to the product $\left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left(\sum_{i=1}^{n} v_i\right) = \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)$.

3. Decomposing the *ER* Index by Income Sources

Assume there are J income sources. The average income μ_i, in population subgroup i, may then be expressed as

$$\mu_i = \sum_{j=1}^{J} \mu_{ij} \quad (5)$$

so that expression (3) may also be written as

$$ER = t' G\left[\sum_{j=1}^{J} s_{.j}\right] + v' G\left[\sum_{j=1}^{J} r_{.j}\right] \quad (6)$$

where $s_{.j}$ is a (n by 1) column vector whose typical element s_{ij} is equal to $v_i \mu_{ij}$ while $r_{.j}$ is a (n by 1) column vector whose typical element r_{ij} is equal to $v_i^{\beta} \mu_{ij}$. Note that the elements s_{ij} in vector $s_{.j}$ and the elements r_{ij} in vector $r_{.j}$ have to be ranked by decreasing mean incomes μ_i.

We may then rewrite (6) as

$$ER = \sum_{j=1}^{J} D_j \quad (7)$$

where D_j, the contribution of income source j to the ER index, is expressed as

$$D_j = \left[t' G s_{.j} + v' G r_{.j}\right] \quad (8)$$

We could also express (8) as

$$D_j = \left[t' G\widetilde{s_{.j}} \frac{t' G s_{.j}}{t' G\widetilde{s_{.j}}} + v' G\widetilde{r_{.j}} \frac{v' G r_{.j}}{v' G\widetilde{r_{.j}}}\right] \quad (9)$$

where $\widetilde{s_{.j}}$ is a (n by 1) column vector whose typical elements $\widetilde{s_{ij}}$, which are equal to $v_i \mu_{ij}$, are ranked in descending order of μ_{ij}, while $\widetilde{r_{.j}}$ is a (n by 1) column vector whose typical elements $\widetilde{r_{ij}}$, which are equal to $v_i^{\beta} \mu_{ij}$, are ranked also in descending order of μ_{ij}.

Note however that

$$\left(t' G\widetilde{s_{.j}} + v' G\widetilde{r_{.j}}\right) = ER_j^A + ER_j^B = ER_j \quad (10)$$

where ER_j is the Esteban and Ray polarization index for income source j, ER_j^A and ER_j^B being its two components.

Let us also define two correlation measures, COR_j^A and COR_j^B, with

$$COR_j^A = \frac{t' G s_{.j}}{t' G\widetilde{s_{.j}}} \quad (11)$$

$$COR_j^B = \frac{v' G r_{.j}}{v' G\widetilde{r_{.j}}} \quad (12)$$

These correlation measures may evidently be positive or negative.

Combining expressions (7)–(12) we derive that

$$ER = \sum_{j=1}^{J} \left\{ \left[ER_j^A COR_j^A \right] + \left[ER_j^B COR_j^B \right] \right\} \quad (13)$$

We therefore conclude that, ceteris paribus,

- The higher ER_j^A, the higher the degree of polarization of the distribution of total income.
- The higher ER_j^B, the higher the degree of polarization of the distribution of total income.
- If COR_j^A is positive, the higher this correlation measure, the higher the degree of polarization of the distribution of total income. However, if it is negative, it will have a negative impact on the overall Esteban and Ray index *ER*.
- Similarly, if COR_j^B is positive, the higher this correlation measure, the higher the degree of polarization of the distribution of total income. However, if it is negative, it will have a negative impact on the overall Esteban and Ray index *ER*[3].

4. A Short Empirical Illustration

In this section, we present a simple empirical illustration, based on EU data from the European Union Statistics on Income and Living Conditions (EU-SILC) data set for the 2016 wave (EUROSTAT 2016). EU-SILC is an international database that consists of comparable, country-specific data. We analyze polarization in the 17 countries with data available for 2016: AT (Austria), BE (Belgium), BG (Bulgaria), EE (Estonia), EL (Greece), ES (Spain), FR (France), HR (Croatia), HU (Hungary), LT (Lithuania), LV (Latvia), PL (Poland), PT (Portugal), RO (Romania), RS (Serbia), SE (Sweden), and SI (Slovenia). The units of analysis are the individuals and the unit of measurement is the household. The measure of income is the total disposable household income. Since a given level of household income corresponds to a different standard of living, depending on the size and composition of the household, we adjust incomes for differences in household size and composition using the "modified OECD" equivalence scale[4]. The latter assigns a value of 1 to the first adult in the household, 0.5 to each remaining adult, and 0.3 to each person younger than 14.

Disposable income includes net income from work, other private income not related to work, pensions and other social transfers. Net money income includes all income sources received by the household and by each of its current members in the year preceding the survey. Social insurance contributions, pay-as-you-earn taxes, and non-money income are not included in this definition of income.

The decomposition of the *ER* polarization index by income sources is based on three income sources:

1. Benefits (benefits) that include: old-age and survivor' benefits, unemployment benefits, sickness benefits, disability benefits, education-related allowances, family/children related allowances, social exclusion not classified elsewhere, housing allowances
2. Income from rental of a property or land, interest, dividends, profit from capital investments in unincorporated business (property and interest)
3. Income available before including sources 1 and 2 (income before)

---

[3] Expression (13) reminds us of the decomposition of the Gini index by income sources (see, Lerman and Yitzhaki 1985) where the contribution of an income source to the overall Gini index is a function of the share of this source in total income, of the Gini index of this source and of the Gini-correlation between this source and total income. In (13) the contribution of an income source to the overall ER index is a function of the two components of the ER index for this source, and of two correlation measures. However the share of the source does not appear. In Appendix A, we provide a more detailed decomposition where the parallel with the traditional decomposition of the Gini index by income sources becomes evident.

[4] For a survey of equivalence scales and related income distribution issues, and some comparisons of scale relativities, see Coulter et al. (1992).

Table A1 in the Appendix A gives, for each of these countries, the average value of these income sources, the average total income and the population size.

Table 1 refers to data in Euros. We give there the value of the *ER* index when the parameter β is equal to 2.5 and when it is equal to 1 (Gini related measure[5]). We also computed, as suggested by Esteban and Ray (1994), the *ER* index with these two values of the parameter β, for the case where the logarithm of income rather than income was the variable under study. Table 1 gives also, when income and not the logarithm of income is used, the relative contributions of the different income sources, to the *ER* index. It appears that the most important (relative) contribution to the value of the *ER* index is that of income before transfer (62.4%) while this source has a share in total income of 70.7%. On the contrary, benefits and 'property income and interest' have a higher relative contribution to the *ER* index (respectively 25.4% and 12.2%) than their share in total income (23.2% and 6.1%). We may also observe that the contributions of these sources to the Gini-related index (parameter β equal to 1) is quite similar to their contributions to the *ER* index (65.6, 24.9, and 9.5%). They actually lie between their contributions to the average total income and to the *ER* index.

When introducing the logarithm of income into the formulation of the ER index with $\beta = 2.5$, we observe that this index is quite small (0.045) when compared to its value (0.577) when $\beta = 1$.

Table 1. Contributions of the income sources to the *ER* index (based on income data in Euros).

Measure Computed	Value for Total Income	Relative Contribution of Income Before	Relative Contribution of Benefits	Relative Contribution of Property Income and Interest
Average income with absolute contribution of income sources	15,634	11,060	3626	948
Average income with relative contribution of income sources	100%	70.70%	23.20%	6.10%
ER with parameter β equal to 2.5 computed on basis of relative incomes (relative contributions of income sources)	0.038	62.4%	25.4%	12.2%
ER with parameter β equal to 1 computed on basis of relative incomes (relative contributions of income sources)	0.645	65.6%	24.9%	9.5%
ER with parameter β equal to 2.5 and logarithms of incomes	0.045			
ER with parameter β equal to 1 (like Gini) and logarithms of incomes	0.577			

Table 2 is similar to Table 1 but here all the computations are derived from PPP income data. While the relative contributions of the three income sources to the average EU PPP income (on the basis of the countries for which data were available) are quite similar to those presented in Table 1, the computation of the *ER* index and of the contributions of the income sources to this index show a somehow different picture. When the parameter β is equal to 2.5, it appears that the *ER* index is lower than in Table 1, whether this index is derived from income data or from the logarithm of incomes. What is more interesting is that there is an important decrease in the relative contribution

[5] When, in expression (1), we divide the income data by the average income and assume that $\beta = 1$, *ER* will equal to twice the traditional Gini index. What is called the absolute Gini index, is actually the product of the Gini index by the mean, so that when $\beta = 1$ and we use absolute incomes and not relative incomes in (1) *ER* will be equal to twice the absolute Gini index.

of income before transfer (from 62.4% to 54.4%) when $\beta = 2.5$ and from 65.6% to 59.6% when $\beta = 1$. On the contrary, there is an increase in the relative contribution of benefits: from 25.4% to 28.6% when $\beta = 2.5$ and from 24.9% to 27.7% when $\beta = 1$. A similar increase is observed for property income and interest since the relative contribution rises from 12.2% to 17.0% when $\beta = 2.5$ and from 9.5% to 12.8% when $\beta = 1$. In short, when using PPP rather than current data, polarization and inequality turn out to be smaller, but the relative contribution of benefits and property income and interest to polarization and inequality rises.

Table 2. Contributions of the income sources to the *ER* index (based on PPP income data).

Measure Computed	Value for Total Income	Relative Contribution of Income Before	Relative Contribution of Benefits	Relative Contribution of Property Income and Interest
Average income with absolute contribution of income sources	17,048	12,233	3892	924
Average income with relative contribution of income sources	100%	71.7%	22.8%	5.4%
ER with parameter β equal to 2.5 computed on basis of relative incomes (relative contributions of income sources)	0.024	54.4%	28.6%	17.0%
ER with parameter β equal to 1 computed on basis of relative incomes (relative contributions of income sources)	0.413	59.6%	27.7%	12.8%
ER with parameter β equal to 2.5 and logarithms of incomes	0.026			
ER with parameter β equal to 1 (like Gini) and logarithms of incomes	0.478			

5. Concluding Comments

This paper has shown how it is possible to express the Esteban and Ray (1994) ER index in matrix form. Such a formulation greatly simplifies the decomposition of this index by income sources. We gave a simple empirical illustration showing that this breakdown gives useful information as to the impact of the different income sources on the polarization of incomes. This illustration was based first on income data in Euros and then on PPP income data. We could also apply the proposed breakdown to an analysis of the polarization of the distribution of wages or earnings. If we estimate a traditional earnings function, we could then easily derive the contribution to the polarization of wages of the explanatory variables of such a function. Indeed, we intend to explore these issues in future empirical work.

Acknowledgments: Elena Bárcena-Martín gratefully acknowledges the financial support provided by the University of Malaga.

Author Contributions: Both authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Similarity between the Decomposition by Income Sources of the Gini Index and of the ER Index

Remember that expression (4) is written as

$$ER = \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left[\tau' G \theta\right] + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left[v' G \eta\right] \tag{A1}$$

where τ' is a (1 by n) row vector, written as $\tau' = \left[\left(\frac{v_1^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}}\right) \dots \left(\frac{v_i^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}}\right) \dots \left(\frac{v_n^{\beta}}{\sum_{i=1}^{n} v_i^{\beta}}\right)\right]$, θ a (n by 1) column vector of the income shares $\left(\frac{\mu_i v_i}{(\sum_{i=1}^{n} \mu_i v_i)}\right)$, η' a (n by 1) row vector whose typical element η_i is written as $\eta_i = \left(\frac{\mu_i v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}}\right)$ and v' a row vector of the population shares.

We can rewrite (A1) as

$$ER = \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left[\tau' G\left(\sum_{j=1}^{J} \theta_j\right)\right] + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left[v' G\left(\sum_{j=1}^{J} \eta_j\right)\right] \tag{A2}$$

where

$$\theta_j = \left(\frac{\mu_{ij} v_i}{(\sum_{i=1}^{n} \mu_i v_i)}\right) = \left(\frac{\mu_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)\left(\frac{(\sum_{i=1}^{n} \mu_{ij} v_i)}{(\sum_{i=1}^{n} \mu_i v_i)}\right) \tag{A3}$$

$$\eta_j = \left(\frac{\mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}}\right) = \left(\frac{\mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}\right)\left(\frac{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}}\right) \tag{A4}$$

Given that the G-matrix is a linear operator we then derive that

$$ER = \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left[\sum_{j=1}^{J} \tau' G \theta_j\right] + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left[\sum_{j=1}^{J} v' G \eta_j\right]$$

$$\leftrightarrow ER = \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left[\sum_{j=1}^{J} \tau' G\left\{\left(\frac{\mu_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)\left(\frac{(\sum_{i=1}^{n} \mu_{ij} v_i)}{(\sum_{i=1}^{n} \mu_i v_i)}\right)\right\}\right] + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left[\sum_{j=1}^{J} v' G\left\{\left(\frac{\mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}\right)\left(\frac{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}}\right)\right\}\right]$$

$$\leftrightarrow ER = \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left\{\left[\sum_{j=1}^{J}\left(\frac{(\sum_{i=1}^{n} \mu_{ij} v_i)}{(\sum_{i=1}^{n} \mu_i v_i)}\right)\tau' G\left(\frac{\mu_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)\right]\right\} + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left\{\left[\sum_{j=1}^{J}\left(\frac{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}}\right) v' G\left(\frac{\mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}\right)\right]\right\}$$

If instead of ranking the incomes μ_{ij} by decreasing values of the incomes μ_i, we rank them by decreasing values of the incomes μ_{ij}, and call $\widetilde{\mu}_{ij}$ this re-ordered vector, we end up with

$$\begin{aligned} ER =& \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\left\{\left[\sum_{j=1}^{J}\left(\frac{(\sum_{i=1}^{n} \mu_{ij} v_i)}{(\sum_{i=1}^{n} \mu_i v_i)}\right)\left[\tau' G\left(\frac{\widetilde{\mu}_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)\right]\right]\left(\left(\frac{\mu_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right) / \left(\frac{\widetilde{\mu}_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)\right)\right\} \\ &+ \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\left\{\left[\sum_{j=1}^{J}\left(\frac{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}}\right)\left[v' G\left(\frac{\widetilde{\mu}_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}\right)\right]\right]\left(\left(\frac{\mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}\right) / \left(\frac{\widetilde{\mu}_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}\right)\right)\right\} \\ \leftrightarrow ER =& \left(\sum_{i=1}^{n} v_i^{\beta}\right)\left(\sum_{i=1}^{n} \mu_i v_i\right)\sum_{j=1}^{J} \alpha_j \beta_j \gamma_j + \left(\sum_{i=1}^{n} \mu_i v_i^{\beta}\right)\sum_{j=1}^{J} \lambda_j \nu_j \rho_j \end{aligned}$$

with

$$\alpha_j = \left(\frac{(\sum_{i=1}^{n} \mu_{ij} v_i)}{(\sum_{i=1}^{n} \mu_i v_i)}\right)$$

$$\beta_j = \tau' G\left(\frac{\widetilde{\mu}_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)$$

$$\gamma_j = \left(\left(\frac{\mu_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right) / \left(\frac{\widetilde{\mu}_{ij} v_i}{(\sum_{i=1}^{n} \mu_{ij} v_i)}\right)\right)$$

$$\lambda_j = \left(\frac{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_i v_i^{\beta}} \right)$$

$$\nu_j = v' G \left(\frac{\widetilde{\mu}_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}} \right)$$

$$\rho_j = \left(\left(\frac{\mu_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}} \right) / \left(\frac{\widetilde{\mu}_{ij} v_i^{\beta}}{\sum_{i=1}^{n} \mu_{ij} v_i^{\beta}} \right) \right)$$

where α_j and λ_j are similar to income shares, β_j and ν_j are components of the *ER* index for income source j and γ_j and ρ_j are 'correlation measures'.

In other words, we have here quite a similar decomposition to that proposed by Lerman and Yitzhaki (1985) for the Gini index.

Table A1. Database.

Country	Mean Total Income	Income Before %	Benefits %	Property and Interest %	Total Population
AT	26,662.48	70.0%	27.3%	2.7%	7,963,391
BE	24,520.2	74.0%	24.6%	1.3%	9,319,177
BG	4164.49	76.4%	21.9%	1.7%	6,235,715
EE	11,043.97	82.3%	16.6%	1.1%	1,113,681
EL	9161.76	75.7%	19.8%	4.5%	8,092,137
ES	16,370.34	71.9%	24.4%	3.6%	42,446,793
FR	25,730.84	65.1%	24.4%	10.5%	55,793,599
HR	6663.01	80.3%	18.3%	1.4%	3,225,726
HU	5474.74	75.6%	23.2%	1.2%	8,332,493
LT	7742.34	81.6%	16.6%	1.8%	2,417,930
LV	8135.06	80.3%	18.6%	1.2%	1,708,676
PL	6912.37	80.9%	18.2%	0.9%	32,623,207
PT	10,892.61	79.1%	17.7%	3.2%	8,183,986
RO	2850.82	82.4%	17.5%	0.1%	15,991,057
RS	3214.21	74.2%	25.0%	0.8%	5,432,579
SE	29,761.2	77.6%	17.6%	4.8%	7,647,944
SI	13,678.07	73.8%	23.5%	2.7%	1,794,388

Country codes: AT (Austria), BE (Belgium), BG (Bulgaria), EE (Estonia), EL (Greece), ES (Spain), FR (France), HR (Croatia), HU (Hungary), LT (Lithuania), LV (Latvia), PL (Poland), PT (Portugal), RO (Romania), RS (Serbia), SE (Sweden), SI (Slovenia).

References

Araar, Abdelkrim. 2008. On the Decomposition of Polarization Indices: Illustrations with Chinese and Nigerian Household Surveys. Working Paper 08-06. Centre Inter-universitaire sur le Risque, les Politiques Economiques et l'Emploi, Université Laval, Québec, Canda.

Bárcena-Martín, Elena, Joseph Deutsch, and Jacques Silber. 2017. On the Decomposition of the Foster and Wolfson Bi-Polarization Index by Income Sources. *Review of Income and Wealth*. [CrossRef]

Coulter, Fiona A. E., Frank A. Cowell, and Stephen P. Jenkins. 1992. Equivalence scales relativities and the extent of inequality and poverty. *Economic Journal* 102: 1067–82. [CrossRef]

Deutsch, Joseph, and Jacques Silber. 2010. Analyzing the Impact of Income Sources on Changes in Bi-Polarization. In *The Measurement of Individual Well-Being and Group Inequalities: Essays in Memory of Z. M. Berrebi*. Edited by Joseph Deutsch and Jacques Silber. London: Routledge Economics, Taylor and Francis Group, pp. 127–52.

Duclos, Jean-Yves, Joan Esteban, and Debraj Ray. 2004. Polarization: Concepts, Measurement, Estimation. *Econometrica* 72: 1737–72. [CrossRef]

Esteban, Joan-Maria, and Debraj Ray. 1994. On the Measurement of Polarization. *Econometrica* 62: 819–51. [CrossRef]

Esteban, Joan, Carlos Gradín, and Debraj Ray. 2007. An Extension of a Measure of Polarization, with an application to the income distribution of five OECD countries. *Journal of Economic Inequality* 5: 1–19. [CrossRef]

Lerman, Robert I., and Shlomo Yitzhaki. 1985. Income Inequality Effects by Income Sources: A New Approach and Applications to the United States. *Review of Economics and Statistics* 67: 151–56. [CrossRef]

Pyatt, Graham. 1976. On the Interpretation and Disaggregation of Gini coefficients. *Economic Journal* 86: 243–55. [CrossRef]

Silber, Jacques. 1989. Factors Components, Population Subgroups and the Computation of the Gini Index of Inequality. *The Review of Economics and Statistics* LXXI: 107–15. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Decomposing the Bonferroni Inequality Index by Subgroups: Shapley Value and Balance of Inequality

Giovanni M. Giorgi [1,*] and Alessio Guandalini [2]

[1] Department of Statistical Sciences, "Sapienza" University of Rome, Piazzale Aldo Moro 5, Rome 00185, Italy
[2] Italian National Institute of Statistics—ISTAT, Via Cesare Balbo 16, Rome 00184, Italy; alessio.guandalini@istat.it
* Correspondence: giovanni.giorgi@uniroma1.it; Tel.: +39-06-4991-0488

Received: 11 December 2017; Accepted: 23 March 2018; Published: 2 April 2018

Abstract: Additive decomposability is an interesting feature of inequality indices which, however, is not always fulfilled; solutions to overcome such an issue have been given by Deutsch and Silber (2007) and by Di Maio and Landoni (2017). In this paper, we apply these methods, based on the "Shapley value" and the "balance of inequality" respectively, to the Bonferroni inequality index. We also discuss a comparison with the Gini concentration index and highlight interesting properties of the Bonferroni index.

Keywords: inequality measurement; Bonferroni index; Gini concentration ratio; decomposition methods; Shapley value; balance of inequality; complex survey data

JEL Classification: D63; C71; I32

1. Introduction

Carlo Emilio Bonferroni (1930) proposed the inequality index B as an alternative to the Gini index R, also referred to as the concentration ratio (Gini 1914). For about half a century, B remained almost forgotten because it was ostracized by Corrado Gini and his followers, who tried to prevent any measures of inequality from overshadowing the concentration ratio R (Giorgi 1998). De Vergottini (1950) proposed an interesting and general formula that nests Bonferroni and Gini indices as special cases.

In the last two decades, B has been revalued and studied for its interesting features. Piesch (1975) and Nygård and Sandström (1981) were the first to investigate B in depth. New and interesting interpretations and extensions of B have been just recently proposed: its welfare implications have been studied by Benedetti (1986), Aaberge (2000), Chakravarty (2007) and Bárcena-Martin and Silber (2013). Giorgi and Crescenzi (2001c) proposed a poverty measure based on B, while other socio-economic aspects have been studied by Bárcena-Martin and Olmedo (2008), Silber and Son (2010), Bárcena-Martin and Silber (2011, 2013), and Imedio Olmedo et al. (2012). The Bonferroni index has also been investigated in fuzzy and reliability frameworks (Giordani and Giorgi 2010; Giorgi and Crescenzi 2001b) and, in particular cases, a Bayesian estimation is followed (Giorgi and Crescenzi 2001a).

An important topic in the literature on inequality measures entails their decomposability. Many contributions are related to the decomposition of R (for a deep investigation see, e.g., Kakwani 1980; Nygård and Sandström 1981; Giorgi 2011a). Tarsitano (1990) introduced several standard results that can be used for the decomposition of B, while Bárcena-Martin and Silber (2013) derived an algorithm that greatly simplifies such a decomposition.

In this field, two main lines of research can be distinguished: decomposition by income sources and by population subgroups. The former has been widely treated, while less attention has been paid to the latter (Giorgi 2011a). The reason lies in the difficulties we face when trying to additively decompose (as in the analysis of variance) inequality indices, including R and B. To overcome such a drawback when R is entailed, Deutsch and Silber (2007) used the so-called "Shapley value", while Di Maio and Landoni (2017) suggested the "balance of inequality" (BOI).

In the present paper, we detail how the Bonferroni index can be decomposed using these methods. We further discuss interesting similarities and differences between R and B and propose a deeper investigation of some properties of B.

The paper is organized as follows: in Section 2, the main properties of Gini and Bonferroni indices are discussed. A brief overview on the inequality indices' decomposition is given in Section 3, while the so-called "Shapley method" and "balance of inequality" (BOI) are detailed in Sections 4 and 5, respectively. We also extend the BOI to provide a decomposition of B. In Section 6, R and B are compared on income data drawn from the 2015 Italian component of the European Survey on Income and Living Conditions (It-SILC). The differences between the two decompositions and the two indices are highlighted in Section 7.

2. The Gini and the Bonferroni Inequality Index

2.1. The Gini Concentration Index

The Gini concentration ratio (Gini 1914), also referred to as the Gini coefficient or the Gini index, is probably the most used index to measure inequality in income distributions. Simplicity, fulfillment of general properties, useful decompositions, the links with the Lorenz curve (Lorenz 1905) and the mean difference (Gini 1912) are just few of the reasons of its widespread use and longevity (see, e.g., Giorgi (1990, 1993, 1998, 1999, 2005, 2011b)).

Among the several ways we may use to define the Gini index (see Giorgi 1992; Yitzhaki 1998), the most useful, for the present purpose, is

$$R = \frac{2\sum_{i=1}^{N} x_i(i-1)}{(N-1)t_x} - 1 \qquad (1)$$

$$0 \leq R \leq 1$$

$$= \frac{2\sum_{i=1}^{N} i\, x_i}{(N-1)t_x} - \frac{(N+1)}{(N-1)} \qquad (2)$$

where N is the population size, and i is the rank, within the observed population, for the generic recipient, arranged in non-decreasing income values. Furthermore, x_i is the income earned by the i-th recipient and $t_x = \sum_{i=1}^{N} x_i$ is the total income in the whole population.

The Gini concentration index is linked to the Lorenz curve (Figure 1). In the discrete case, the Lorenz curve is the polygonal line connecting points with coordinates given by the cumulative proportion of recipients, arranged in non-decreasing values of income, $p_i = i/N$, and the corresponding share of income, $q_i = \sum_{j=1}^{i} x_j/t_x$. In the case of perfect equality, the Lorenz curve corresponds to the egalitarian line. In the case of maximum concentration, the Lorenz curve is defined by linking coordinate points (0,0), ($\frac{N-1}{N}$,0), (1,1).

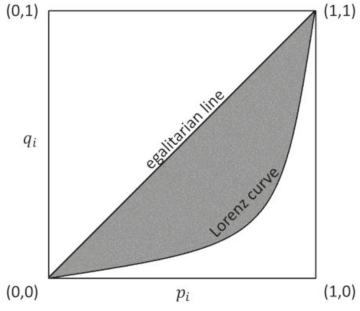


Figure 1. An example of the Lorenz curve in the continuous case (i.e., N goes to infinity).

The Gini concentration index is equal to the ratio between the Lorenz area—the area between the Lorenz curve and the egalitarian line—and the Lorenz area in case of maximum concentration—the area of the triangle defined by the points (0,0), ($\frac{N-1}{N}$,0), (1,1), (Nygård and Sandström 1981, pp. 266–71). As N goes to infinity, the quantity $\frac{N-1}{N}$ goes to 1 and the Lorenz area in the case of maximum concentration approaches 1/2. Then, R is twice the area between the Lorenz curve and the egalitarian line (Nygård and Sandström 1981, p. 240).

2.2. The Bonferroni Inequality Index

Bonferroni (1930) defined the inequality index as a function of partial means:

$$B = \frac{1}{N-1} \sum_{i=1}^{N} \frac{(\mu - \mu_i)}{\mu}, \tag{3}$$

where $0 \leq B \leq 1$, and

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i \quad \mu_i = \frac{1}{i} \sum_{j=1}^{i} x_j \quad i = 1, 2, \ldots, N$$

denote the general and the partial means for units sorted in non-decreasing order with respect to the variable of interest, X.[1]

The B index gives a higher weight to units with lower income (see, e.g., De Vergottini 1950, pp. 318–19; Pizzetti 1951, p. 302). For this reason, B is more sensitive to lower levels in the distribution (see, e.g., Giorgi and Mondani 1995).

The Bonferroni index is linked to the Bonferroni curve (Figure 2) which is obtained by plotting the cumulative proportion of recipients, arranged in non-decreasing values of income, versus the corresponding ratio between partial mean and total mean (μ_i/μ).

[1] In expression (3) the summation is limited to $N-1$ and then divided by $N-1$. This formulation is different from the one used in other papers mentioned in the Introduction (where the summation is up to N the division by N is used). Of course, increasing N, $1/(N-1) \approx 1/N$ and the last term in the summation is null.

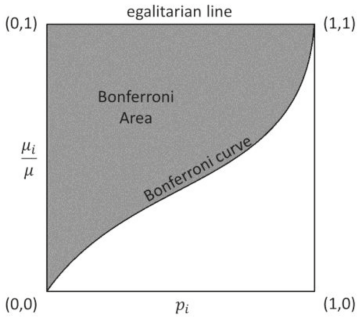


Figure 2. An example of the Bonferroni curve in the continuous case (i.e., N goes to infinity).

The polygonal line joining the points $(p_i, \mu_i/\mu)$ is the Bonferroni curve. If all the recipients in the population have the same income (i.e., equal to μ), the Bonferroni curve coincides with the egalitarian line that joins the coordinate points (0,1), (1,1). If just a recipient owns the total amount of X, the Bonferroni curve is the broken line joining the points (0,0), $(\frac{N-1}{N},0)$, $(\frac{N-1}{N},1)$.

The value of the Bonferroni index is equal to the ratio between the Bonferroni area—the area between the Bonferroni curve and the egalitarian line—and the Bonferroni area in the case of maximum concentration—the area of the quadrangle defined by the points (0,0), $(\frac{N-1}{N},0)$, $(\frac{N-1}{N},1)$, (0,1). As N goes to infinity, the quantity $\frac{N-1}{N}$ goes to 1 and the Bonferroni area in case of maximum concentration is equal to 1. Then, the value of B coincides with the Bonferroni area (Giorgi and Crescenzi 2001b, pp. 572–73).

3. A Brief Overview on Inequality Index Decomposition

When we consider the decomposition of inequality indices, two main lines of research can be distinguished: decomposition by income sources and by population subgroups (for a comprehensive survey on the subject see, e.g., Giorgi 2011a).

The decomposition by income sources is based on the hypothesis that the total income is the sum of several components, such as wages, salaries, capital incomes, etc. Therefore, the contribution of each source to the overall inequality can be identified. The decomposition by income sources is appealing since the inequality indices can be exactly decomposed into separate components, each one referring to a given factor. Fields (1979a, 1979b) derived the contribution of each source to R via so called Factor Inequality Weight (FIW). With slight changes, this method can be adapted to decompose B as follows:

$$B = \sum_{j=1}^{k} h_j w_j B_j$$

where μ_j, $h_j = \mu_j/\mu$ and B_j are, respectively, the mean income, the share and the value of B computed for the j-th factor (see, e.g., Tarsitano 1990, p. 236). The w_j is the weight of the j-th source which can be referred to as the Bonferroni correlation. In fact, it has the same meaning of the Gini correlation in FIW decomposition. The Bonferroni correlation reflects the degree of concordance between the log-rank ordering of units with respect to the j-th income source and the corresponding log-rank order for the total income. In other words, the overall inequality, measured through B, depends on the degree of inequality in the distribution of each factor (B_j), the importance of the factor on the total income (h_j) and the amount of agreement between the different rankings (w_j).

The decomposition by population subgroups aims at exploring the contribution of individual features such as age, sex, level of education, geographical area, etc., to total inequality (for a deeper investigation on this topic seeDeutsch and Silber 1999; Mussard et al. 2006). A first attempt has been proposed by Bhattacharya and Mahalanobis (1967), who tried to decompose R by subgroups via an

approach based on the analysis of variance. However, R cannot be additively decomposed into the sum of between and within components. Mehran (1975) showed that R can be decomposed into the sum of within and across components. The difference between the across and the between components is in the interaction component; this is "a measure of the extent of income domination of one group over the other apart from the differences between their mean incomes" (see also Ferrari and Rigo 1987).

As the concentration ratio R, the Bonferroni index B can be completely decomposed by the sum of three terms:

$$B = B_w + B_b + B_i \tag{4}$$

where B_w is the within component, B_b is the between component and B_i is the interaction component that accounts for the degree of overlap between the income distributions in the different subgroups (for this reason it is also referred to as the overlapping component)[2]. Therefore, also B cannot be additively decomposed (see, e.g., Shorrocks 1980).

4. The Shapley Decomposition

Deutsch and Silber (2007) used the Shapley decomposition introduced, in this field, by Shorrocks (1999), to solve the problem of additive decomposition of R by population subgroups. They derived the impact of four components: inequality within subgroups (w), inequality between subgroups (b), ranking (r) and relative size in each subgroup (n).

The Shapley decomposition is based on the well-known concept of Shapley value in cooperative game theory (Shapley 1953). The idea of the Shapley value is to compute the value of a function considering all the possible combinations of factors. When such a decomposition is applied to inequality indices, and the factors are considered as symmetrical, it allows to derive the expected marginal contribution of each factor to inequality. Moreover, the contributions sum exactly to the amount of the inequality index considered (Shorrocks (1999, 2013)). To decompose B, we consider the same factors (i.e., w, b, r, n), used by Deutsch and Silber (2007) for R.

Let us assume that a population P is partitioned into J subgroups s_j ($j = 1, \ldots, J$) where x_{ji} is the income of the i-th recipient ($i = 1, \ldots, N_j$) in subgroup $j = 1, \ldots, J$. A given inequality measure I (for instance, R or B) can be seen as a function of the observed incomes, $I = f\left(x_{11}, \ldots, x_{1N_1}, \ldots, x_{j1}, \ldots, x_{jN_j}, \ldots, x_{J1}, \ldots, x_{JN_J}\right)$.

In the general case, we may consider within subgroups inequality ($x_{ji} \neq \mu_j$), between subgroups inequality ($\mu_j \neq \mu$) and differences in both the size among the subgroups ($f_j \neq 1/J$, where $f_j = \frac{N_j}{N}$) and the rank of recipients $\left(r = r_{ij}\right)$. Therefore, the overall inequality can be written as a function of such factors:

$$I\left(\left(x_{ji} \neq \mu_j\right), \left(\mu_j \neq \mu\right), \left(f_j \neq 1/J\right), \left(r = r_{ij}\right)\right)$$

The Shapley decomposition may help us derive the marginal impact of each factor measuring the difference in the value of the inequality index corresponding to the observed situation and the reference one, where the income does not change with the factor. Just to give an example, the impact of within subgroups inequality (w), is derived by comparing the situations where the incomes of recipients in a given subgroup are different ($x_{ji} \neq \mu_j$), to the case when all the recipients in that subgroup have the same income ($x_{ji} = \mu_j$). To compute the impact of inequality between subgroups (b), we compare the case when the mean of incomes is different between subgroups ($\mu_j \neq \mu$) and the case when the average income is constant across subgroups ($\mu_j = \mu$). To obtain $\mu_j = \mu$ a kind of standardization is applied and x_{ji} is replaced by $x_{ji}\,\frac{\mu}{\mu_j}$. To measure the effect of the differences in size (n), we compare the case when the subgroups have different sizes ($f_j \neq 1/J$) to the case when the sizes are equal ($f_j = 1/J$). To make the subgroups have the same size, the least common multiple (lcm) is

[2] For the expressions of B_w, B_b and B_i in the case of income classes, see Tarsitano (1990), while for the matrix decomposition of B see Bárcena-Martin and Silber (2013).

calculated for the sizes of the analyzed subgroups and the values x_{ji} are repeated *lcm* times; this leads to equality in size between the subgroups. When applying such an approach to B, the objection is usually raised that B, as opposed to R, does not satisfy the Dalton (1925) principle of being replication invariant. However, using the simulation study reported in Appendix A, we may show that the effect of replications becomes negligible for B when the population size is greater than 1000 units. According to this feature, B can be defined as being an 'asymptotically replication invariant'.

Finally, to derive the effect of ranking (r), we compare the case when the recipients are sorted by their income $\left(r = r_{ij}\right)$ to the case when we first sort the subgroups on the basis of their average income, μ_j, and then the recipients by their income within each subgroup $\left(r = r_{ji}\right)$.

The marginal impact (SV) of each factor on the generic index I (either R or B) can be derived by computing the following weighted means of the index when, from time to time, the effect of components is removed.

$$\begin{aligned} SV_w = {} & \tfrac{1}{4}(I - I_w) + \tfrac{1}{12}[(I_b - I_{wb}) + (I_n - I_{wn}) + (I_r - I_{wr})] \\ & + \tfrac{1}{12}[(I_{bn} - I_{wbn}) + (I_{br} - I_{wbr}) + (I_{rn} - I_{wrn})] + \tfrac{1}{4}(I_{bnr} - I_{wbnr}) \end{aligned} \tag{5}$$

$$\begin{aligned} SV_b = {} & \tfrac{1}{4}(I - I_b) + \tfrac{1}{12}[(I_w - I_{wb}) + (I_n - I_{bn}) + (I_r - I_{br})]SV_b \\ & + \tfrac{1}{12}[(I_{wn} - I_{wbn}) + (I_{wr} - I_{wbr}) + (I_{rn} - I_{bnr})] + \tfrac{1}{4}(I_{wnr} - I_{wbnr}) \end{aligned} \tag{6}$$

$$\begin{aligned} SV_n = {} & \tfrac{1}{4}(I - I_n) + \tfrac{1}{12}[(I_w - I_{wn}) + (I_b - I_{bn}) + (I_r - I_{nr})]SV_n \\ & + \tfrac{1}{12}[(I_{wb} - I_{wbn}) + (I_{wr} - I_{wnr}) + (I_{br} - I_{bnr})] + \tfrac{1}{4}(I_{wbr} - I_{wbnr}) \end{aligned} \tag{7}$$

$$\begin{aligned} SV_r = {} & \tfrac{1}{4}(I - I_r) + \tfrac{1}{12}[(I_w - I_{wr}) + (I_b - I_{br}) + (I_n - I_{nr})]SV_r \\ & + \tfrac{1}{12}[(I_{wb} - I_{wbr}) + (I_{wn} - I_{wnr}) + (I_{bn} - I_{bnr})] + \tfrac{1}{4}(I_{wbn} - I_{wbnr}) \end{aligned} \tag{8}$$

In expressions (5)–(8) by the subscript of I we denote the factor that has been removed. For instance, I_w is the index computed when the component of within inequality (w) has been removed, that is $I_w = I\big(\left(x_{ji} = \mu_j\right), \left(\mu_j \neq \mu\right), \left(f_j \neq 1/J\right), \left(r = r_{ij}\right)\big)$. Furthermore, $I_{wbr} = I\big(\left(x_{ji} = \mu_j\right), \left(\mu_j = \mu\right), \left(f_j \neq 1/J\right), \left(r = r_{ji}\right)\big)$ is the index computed when component of within inequality (w), between inequality (b) have been removed and the recipients are ranked first by the average income of the subgroup they belong and then with respect to their income[3].

A Numerical Illustration

To illustrate, we consider a population composed by 10 recipients with income 2, 6, 10, 18, 20, 25, 30, 50, 55, and 84. Let us assume that recipients with income 2, 6 and 25 belong to subgroup A, those with income 10, 20 and 84 to subgroup B and, last, those with income 18, 30, 50 and 55 to subgroup C.

Since this is just an illustrative example of the application of the Shapley decomposition, the replication invariance principle is overlooked. We should remark that, in some cases, when we remove w, the corresponding value of B can be negative. It occurs when there is a negative correlation between mean income and mean rank (Frick and Goebel 2007, p. 10). In fact, in these extreme cases, when sorting the income distribution in decreasing order, as shown by Rao (1969, p. 245), R is equal to $-R$, and the same occurs for B.

Table 1 shows all the scenarios obtained by removing factors separately, in pairs, in set of three and all together. Furthermore, the corresponding income distribution and the values of R and B are also presented. We report in Table 2 the marginal contributions for each factor (SV) derived using expressions (5)–(8).

[3] In expressions (5)–(8), $I_{wbnr} = 0$ because all the inequality factors have been removed.

Table 1. Gini concentration ratio (R) and Bonferroni inequality index (B) in different scenarios according to factors that have been removed. Illustrative example on the income of 10 recipients from three different subgroups: A = {2, 6, 25} and B = {10, 20, 84}, C = {18, 30, 50, 55}.

Removed Factor		Income Distribution	I	
			R	B
1	–	2, 6, 10, 18, 20, 25, 30, 50, 55, 84	0.490	0.609
2	*w*	11, 11, 38, 38.25, 38, 11, 38.25, 38.25, 38.25, 38	0.151	0.252
3	*b*	5.45, 16.36, 7.89, 14.12, 15.79, 68.18, 23.53, 39.22, 43.14, 66.32	0.360	0.475
4	*n*	2, 2, 2, 2, 6, 6, 6, 10, 10, 10, 10, 18, 18, 18, 20, 20, 20, 20, 25, 25, 25, 25, 30, 30, 30, 50, 50, 50, 55, 55, 55, 84, 84, 84, 84	0.474	0.612
5	*r*	2, 6, 25, 10, 20, 84, 18, 30, 50, 55	0.333	0.481
6	*wb*	30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30	0.000	0.000
7	*wn*	11, 11, 11, 11, 11, 11, 11, 11, 11, 38, 38, 38, 38.25, 38.25, 38.25 38, 38, 38, 38, 11, 11, 11, 11, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38, 38, 38	0.167	0.283
8	*wr*	11, 11, 11, 38, 38, 38, 38.25, 38.25, 38.25, 38.25	0.212	0.331
9	*bn*	5.45, 5.45, 5.45, 5.45, 16.36, 16.36, 16.36, 16.36, 27.27, 7.89, 7.89, 7.89, 14.12, 14.12, 14.12, 15.79, 15.79, 15.79, 15.79, 68.18, 68.18, 68.18, 68.18, 23.53, 23.53, 23.53, 39.22, 39.22, 39.22, 43.14, 43.14, 43.14, 65.88, 66.32, 66.32, 66.32	0.334	0.466
10	*br*	5.45, 16.36, 68.18, 7.89, 15.79, 66.32, 14.12, 23.53, 39.22, 43.14	0.128	0.233
11	*nr*	2, 2, 2, 2, 6, 6, 6, 6, 25, 25, 25, 25, 10, 10, 10, 10, 20, 20, 20, 20, 84, 84, 84, 84, 18, 18, 18, 30, 30, 30, 50, 50, 50, 55, 55, 55	0.331	0.501
12	*wbn*	30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30	0.000	0.000
13	*wbr*	30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30	0.000	0.000
14	*wnr*	11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 38, 38, 38, 38, 38, 38, 38, 38, 38, 38, 38, 38, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25, 38.25	0.214	0.343
15	*bnr*	5.45, 5.45, 5.45, 5.45, 16.36, 16.36, 16.36, 16.36, 68.18, 68.18, 68.18, 68.18, 7.89, 7.89, 7.89, 7.89, 15.79, 15.79, 15.79, 15.79, 66.32, 66.32, 66.32, 66.32, 14.12, 14.12, 14.12, 23.53, 23.53, 23.53, 39.22, 39.22, 39.22, 43.14, 43.14, 43.14	0.127	0.259
16	*wbnr*	30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30, 30	0.000	0.000

Note: w = inequality within; b = inequality between; n = size; r = ranking.

Table 2. Marginal impact of factors on Gini concentration ratio (R) and Bonferroni inequality index (B). Illustrative example on the income of 10 recipients from three different population subgroups: A = {2, 6, 25} and B = {10, 20, 84}, C = {18, 30, 50, 55}.

Factor	Contribution to R		Contribution to B	
	SV	%	*SV*	%
within inequality	0.230	47.02	0.305	50.07
between inequality	0.176	35.90	0.244	40.13
size	0.005	1.00	−0.007	−1.22
ranking	0.079	16.07	0.067	11.02
Total	0.490	100.00	0.609	100.00

5. The Balance of Inequality Approach

The Balance of Inequality, proposed by Di Maio and Landoni (2017), is an alternative approach that, as in the Shapley method, helps solve the problem of additive decomposition of R by population subgroups. They use the center mass (or barycenter) to derive a measure of inequality, thus giving a physical interpretation to the inequality measure.

The barycenter of the income distribution, in which in abscissa is the income x_i and in ordinate is the ranking minus one $(i-1)$, is defined by the following expression

$$_Rb = \frac{\sum_{i=1}^{N} x_i(i-1)}{\sum_{i=1}^{N} x_i}.$$

It is equal to $(N-1)/2$ in the case of perfect equality, while it is equal to $N-1$ in the case of maximum inequality. Di Maio and Landoni (2017, p. 12) proceeded to normalize the barycenter and obtained, after a little algebra, the *BOI*:

$$BOI = \frac{\frac{\sum_{i=1}^{N} x_i(i-1)}{\sum_{i=1}^{N} x_i} - \frac{N-1}{2}}{(N-1) - \frac{N-1}{2}}$$

$$= \frac{2\sum_{i=1}^{N} i\, x_i}{t_x(N-1)} - \frac{N+1}{N-1} = {}_RBOI. \tag{9}$$

Expression (9) corresponds to the Gini concentration index (1) and, for this reason, we will refer to expression (9) as ${}_RBOI$ in the following. They show that ${}_RBOI$ (9), and therefore the Gini concentration ratio, can be decomposed by considering four factors. Besides those already seen in the previous paragraph—the inequality within and between population subgroups—*BOI* helps derive the impact on the inequality value due to asymmetry and irregularity of subgroups[4]. A population (or a subgroup) is symmetrical if the distribution of the analyzed variable is symmetrical with respect to its center. Furthermore, it is regular if the distance between two adjacent individuals in the population or in the subgroup is constant. A regular population (or a subgroup) is also symmetrical.

The ${}_RBOI$ for the Gini index can therefore be decomposed as

$${}_RBOI = \sum_{j=1}^{J} \frac{t_{xj}}{t_x}\left[\frac{{}_Rb_j^{*1} - {}_Rb_j^{*0}}{\frac{N-1}{2}}\right] {}_RBOI_j + \left(\sum_{j=1}^{J} \frac{t_{xj}}{t_x}\left[\frac{{}_Rb_j^{*0}}{\frac{N-1}{2}}\right]\right) - 1$$

$$+ \sum_{j=1}^{J} \frac{t_{xj}}{t_x}\left[\frac{{}_Rb_j^{*1} - {}_Rb_j^{*0}}{\frac{N-1}{2}}\right] ({}_RAE_j) + \sum_{j=1}^{J} \frac{t_{xj}}{t_x}\left[\frac{{}_Rb_j^{*1} - {}_Rb_j^{*0}}{\frac{N-1}{2}}\right] ({}_RIE_j) \tag{10}$$

where t_{xj} is the total income in the j-th subgroup; equivalently, we may define the *BOI* in the j-th subgroup via the following expression

$${}_RBOI_j = \frac{2\sum_{i\epsilon s_j} k\, x_i}{t_x\,(N_j-1)} - \frac{N_j+1}{N_j-1}$$

where k is the rank of the i-th recipient in subgroup j. Furthermore, ${}_Rb_j^{*1} = \max i \in s_j - 1$ is the barycenter of the subgroup in the population in case of perfect inequality, and ${}_Rb_j^{*0} = \frac{1}{N_j}\sum_{i\,\in\, s_j}(i-1)$ is the barycenter of the subgroup in the population in the case of perfect equality. In this context, ${}_RAE_j = {}_RBOI_j^* - {}_RBOI_{j\,sym}^*$ represents the asymmetry effect and ${}_RIE_j = {}_RBOI_{j\,sym}^* - {}_RBOI_j$ the irregularity effect where

$${}_RBOI_j^* = \frac{\frac{1}{t_{x_j}}\sum_{i\,\epsilon\, s_j} i\, x_i - \frac{1}{N_j}\sum_{i\,\epsilon\, s_j} i}{\max i \in s_j - \frac{1}{N_j}\sum_{i\,\epsilon\, s_j} i}$$

is the ${}_RBOI$ index for the j-th subgroup, while

$${}_RBOI_{j\,sym}^* = \frac{2}{t_{x_j}(\max i \in s_j - \min i \in s_j)} \sum_{i\,\epsilon\, s_j}\left(i - \frac{\min i \in s_j + \max i \in s_j}{2}\right) x_i$$

is the ${}_RBOI$ index for the j-th subgroup, in the case of symmetrical subgroups.

---

[4] Di Maio and Landoni (2017) consider the asymmetry and the irregularity as a unique factor but, to investigate the differences between *R* and *B*, it could be useful to consider them separately.

The first component in expression (10) is the weighted average of the within subgroup inequality, the second is the inequality between subgroups, the third and the fourth are the weighted average of the effects of asymmetry and irregularity of the distribution in each subgroup, respectively.[5]

The extension of the *BOI* methodology to the Bonferroni inequality index (B) requires that we consider a different representation of the income distribution. Let us consider the couple with the income on the abscissa and $\left(\frac{1-l_i}{\mu}\right)$ on the ordinate, where $l_i = \sum_{i=1}^{N} \frac{1}{i}$ and $\sum_{i=1}^{N} l_i = N$. The barycenter of this distribution is

$$ {}_{B}b = \frac{\sum_{i=1}^{N} x_i \left(\frac{1-l_i}{\mu}\right)}{\sum_{i=1}^{N} x_i}. $$

It is zero in the case of perfect equality and equal to $(N-1)/\mathrm{t}_x$ in the case of maximum inequality. Normalizing the barycenter, as before, and using a little algebra we obtain the expression of B in expression (3).

$$ {}_{B}BOI = \frac{\sum_{i=1}^{N} x_i (1-l_i)}{\mu(N-1)} \,. \tag{11} $$

This enable us to apply the *BOI* approach also to B. Expression (11) can be written as

$$ {}_{B}BOI = \sum_{j=1}^{J} \frac{t_{xj}}{t_x} \left[\frac{{}_{B}b_j^{*1} - {}_{B}b_j^{*0}}{\frac{N-1}{t_x}} \right] {}_{B}BOI_j + \sum_{j=1}^{J} \frac{t_{xj}}{t_x} \left[\frac{{}_{B}b_j^{*0}}{\frac{N-1}{t_x}} \right] $$

$$ + \sum_{j=1}^{J} \frac{t_{xj}}{t_x} \left[\frac{{}_{B}b_j^{*1} - {}_{B}b_j^{*0}}{\frac{N-1}{t_x}} \right] \left({}_{B}AE_j \right) + \sum_{j=1}^{J} \frac{t_{xj}}{t_x} \left[\frac{{}_{B}b_j^{*1} - {}_{B}b_j^{*0}}{\frac{N-1}{t_x}} \right] \left({}_{B}IE_j \right) \,. $$

Equivalently

$$ {}_{B}BOI_j = \frac{\sum_{i=1}^{N} x_k (1-l_k)}{\mu_j \, (N_j - 1)} $$

denotes the *BOI* for the j-th subgroup with size N_j, $\mu_j = \sum_{i \,\in\, s_j} x_k / N_j$, $l_k = \sum_{k=1}^{N_j} \frac{1}{k}$ and $\sum_{k=1}^{N_j} l_k = N_j$ where k is the rank of recipients in subgroup j. Furthermore,

$$ {}_{B}b_j^{*1} = \left(1 - \max_{i \,\in\, s_j} \; l_i \right) / \mu \,, $$

where $\max_{i \,\in\, s_j} \; l_i$ is the value of l_i corresponding to the recipients with the highest income in subgroup s_j, and ${}_{B}b_j^{*1}$ denotes the barycenter of the subgroup in the population in case of perfect inequality. The barycenter of the subgroup in the case of perfect equality is

$$ {}_{B}b_j^{*0} = \left(1 - \frac{\sum_{i \,\in s_j} l_i}{N_j} \right) / \mu \,. $$

[5] For more detail on ${}_{R}BOI$ see Di Maio and Landoni (2017).

As above, ${}_BAE_j = {}_BBOI_j^* - {}_BBOI_{j\,sym}^*$ represents the effect of asymmetry while ${}_BIE_j = {}_BBOI_{j\,sym}^* - {}_BBOI_j$ denotes the effect of irregularity on B. The BOI index for subgroup s_j is equal to

$${}_BBOI_j^* = \frac{\frac{\sum_{i\,\epsilon s_j} l_i}{N_j} - \frac{\sum_{i\,\epsilon s_j} x_i\, l_i}{t_{xj}}}{\frac{\sum_{i\,\epsilon s_j} l_i}{N_j} - \max_{i\,\in s_j} l_i}$$

while the BOI index for a symmetrical subgroup is given by

$${}_BBOI_{j\,sym}^* = \frac{2}{t_x\left(\min_{i\,\in s_j} l_i - \max_{i\,\in s_j} l_i\right)} \sum_{i\,\epsilon\, s_j} \left(\frac{\min_{i\,\in s_j} l_i + \max_{i\,\in s_j} l_i}{2} - l_i \right) x_i \,.$$

A Numerical Illustration

Let us consider the same population of 10 recipients we have already discussed in Section 4. We report in Table 3 the contribution of each factor obtained via the balance inequality approach.

Table 3. Balance of inequality decomposition for the Gini concentration ratio (R) and the Bonferroni inequality index (B). Illustrative example on the income of 10 recipients from three different population subgroups: A = {2, 6, 25} and B = {10, 20, 84}, C = {18, 30, 50, 55}.

Factor	Contribution to R		Contribution to B	
	${}_RBOI$	%	${}_BBOI$	%
within inequality	0.256	52.38	0.365	59.95
between inequality	0.151	30.86	0.252	41.42
asymmetry	−0.007	−1.35	−0.025	−4.10
irregularity	0.010	2.05	0.017	2.73
Total	0.490	100.00	0.609	100.00

By looking at this illustrative example, some preliminary results can be derived. In both cases, the higher contribution to the overall inequality corresponds to the within factor, followed by the between one and other factors. However, we may observe some differences when comparing the current decomposition to the Shapley decomposition in Table 2. The impact of between inequality on R is lower when measured with the BOI (30.86% vs. 35.90%), while the impact of within inequality is higher (52.38% vs. 47.02%). On the other hand, when we consider B, the values of between inequality are very similar (41.42% vs. 40.13%), while the difference for the within inequality is substantial (59.95% vs. 50.07%). For both indices, asymmetry reduces inequality: this issue is more evident when looking at the decomposition of R rather than the one of B.

6. An Application to the Italian Income Distribution

The Shapley decomposition of the Gini concentration ratio (R) and the Bonferroni index (B) has been applied to income data collected in 2015 by the Italian component of the European Survey on Income and Living Conditions (It-SILC, Istat 2015). The Eu-SILC is a yearly survey carried out by European countries according to the European Regulation n. 1177/2003. Its main aim is to provide data on income, poverty and social exclusion. The 2015 Italian sample is a two-stage sample of municipalities, stratified by population size, and households. The sample size is composed by 17,985 household and 36,602 individuals.

We consider the Italian households as divided into three subgroups, represented by the main geographical areas: North, Center and South. Table 4 shows some explanatory statistics on the distribution of household income for the whole population and the subgroups.

The inequality measures have been computed on the distribution of household income. The incomes have not been equivalized to account for the different households' size. The values of R have been estimated using the expression of the sampling estimator defined by Osier (2009, p. 169), while B has been estimated using the expression of the sampling estimator derived in Giorgi and Guandalini (2013, p. 154). The BOI values, for R and B, have been computed through a plug-in estimator.

Looking at Table 4, we observe that $R = 0.367$ and $B = 0.462$ for the whole population. North and Center have quite a similar situation. While in South the incomes are lower, and the inequality is higher. In the three subgroups, but also at the national level, there is a strong positive asymmetry in the income distribution. The asymmetry is greater in the South when compared to the other geographical areas.

In Table 5, R has been decomposed using the Shapley decomposition (as shown in Section 4) and the balance of inequality (as shown in Section 5). As for the Shapley decomposition, it is important to point out that the sample size for all the subgroups is larger than 4000; therefore, the Dalton principle of replication invariance can be considered as (at least approximately) satisfied also for B.

The impact of the different factors obtained by the decomposition methods are reported in Table 5. For each component and decomposition, the corresponding confidence interval, estimated via nonparametric bootstrap ($M = 500$ samples), are reported.

The plug-in estimators based on BOI are biased. The bias is negligible for ${}_{R}BOI$, while it is more evident for ${}_{B}BOI$. However, this does not affect the comparison between the decomposition methods and the two indices, since, in any case, the bias does not change the balance of power between factors considered obtained via the balance of inequality.

Table 4. Some explanatory statistics on average Italian household income distribution by three subgroups (North, Center and South). Source: It-SILC, Italy 2015.

Geographical Area	Households		First Quartile	Median	Mean	Third Quartile	Fisher Asymmetry Coefficient	R	B
	Sample Size	Population Size							
North	8922	12,294,699	25,809	39,180	47,621	59,749	4.273	0.346	0.439
Center	4223	5,295,623	23,114	36,459	44,626	56,524	2.379	0.360	0.457
South	4840	8,185,550	16,939	26,617	32,561	40,400	10.861	0.372	0.482
Italy	17,985	25,775,872	22,007	34,199	42,223	53,480	5.143	0.367	0.462

Both decompositions identify the within inequality as a very important factor. Under the Shapley decomposition, it accounts for more than 60% of the whole inequality, both for R and B. Ranking is more important than between inequality (20% versus 14%), since the subgroups are strongly overlapped. Finally, subgroup size plays a minor role. Under the Shapley decomposition, the magnitude of factors is similar for both the analyzed indices. However, when we consider B the impact of within inequality is higher while that of ranking is lower than for R. The importance of differences among subgroups in size is negligible when we consider R, since it is population size independent, while the same is different from zero in B, even if not that high.

Under the BOI decomposition, within inequality accounts for more than 80% of the whole inequality for both the analyzed indices, even if its role is slightly more evident in R. Unlike the Shapley decomposition, the two indices show a different "hierarchy" of factors when we look at the corresponding impact. When we consider R, the most important factor is the within inequality followed by the between inequality. The contribution of asymmetry and irregularity is almost negligible. On the contrary, when we look at B, the asymmetry is the most important factor followed by the within inequality and the irregularity (with a negative sign). Between inequality plays a minor role.

It is important to point out that the combined effect of asymmetry and irregularity has opposite signs on the two indices ($0.55 - 2.07 = -1.52\%$ for R and $89.13 - 78.00 = 11.13\%$ for B). This is probably

due to the indices' sensitivity to different levels of the income distribution. As remarked above, B is more sensitive to lower values (left tail of the distribution), while R is more sensitive to the central values of the distribution. Moreover, the high value for the impact of asymmetry and irregularity when we consider B can be due to the asymmetry in the income distribution, as already stated, but also to an indirect effect of population size. In fact, as opposed to the numerical illustration in Section 5, the contribution of asymmetry and irregularity to B is higher in the It-SILC data due to the larger population size and asymmetry.

The two decomposition methods are deeply different. The Shapley decomposition represents a more general tool which can be used to decompose not only inequality measures and not only by the four factors we have considered here. It can be modified by considering a different (lower or higher) number of factors. The BOI is more similar to a standard decomposition approach, since it is less customizable. In fact it is possible to decompose the index by within inequality, between inequality, asymmetry and irregularity only.

Table 5. Shapley and Balance of inequality decompositions for the Gini concentration ratio (R) and the Bonferroni inequality index (B). Application to Italian household income distribution by three population subgroups (North, Center and South). Source: It-SILC, Italy 2015.

Factor	Contribution to R		Contribution to B	
	Absolute Value	**%**	**Absolute Value**	**%**
		Shapley decomposition		
within inequality	0.2348 [0.2281, 0.2415]	63.99 [62.16, 65.80]	0.3065 [0.2956, 0.3174]	66.30 [63.94, 68.66]
between inequality	0.0530 [0.0439, 0.0621]	14.44 [11.95, 16.93]	0.0626 [0.0522, 0.0730]	13.55 [11.28, 15.80]
size	−0.0002 [−0.0037, 0.0033]	−0.05 [−1.00, 0.89]	0.0090 [0.0046, 0.0134]	1.95 [0.99, 2.90]
ranking	0.0793 [0.0700, 0.0886]	21.62 [19.08, 24.13]	0.0841 [0.0757, 0.0925]	18.20 [16.37, 20.01]
Total	0.3670 [0.3568, 0.3772]	100.00	0.4623 [0.4505, 0.4741]	100.00
		Balance of Inequality (BOI)		
within inequality	0.3483 [0.3384, 0.3582]	94.98 [94.89, 95.02]	0.4007 [0.3908, 0.4106]	81.07 [79.07, 83.09]
between inequality	0.0239 [0.0182, 0.0296]	6.54 [5.11, 7.85]	0.0385 [0.0293, 0.0477]	7.79 [6.05, 9.65]
asymmetry	0.0020 [0.0008, 0.0032]	0.55 [0.24, 0.84]	0.4405 [0.4393, 0.4417]	89.13 [90.78, 89.36]
irregularity	−0.0076 [−0.0095, −0.0057]	−2.07 [−2.65, −1.52]	−0.3855 [−0.3874, −0.3836]	−78.00 [−80.04, −77.62]
Total	0.3668 [0.3566, 0.3770]	100.00	0.4942 [0.3908, 0.4106]	100.00

Note: Bootstrap confidence interval at 95% in squared brackets.

Some Considerations on the Shapley Decomposition and the Balance of Inequality

The numerical examples and the application to real data show that the two decomposition methods point out different aspects of the inequality indices. The Shapley decomposition is more sensitive to the ranking in the income distribution, while the BOI decomposition is more influenced by the shape of the distribution.

Looking at the behavior of the two indices with respect to the adopted decomposition, it is possible to draw some interesting conclusions. Since R and B adopt a similar ranking system, we cannot observe substantial differences when considering the Shapley decomposition; however, since the indices have

different sensitivity to different portions of the distribution, asymmetry and irregularity often play a crucial role in the BOI decomposition, and this may lead to different results.

Finally, as using more synthetic indices can help us highlight differences between socio-economic reality and political significance of inequality (Piketty 2014, p. 156); using more than one decomposition may help focus on different aspects and factors of inequality.

7. Conclusions and Further Research

An important topic on inequality measures is their decomposability. Two main lines of research can be identified: decomposition by income sources and by population subgroups. Some indices, such as the Gini concentration ratio R (Gini 1914) and the Bonferroni inequality index B (Bonferroni 1930) are not additively decomposable by population subgroups. To overcome this drawback, Deutsch and Silber (2007) proposed the so-called "Shapley value", and Di Maio and Landoni (2017) suggested the "balance of inequality" (BOI) approach to decompose the Gini concentration ratio (R).

In this paper, we have discussed the Shapley decomposition for the Bonferroni inequality index (B). Furthermore, we also show how the balance of inequality can be extended to B. The two indices have been estimated on real data from the 2015 Italian component of the European Survey on Income and Living Conditions (It-SILC) and the two decomposition methods have been considered in this context.

The results of the application highlights that the features of each subpopulation, such as homogeneity within (denoted by the component of within inequality), and the difference in subpopulation size, have higher influence on B than on R. Furthermore, B seems to be more sensitive to asymmetry and irregularity in the observed distribution and the population size.

The two decomposition methods focus on different aspects of the distribution. The Shapley value reflects the ranking in the income distribution, while the BOI is mainly influenced by the shape of the distribution. For these reasons, the two indices have a similar behavior under the Shapley decomposition, as their ranking system is similar, while they may show a completely different "hierarchy" of factors under the balance of inequality decomposition.

The results of our research also suggest the possibility of supplementing the measure of overall inequality through indices with different sensitivity to different parts of the income distribution, trying to answer, at least in part, the possible disadvantages in using a single index (Osberg 2017). This follows, in our view, the path suggested by Piketty (2014, p. 156). Piketty proposed to use different indices to account for the differences between socio-economic reality and political significance of inequality in different parts of the income distribution. In the same way, the use of different kinds of decompositions can help to focus on different aspects and factors of inequality. In this perspective, further studies could focus on the extension of the BOI approach to other indices.

Acknowledgments: The authors would like to thank the Editors and two anonymous reviewers for their valuable comments and suggestions.

Author Contributions: The authors contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Let us assume to have a population with a vector of income $x = (x_1, x_2, x_3)$. Furthermore, let us assume to repeat a finite number of times the income in x and define a vector $y = (x_1, \ldots, x_1, x_2, \ldots, x_2, x_3, \ldots, x_3)$. If R is computed on x and on y, $R(x) = R(y)$, that is, R satisfies the Dalton principle of replication invariance.

When computing B on x and y, usually $B(x) \neq B(y)$. Therefore, this is generally intended to show that B does not satisfy the Dalton principle of replication invariance. However, this holds for small population sizes (i.e., dimension of x). In fact, it is possible to prove that the difference between $B(y)$ and $B(x)$ becomes quickly negligible as the population size increases.

The departure of B from the Dalton principle of replication invariance can be influenced by three factors: population size, level of concentration and number of replication.

In Table A1, we present the results of a small simulation study. The income for units belonging to twelve populations which differ by size and level of concentration of the corresponding income distribution have been generated from a log-normal distribution. We have considered four values for the population sizes (10, 100, 1000 and 10,000) and three levels of concentration for the corresponding income: low, medium and high, that is about $R \cong 0.20$, 0.50 and 0.80 respectively.

For each population, the B index has been computed. Then, the incomes have been replicated 2, 10 and 100 times and B has been computed also for the populations with the replicated incomes.

Looking at Table A1, it is clear that B does not satisfy the Dalton principle of replication invariance. In fact, the relative differences the value of B for populations without replicated incomes and for population with replicated incomes are all non-zero. However, it is possible to note that, generally, the differences are larger when the replications refer to a population with a higher concentration of income distribution. Furthermore, increasing the times of replications contributes to increase the difference between the values of B, while, instead, increasing the population size leads to differences going quickly to 0. In all the cases, the differences are negligible to the third decimal place when n is greater than 1000. Therefore, it is possible to state that B is asymptotically replication invariant.

Table A1. Values and relative differences of the Bonferroni index (B) computed for a population with incomes generated by a log-normal distribution and for the same population with incomes replicated 2, 10 and 100 times. For different population sizes (10, 100, 1000, 10,000) and for different concentration of incomes (R $\cong$ 0.20, 0.50 and 0.80).

Population Size	B				Relative Difference		
	Number of Replications				Number of Replications		
	No Replication	2	10	100	2	10	100
	(a)	(b)	(c)	(d)	(b − a)/a	(c − a)/a	(d − a)/a
			Low level of concentration ($R \cong 0.20$)				
10	0.30789	0.30394	0.30183	0.30147	−0.01285	−0.01971	−0.02085
100	0.29684	0.29692	0.29700	0.29702	0.00025	0.00054	0.00061
1000	0.29289	0.29291	0.29292	0.29292	0.00006	0.00011	0.00012
10,000	0.28858	0.28859	0.28860	0.28860	0.00002	0.00004	0.00004
			Medium level of concentration ($R \cong 0.50$)				
10	0.68310	0.67073	0.66296	0.66145	−0.01812	−0.02949	−0.03170
100	0.64458	0.64382	0.64323	0.64310	−0.00119	−0.00210	−0.00230
1000	0.63418	0.63411	0.63405	0.63404	−0.00011	−0.00019	−0.00021
10,000	0.62732	0.62732	0.62731	0.62731	−0.00001	−0.00002	−0.00002
			High level of concentration ($R \cong 0.80$)				
10	0.94323	0.91843	0.90107	0.89744	−0.02630	−0.04470	−0.04855
100	0.88648	0.88452	0.88298	0.88263	−0.00221	−0.00395	−0.00434
1000	0.88150	0.88131	0.88116	0.88113	−0.00022	−0.00039	−0.00043
10,000	0.87653	0.87651	0.87649	0.87649	−0.00002	−0.00004	−0.00004

References

Aaberge, Rolf. 2000. Characterizations of Lorenz Curves and Income Distributions. *Social Choice and Welfare* 17: 639–53. [CrossRef]

Bárcena-Martin, Elena, and Luis J. Imedio Olmedo. 2008. The Bonferroni, Gini and De Vergottini Indices. Inequality, Welfare and Deprivation in the European Union in 2000. *Research on Economic Inequality* 16: 231–57.

Bárcena-Martin, Elena, and Jacques Silber. 2011. On the Concepts of Bonferroni Segregation Index Curve. *Rivista Italiana di Economia, Demografia e Statistica* 62: 57–74.

Bárcena-Martin, Elena, and Jacques Silber. 2013. On the Generalization and Decomposition of the Bonferroni Index. *Social Choice and Welfare* 41: 763–87. [CrossRef]

Benedetti, Carlo. 1986. Sulla Interpretazione Benesseriale di Noti Indici di Concentrazione e di altri. *Metron* 45: 421–29.

Bhattacharya, N., and B. Mahalanobis. 1967. Regional Disparities in Household Consumption in India. *Journal of the American Statistical Association* 62: 143–61. [CrossRef]

Bonferroni, Carlo E. 1930. *Elementi Di Statistica Generale*. Firenze: Libreria Seber.

Chakravarty, Satya R. 2007. A Deprivation-Based Axiomatic Characterization of the Absolute Bonferroni Index of Inequality. *Journal of Economic Theory* 5: 339–51. [CrossRef]

Dalton, Hugh D. 1925. *Some Aspects of the Inequality of Incomes in Modern Communities*. London: Routledge.

De Vergottini, Mario. 1950. Sugli Indici di Concentrazione. *Statistica* 10: 445–54.

Deutsch, Joseph, and Jacques Silber. 1999. Inequality Decomposition by Population Subgroups and the Analysis of Interdistributional Inequality. In *Handbook on Income Inequality Measurement*. Edited by Silber Jacques. Boston: Kluwer Academic Publisher, vol. 71, pp. 363–97.

Deutsch, Joseph, and Jacques Silber. 2007. Decomposing Income Inequality by Population Subgroups: A Generalization. In *Research on Economic Inequality: Inequality and Poverty*. Edited by Bishop John and Amiel Yoram. Berlin: Springer, vol. 14, pp. 237–53.

Di Maio, Giorgio, and Paolo Landoni. 2017. The Balance of Inequality: A Rediscovery of The Gini's R Concentration Ratio and a New Inequality Decomposition by Population Subgroups Based on Physical Rationale. Paper presented at Seventh Meeting of The Society for the Study of Economic Inequality (Ecineq), New York City, NY, USA, July 17–19.

Ferrari, Guido, and Pietro Rigo. 1987. Sulla Scomposizione del Rapporto di Concentrazione di Gini. In *La Distribuzione Personale del Reddito: Problemi di Formazione, di Ripartizione e di Misurazione*. Edited by Zenga Michele. Milano: Vita e Pensiero, pp. 347–63.

Fields, Gary S. 1979a. Income Inequality in Urban Colombia: A Decomposition Analysis. *Review of Income and Wealth* 25: 327–41. [CrossRef]

Fields, Gary S. 1979b. Decomposing LDC Inequality. *Oxford Economic Papers* 31: 437–59. [CrossRef]

Frick, Joachim R., and Jan Goebel. 2007. Regional Income Stratification in Unified Germany Using a Gini Decomposition Approach. Discussion paper. Berlin, Germany: Germany Institute for Economich Research, 1–32.

Gini, Corrado. 1912. Studi Economico-Giuridici della Facoltà di Giurisprudenza della Regia Università di Cagliari. In *Variabilità e Mutabilità: Contributo Allo Studio Delle Distribuzioni e Delle Relazioni Statistiche*. Bologna: Cuppini, vol. 3.

Gini, Corrado. 1914. Sulla Misura della Concentrazione e della Variabilità dei Caratteri. *Atti Del Reale Istituto Veneto Di Scienze, Lettere ed Arti* 73: 1203–48, English Translation In *Metron* 2005, 63: 3–38.

Giordani, Paolo, and Giovanni M. Giorgi. 2010. A Fuzzy Logic Approach to Poverty Analysis Based on the Gini and Bonferroni Inequality Indices. *Statistical Methods and Applications* 19: 587–607. [CrossRef]

Giorgi, Giovanni M. 1990. Bibliographic Portrait of the Gini Concentration Ratio. *Metron* 48: 183–221.

Giorgi, Giovanni M. 1992. *Il Rapporto di Concentrazione di Gini. Genesi, Evoluzione ed una Bibliografia Commentata*. Siena: Libreria Editrice Ticci.

Giorgi, Giovanni M. 1993. A Fresh Look at the Topical Interest of the Gini Concentration Ratio. *Metron* 51: 83–98.

Giorgi, Giovanni M. 1998. Concentration Index, Bonferroni. In *Encyclopedia of Statistical Sciences*. Updated Series; Edited by Kotz Samuel, Read Campbell B. and Banks David L. New York: Wiley-Intersciences, vol. 2, pp. 141–46.

Giorgi, Giovanni M. 1999. Income Inequality Measurement: The Statistical Approach. In *Handbook on Income Inequality Measurement*. Edited by Silber Jacques. Boston: Kluwer Academic Publishers, pp. 245–60.

Giorgi, Giovanni M. 2005. Gini's Scientific Work: An Evergreen. *Metron* 63: 299–315.

Giorgi, Giovanni M. 2011a. The Gini Inequality Index Decomposition, an Evolutionary Study. In *The Measurement of Individual Well-Being and Group Inequality: Essay In Memory Of Z.M. Berrebi*. London: Routledge, pp. 185–218.

Giorgi, Giovanni M. 2011b. Corrado Gini: The Man and the Scientist. *Metron* 69: 1–28. [CrossRef]

Giorgi, Giovanni M., and Michele Crescenzi. 2001a. Bayesian Estimation of the Bonferroni Index in a Pareto-Type I Population. *Statistical Methods and Applications* 10: 41–48. [CrossRef]

Giorgi, Giovanni M., and Michele Crescenzi. 2001b. A Look at the Bonferroni Inequality Measure in a Reliability Framework. *Statistica* 61: 571–83.

Giorgi, Giovanni M., and Michele Crescenzi. 2001c. A Proposal of Poverty Measures Based on the Bonferroni Inequality Index. *Metron* 59: 3–15.

Giorgi, Giovanni M., and Alessio Guandalini. 2013. A Sampling Estimator of the Bonferroni Inequality Index. *Rivista Italiana di Economia, Demografia e Statistica* 67: 151–58.

Giorgi, Giovanni M., and Riccardo Mondani. 1995. Sampling Distribution of Bonferroni Inequality Index from an Exponential Population. *Sankhya* 57: 10–18.

Imedio Olmedo, Luis J., Elena Bárcena-Martin, and Encarnación M. Parrado-Gallardo. 2012. Income Inequality Indices Interpreted as Measures of Relative Deprivation/Satisfaction. *Social Indicator Research* 109: 471–91. [CrossRef]

Istat. 2015. Indagine Sulle Condizioni di Vita (UDB IT—SILC). Available online: https://www.istat.it/it/archivio/4152 (accessed on 4 December 2017).

Kakwani, Nanak C. 1980. *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. Oxford: Oxford University Press.

Lorenz, Max O. 1905. Method of Measuring the Concentration of Wealth. *Publication of the American Statistical Association* 9: 209–19. [CrossRef]

Mehran, Farhad. 1975. A Statistical Analysis of Income Inequality Based on a Decomposition of the Gini Index. *Bulletin of the International Statistical Institute* 46: 145–50, Contributed Paper, 40th Session, Warsaw, Poland.

Mussard, Stéphane, Françoise Seyte, and Michel Terrazza. 2006. La Décomposition de l'Indicateur de Gini en Sous Groupes: Une Revue de la Littérature. GRÉDI Working paper 06-11. Sherbrooke, QC, Canada: Université De Sherbrooke.

Nygård, Fredrik, and Arne Sandström. 1981. *Measuring Income Inequality*. Stockholm: Almqvist & Wiksell International.

Osberg, Lars. 2017. On the Limitations of Some Current Usages of the Gini Index. *Review of Income and Wealth* 63: 574–84. [CrossRef]

Osier, Guillaume. 2009. Variance Estimation for Complex Indicators of Poverty and Inequality Using Linearization Techniques. *Survey Research Methods* 3: 167–95.

Piesch, Walter. 1975. *Statistische Konzentrationsmasse*. Tübingen: J.B.C. Mohr (Paul Siebeck).

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge: Harvard University Press.

Pizzetti, Ernesto. 1951. Relazioni fra Indici di Concentrazione. *Statistica* 11: 294–316.

Rao, V. 1969. Two Decompositions of Concentration Ratio. *Journal of the Royal Statistical Society* 132: 418–25. [CrossRef]

Shapley, Lloyd. 1953. A Value for N-Person Games. In *Contributions to the Theory of Games (AM-28)*. Edited by Kuhn Harold W. and Tucker Albert W. Princeton: Princeton University Press, vol. 2, pp. 307–18.

Shorrocks, Anthony F. 1980. The Class of Additively Decomposable Measures. *Econometrica* 48: 613–25. [CrossRef]

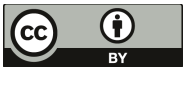
Shorrocks, Anthony F. 1999. Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value. Essex, UK: Department of Economics, University of Essex.

Shorrocks, Anthony F. 2013. Decomposition Procedures for Distributional Analysis: A Unified Framework Based on the Shapley Value. *The Journal of Economic Inequality* 11: 99–126. [CrossRef]

Silber, Jacques, and Hyun Son. 2010. On the Link between the Bonferroni Index and the Measurement of Inclusive Growth. *Economics Bulletin* 30: 421–28.

Tarsitano, Agostino. 1990. The Bonferroni Index of Income Inequality. In *Income and Wealth Distribution, Inequality and Poverty*. Edited by Dagum Camilo and Zenga Michele. Berlin: Springer, pp. 228–42.

Yitzhaki, Shlomo. 1998. More than a dozen alternative ways of spelling Gini. *Research on Economic Inequality* 8: 13–30.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Inequality and Poverty When Effort Matters

Martin Ravallion

Department of Economics, Georgetown University, Washington, DC 20057, USA; mr1185@georgetown.edu

Received: 25 August 2017; Accepted: 23 October 2017; Published: 6 November 2017

Abstract: On the presumption that poorer people tend to work less, it is often claimed that standard measures of inequality and poverty are overestimates. The paper points to a number of reasons to question this claim. It is shown that, while the labor supplies of American adults have a positive income gradient, the heterogeneity in labor supplies generates considerable horizontal inequality. Using equivalent incomes to adjust for effort can reveal either higher or lower inequality depending on the measurement assumptions. With only a modest allowance for leisure as a basic need, the effort-adjusted poverty rate in terms of equivalent incomes rises.

Keywords: equivalent income; welfare; inequality; poverty; labor supply

JEL Classification: D31; D63

1. Introduction

Disparities in levels of living reflect, to some degree, differences in personal efforts. While views that many people believe that effort plays a role. In a 2014 opinion poll of the American public, about one third of respondents viewed poverty as stemming from a lack of effort by poor people while a similar proportion believed that the rich were rich simply because they worked harder (Pew Research Center 2014). Though it is not often made explicit, it is at least implicit in these views that the differences in effort reflect differences in personal aversion to work—differences in preferences over effort versus consumption. In the simplest expression of this view, poor people are deemed to be poor because they are lazy.

The standard model of consumption-leisure choice does not imply that a person with lower income will chose to work less, although certain restricted forms of the model do have that property.[1] If poorer people do tend to work less, then it is theoretically possible that there is equality of welfare even when there is considerable inequality based on observed incomes.[2] While that theoretical possibility may be dismissed as unlikely, there appears to be a widely accepted view that there is less inequality and poverty than suggested by observed incomes. For example, Bourguignon (2015, p. 61) writes that " . . . correcting inequality in standards of living for disparities in hours worked between households would result in lower estimates of inequality".

This paper aims to assess the validity of that claim. One can say that "effort matters" in this context when it affects welfare (negatively) and it varies at given income. The paper explores the implications for the measurement of inequality and poverty amongst adults.[3] The starting point is to

[1] As is well known, a source of ambiguity is that there are opposing income and substitution effects of higher wage rates on labor supply (assuming that leisure is a normal good). Higher unearned income will reduce work effort. The direction of the relationship with total income is unclear on theoretical grounds.

[2] See, for example, Allingham (1972) comment on Atkinson (1970).

[3] Of course, effort is only one aspect of the debates about inequality numbers; for example, there are also issues about price indices and equivalence scales. Note also that practitioners are on safer ground in measuring inequality amongst children for whom personal effort is not yet an issue. Here the concern is about inequality among adults.

note that some concept of individual welfare is implicit in any assessment of whether one person is better off than another. This is taken for granted in measuring "real income", such as when deflating nominal incomes for cost-of-living differences or adjusting for demographic heterogeneity using equivalence scales. But it is no less compelling when welfare depends on effort. While there may be constraints (such as labor-market frictions) on the scope for freely choosing one's effort, a significant degree of choice can be exercised by most people. Presumably the reason people who think that income inequality is largely due to different efforts are not so troubled by that inequality is that they think there is little or no underlying inequality in welfare; the inequality reflects personal choices.[4]

The nub of the matter then is that the way inequality is being assessed in practice does not use a valid money-metric of welfare when effort matters. As long as people care about effort and it varies, observed incomes do not identify how welfare varies and so they are a questionable basis for assessing inequality of outcomes or opportunities. Nor is the use of predicted income based on circumstances (as has become popular in the recent literature on measuring inequality of opportunity) welfare consistent, as will be explained later. Recognizing that people take responsibility for their efforts, given their circumstances, leads one to ask how a true money-metric of welfare—reflecting the disutility of effort—varies. It has long been known that one can in principle measure income in a welfare-consistent way, as the monetary equivalent of utility.[5] However, the implications for inequality are far from obvious. Those who claim that high (low) incomes largely reflect high (low) effort will expect to see a systematic positive relationship between effort and income, which will attenuate the welfare disparities suggested by observed incomes, as Bourguignon (2015) claims in the quote above. Against this view, people in disadvantaged circumstances may be encouraged to make greater effort to compensate.

However, a key message of this paper is that, when effort matters, these vertical differences in how effort varies with income are not sufficient to predict the impact on inequality. Alongside the vertical differences, there is also heterogeneity in work effort at given income, reflecting differences in (*inter alia*) wage rates (or skills) and preferences. While there may be a tendency for poorer people to work less (although that is an empirical question), that is unlikely to always be true; anecdotal observations can point to both hard working poor people and the "idle rich". When two people with the same observed income make different efforts to derive that income, adjusting for the disutility of effort implies higher inequality between them. This horizontal effect mitigates the systematic effect on welfare inequality of vertical differences stemming from a positive relationship between income and mean effort. Heterogeneity in preferences can either magnify this horizontal effect, to the extent that people who work more value leisure more, or mitigate it, if work and leisure preferences are related in the opposite way.

A related issue arises in the context of measuring poverty. Here an appealing principle is that one should set the poverty line consistently with the metric used to assess who is poor. For example, if one uses total income or consumption expenditure one would not want the poverty line to exclude any major component of consumption, such as non-food goods.[6] Similarly, if one allows for the disutility of work in assessing welfare by adding the imputed value of leisure then one should include an allowance for leisure as a basic need when setting the poverty line. It would surely make little sense to say that, on allowing for effort, the poverty rate has fallen if one has used the same poverty line as for observed incomes ignoring effort.

---

[4] This is an instance of a more general point that is well understood in welfare economics, namely that inequality of income need not imply inequality of welfare. Heterogeneity in preferences further complicates matters.

[5] There have been a number of applications of the idea of money-metric utility to distributional analysis, including King (1983), Jorgenson and Slesnick (1984), Blundell et al. (1988), Apps and Savage (1989), Kanbur and Keen (1989). Also see the discussions in Slesnick (1998).

[6] The economic arguments for assuring such consistency are reviewed in Ravallion (2016, Part 2).

The upshot is that even if it is in fact true that higher income people tend to work harder it does not follow that there is less inequality or poverty than observed incomes suggest. The paper elaborates the above points and illustrates their relevance to assessments of the extent of inequality and poverty in the U.S. in 2013. To abstract from the thorny issues of setting demographic scales and other issues of interpersonal comparisons of welfare, the paper focuses on single adults without disabilities. This could well be biasing the study's results toward underestimating the effects on inequality measures of ignoring heterogeneity in effort, on the presumption that allowing for demographic differences between households would add to the heterogeneity.

The paper's principle finding is that the claim that inequality and poverty measures are being overstated given that higher-income workers tend to work more (which is confirmed empirically) is not robust to allowing for heterogeneity in work effort at given income. Allowing for heterogeneity consistently with the data and assuming full optimization suggest that there is higher inequality, though largely among the three or four upper-income deciles. This finding is sensitive to a number of methodological choices. A seemingly plausible regression-based trimming of the extremes in the data used to infer the preferences suggests that standard inequality measures are quite robust to adjusting for effort using welfare-consistent equivalent incomes that respect individual preferences.

Poverty measures are less robust, but the impact of allowing for heterogeneity goes in the opposite direction to the arguments often made. As long as one includes a modest allowance for leisure in the poverty bundle—to assure consistency between how the line is measured and how welfare is assessed—poverty measures rise on adjusting for effort. With the trimmed series, it takes only a very small allowance for leisure as a basic need to overturn the claim that allowing for effort implies less poverty in terms of welfare than raw income data suggest.

Three responses can be anticipated. First, the concern identified here applies to any situation in which income is used to measure welfare, which also depends on personal choices that matter independently of income. That is true. The present focus is nonetheless justified given that effort has been so widely acknowledged as a source of inequality that needs to be treated differently to inequalities stemming from circumstances.

Second, one might be uncomfortable with the welfarist perspective, in which personal utilities are the basis for judgements about inequality and social welfare. However, it would surely be hard to defend a view that (on the one hand) people take responsibility for their effort but (on the other hand) the degree of their effort has no bearing on how their welfare should be assessed. Rejecting the view that utility is the sole metric of welfare does not justify ignoring the differences in the efforts taken to make a living.

Third, it may be argued that one can still be justifiably interested in measuring inequality in terms of incomes, ignoring the disutility of the effort in deriving those incomes. Such inequality is a well-recognized parameter in how we assess social progress. Without disputing this point, it seems that measurement practices should take seriously the concerns that have been raised about the relevance of such measures when efforts and preferences vary. It remains an empirical question just how much these concerns matter.

The next section discusses how effort has been treated in the literature. Section 3 draws out some theoretical implications of behavioral responses for measuring inequality of outcomes or opportunities, allowing better circumstances to either encourage or discourage effort. Section 4 outlines at a simple parametric model, which is implemented on U.S. data, and discusses the results. A concluding discussion is found in Section 5.

2. Antecedents in the Literature

It has been argued in some quarters that inequalities stemming from effort do not have the same ethical salience as those stemming from circumstances beyond an individual's control. For example, Checchi and Peragine (2010, p. 430) argue that " ... existing surveys show that most people judge income inequalities arising from different levels of effort as less objectionable than those due to

exogenous circumstances". This view has influenced social policy making. For example, antipoverty policies in America and elsewhere have often identified the "undeserving poor" as those who are judged to be poor for lack of effort.[7] "Bad behaviors" creating "choice-based poverty" are also seen by some observers as a source of exaggerated concerns about inequality.[8] Those who take the alternative view—that it is really differing circumstances that divide the "rich" from the "poor"—tend to find the inequality far more troubling, and are more demanding of a policy response. (In the same PEW Research Center poll mentioned in the Introduction, about 50% of respondents felt that circumstances/advantages were the main reason for poverty and inequality.)

Prevailing measures of inequality and poverty largely ignore differences in effort. The measures found in practice treat two people with the same income (or consumption) equally even if one of them must work hard to obtain that income while the other is idle. Nor are differences in preferences addressed by standard measures, recognizing that the disutility of effort almost surely depends on personal circumstances. Thus, there is a disconnection between the social-policy debates on poverty and inequality and prevailing measurement practices.

While the vast bulk of the applied literature on measuring inequality has ignored effort heterogeneity, one can find exceptions in three distinct places in the literature. All three will have a role in this paper's subsequent analysis.

First, there is the idea of a "potential wage" (Champernowne and Cowell 1998), also called "full-time equivalent income" and "standard income" (Kanbur and Keen 1989).[9] I will use the term "full income".[10] The idea is that one measures income as if every able-bodied adult worked some standard number of hours, such as a full-time job. Assuming that everyone is free to work as much or as little as they like, if someone has an observed income below the poverty line but could in principle avoid this by working full time then she is not deemed to be poor by the full income approach. (Of course, the welfare interpretation is different if the person is physically unable to work full time, or is rationed in the labor market such that she cannot find the stipulated standard amount of work.) While full income is often used in business and labor studies when comparing full-time and part-time workers, it has only rarely been used in measuring inequality (an example is found in Salverda et al. 2014). The concept can be useful in quantifying the contribution of different levels of employment by income group to inequality.

Second, there is a strand of the literature that uses the concept of a money-metric of utility. An example is the concept of "equivalent income" (King 1983), given by the income that yields the actual utility level (dependent on the person's own effort, income and preferences) at fixed reference values. Unlike full income, this delivers a valid welfare metric.[11] Empirical contributions in the context of labor supply include Blundell et al. (1988) and Apps and Savage (1989). Bargain et al. (2013) and Decoster and Haan (2015) use somewhat different monetary measures of welfare in making comparisons across countries.[12]

---

7 This is an old idea, but in modern times it became prominent in Katz (1987) critique of American antipoverty policy. See Ravallion (2016, Part 1) on the history of economic thought on antipoverty policy. Also see Gans (1995, chp. 1) discussion of the history of derogatory labels for poor people.

8 For example, with reference to the U.S., Stein (2014) argues that: "There is an immense amount of income inequality here and everywhere. I am not sure why that is a bad thing. Some people will just be better students, harder working, more clever, more ruthless than other people". Stein goes on to claim that long-term poverty reflects "poor work habits". Also see the debate between Eichelberger (2014) and Williamson (2014) on the proposition that "poor people are lazy".

9 Champernowne and Cowell (1998) only give passing reference to the idea, and do not develop its implications. Kanbur and Keen (1989) discuss its use in the context of inequality and taxation. The concept of "full-time equivalent income" is found in business and labor studies; see, for example, the online Business Dictionary.

10 This is not the same concept of full income found in Becker (1965), which includes the imputed value of the entire time endowment.

11 This is shown by Kanbur and Keen (1989) in the context of heterogeneous effort though the point is more general.

12 The measures include Pencavel (1977) real wage metric, given by wage rate equivalent of the actual utility level at fixed values of other factors, including unearned income, and an analogous "rent metric" given by the unearned income equivalent of utility. A useful overview of the various measures possible can be found in Preston and Walker (1999). An earlier empirical application of the real wage index idea can be found in Coles and Harte-Chen (1985).

The third relevant strand of the literature focuses on inequality of opportunities (IOP). There is a (rapidly expanding) literature on measuring IOP, giving an explicit recognition of the role of effort in determining incomes. The usual theoretical starting point is Roemer (1998) argument that income depends on both circumstances and personal efforts, such as labor supply. (Examples of relevant circumstances are parental income and parental education.) Income inequalities due to differing efforts are not seen as having ethical or policy salience although it is arguably a big step to say that we should not be concerned about inequalities stemming from different efforts if only because such inequalities today can generate troubling inequalities of opportunity tomorrow. Motivated by Roemer's formulation, there have been many attempts to measure IOP.[13] However, while "effort" figures prominently in the theory of IOP, it has been largely ignored in the empirical studies of IOP. Equality of opportunity is deemed to prevail if observed incomes do not vary with observed circumstances.[14] The main empirical approach to IOP measurement in the literature focuses on an estimate of the reduced-form equation for income, solving out effort.[15] As we will see, the predicted values from this reduced-form for income as a function of observed circumstances is not a valid welfare-metric when effort is a matter of personal choice.

3. Inequality of What? Observed versus Equivalent Incomes

In motivating existing measures of income inequality (whether in outcomes or opportunities) one might start by assuming that utility depends solely on income, and is some inter-personally constant function of income. Effort may matter for income, but there will be no interior solution for effort; everyone will work as hard as is humanly possible. While circumstances may still influence a person's maximum effort, this model is clearly unrealistic. It is also too simple to capture the way effort has been widely seen as a matter of personal choice and responsibility in policy debates.

Instead, following the long-standing approach in labor economics, utility is taken to be a function of effort (denoted x_i for person $I = 1, \ldots, n$) as well as total personal income (y_i), entering negatively and positively respectively.[16] (The relevant income concept for welfare is normally taken to be net of taxes. Here we can "solve out" taxes by treating them as a function of gross income.) There are two sources of heterogeneity. The first is in the circumstances relevant to income, denoted c_i. Second there is also heterogeneity in preferences, represented by an indexing of utility functions. We can write the utility function as $u_i(y_i, x_i)$ while income is:

$$y_i = y(x_i, c_i) \tag{1}$$

The function y is taken to be increasing in both arguments. Define:

$$\widetilde{u}_i(x_i, c_i) \equiv u_i[y(x_i, c_i), x_i]$$

---

[13] Contributions include Bourguignon et al. (2007), Barros et al. (2009), Checchi and Peragine (2010), Trannoy et al. (2010), Ferreira and Gignoux (2011), Ferreira et al. (2011), Hassine (2012), Marrero and Rodriguez (2012), Singh (2012) and Brunori et al. (2013). Also see the broader discussions in Pignataro (2011), Roemer (2014), Roemer and Trannoy (2015) and Ferreira and Peragine (2015).

[14] This is sometimes called "ex-ante" equality; "ex-post" equality requires equal reward for equal effort; see the discussion in Fleurbaey and Peragine (2013). For example, if someone starting out with a disadvantage in terms of her ability to generate income can make up the difference by hard work then one would surely be reluctant to say that there is no remaining inequality of opportunity; while the income difference according to circumstances may have vanished (no ex ante inequality), the difference in welfare remains (ex post inequality).

[15] This is explicit in Bourguignon et al. (2007), Trannoy et al. (2010) and Ferreira and Gignoux (2011), but implicit in most of the literature. Ferreira and Peragine (2015) claim that the method has been applied to at least 40 countries.

[16] Effort is bounded, but this is not made explicit for now since attention is confined to interior solutions for effort. (In the parametric model in Section 4 a time constraint will be explicit.)

It is assumed that:[17]

$$\widetilde{u}_{xx}(x_i, c_i) = u_y y_{xx} + y_x^2 u_{yy} + 2y_x u_{yx} + u_{xx} < 0$$

Effort is taken to be a matter of personal choice. The interior solution requires that:

$$\widetilde{u}_x(x_i, c_i) = u_y(y_i, x_i) y_x(x_i, c_i) + u_x(y_i, x_i) = 0 \tag{2}$$

The chosen effort (solving (1) and (2)) depends on circumstances and preferences, which we can write as $x_i = x_i(c_i)$.[18] The reduced-form equation for income can be written as:[19]

$$\widetilde{y}_i(c_i) \equiv y[x_i(c_i), c_i] \tag{3}$$

The corresponding regression specification in the literature typically takes the form:

$$y_i = \beta_0 + \beta_1 c_i + \varepsilon_i \tag{4}$$

where ε is treated as a zero-mean error term uncorrelated with circumstances ($E(\varepsilon_i | c_i) = 0$). Heterogeneity in preferences is relegated to the error term.[20]

When measuring inequality (or poverty) we typically aim to assure that the monetary metric of welfare is "real", which is normally identified by consistency with a model of utility. This is implemented using cost-of-living indices and equivalence scales or (more generally) equivalent income functions. The appeal of welfare consistency is no less obvious when effort matters. We are presumably concerned with how welfare varies with circumstances. However, on noting that utility is $u_i(\widetilde{y}_i(c_i), x_i)$ it is immediately evident that $\widetilde{y}_i(c_i)$ is only a valid monetary metric of welfare if effort is constant or does not matter to welfare. These must be deemed extremely strong assumptions. Similar comments apply to full income. Re-write (1) in the usual separable form:

$$y(x_i, c_i) = w(c_i) x_i + \pi(c_i) \tag{5}$$

The notation recognizes explicitly that circumstances determine the wage rate and unearned income, denoted $w(c_i)$ and $\pi(c_i)$ respectively. Suppose that all those working less than the stipulated standard hours (x^s) are able to make up the gap at their current average wage rate; there is no change for those working at or above x^s. Then full income is:

$$y_i^s \equiv w_i \max(x^s, x_i) + \pi_i \tag{6}$$

It can be readily shown that y_i^s is not a valid welfare metric (Kanbur and Keen 1989).[21]

---

[17] Subscripts for person i are dropped in places to simplify the notation. Twice differentiability is assumed when convenient. Subscripts are used for partial derivatives, in obvious notation. When convenient for the exposition, c and x are treated as continuous scalars (such as parental income and labor supply respectively), but they are vectors in reality and with discrete elements.

[18] Notice that this model is static, in that all effort is a current choice. In extending to a dynamic model one might postulate that there are also current gains from past efforts, which are taken as exogenous to choices about current effort. (An example is past effort at school versus current labor supply given schooling.)

[19] This is explicit in Bourguignon et al. (2007), Trannoy et al. (2010) and Ferreira and Gignoux (2011), but implicit in most of the literature.

[20] Of course, in practice ε also includes unobserved circumstances and measurement errors. A discussion of the econometric issues in specifying and estimating Equation (4) can be found in Ramos and Van de gaer (2016).

[21] A similar comment applies to the use of the wage rate as a metric of welfare.

We are after a money metric of utility, i.e., an income metric for a given person with given preferences that is a strictly increasing function of that person's attained utility, as judged by that person. The required concept is the equivalent income, y_i^*, defined by:

$$u_i(y_i^*, \overline{x}) = u_i(\widetilde{y}_i(c_i), x_i) \tag{7}$$

Thus, the equivalent income is the money income one would need to attain one's actual utility at a fixed reference level of effort, $\overline{x}$. The implied value of y_i^* is a monotonic increasing function of utility, although the precise function differs according to idiosyncratic preferences. By this approach, one measures the income inequality between two people, A and B, by comparing the income that A needs to attain A's actual utility, as judged by A's preferences, with that needed by B, judged by B's preferences, when both make the same level of effort. In general, the value of y_i^* will depend on the choice of the reference level of effort, $\overline{x}$. The empirical work will examine sensitivity to that choice.

On inverting the utility function (with the inverse w.r.t. income denoted u^{-1}) it is evident from (7) that:[22]

$$y_i^* = u_i^{-1}[u_i(\widetilde{y}_i(c_i), x_i(c_i)); \overline{x}] = f_i(c_i) \tag{8}$$

It is readily verified that better circumstances (meaning that $y_c > 0$) yield higher equivalent income. (Applying the envelope theorem, $f_c = y_c u_y / u_{y^*} > 0$.)

Whether there is more or less inequality in the equivalent income space than for observed incomes depends on the properties of the utility function and how both efforts and preferences vary across the population. We cannot determine the outcome solely by looking at how effort varies with observed income. One might find that mean effort (forming an expectation over the distribution of the preference parameters) rises with income, yet the variance in effort and preferences entails higher inequality of equivalent income than observed income. Indeed, one can readily construct examples in which mean effort is a non-decreasing function of income but the horizontal heterogeneity in effort at given income implies unambiguously higher inequality in the welfare space.

To illustrate, suppose that there are three income levels, $y = (1, 1, 2)$, with corresponding efforts $x = (0, 1, 1)$ and that welfare is $y - \alpha x$ for a preference parameter α with $0 < \alpha < 1$. Then the Lorenz curve for $y - \alpha x$ shifts out relative to that for y for the poorest two-thirds, but is unchanged for the top third.[23] For all measures satisfying the usual transfer axiom, inequality is higher (or no-lower) for welfare over this range of the preference parameter.[24] Higher poverty rates are also possible for some poverty lines and parameter values; for example, if the poverty line is 0.9 then nobody is income poor but 1/3 are welfare poor for all $\alpha \geq 0.1$.[25] While this is only one example, it suffices to disprove that welfare inequality is necessarily lower than income inequality when richer people tend to work harder.

Since nothing very general can be said in theory, the effect on measured inequality of adjusting for effort will be treated as an empirical question to be taken up in the next section.

4. An Empirical Analysis

The following example only aims to illustrate the sensitivity of inequality and poverty measures to addressing the concerns raised above. The empirical example will suffice to show that the kind of example given above—whereby welfare inequality is even higher than income inequality even when effort tends to rise with income—can be found in reality. And it will also illustrate that allowing for effort in a welfare-consistent way implies higher poverty measures. The discussion focuses solely

---

[22] In obvious notation and subsuming $\overline{x}$ in the definition of the equivalent-income function f.

[23] The interior points on the income Lorenz curve, $L(p)$, are $L(1/3) = 0.25$ and $L(2/3) = 0.5$, while those for the welfare Lorenz curve are $L(1/3) = (1 - \alpha)/(4 - 2\alpha) < 0.25$ and $L(2/3) = 0.5$. (Note that the two people with lowest incomes are re-ranked when one switches to the welfare space.)

[24] This claim uses the well-known Lorenz dominance condition (Atkinson 1970).

[25] This assumes a common poverty line; the empirical work will relax this to allow for leisure as a basic need.

on effort through labor supply. To keep things simple, the utility function is assumed to have the Cobb-Douglas functional form.

Any direct welfare effects of circumstances that are not evident in income or labor supply are ignored. This limitation is likely to be especially salient for disabilities and demographic effects on welfare due to differing numbers of children and family sizes.[26] In recognition of this concern, the analysis here is only done for a specific family type, namely single-person households, and excludes those with any (self-reported) disability. Thus a number of thorny issues of inter-household distribution, setting equivalence scales and making inter-personal welfare comparisons between those with and without disabilities are swept aside for the present purpose.

Data: The data are from the Annual Social and Economic Supplement of the Current Population Survey (CPS) for the U.S. for 2014 (with reference to incomes for 2013).[27] The analysis is confined to the roughly 6000 single-person households in the 2014 CPS.

Labor supply is measured by average hours of work per week in 2013.[28] The mean is 39 h (with a median is 40 h). The range in hours worked is from nearly zero to 99 h. Table 1 provides some key summary statistics and Figure 1 plots log hours worked per week in the last year against log total pre-tax income.[29] Mean labor supply for those with an income under $20,000 (the poorest 16%) is 30 h, while it falls to 26 h for those living under $15,000 (the poorest 8%) (Table 1). We see that mean (log) labor supply rises with income up to a certain point then levels off for the upper 30% or so (Figure 1).

Table 1. Summary statistics.

Income Cut-off (z)	% of Sample	Mean Hours of Work per Week ($\overline{h}_z$)	Mean Wage Rate ($/Hour) ($\overline{w}_z$)	Mean Income ($/Week) ($\overline{y}_z$)	% of Income Gap Covered by Working Average Hours per Week ($\frac{100(39.26-\overline{h}_z)\overline{w}_z}{(1048.26-\overline{y}_z)}$)	Extra Hours per Week to Reach Mean Income ($\frac{1048.26-\overline{y}}{\overline{w}_z}$)
10,000	4.03	23.66	5.62	119.15	9.44	165.32
15,000	8.31	26.35	7.10	177.20	10.52	122.68
20,000	15.11	29.56	8.26	244.96	9.97	97.25
25,000	22.67	31.64	9.38	304.60	9.61	79.28
30,000	29.66	33.00	10.28	354.90	9.28	67.45
35,000	38.20	34.50	11.40	411.69	8.52	55.84
Median	50.00	35.81	12.92	487.84	7.95	43.38
Maximum	100.00	39.26	24.09	1048.26	n.a.	0.00

Note: The median is $42,010. Means are calculated for all sample points up to z.

[26] This relates to the long-standing problem of inferring welfare from observed demand or supply behavior across demographically heterogeneous households (Pollak and Wales 1979; Browning 1992).

[27] The CPS data were accessed through the University of Minnesota's IPUMS-CPS site.

[28] This is obtained by multiplying reported weeks of work in the last year by reported average hours of work per week then dividing by 52.

[29] Recall that pre-tax income (y) is the relevant concept in the model in Section 2 in which taxes are solved-out, assuming that they are some function of y. Also note that the CPS does not ask for taxes paid so imputations of uncertain reliability are required.

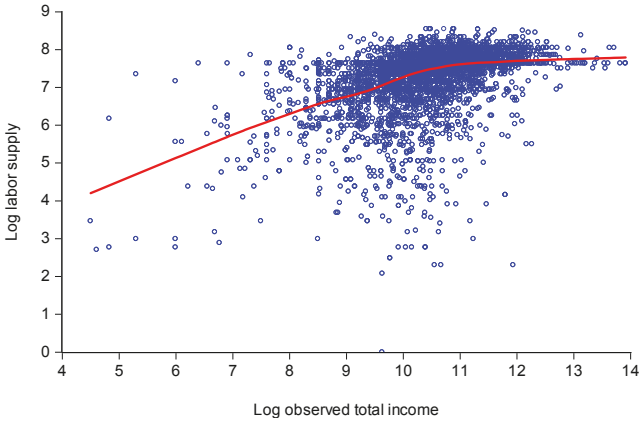


Figure 1. Labor supply plotted against total income for U.S. single adults in 2013. Note: The regression line is the "nearest neighbor" smothered scatter plot using a locally-weighted quadratic function. The overall quadratic regression (with White standard errors in parentheses) is: $\ln x_i = \underset{(0.994)}{-6.873} + \underset{(0.143)}{1.712} \ln y_i - \underset{(0.009)}{0.069} \ln y_i^2 + \hat{\varepsilon}_i R^2 = 0.206$; $n = 5863$.

While there is an income gradient in labor supplies, it does not appear to be large enough to plausibly account for much of the income disparities. For example, the average hourly wage rate of those with income less than \$20,000 is \$8.26. Ten hours extra work at this wage rate would only make up 10% of the gap between the average income of this group and the overall mean income.[30] Looked at a different way, this group of workers would have to work almost 100 h per week extra to reach mean income—equivalent to three full-time jobs. (Table 1 gives these calculations for various income cut-offs.)

While the income gradient in hours worked based on the regression function in Figure 1 does not seem especially steep, the pattern suggests that the partial effect of adjusting for effort as forgone leisure will go some way toward attenuating overall inequality in observed incomes. However, the large variance in labor supply at given income, especially at middle income levels evident in Figure 1 also comes into play. This "horizontal" effect is inequality increasing.

To see the net effect, consider first the measure of full income in which the standard for labor supply is set at 39 h. The assumption that the current wage can be maintained is questionable; to make up the hours, some may well have to switch to lower-paying jobs or incur prohibitively high personal costs of supplying the extra effort. So this simulation could well over-estimate the impact.

Figure 2 plots the full income against observed income (both in logs). There are some large proportionate gains, although they are spread through the income range. The first two rows of Table 2 give inequality measures for observed incomes and the full incomes. The full-time worker simulation brings down all three inequality measures. Figure 3 gives the Lorenz curves; there is not strict dominance, although the overlap does not happen until the 98th percentile.

When we come to incorporate effort in a welfare-consistent measure of income, this horizontal effect will again become important although then it will also interact with preferences. The net effect on measured inequality is thus an empirical issue to which we turn after describing the parametric model to be used.

[30] The overall mean weekly income of the sample is \$1048, while the mean weekly income of those living below \$20,000 per annum is \$245.

Table 2. Inequality measures for U.S. working singles without disabilities.

Income Concept	Gini Index	Mean Log Deviation (MLD)	Robin Hood Index
Observed income	0.402	0.296	0.284
Full income	0.387	0.262	0.275
Equivalent income without trimming extreme values	0.421	0.310	0.299
Equivalent income trimming extreme values	0.385	0.272	0.272

Note: Full incomes are calculated by assuming that all those working less than the mean hours of 39 per week were to work those hours at the same wage rate as at present. The equivalent incomes are explained in the text. The Gini index is half the average absolute difference between all pairs of incomes, expressed as a proportion of the mean. MLD is given by the mean of the log of the ratio of the overall mean income to individual income. Robin Hood index is the maximum vertical difference between the diagonal and the Lorenz curve, interpretable as the fraction of total income that one would need to take away from the richer half and give to the poorer half to assure equality.

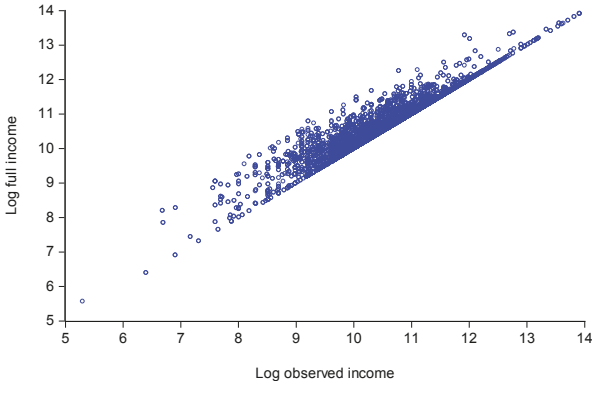


Figure 2. Plot of full incomes against observed incomes. Note: The full incomes are calculated by assuming that all those working less than average hours were to work average hours at the same wage rate as at present.

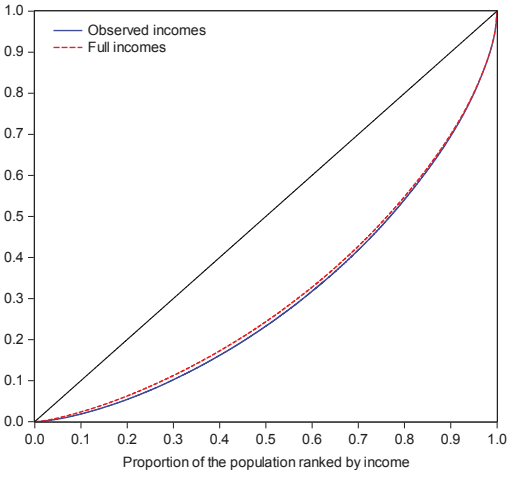


Figure 3. Lorenz curves for observed incomes and "full-employment" incomes.

Parametric model: In implementing an empirical model of income as a function of circumstances and effort the literature has often assumed a functional form that is additively-separable between effort and circumstances. However, it would clearly be questionable to assume that the marginal returns to circumstances are independent of effort. Indeed, in thinking about the economics one is drawn to postulate that the returns to effort (the wage rate when effort is simply labor supply) depend on circumstances—creating a natural interaction effect.

To consider the implications further, let us again write Equation (1) in the form of Equation (5). The values of $w(c_i)$ and $\pi(c_i)$ are the key parameters of effort choice. There are many possible assumptions one might make about preferences, and the results may well depend on the choice made. For the purpose of this example, a simple Cobb-Douglas representation is assumed, such that effort maximizes a utility function of the form:

$$u(y_i, x_i, \alpha_i) = \ln y_i + \alpha_i \ln(t - x_i) \tag{9}$$

where t is the total time available (so that $t - x_i$ is leisure time). The heterogeneity in preferences is taken to be fully captured by the differences in the α_i's. The (log) equivalent income is:

$$\ln y_i^* = \ln y_i + \alpha_i \ln\left(\frac{t - x_i}{t - \overline{x}}\right) \tag{10}$$

Note that $y_i^* \geq (\leq) y_i$ as $x_i \leq (\geq)\overline{x}$. Optimal labor supply requires $\alpha_i = w(c_i)(t - x_i)/y_i$; the latter is called here the leisure ratio (the ratio of the imputed value of leisure to income). Mean labor supply of 39 h per week is used as the reference, though sensitivity to this choice is discussed below.

Comparison of the empirical income inequality measures: There are a number of possible scenarios of interest for the parameters and data. It may be expected that the presence of the relatively few low labor supplies in Figure 1 will exaggerate the extent of inequality in equivalent incomes. To address this concern the following analysis is restricted to those households who worked for money at least one day (8 h) per week on average over 2013. This cuts out about 200 households.[31] The available time for work or leisure is set at 100, leaving out about 10 h per day. This seems reasonable.

In allowing the preference parameter to vary, one possibility is to assume that everyone in the survey has freely chosen their ideal labor supply, and to set $\alpha_i = w(c_i)(t - x_i)/y_i$ for all i. Results are given for this case, but it is questionable given the existence of labor-market frictions, whereby some survey respondents had too little leisure, and some too much, relative to their ideals. Setting the parameter to accord exactly with the leisure ratios in the survey data may be considered to produce an implausibly large variance. The spread of leisure ratios is evident in Figure 4. While the spread of empirical leisure ratios undoubtedly reflects labor-market frictions, measurement errors are also likely to be playing a role.

As an alternative, some degree of smoothing of the empirical leisure ratios is considered. For this purpose, the idiosyncratic preferences are set at the predicted values based on a regression of $\ln[w(c_i)(t - x_i)/y_i]$ on a quadratic function of the log wage rate, log unearned income (+\$1) (with their interactions) and a vector of observed circumstances from the CPS related to gender, age, race, place of birth, whether parents were born in the U.S. (Unfortunately, the data source does not include other information about parents, such as their education.) Age enters as the deviation from the median of 49 years. The left-out group for the dummy variables comprises white, native-born, males of 49 years of age with parents born in the U.S.; 25% of the sample is in this group. The Appendix A Table A1 gives the regression for the leisure ratio. Figure 4 gives the densities of the predicted leisure ratio, showing how this trims the extreme values.

[31] As noted, those reporting any disability affecting work or any difficulty (seeing, hearing, remembering, mobility, personal care) are excluded from the main analysis reported here. 5% of the sample reported a disability affecting their work.

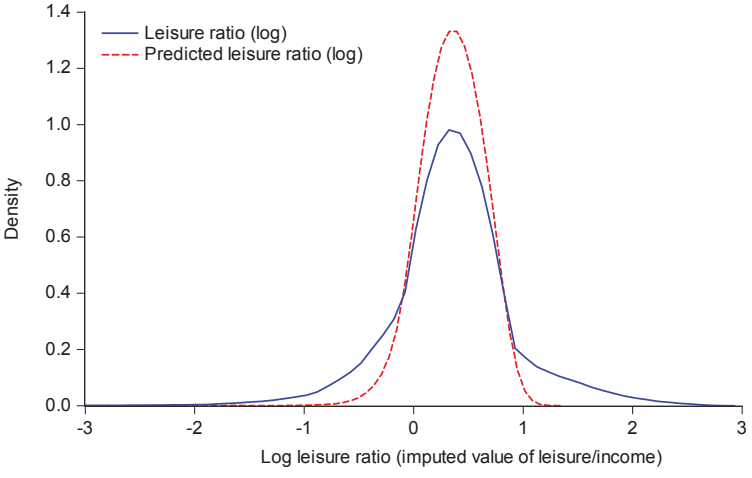


Figure 4. Kernel densities of the log leisure ratio.

We can now calculate the equivalent incomes. Figure 5 gives the kernel density functions for log observed income and log equivalent income, using both the actual leisure ratios and the trimmed ratios (using the aforementioned predicted values). Using the trimmed preferences, the effect of the adjustment for effort is to attenuate both tails, and bring the mode down slight. Without the trimming (using actual leisure ratios) we only see the attenuation at the bottom tail (roughly speaking implying less poverty), though we still see the fall in the mode.

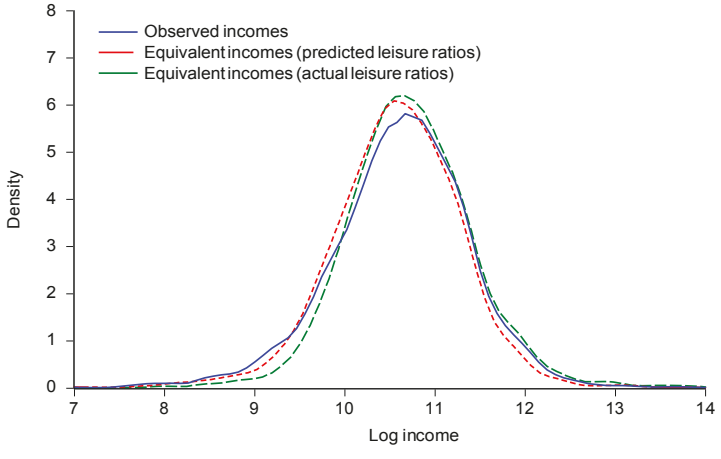


Figure 5. Kernel density functions for log incomes.

The effect of trimming the extremes in the preference parameter can be seen in Figure 6, which plots (log) equivalent income using the predicted leisure shares against those using the actual shares. As expected (based on Figure 4) there is a marked increase in the variance, especially around the middle. The Gini index rises to 0.421 (Table 2).

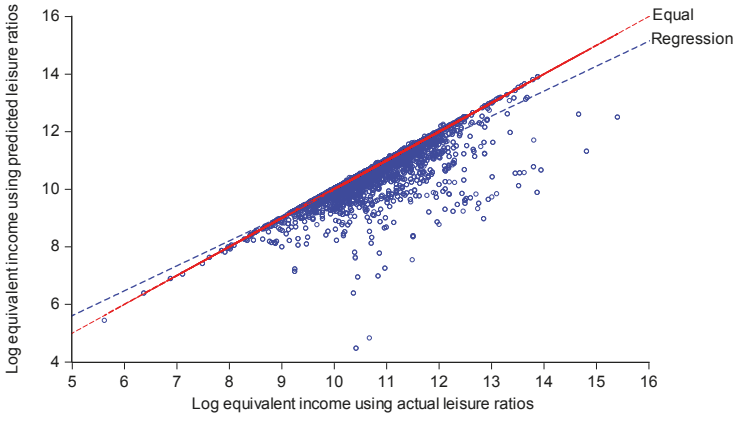


Figure 6. Effect of trimming the preference parameters.

Figure 7 plots log equivalent income (using predicted leisure ratios) against log observed income. The Figure also gives the regression lines, which have slopes that are significantly less than unity.[32] In other words, the adjustment for effort tends to raise (lower) equivalent incomes for the poor (rich). Equivalent incomes are also highly correlated with full incomes, again using a full-time job as the standard; in logs one finds that $r = 0.924$.

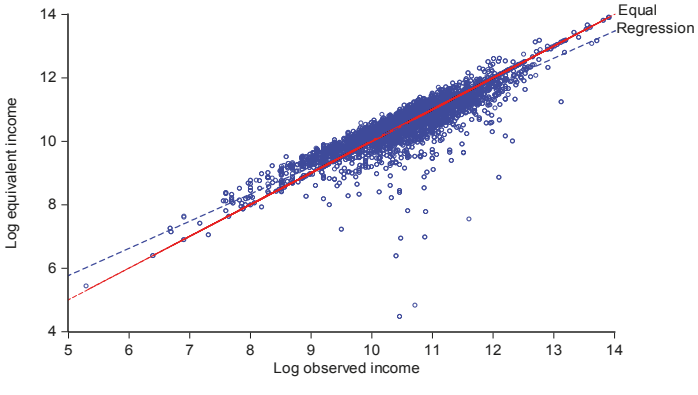


Figure 7. Plot of log equivalent income against log observed income. Note: Equivalent incomes based on predicted leisure ratios.

Table 2 also provides the same inequality indices for equivalent incomes and Figure 8 gives the Lorenz curves. On adjusting for effort without trimming the extremes of the preference parameters, the variance in the latter generates a marked outward shift in the Lorenz curve for the upper half; for the lower half the Lorenz curves are virtually indistinguishable, although there is not Lorenz dominance (so the ranking is not robust to the choice of inequality measure). The level of inequality falls when one adjusts for effort using the trimmed preference parameters. However, the effect is clearly very small.

[32] The regression coefficient is 0.856 (White s.e. = 0.006).

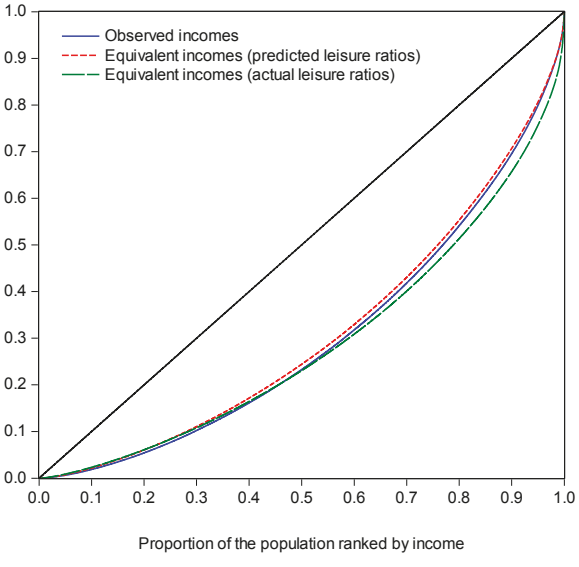


Figure 8. Lorenz curves for observed and equivalent incomes.

As already noted, the choice of reference alters equivalent income. Lowering (increasing) the reference level of effort increases (reduces) measured inequality. For example, using $\overline{x} = 30$ h per week (instead of the mean of 39) yields a Gini index for the equivalent incomes with trimming of 0.389. Using $\overline{x} = 50$ h per week one gets a Gini index of 0.378.

Poverty measures: Table 3 gives poverty rates based on observed incomes for two illustrative income poverty lines, namely \$15,000 and \$20,000 per year. The poverty rates are 8% and 17% respectively. The table also gives the poverty rates using full income and equivalent income (with and without the trimming). Using the same nominal line, the poverty rates fall by similar amounts for full income and equivalent income without trimming, but bounce back to values very close to those for unadjusted incomes when the data are smoothed.

Table 3. Poverty measures for U.S. working singles without disabilities.

	Income Poverty Line	
	\$15,000	**\$20,000**
Observed income	0.083	0.165
Full income	0.046	0.115
Equivalent income without trimming extreme values		
No basic need for leisure	0.045	0.103
Basic need = 10 h/week	0.081	0.155
Basic need = 20 h/week	0.129	0.216
Equivalent income trimming extreme values		
No basic need for leisure	0.082	0.158
Basic need = 10 h/week	0.133	0.219
Basic need = 20 h/week	0.191	0.283

Note: The basic need for leisure is valued at \$7 per hour. The poverty lines allowing for a basic need for leisure of 10 h per week are \$18,640 (for the \$15,000 income poverty line) and \$23,640 (for \$20,000). Allowing for a basic need for leisure of 20 h per week the corresponding lines are \$22,280 and \$27,280. (Also see notes to Table 2.)

However, to calculate poverty rates based on equivalent incomes it is compelling to adjust the poverty line consistently with that metric of welfare (as discussed in the introduction). Table 3 also gives poverty rates for two indicative allowances for leisure as a basic need, namely 10 and 20 h per week, each valued at \$7 per hour (the average wage of those with incomes under \$15,000 per year). These are not particularly generous allowances; on average (in 2015), the U.S. population over 15 years spent 36 h per week in leisure activities (Bureau of Labor Statistics 2016). So the figure of 20 h is only a little more than half the mean. However, while these choices can be questioned, the aim here is to assess sensitivity to allowing for leisure as a basic need. Using the unsmoothed data, one finds that even a seemingly modest allowance for leisure as a basic need of a little over 10 h per week is enough to obtain higher poverty rates using equivalent incomes; at a basic need of 20 h of leisure per week, the poverty rates rise to 26% and 35% for basic lines of \$15,000 and \$20,000 respectively. For the smoothed data, even a very small allowance for leisure of two hours per week is sufficient to yield a higher poverty rate for equivalent incomes than observed incomes.[33]

Covariates of income: To throw some light on implications for the structure of inequality and poverty, Table 4 gives regressions of log observed income and log equivalent income (with and without trimming the preference parameters using the predicted leisure ratios) against the same set of variables describing circumstances used in predicting the leisure share. The regressions are very similar. The female income differential is halved when one adjusts for labor supply, though it remains significant.[34] There are small differences in the effects of race and place of birth.[35] Some of these effects may well be confounded by differences in unemployment rates by gender or race, and labor-market discrimination.

Table 4. Testing for inequality of opportunity for U.S. working singles without disabilities.

	(1)			(2)			(3)		
	Log Observed Income			Log Equivalent Income (Predicted Leisure Ratios)			Log Equivalent Income (Actual Leisure Ratios)		
	Coeff.	s.e.	Prob.	Coeff.	s.e.	Prob.	Coeff.	s.e.	Prob.
Constant	10.842	0.019	0.000	10.727	0.019	0.000	10.847	0.018	0.000
Female	−0.107	0.021	0.000	−0.053	0.020	0.007	−0.054	0.019	0.005
Age-49 *	0.007	0.001	0.000	0.008	0.001	0.000	0.008	0.001	0.000
(Age-49) squared *	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.057
Race: Black	−0.224	0.026	0.000	−0.182	0.024	0.000	−0.189	0.024	0.000
Race: Black mixed	−0.142	0.117	0.223	−0.083	0.105	0.427	−0.108	0.109	0.321
Race: Am. Indian	−0.261	0.086	0.002	−0.227	0.079	0.004	−0.221	0.081	0.007
Race: Asian	0.152	0.069	0.028	0.149	0.063	0.019	0.198	0.065	0.002
Race: Other	−0.083	0.097	0.389	−0.106	0.087	0.225	−0.103	0.089	0.251
Hispanic	−0.162	0.037	0.000	−0.134	0.036	0.000	−0.127	0.034	0.000
Born US Oth.Terr.	−0.138	0.247	0.577	−0.145	0.223	0.514	−0.284	0.206	0.168
Born Central Am.	−0.724	0.197	0.000	−0.668	0.176	0.000	−0.667	0.157	0.000
Born Caribbean	−0.435	0.203	0.032	−0.430	0.183	0.019	−0.474	0.166	0.004
Born S. America	−0.311	0.215	0.149	−0.342	0.196	0.082	−0.438	0.178	0.014
Born N. Eur.	0.229	0.235	0.331	0.186	0.209	0.375	0.066	0.192	0.731
Born Western Eur.	−0.052	0.276	0.850	−0.118	0.248	0.633	−0.120	0.241	0.618
Born C-East Eur.	−0.249	0.206	0.226	−0.300	0.184	0.103	−0.410	0.163	0.012
Born East Asia	−0.314	0.212	0.139	−0.284	0.190	0.134	−0.309	0.175	0.078
Born SE Asia	−0.548	0.228	0.016	−0.594	0.212	0.005	−0.655	0.191	0.001
Born SW Asia	−0.143	0.226	0.526	−0.210	0.213	0.326	−0.299	0.186	0.108
Born Middle East	0.096	0.267	0.719	−0.021	0.246	0.932	0.088	0.244	0.717
Born Africa	−0.185	0.204	0.365	−0.283	0.185	0.127	−0.302	0.171	0.077
Foreign born	0.260	0.187	0.165	0.289	0.167	0.084	0.332	0.148	0.025
Foreign: Dad	0.106	0.059	0.073	0.087	0.056	0.117	0.083	0.058	0.152

[33] With two hours per week of leisure the poverty rate using the smoothed data is 9.1% using the \$15,000 income line and 16.9% using \$20,000.

[34] The data do not include work done within the home, though this is probably similar by gender in the sample of single adults.

[35] For example, the negative income effects of being born in South America or Center-Eastern Europe become somewhat larger (and statistical significant) using equivalent incomes based on the actual leisure ratios.

Table 4. *Cont.*

	(1)			(2)			(3)		
	Log Observed Income			Log Equivalent Income (Predicted Leisure Ratios)			Log Equivalent Income (Actual Leisure Ratios)		
	Coeff.	**s.e.**	**Prob.**	**Coeff.**	**s.e.**	**Prob.**	**Coeff.**	**s.e.**	**Prob.**
Foreign: Mom	0.158	0.074	0.034	0.173	0.062	0.006	0.228	0.068	0.001
Foreign: Both	0.132	0.056	0.018	0.119	0.051	0.020	0.087	0.050	0.083
N	5633			5633			5633		
R^2	0.088			0.077			0.068		
S.E. of regression	0.750			0.714			0.698		
Mean dep. var.	10.610			10.569			10.724		
F-statistic	21.740			18.600			16.373		
Prob (F-statistic)	0.000			0.000			0.000		

Note: White standard errors (s.e.). * coefficients scaled up by 100.

5. Conclusions

One often hears that high incomes are simply the reward for greater effort, and poverty reflects laziness, with the implication that there is less inequality and poverty than we think. Accepting that effort choice is a key factor in assessing inequality and that richer people tend to work more, this paper has shown that it is far from obvious that allowing for the disutility of effort implies less inequality or poverty.

If one takes seriously the idea that effort comes at a cost to welfare then it is clear that prevailing approaches are not using a valid monetary measure of welfare. While this much is obvious enough, the likely heterogeneity in effort must also be brought into the picture. Then the distributional outcome is far from obvious. It may be granted that average effort rises with income, but there is also a variance in effort at given income. The implications for measuring inequality and poverty stem from both the vertical differences (in how mean effort varies with income) and the horizontal differences (in how effort varies at given income).

It is unclear on a priori grounds what effect adjusting for effort in a welfare-consistent way will have on standard measures. There are both empirical and conceptual issues. The implications for measurement of taking effort seriously depend crucially on the behavioral responses to unequal opportunities, and not all of those responses are readily observable. Measures with a clearer welfare-economic interpretation call for data on efforts, for which existing surveys are limited to a subset of the dimensions of effort.

While acknowledging these limitations, the paper has provided illustrative calculations for American working singles without disabilities. A positive income gradient in labor supply is evident in the data. This gradient accounts for very little of the income gap between the poorest third (say) and the overall mean. The fact that poorer workers work less appears to contribute rather little to overall inequality in observed incomes. However, the considerable heterogeneity in effort at given incomes imparts a large horizontal element to inequality measures that adjust for effort consistently with behavior. On calculating distributions of welfare-consistent equivalent incomes to allow for this heterogeneity, the paper finds higher measures of inequality than for observed (unadjusted) incomes. Contrary to the common view, the prevailing practice of ignoring differences in effort understates inequality. It can be acknowledged, however, that some of the apparent heterogeneity in leisure preferences seen in the data is deceptive given likely rationing and measurement errors. When one smooths using predicted leisure shares based on covariates one finds a modest drop in the measured levels of inequality on adjusting for effort. Adjusting for effort does not appear to make much difference in the structure of inequality, as indicated by regressions using a set of circumstances related to gender, age, race and place of birth.

The implications for measures of poverty depend crucially on whether one sets the poverty line consistently with the welfare metric. If one does not do so, then poverty rates are lower using equivalent incomes although this essentially vanishes when one smooths the data. However, these

comparisons are arguably deceptive since one is not setting the poverty line consistently with how one is assessing welfare. To correct for this, one needs to include a normative allowance for leisure as a basic need in setting the poverty line. On introducing even a modest allowance valued at a low wage rate, one finds higher poverty rates when one adjusts for effort. If half the average amount of leisure taken by American adults is deemed to be a basic need then the poverty rate based on equivalent incomes, adjusted for effort, is nearly twice as high as that based on observed incomes.

Whether one accepts all the assumptions underlying these calculations is an open question. However, it is clear from this study that it should not be presumed that allowing for effort in a way that is broadly consistent with behavior would substantially attenuate the disparities suggested by standard data sources on income inequality or poverty.

Acknowledgments: The comments of Tony Atkinson, Kristof Bosmans, Francois Bourguignon, Denis Cogneau, Quy-Toan Do, Raquel Fernandez, Francisco Ferreira, Garance Genicot, Jèrèmie Gignoux, Ravi Kanbur, Erwin Ooghe, John Rust, Elizabeth Savage, Dominique van de Walle and the journal's two anonymous referees are gratefully acknowledged. The author thanks Naz Koont for very capable assistance in assembling the data files for Section 4.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Regression used to predict the leisure ratio to trim the extremes in allowing for idiosyncratic preferences.

	Log Leisure Ratio		
	Coeff.	**SE**	**Prob.**
Constant	0.256	0.056	0.000
Log wage rate	0.150	0.031	0.000
Log wage rate squared	−0.046	0.005	0.000
Log unearned income (+1)	−0.073	0.009	0.000
Log unearned income squared	−0.007	0.001	0.000
Log wage x log unearned income	0.036	0.002	0.000
Female	0.087	0.015	0.000
Age-49 *	0.000	0.058	0.992
(Age-49) squared *	0.024	0.003	0.000
Race: Black	0.075	0.019	0.000
Race: Black mixed	0.087	0.084	0.303
Race: American Indian	0.061	0.066	0.355
Race: Asian	0.020	0.052	0.705
Race: Other	−0.008	0.058	0.885
Hispanic	0.063	0.026	0.015
Born US Other Territories	−0.114	0.152	0.452
Born Central America	0.037	0.123	0.762
Born Caribbean	−0.042	0.128	0.746
Born South America	−0.107	0.138	0.438
Born Northern Europe	−0.161	0.164	0.325
Born Western Europe	−0.108	0.153	0.480
Born Central or Eastern Europe	−0.126	0.136	0.352
Born East Asia	0.036	0.136	0.789
Born SE Asia	−0.055	0.142	0.698
Born SW Asia	−0.144	0.150	0.337
Born Middle East	−0.136	0.170	0.423
Born Africa	−0.174	0.133	0.192
Foreign born	0.084	0.116	0.472
Foreign: Dad	−0.028	0.048	0.569
Foreign: Mom	0.023	0.051	0.660
Foreign: Both	−0.042	0.039	0.282
N	5633		
R^2	0.122		
S.E. of regression	0.529		
Mean dep. var.	0.348		
F-statistic	25.962		
Prob (F-statistic)	0.000		

Note: White standard errors (SE). * coefficients scaled up by 100.

References

Allingham, Michael. 1972. The Measurement of Inequality. *Journal of Economic Theory* 5: 163–69. [CrossRef]

Apps, Patricia, and Elizabeth Savage. 1989. Labour Supply, Welfare Rankings and the Measurement of Inequality. *Journal of Public Economics* 39: 335–64. [CrossRef]

Atkinson, Anthony B. 1970. On the Measurement of Inequality. *Journal of Economic Theory* 2: 244–63. [CrossRef]

Bargain, Olivier, Andre Decoster, Mathias Dolls, Dirk Neumann, Andreas Peichl, and Sebastian Siegloch. 2013. Welfare, Labor Supply and Heterogeneous Preferences: Evidence for Europe and the US. *Social Choice and Welfare* 41: 789–817. [CrossRef]

Barros, Ricardo Paes de, Francisco H. G. Ferreira, J. Molinas Vega, and J. Saavedra Chanduvi. 2009. *Measuring Inequality of Opportunities in Latin America and the Caribbean*. Washington: The World Bank.

Becker, Gary. 1965. A Theory of the Allocation of Time. *The Economic Journal* 75: 493–517. [CrossRef]

Blundell, Richard, Costas Meghir, Elizabeth Symons, and Ian Walker. 1988. Labor Supply Specification and the Evaluation of Tax Reforms. *Journal of Public Economics* 36: 23–52. [CrossRef]

Bourguignon, François. 2015. *The Globalization of Inequality*. Princeton: Princeton University Press.

Bourguignon, François, Francisco Ferreira, and Marta Menéndez. 2007. Inequality of Opportunity in Brazil. *Review of Income and Wealth* 53: 585–618. [CrossRef]

Browning, Martin. 1992. Children and Household Economic Behavior. *Journal of Economic Literature* 30: 1434–75.

Brunori, Paolo, Francisco Ferreira, and Vito Peragine. 2013. Inequality of Opportunity, Income Inequality and Economic Mobility: Some international comparisons. In *Getting Development Right*. Edited by Eva Paus. New York: Palgrave Macmillan, chp. 5.

Bureau of Labor Statistics. 2016. *American Time Use Survey*; Washington: United States Department of Labor.

Champernowne, David, and Frank Cowell. 1998. *Economic Inequality and Income Distribution*. Cambridge: Cambridge University Press.

Checchi, Daniele, and Vito Peragine. 2010. Inequality of Opportunity in Italy. *Journal of Economic Inequality* 8: 429–50. [CrossRef]

Coles, Jeffrey L., and Paul Harte-Chen. 1985. Real Wage Indices. *Journal of Labor Economics* 3: 317–36. [CrossRef]

Decoster, Andre, and Peter Haan. 2015. Empirical Welfare Analysis with Preference Heterogeneity. *International Tax and Public Finance* 22: 224–51. [CrossRef]

Eichelberger, Erika. 2014. Debunking the Attempted Debunking of Our 10 Poverty Myths, Debunked. *Mother Jones*, March 28.

Ferreira, Francisco, and Jèrèmie Gignoux. 2011. The Measurement of Inequality of Opportunity: Theory and an Application to Latin America. *Review of Income and Wealth* 57: 622–57. [CrossRef]

Ferreira, Francisco, and Vito Peragine. 2015. Individual Responsibility and Equality of Opportunity. In *Handbook of Well Being and Public Policy*. Edited by M. Adler and M. Fleurbaey. Oxford: Oxford University Press.

Ferreira, Francisco, Jèrèmie Gignoux, and Meltem Aran. 2011. Measuring Inequality of Opportunity with Imperfect Data: The Case of Turkey. *Journal of Economic Inequality* 9: 651–80. [CrossRef]

Fleurbaey, Marc, and Vito Peragine. 2013. Ex Ante versus Ex Post Equality of Opportunity. *Economica* 80: 118–30. [CrossRef]

Gans, Herbert. 1995. *The War against the Poor*. New York: Basic Books.

Hassine, Nadia Belhaj. 2012. Inequality of Opportunity in Egypt. *World Bank Economic Review* 26: 265–95. [CrossRef]

Jorgenson, Dale W., and Daniel Slesnick. 1984. Aggregate Consumer Behavior and the Measurement of Inequality. *Review of Economic Studies* 60: 369–92. [CrossRef]

Kanbur, Ravi, and Michael Keen. 1989. Poverty, Incentives, and Linear Income Taxation. In *The Economics of Social Security*. Edited by Andrew Dilnot and Ian Walker. Oxford: Oxford University Press.

Katz, Michael B. 1987. *The Undeserving Poor: From the War on Poverty to the War on Welfare*. New York: Pantheon Books.

King, Mervyn A. 1983. Welfare Analysis of Tax Reforms Using Household Level Data. *Journal of Public Economics* 21: 183–14. [CrossRef]

Marrero, Gustavo, and Juan Gabriel Rodriguez. 2012. Inequality of Opportunity in Europe. *Review of Income and Wealth* 58: 597–21. [CrossRef]

Pencavel, John. 1977. Constant-Utility Index Numbers of Real Wages. *American Economic Review* 67: 91–100.

Pew Research Center. 2014. *Most See Inequality Growing, but Partisans Differ over Solutions*. A Pew Research Center/USA TODAY Survey. Washington: Pew Research Center.

Pignataro, Giuseppe. 2011. Equality of Opportunity: Policy and Measurement Paradigms. *Journal of Economic Surveys* 26: 800–34. [CrossRef]

Pollak, Robert, and Terence Wales. 1979. Welfare Comparison and Equivalence Scale. *American Economic Review* 69: 216–21.

Preston, Ian, and Ian Walker. 1999. Welfare Measurement in Labour Supply Models with Nonlinear Budget Constraints. *Journal of Population Economics* 12: 343–61. [CrossRef]

Ramos, Xavier, and Dirk Van de gaer. 2016. Approaches to Inequality of Opportunity: Principles, Measures and Evidence. *Journal of Economic Surveys* 30: 855–83. [CrossRef]

Ravallion, Martin. 2016. *The Economics of Poverty. History, Measurement, Policy*. New York: Oxford University Press.

Roemer, John. 1998. *Equality of Opportunity*. Cambridge: Harvard University Press.

Roemer, John E. 2014. Economic Development as Opportunity Equalization. *World Bank Economic Review* 28: 189–209. [CrossRef]

Roemer, John, and Alain Trannoy. 2015. Equality of Opportunity. In *Handbook of Income Distribution Volume 2*. Edited by A. B. Atkinson and F. Bourguignon. Amsterdam: North Holland.

Salverda, Weimer, Marloes de Graaf-Zijl, Christina Haas, Bram Lancee, and Natascha Notten. 2014. The Netherlands: Policy-Enhanced Inequalities Tempered by Household Formation. In *Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries' Experiences*. Edited by Brian Nolan, Wiemer Salverda, Daniele Checchi, Ive Marx, Abigail McKnight, István György Tóth and Herman G. van de Werfhorst. Oxford: Oxford University Press.

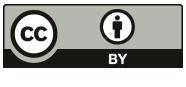
Singh, Ashish. 2012. Inequality of Opportunity in Earnings and Consumption Expenditure: The Case of Indian Men. *Review of Income and Wealth* 58: 79–106. [CrossRef]

Slesnick, Daniel. 1998. Empirical Approaches to the Measurement of Welfare. *Journal of Economic Literature* 36: 2108–65.

Stein, Ben. 2014. Poverty and Income Inequality. One really has Nothing to do with the Other. *The American Spectator*, April 4.

Trannoy, Alain, Sandy Tubeuf, Florence Jusot, and Marion Devaux. 2010. Inequality of Opportunities in Health in France: A First Pass. *Health Economics* 19: 921–38. [CrossRef] [PubMed]

Williamson, Kevin. 2014. Debunker Debunked. *National Review*, March 26.



© 2017 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Decomposing Wage Distributions Using Recentered Influence Function Regressions

Sergio P. Firpo [1], Nicole M. Fortin [2,*] and Thomas Lemieux [2]

[1] Insper Institute of Education and Research, R. Quatá, 300, São Paulo–SP 04546-042, Brazil; firpo@insper.edu.br

[2] Vancouver School of Economics, University of British Columbia, 6000 Iona Drive, Vancouver, BC V6T 1L4, Canada; thomas.lemieux@ubc.ca

* Correspondence: nicole.fortin@ubc.ca

Received: 31 December 2017; Accepted: 9 May 2018; Published: 25 May 2018

Abstract: This paper provides a detailed exposition of an extension of the Oaxaca-Blinder decomposition method that can be applied to various distributional measures. The two-stage procedure first divides distributional changes into a wage structure effect and a composition effect using a reweighting method. Second, the two components are further divided into the contribution of each explanatory variable using recentered influence function (RIF) regressions. We illustrate the practical aspects of the procedure by analyzing how the polarization of U.S. male wages between the late 1980s and the mid 2010s was affected by factors such as de-unionization, education, occupations, and industry changes.

Keywords: decomposition methods; RIF-regressions; wage inequality

JEL Classification: C18; J31

1. Introduction

The ongoing growth in wage inequality in the United States and several other countries over the past thirty-five years has generated a resurgence of interest for distributional issues and methods to analyze these issues. There is also a sizeable literature looking at wages differentials between subgroups that goes beyond simple mean comparisons. More generally, there is increasing interest in distributional impacts of various programs or interventions. In all these cases, the key question of economic interest is which factors account for changes (or differences) in distributions. For example, did wage inequality increase because education or other wage setting factors became more unequally distributed, or because the return to these factors changed over time?

In response to these important questions, several decomposition procedures have been suggested to untangle the sources of changes or differences in wage distributions. In Fortin et al. (2011), we reviewed the traditional Oaxaca-Blinder (OB) decomposition method and several of its extensions in the context of the treatment effect literature to highlight the advantages and disadvantages of different methodologies. The goal of the current paper is to provide a detailed and updated exposition of an extension to the OB decomposition that relies on recentered influence function (RIF) regressions (Firpo et al. 2009) [FFL, thereafter] to estimate the effect of covariates on inequality measures, such as percentile differences and ratios, the variance of log wages, or the Gini coefficient.[1] Relative to several procedures proposed recently (Machado and Mata 2005; Melly 2005; Chernozhukov et al. 2013) [CFM, thereafter], this method has the advantage of allowing general distributional measures to be

[1] Recentered influence functions have since been derived for a host of inequality measures by Essama-Nssah and Lambert (2012).

decomposed non-sequentially in the same way means can be decomposed using the conventional OB method. The methodology has been applied in a number of different settings where the object of interest is the unconditional distribution of outcomes.[2]

As is well known, the OB procedure provides a way of: (1) decomposing changes or differences in mean wages into a wage structure effect and a composition effect; and (2) further dividing these two components into the contribution of each covariate. The main problem with sequential decomposition methods is that they cannot be used to divide the composition effect into the role of each covariate in a way that is independent of the order of the decomposition. Thus, while it is natural to ask to what extent changes in the distribution of education have contributed to the growth in wage inequality, this particular question has not been answered in the literature for lack of available decomposition methods. In contrast, this question is straightforward to answer in the case of the mean using a OB decomposition.

In this paper, we focus on a two-stage procedure that can be used to perform OB type decompositions on any distributional measure, and not only the mean. The first stage consists of decomposing the distributional statistic of interest into a wage structure and a composition component using a reweighting approach, where the weights are either parametrically or non-parametrically estimated. As in the related program evaluation literature, we show that ignorability and common support are key assumptions required to identify separately the wage structure and composition effects. Provided that these assumptions are satisfied, the underlying wage setting model can be as general as possible. The idea of the first stage is thus very similar to DiNardo et al. (1996). Here, we clarify the assumptions required for the identification of distributional statistics besides the mean by drawing a parallel with the program evaluation (treatment effect) literature.

In the second stage, we further divide the wage structure and composition effects into the contribution of each covariate, just as in the usual OB decomposition. This is done using the regression-based method proposed by FFL to estimate the effect of changes in covariates on any distributional statistics such as inter-quartile ranges, the variance, or the Gini coefficient.

The method developed in FFL replaces the dependent variable of a regression by the corresponding recentered influence function (RIF) for the distributional statistics of interest. The influence function, also known as Gâteaux (1913) derivative, is a widely used concept in robust statistics and is easy to compute. Using the fact that the expected value of the influence function is equal to zero and the law of iterated expectations, we can express the distributional statistic of interest as the average of the conditional expectation of the RIF given the covariates. As in FFL, we call these conditional expectations RIF-regressions.

Average derivatives computed using the RIF-regressions yield the partial effect of a small location shift in the distribution of covariates on the distributional statistic of interest. FFL call this parameter *Unconditional Partial Effect* (UPE), which for the special case of quantiles become the *Unconditional Quantile Partial Effect* (UQPE). By approximating the conditional expectations by linear functions, the coefficients of these RIF-regressions indicate by how much the functional (e.g., the quantile) of the marginal outcome distribution is affected by an infinitesimal shift to the right in the distribution of the regressors.

Because the UPE parameter corresponds to the effect of infinitesimal shift in the distribution of regressors, it approximates well small changes in that distribution, but not necessarily large changes. For known changes in the distribution of covariates (e.g., between two time periods), one can easily compute the associated total change in the functional of the outcome distribution

---

[2] Eeckhout et al. (2014) compare the CFM approach to the RIF-regressions approach to decompose the skill distributions across large and small cities in terms of education, occupations, and industries, focusing on the bottom and top decile. Bento et al. (2017) provide a useful comparison of local kernel regressions, conditional quantile regressions, and RIF regressions in the context of a Monte-Carlo simulation of the effect of fuel economy standards on the distribution of vehicle weight.

of interest. Rothe (2012) proposes statistical inference for that case.[3] Both Rothe (2012) and CFM compute the conditional CDF (cumulative distribution function) of the outcome given covariates in the first step. This adds a computationally intensive layer of estimation, since one needs to calculate the entire conditional CDF, even if only interested in one single quantile of the marginal outcome distribution. By contrast, our approach requires only one OLS regression, which is very attractive from a computational standpoint. Finally, even though we end up performing bootstrap-based inference in our empirical application, we show in the Appendix B that the analytical formulas for the standard errors of the reweighting estimates can be derived.

The main advantage of using the RIF-regression method in a Oaxaca-Blinder type decomposition is that it provides a linear approximation of highly non-linear functionals, such as the quantiles or the Gini coefficient. Nevertheless, its simplicity comes at a cost. As pointed out by Rothe (2015), the impact of changes in the distribution of covariates on some non-linear functionals may be poorly approximated by RIF-regressions. Thus, approximation errors are a by-product of the method and they should always be reported in the decomposition results, as we do in our empirical analysis below.

We illustrate how our procedure works in practice by looking at changes in the distribution of male wages in the United States between the late 1980s and the mid 2010s. This period is quite interesting from a distributional point of view as inequality increased in the top end of the wage distribution, but decreased in the low end of the distribution, a phenomenon that Autor et al. (2006) referred to as the polarization of the U.S. labor market. We use our method to investigate the source of change in the wage distribution by decomposing the changes at various wage quantiles. The results indicate that no single factor appears to be able to fully explain the polarization of the wage distribution. De-unionization accounts for some of the decreasing wage inequality at the low end and increasing inequality at the top end. The continuing growth in returns to education, especially at a level above high school, is the most important source of growth in top-end inequality. Changes in the occupational structure of the workforce helps account for the polarization of wages, but these wage changes are mostly offset by changes in the effect of industry at the upper end of the distribution. This explains why, despite convincing evidence that the "routinization of jobs" had substantial impact of the polarization of employment, its effects of wage polarization has been more difficult to identify directly (e.g., Autor and Dorn 2013). Our results suggest that the wage decline in "routine occupations" (Autor et al. 2003), such as production jobs in the manufacturing sector, has been compensated by increases in the primary sector (e.g., mining, oil and gas, etc.), the distribution sector (transportation and wholesale) and in the services sector. Potentially offsetting effects underline the need for the proposed approach that can "run horse races" between different sets of factors. However, increases at the lower end appear to be attributable to changes in minimum wages, which we do not model here.[4]

The remainder of the paper is organized as follows. Section 2 discusses the decomposition problem and reviews the strengths and weaknesses of existing procedures. The identification of the proposed decomposition procedure is presented in Section 3. Section 4 discusses estimation and inference, and illustrates how the decomposition methodology works in the case of quantiles, the variance, and the Gini coefficient. Section 5 provides an empirical application of the methodology to the changes in the distribution of male wages in the United States between the late 1980s and the mid 2000s.

2. The Decomposition Problem and Shortcomings of Existing Methods

Before presenting our method in detail, it is useful to first review the case of the mean for which the standard OB method is very well known. To simplify the exposition, we will work with the case

[3] See also Rothe (2010).

[4] The federal minimum wage has declined substantially (in real terms) over time and is now superseeded by higher state minimum wages in most states. As a result, the effect of state and federal minimum wages would need to be modeled over of a range of wages. This task is beyond the scope of the current paper.

where the outcome variable, Y, is the wage, although our approach can be used for any other outcome variable. The OB method can be used to divide a difference in mean wages between two groups, or overall mean wage gap, into a composition effect linked to differences in covariates between the two groups, and a wage structure effect linked to differences in the return to these covariates between the two groups. The two groups are labeled as $t = 0, 1$. In the original papers by Oaxaca (1973) and Blinder (1973), the two groups used were either men and women, or blacks and whites. More generally, the two groups can be a control and a treatment group, or similar groups of individuals at two points in time, as in the wage inequality literature.

We first review how the OB decomposition provides a straightforward way of dividing up the contribution of each covariate in a composition and a wage structure effect. Focusing on differences in the wage distributions of two groups, 1 and 0, for a worker i, let Y_{1i} be the wage that would be paid in Group 1, and Y_{0i} the wage that would be paid in Group 0. Since a given individual i is only observed in one of the two groups, we either observe Y_{1i} or Y_{0i}, but never both. Therefore, for each i, we can define the observed wage, Y_i, as $Y_i = Y_{1i} \cdot T_i + Y_{0i} \cdot (1 - T_i)$, where $T_i = 1$ if individual i is observed in Group 1, and $T_i = 0$ if individual i is observed in group 0. There is also a vector of covariates $X \in \mathcal{X} \subset \mathbb{R}^K$ that we can observe in both groups.

In the standard OB decomposition, one assumes a linear functional form. In other words, one writes

$$Y_{ti} = X_i' \beta_t + \varepsilon_{ti}, \qquad \text{for} \quad t = 0, 1,$$

where $\mathbb{E}[\varepsilon_{ti} | X_i, T = t] = 0$.

Define the overall mean wage gap as $\Delta_O^\mu = \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$, and consider dividing the overall mean gap into a wage structure effect and a composition effect. Averaging over X, the mean wage gap Δ_O^μ can be written as

$$\begin{aligned} \Delta_O^\mu &= \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0] \\ &= \mathbb{E}[\mathbb{E}(Y|X, T = 1)|T = 1] - \mathbb{E}[\mathbb{E}(Y|X, T = 0)|T = 0] \\ &= \mathbb{E}\left[X|T = 1\right]' \beta_1 + \mathbb{E}\left[\varepsilon_1 | T = 1\right] - \left(\mathbb{E}\left[X|T = 0\right]' \beta_0 + \mathbb{E}\left[\varepsilon_0 | T = 0\right]\right), \end{aligned}$$

where $\mathbb{E}\left[\varepsilon_t | T = t\right] = 0$ because $\mathbb{E}\left[\varepsilon_t | X, T = t\right] = 0$, so the expression reduces to $\Delta_O^\mu = \mathbb{E}\left[X|T = 1\right]' \beta_1 - \mathbb{E}\left[X|T = 0\right]' \beta_0$. Thus, by adding and subtracting $\mathbb{E}\left[X|T = 1\right]' \beta_0$ we get

$$\Delta_O^\mu = \underbrace{\mathbb{E}\left[X|T = 1\right]' (\beta_1 - \beta_0)}_{\Delta_{S,OB}^\mu} + \underbrace{\left(\mathbb{E}\left[X|T = 1\right] - \mathbb{E}\left[X|T = 0\right]\right)' \beta_0}_{\Delta_{X,OB}^\mu}.$$

The first term in the equation is the wage structure effect, $\Delta_{S,OB}^\mu$, while the second term is the composition effect, $\Delta_{X,OB}^\mu$. Note that the reference group used to compute the wage structure effect here is the Group 0, though the decomposition could also be performed using Group 1 instead as the reference group. The wage structure and composition effects can also be written in terms of sums over the explanatory variables

$$\begin{aligned} \Delta_{S,OB}^\mu &= \sum_{k=1}^{K} \mathbb{E}\left[X^k | T = 1\right] (\beta_{1,k} - \beta_{0,k}), \\ \Delta_{X,OB}^\mu &= \sum_{k=1}^{K} \left[\mathbb{E}\left[X^k | T = 1\right] - \mathbb{E}\left[X^k | T = 0\right]\right] \beta_{0,k}, \end{aligned}$$

where X^k and $\beta_{t,k}$ represent the kth element of X and β_t, respectively. This provides a simple way of dividing $\Delta_{S,OB}^\mu$ and $\Delta_{X,OB}^\mu$ into the contribution of a single covariate or a group of covariates as needed.

Because of the linearity assumption, the OB decomposition is very easy to compute in practice. It can be estimated by replacing the parameter vectors β_t by their OLS estimates, and replacing the expected value of the covariates $\mathbb{E}[X \mid T = t]$ by the sample averages.

There are nonetheless some important limitations to the standard OB decomposition. A well-known difficulty discussed by Oaxaca and Ransom (1999) and Gardeazabal and Ugidos (2004) is that the contribution of each covariate to the wage structure effect, $\mathbb{E}\left[X^k | T = 1\right][\beta_{1,k} - \beta_{0,k}]$, is sensitive to the choice of the base group.[5]

A second limitation discussed by Barsky et al. (2002) is that the OB decomposition provides consistent estimates of the wage structure and composition effect only under the assumption that the conditional expectation is linear.[6] One possible solution to the problem is to estimate the conditional expectation using non-parametric methods. Another solution proposed by Barsky et al. (2002) is to use a (non-parametric) reweighting approach as in DiNardo et al. (1996) to perform the decomposition.[7] The advantage of this solution is that it can be applied to more general distributional statistics. The disadvantage of both solutions, however, is that they do not provide direct ways, in general, of further dividing the contribution of each covariate to the wage structure and composition effects.[8]

Currently available methods, such as DiNardo et al. (1996), can be used to compute the overall wage structure and composition effects for various distributional statistics. We build on this in the current paper by suggesting to estimate these two overall effects using a reweighting procedure. Available methods are much more limited, however, when it comes to further dividing the wage structure and, especially, the composition effect into the contribution each covariate. The main contribution of the paper is to explain how a simple regression-based procedure to remedy this shortcoming building on recent work by FFL.

3. Identification of General Composition and Structure Effects

3.1. Wage Structure and Composition Effects

Following the treatment effect literature (Rosenbaum and Rubin 1983, Heckman 1990, Heckman and Robb 1985, 1986), we focus on differences in the wage distributions between two groups, 1 and 0. Suppose we could observe a random sample of $N = N_1 + N_0$ individuals, where N_1 and N_0 are the number of individuals in each group and we index individuals by $i = 1, \ldots, N$. We define the probability that an individual i is in Group 1 as p, whereas the conditional probability that an individual i is in Group 1 given $X = x$, is $p(x) = \Pr[T = 1 | X = x]$, sometimes simply called the propensity score.

---

[5] Consider, for instance, the contribution of increasing returns to education to changes in mean wages over time in the case where workers are either high school graduates or college graduates. In the case where high school is the base group, $X_{i,k}$ is a dummy variable indicating that the worker is a college graduate, and $\beta_{0,k}$ and $\beta_{1,k}$ are the effect of college on wages in years $t = 0$ and 1. If returns to college increase over time ($\beta_{1,k} - \beta_{0,k} > 0$), then the contribution of education to the wage structure effect, $\overline{X}_{1,k}[\beta_{1,k} - \beta_{0,k}]$, is positive, where $\overline{X}_{1,k}$ is the share of college graduates. If we use instead college as the base group, then $\overline{X}_{1,k}[\beta_{1,k} - \beta_{0,k}]$ is negative, where $\overline{X}_{1,k}$ represents the share of high school ($\overline{X}_{1,k} = 1 - \overline{X}_{1,k}$) and $\beta_{t,k}$ represents the effect of high school ($\boldsymbol{\beta_{t,k} = -\beta_{t,k}}$). Thus, whether changes in returns to schooling contribute positively or negatively to the change in mean wages critically depends on the choice of the base group.

[6] As we show below, our goal is to estimate a counterfactual mean wage that would prevail if workers in Group 1 were paid under the wage structure of Group 0. Under the linearity assumption, this is equal to $\mathbb{E}[X \mid T = 1]' \beta_0$, a term that appears in both the wage structure and composition effect. The problem is that, when linearity does not hold, the counterfactual mean wage is not be equal to $\mathbb{E}[X \mid T = 1]' \beta_0$.

[7] Kline (2011) notes that, if the reweighting factor is linear in the covariates, the OB decomposition will yield a valid estimate of the counterfactual mean even if the conditional expectation is not linear in the covariates.

[8] We discuss the case of reweighting in more detail below. In the case where the conditional expectation $\mathbb{E}(Y_i | X_i, T = t)$ is estimated non-parametrically, a whole different procedure would have to be used to separate the wage structure into the contribution of each covariate. For instance, average derivative methods could be used to estimate an effect akin to the β coefficients used in standard decompositions. Unfortunately, these methods are difficult to use in practice, and would not be helpful in dividing up the composition effect into the contribution of each individual covariate.

Wage determination depends on some observed components X_i and on some unobserved components $\varepsilon_i \in \mathbb{R}^m$ through the wage structure functions

$$Y_{ti} = g_t(X_i, \varepsilon_i), \qquad \text{for} \quad t = 0, 1 \tag{1}$$

where $g_t(\cdot,\cdot)$ are unknown real-valued mappings: $g_t : \mathcal{X} \times \mathbb{R}^m \to \mathbb{R}^+ \cup \{0\}$. As we are not imposing any distribution assumption or specific functional form, writing Y_1 and Y_0 in this way does not restrict the analysis in any sense. We will however assume that (T, X, ε), or equivalently (Y, T, X), have an unknown joint distribution but that is far from being restrictive.

From observed data on (Y, T, X), we can non-parametrically identify the distributions of $Y_1|T = 1 \overset{d}{\sim} F_1$ and $Y_0|T = 0 \overset{d}{\sim} F_0$. Without further assumptions, however, we cannot identify the counterfactual distribution of $Y_0|T = 1 \overset{d}{\sim} F_C$. The counterfactual distribution F_C is the one that would have prevailed under the wage structure of Group 0, but with the distribution of observed and unobserved characteristics of Group 1. For the sake of completeness, we consider also the conditional distributions $Y_1|X, T = 1 \overset{d}{\sim} F_{1|X}$, $Y_0|X, T = 0 \overset{d}{\sim} F_{0|X}$ and $Y_0|X, T = 1 \overset{d}{\sim} F_{C|X}$.

We typically analyze the difference in wage distributions between Groups 1 and 0 by looking at some functionals of these distributions. Let ν be a functional of the conditional joint distribution of $(Y_1, Y_0)|T$, that is $\nu : \mathcal{F}_{\succeq} \to \mathbb{R}$, and $\mathcal{F}_{\succeq}$ is a class of distribution functions such that $F \in \mathcal{F}_{\succeq}$ if $\|\nu(F)\| < +\infty$. The difference in the νs between the two groups is called here the *ν-overall wage gap*, which is basically the difference in wages measured in terms of the distributional statistic ν:[9]

$$\Delta_O^\nu = \nu(F_1) - \nu(F_0) = \nu_1 - \nu_0. \tag{2}$$

We can use the fact that the distribution of X is not the same across groups to decompose Equation (2) into two parts:

$$\Delta_O^\nu = (\nu_1 - \nu_C) + (\nu_C - \nu_0) = \Delta_S^\nu + \Delta_X^\nu \tag{3}$$

where the second term Δ_X^ν reflects the effect of differences in the distribution of X.

The first term of the sum, Δ_S^ν, will reflect changes in the $g_t(\cdot,\cdot)$ functions only if we are able to fix the distribution of observables and unobservables as the one prevailing for Group 1, that is, the distribution of $(X, \varepsilon)|T = 1$. For that to be true, ν_C will be a functional evaluated at that distribution. This holds under the following assumptions: *Ignorability* and *Overlapping Support*.

The *Ignorability Assumption* has become popular in empirical research following a series of papers by Rubin and coauthors and by Heckman and coauthors.[10] In the program evaluation literature, this assumption is sometimes called *unconfoundedness* and allows identification of the treatment effect on the treated sub-population.

Assumption 1. [***Ignorability***]*: Let (T, X, ε) have a joint distribution. For all x in $\mathcal{X}$: ε is independent of T given $X = x$.*

The *Ignorability* assumption should be analyzed in a case-by-case situation, as it is more plausible in some cases than in others. In our case, it states that the distribution of the unobserved explanatory factors in the wage determination is the same across Groups 1 and 0, once we condition on a

[9] We sometimes refer to the functional $\nu(F_Z)$ simply as ν_Z. In the Oaxaca–Blinder decomposition discussed earlier, the parameter ν equals the mean ($\nu = \mu$) and Δ_O^ν is the total difference in mean wages.

[10] See, for instance, Rosenbaum and Rubin (1983, 1984), Heckman et al. (1997) and Heckman et al. (1998).

vector of observed components.[11] Now, consider the following assumption about the support of the covariates distribution:

Assumption 2. [***Overlapping Support***]: *For all x in $\mathcal{X}$, $p(x) = \Pr[T = 1|X = x] < 1$. Furthermore,* $\Pr[T = 1] > 0$.

The *Overlapping Support* assumption requires that there be an overlap in observable characteristics across groups, in the sense that there is no value of x in $\mathcal{X}$ such that it is only observed among individuals in Group 1.[12] Under these two assumptions, we are able to identify the parameters of the counterfactual distribution of $Y_0|T = 1 \overset{d}{\sim} F_C$. To see how the identification result works, let us define first three relevant weighting functions:

$$\omega_1(T) \equiv \frac{T}{p} \qquad \omega_0(T) \equiv \frac{1-T}{1-p} \qquad \omega_C(T, X) \equiv \left(\frac{p(X)}{1-p(X)}\right) \cdot \left(\frac{1-T}{p}\right).$$

The first two reweighting functions transform features of the marginal distribution of Y into features of the conditional distribution of Y_1 given $T = 1$, and of Y_0 given $T = 0$. The third reweighting function transforms features of the marginal distribution of Y into features of the counterfactual distribution of Y_0 given $T = 1$. We are now able to state our first identification result:[13]

Result 1. [***Inverse Probability Weighting***]:
Under Assumptions 1 and 2:
(i)

$$F_t(y) = \mathbb{E}[\omega_t(T) \cdot \mathbb{1}\{Y \leq y\}] \qquad t = 0, 1$$

(ii)

$$F_C(y) = \mathbb{E}[\omega_C(T, X) \cdot \mathbb{1}\{Y \leq y\}]$$

Identification of F_C implies identification of $\nu(F_C)$ and therefore of Δ_S^ν and Δ_X^ν. Furthermore, because of the ignorability assumption, we know that differences between the conditional distributions of $(X, \varepsilon)|T = 1$ and $(X, \varepsilon)|T = 0$ correspond only to differences in the conditional distributions $F_{X|T=1}$ and $F_{X|T=0}$. Thus, Δ_X^ν will only reflect changes in distribution of X. We state these results more precisely below.

Result 2. [***Identification of Wage Structure and Composition Effects***]:
Under Assumptions 1 and 2:
(i) Δ_S^ν, Δ_X^ν are identifiable from data on (Y, T, X);
(ii) if $g_1(\cdot, \cdot) = g_0(\cdot, \cdot)$ then $\Delta_S^\nu = 0$;[14]
(iii) if $F_{X|T=1} = F_{X|T=0}$, then $\Delta_X^\nu = 0$

In Result 2, the identification of Δ_S^ν and Δ_X^ν follows from the fact that these quantities can be expressed as functionals of the distributions obtained by weighting the observations with the inverse probabilities of belonging to Group 0 or 1 given T, as stated in Result 1. Note that the non-parametric identification of either the wage determination functions $g_1(\cdot, \cdot)$ and $g_0(\cdot, \cdot)$, or the distribution function

[11] This rules out selection into Group 1 or 0 based on unobservables.
[12] This is not a restrictive assumption when looking at changes in the wage distribution over time. Problems could arise, however, in gender wage gap decompositions where some of the detailed occupations are only held by men or by women.
[13] See also Firpo and Pinto (2016).
[14] Note that, even if $g_1(\cdot, \varepsilon) = h_1(\varepsilon)$ and $g_0(\cdot, \varepsilon) = h_0(\varepsilon)$, the result from Result 2 is unaffected. The intuition is that, since (X, ε) have a joint distribution, we can use the available information on that distribution to reweight the effect of the ε's on Y.

of ε are not necessary for the effects Δ_S^ν and Δ_X^ν to be identified. Therefore, methods based on conditional mean restrictions (the OB decomposition approach) and methods based on conditional quantile restrictions (the Machado and Mata (2005) approach) are based on too strong identification conditions that can be easily relaxed if we are simply interested in the terms Δ_S^ν and Δ_X^ν.

Part (*ii*) of Result 2 also states that, when there are no group differences in the wage determination functions, then we should find no wage structure effects. Part (*iii*) states that, if there are no group differences in the distribution of the covariates, there will be no composition effects.

Finally, it is interesting to relate these general results to the OB decomposition. Given the functional form assumptions of OB, the conditional mean zero expectation of ε and ignorability assumption, it follows that $\mathbb{E}\left[X|T=1\right]'\beta_0$ equals μ_C, the counterfactual mean or the expectation of Y_0 given $T=1$:

$$\begin{aligned}\mu_C = E[Y_0|T=1] &= E[g_0(X,\varepsilon)|T=1] = E[E(g_0(X,\varepsilon)|X,T=1)|T=1] \\ &= E[E(g_0(X,\varepsilon)|X,T=0)|T=1] \\ &= E[X|T=1]'\beta_0 + E[E(\varepsilon_0|X,T=0)|T=1] \\ &= E[X|T=1]'\beta_0\end{aligned}$$

In the following subsection, we show how one can generalize other features of the OB decomposition using a regression based approach, the RIF Regression.

3.2. The RIF Regressions

One important goal of the desired approach, as discussed in Section 2, is to apportion the wage structure and composition effects into the contribution of each individual covariate. To do so, we use the method proposed by FFL to compute partial effects of changes in distribution of covariates on a given functional of the distribution of $Y_t|T$. The method works by providing a linear approximation to a non-linear functional of the distribution. Thus, through collecting the leading term of a von Mises (1947) expansion, FFL approximate those non-linear functionals by expectations, which are linear functionals or statistics of the distribution. Finally, that linearization method allows one to apply the law of iterated expectations to the distributional statistics of interest and thus to compute approximate partial effects of changes in the distribution of each single covariate on the functional of interest.

The details of the method are summarized as follows. Consider again a general functional $\nu = \nu\left(F\right)$. Recall the definition of the influence function (Hampel 1974), IF, introduced as a measure of robustness of ν to outlier data when F is replaced by the empirical distribution: $\text{IF}(y;\nu,F) = \lim_{\epsilon\to 0}\left(\nu(F_\epsilon) - \nu(F)\right)/\epsilon$, where $F_\epsilon(y) = (1-\epsilon)F + \epsilon\delta_y$, $0 \le \epsilon \le 1$ and where δ_y is a distribution that only puts mass at the value y. It can be shown that, by definition, $\int_{-\infty}^{\infty} \text{IF}(y;\nu,F)\, dF(y) = 0$.

We use a recentered version of the influence function $\text{RIF}(y;\nu,F) = \nu(F) + \text{IF}(y;\nu,F)$ that has an expectation equal to the original ν :

$$\int \text{RIF}(y;\nu,F)\cdot dF\left(y\right) = \int \left(\nu(F) + \text{IF}(y;\nu,F)\right)\cdot dF\left(y\right) = \nu(F). \tag{4}$$

Letting $\nu_t = \nu(F_t)$ and $\nu_C = \nu(F_C)$, we can therefore write the distributional statistics ν_1, ν_0, and ν_C as the expectations: $\nu_t = \mathbb{E}\left[\text{RIF}(Y_t;\nu,F_t) \mid T=t\right]$, $t=0,1$ and $\nu_C = \mathbb{E}\left[\text{RIF}(Y_0;\nu,F_C) \mid T=1\right]$. Using the law of iterated expectations, the distributional statistics can also be expressed in terms of expectations of the conditional recentered influence functions

$$\nu(F) = \int \mathbb{E}\left[\text{RIF}(Y;\nu,F)|X=x\right]\cdot dF_X(x).$$

Letting the so-called RIF-regressions be written as $m_t^\nu(x) \equiv E[\text{RIF}(Y_t;\nu_t,F_t)|X,T=t]$, for $t=0,1$, and $m_C^\nu(x) \equiv E[\text{RIF}(Y_0;\nu_C,F_C)|X,T=1]$, we have

$$\nu_t = \mathbb{E}\left[m_t^\nu(X) \mid T=t\right], \quad t=0,1 \quad \text{and} \quad \nu_C = \mathbb{E}\left[m_C^\nu(X) \mid T=1\right]. \tag{5}$$

It follows that Δ_S^ν and Δ_X^ν can be rewritten as:

$$\begin{aligned}\Delta_S^\nu &= \mathbb{E}\left[m_1^\nu(X) \mid T=1\right] - \mathbb{E}\left[m_C^\nu(X) \mid T=1\right],\\ \Delta_X^\nu &= \mathbb{E}\left[m_C^\nu(X) \mid T=1\right] - \mathbb{E}\left[m_0^\nu(X) \mid T=0\right].\end{aligned}$$

As is well known, in the case of the mean, the influence function at point y is its deviation from the mean and, therefore, the recentered influence function of the mean is simply the point y itself

$$\begin{aligned}\text{IF}(y;\mu_t,F_t) &= \lim_{\epsilon\to 0}\frac{[(1-\epsilon)\cdot\mu_t + \epsilon\cdot y - \mu_t]}{\epsilon} = y-\mu_t, &(6)\\ \text{RIF}(y;\mu_t,F_t) &= \text{IF}(y;\mu_t,F_t)+\mu_t = y. &(7)\end{aligned}$$

As a result, the RIF-regression coefficients in the case of the mean are identical to standard regression coefficients of Y on X used in the OB decomposition (β_t above), and we have

$$\begin{aligned}\gamma_t^\mu &= \left(E[\omega_t(T)XX']\right)^{-1}\cdot E[\omega_t(T)XY], \quad t=0,1\\ \gamma_C^\mu &= \left(E[\omega_C(T,X)XX']\right)^{-1}\cdot E[\omega_C(T,X)XY],\end{aligned}$$

where $\gamma_t^\mu = \beta_t$, and

$$\begin{aligned}\Delta_S^\mu &= \mathbb{E}\left[X, T=1\right]'\cdot\left(\gamma_1^\mu - \gamma_C^\mu\right), &(8)\\ \Delta_X^\mu &= \left(\mathbb{E}\left[X|T=1\right] - \mathbb{E}\left[X|T=0\right]\right)'\cdot\gamma_0^\mu + R^\mu, &(9)\end{aligned}$$

where R^μ is an approximation error. When the linearity and zero conditional mean assumption of the OB decomposition are satisfied, it follows that $\gamma_C^\mu = \gamma_0^\mu$ and $R^\mu = 0$, as seen in the end of the previous subsection. Our decomposition is then identical to the OB decomposition. However, when these conditions are not satisfied the two decompositions are different.

In general, there is no particular reason to expect the conditional expectations $m_t^\nu(X)$ and $m_C^\nu(X)$ to be linear in X. As a matter of convenience and comparability with OB decompositions, it is nonetheless useful to consider the case of the linear specification. To be more precise, consider the linear projections (indexed by L) $m_L^\nu(x)$

$$m_{t,L}^\nu(x) = x'\gamma_t^\nu \qquad \text{and} \quad m_{C,L}^\nu(x) = x'\gamma_C^\nu,$$

where

$$\begin{aligned}\gamma_t^\nu &= \left(\mathbb{E}\left[XX' \mid T=t\right]\right)^{-1}\cdot\mathbb{E}\left[\text{RIF}(Y_t;\nu_t,F_t)X \mid T=t\right], \quad t=0,1,\\ \gamma_C^\nu &= \left(\mathbb{E}\left[XX' \mid T=1\right]\right)^{-1}\cdot\mathbb{E}\left[\text{RIF}(Y_0;\nu_C,F_C)X \mid T=1\right].\end{aligned}$$

As is well known, even though linear projections are only an approximation for the true conditional expectation, the expected approximation error is zero, so that:

$$\begin{aligned}\mathbb{E}\left[m_{t,L}^\nu(X) \mid T=t\right] &= \mathbb{E}\left[m_t^\nu(X) \mid T=t\right] \quad t=0,1\\ \text{and} \quad \mathbb{E}\left[m_{C,L}^\nu(X) \mid T=1\right] &= \mathbb{E}\left[m_C^\nu(X) \mid T=1\right].\end{aligned}$$

We can thus rewrite Δ_S^ν and Δ_X^ν as:

$$\Delta_S^\nu = \mathbb{E}\left[X|T=1\right]'\left(\gamma_1^\nu - \gamma_C^\nu\right), \tag{10}$$

$$\Delta_X^\nu = \mathbb{E}\left[X|T=1\right]'\gamma_C^\nu - \mathbb{E}\left[X|T=0\right]'\gamma_0^\nu, \tag{11}$$

which generalizes the OB decomposition to any distributional statistic through the projection of its recentered influence function onto the covariates. Note that, under an additional assumption that $m_{t,L}^\nu(\cdot) = m_t^\nu(\cdot)$ and $m_{C,L}^\nu(\cdot) = m_C^\nu(\cdot)$, that is, if the conditional expectation is indeed linear in x, then $\gamma_0^\nu = \gamma_C^\nu$. In the case of the mean ($\nu = \mu$), it then follows that the equations above reproduce exactly the OB decomposition.

It is important to note that the case of the mean is quite unique because the recentered influence function does not depend on the distribution F, i.e., $\text{RIF}(y;\mu,F) = \text{IF}(y;\mu,F) + \mu = y$. The lack of dependence on F is due to the fact that the influence function is a linear approximation that is exact in the case of the mean. For other distributional statistics, the approximation (or specification) error R is due to two separate factors. First, as in the case of the mean the conditional expectation of $\text{RIF}(y;\nu,F)$ given X may not be linear in X. Second, both the RIF and the projection coefficients γ depend on the distribution F. Thus, for more general distributional statistics, $\gamma_0^\nu = \gamma_C^\nu$ will not generally hold regardless of whether the conditional expectation is linear or not. As a result, we should expect to have a non-zero approximation error (see Equation (12)) for distributional statistics besides the mean, although how large the error is remains an empirical question.

3.3. Interpreting the Decomposition

We have just shown that, under a linearity assumption, the decomposition based on RIF-regressions is similar to a standard OB decomposition. We now go beyond this simple analogy to define more explicitly what we mean by the contribution of each single covariate to the wage structure and composition effects.

3.3.1. Composition Effects

FFL show that RIF-regression estimates can either be used to estimate the effect of a "small change" of the distribution of X on ν, or to provide a first-order approximation of a larger change of the distribution of X on ν. The latter effect, that FFL call a "policy effect" , is what concerns us here. In fact, the composition effect Δ_X^ν exactly corresponds to FFL's policy effect, where the " policy" consists of changing the distribution of X from its value at $T=0$ to its value at $T=1$ (holding the wage structure constant).

For the sake of simplicity, we continue to work with the linear specification introduced in Section 3.2. As it turns out, FFL show that, in the case of quantiles, using a linear specification for RIF-regressions generally yields very similar estimates to more flexible methods allowing for non-linearities.[15] We nonetheless discuss below the consequences of the linearity assumption for the interpretation of the results.

An explicit link with the results of FFL concerning policy effects is obtained by rewriting the composition effects as

$$\Delta_X^\nu = \left(\mathbb{E}\left[X|T=1\right] - \mathbb{E}\left[X|T=0\right]\right)'\gamma_0^\nu + R^\nu. \tag{12}$$

[15] This finding is closely linked to the well-known fact that estimates of marginal effects estimated using a linear probability model tend to be very similar, in practice, to those obtained using a probit, logit, or another flexible non-linear discrete response model.

where $R^{\nu} = \mathbb{E}[X|T=1]'\left(\gamma_C^{\nu} - \gamma_0^{\nu}\right)$. The first term in Equation (12) is now similar to the standard OB type composition effect, and can be rewritten in terms of the contribution of each covariate as

$$\sum_{k=1}^{K}\left(\mathbb{E}\left[X^k|T=1\right] - \mathbb{E}\left[X^k|T=0\right]\right)\gamma_{0,k}^{\nu}.$$

Each component of this equation can be interpreted as the " policy effect" of changing the distribution of one covariate from its $T=0$ to $T=1$ level, holding the distribution of the other covariates unchanged.

As discussed earlier, the second term in Equation (12), R^{ν}, is the approximation error linked to the fact that FFL's regression-based procedure only provides a first-order approximation to the composition effect Δ_X^{ν}. In practice, it can be estimated as the difference between the reweighting estimate of the composition effect, $\nu_C - \nu_0$, and the estimate of $\left(\mathbb{E}[X|T=1] - \mathbb{E}[X|T=0]\right)'\gamma_0^{\nu}$ obtained using the RIF-regression approach. When the latter approach provides an accurate (first-order) approximation of the composition effect, the error should be small. Looking at the magnitude of the error thus provides a specification test of FFL's regression-based procedure.

Note that using a linear specification for the RIF-regression instead of a general function $m^{\nu}(X) = \mathbb{E}[\text{RIF}(Y;\nu_t,F_t) \mid X]$ simply changes the interpretation of the specification error R^{ν} by adding an error component linked to the fact that a potentially incorrect specification may be used for the RIF-regression. We nonetheless suggest using the linear specification in practice for three reasons. First, we get an approximation error anyway since FFL's procedure only gives a first-order approximation to the impact of "large" changes in the distribution of X. Second, the linear specification does not affect the overall estimates of the wage structure and composition effects that are obtained using the reweighting procedure. Third, using a linear specification has the advantage of providing a much simpler interpretation of the decomposition, as in the OB decomposition. Our suggestion is thus to use the linear specification but also look at the size of the specification error to make sure that the FFL approach provides an accurate enough approximation for the problem at hand.[16]

3.3.2. Wage Structure Effect

The wage structure effect in Equation (11), $\Delta_S^{\nu} = \mathbb{E}[X|T=1]'\left(\gamma_1^{\nu} - \gamma_C^{\nu}\right)$, already looks very much like the usual wage structure effect in a standard OB decomposition. One important difference relative to the OB decomposition is that the coefficient γ_C^{ν} (the regression coefficient when the Group 0 data are reweighted to have the same distribution of X as Group 1) is used instead of γ_0^{ν} (the unadjusted regression coefficient for Group 0). The reason for using γ_C^{ν} instead of γ_0^{ν} is that the difference $\gamma_1^{\nu} - \gamma_C^{\nu}$ solely reflects differences between the wage structures $g_1(\cdot)$ and $g_0(\cdot)$, while the difference $\gamma_1^{\nu} - \gamma_0^{\nu}$ may be contaminated by differences in the distribution of X between the two groups.

In conventional regression analysis, the main reason why OLS estimates may depend on the distribution of X is that, when the conditional expectation of Y given X is non-linear, OLS minimizes a specification error that itself depends on the distribution of X (White 1980). An additional issue in our context is that for distribution statistics besides the mean, the recentered influence function $\text{RIF}(Y;\nu,F)$ depends on the distribution of Y (F). Changing the distribution of X changes the distribution of Y and, thus, the value of $\text{RIF}(Y;\nu,F)$ for a given value of Y. This also affects the coefficients in a regression of $\text{RIF}(Y;\nu,F)$ on X since we are no longer using the same RIF on the left hand side of the regression. As just discussed, this important problem can be addressed by estimating γ_C^{ν} in the reweighted sample,

[16] In the case of the mean, another rationale for using a linear model comes from Kline (2011), who notes that the OB decomposition remains valid even when the regression function is non-linear as long as the reweighting factor ω_C is well approximated by a linear odds ratio model. Unfortunately, this property does not hold for distributional statistics besides the mean.

which insures that the difference $\gamma_1^\nu - \gamma_C^\nu$ only reflects differences between the wage structures $g_1(\cdot)$ and $g_0(\cdot)$.

Another limitation of OB decompositions that also applies here is that the contribution of each covariate to the wage structure effect is sensitive to the choice of a base group. There is, unfortunately, no simple solution to this problem.[17] To see this, rewrite the wage structure effect

$$\begin{aligned} \Delta_S^\nu &= \nu_1 - \nu_C \\ &= [(\nu_1 - \nu_{B1}) - (\nu_C - \nu_{BC})] + (\nu_{B1} - \nu_{BC}), \end{aligned} \tag{13}$$

where ν_{B1} is the distributional statistic in an arbitrary "base group" under the wage structure $g_1(\cdot,\cdot)$, while ν_{BC} is the distributional statistic for the same base group under the wage structure $g_0(\cdot,\cdot)$. The term $\nu_1 - \nu_{B1}$ represents the "policy effect" of changing the distribution of X from its value in the base group to its $T = 1$ value under the wage structure $g_1(\cdot,\cdot)$, while $\nu_C - \nu_{BC}$ represents the corresponding policy effect under the wage structure $g_0(\cdot,\cdot)$. Since there is no dispersion in X in a base group of workers with similar characteristics, switching to the actual distribution of X will typically result in more wage dispersion. The overall wage structure effect is, thus, equal to the difference in the dispersion enhancing effect under $g_1(\cdot,\cdot)$ and $g_0(\cdot,\cdot)$, respectively, plus a "residual" difference in the distributional statistic in the base group, $\nu_{B1} - \nu_{BC}$. Unless this residual change is invariant to the choice of the base group, the contribution of each covariate to the wage structure will be sensitive to the choice of base group.

4. Estimation and Inference

In this section, we discuss how to estimate the different elements of the decomposition introduced in the previous section: ν_1, ν_0, ν_C, γ_1, γ_0 and γ_C. For ν_1, ν_0, γ_1 and γ_0, the estimation is very standard because the distributions F_1, and F_0, are directly identified from data on (Y, T, X). The distributional statistic ν_1, ν_0 can be estimated as their sample analogs in the data, while γ_1 and γ_0 can be estimated using standard least square methods. In contrast, the estimation of ν_C and γ_C requires first estimating the weighting function $\omega_C(T, X)$. We present two common methods—parametric and non-parametric—to estimate $\omega_C(T, X)$.

We discuss separately the estimation of the first and second stages of the decomposition. The first stage relies on a reweighting procedure, while the second stage is based on the estimation of RIF-regressions. We only present the general lines of the estimation procedure in this section. Proofs and details about the parametric and non-parametric procedure to estimate $\omega_C(T, X)$, and the asymptotic behavior of these estimators are discussed in the Appendix B and in Firpo and Pinto (2016). Finally, we show how the estimation procedure can be applied to the specific cases of the quantiles, interquantile ranges, variance and the Gini coefficient.

4.1. First Stage Estimation

The first step of the estimation procedure consists of estimating the weighting functions $\omega_1(T)$, $\omega_0(T)$ and $\omega_C(T, X)$. Then, the distributional statistics ν_1, ν_0, ν_C are computed directly from the appropriately reweighted samples. Details of the estimation procedure are presented in the Appendix B and in Firpo and Pinto (2016).

---

[17] In the case of the mean, several procedures have been suggested as potential solutions to the base group problem. They typically involve creating an artificial base group with the average observed characteristics in the population (see, e.g., Yun 2005). As this choice is as arbitrary as other choices of base group, and arguably harder to interpret, especially across studies, it does not really solve the base group problem. See Fortin et al. (2011) for a more complete discussion. In Footnote 29, we also discuss some issues with previous attempts (Firpo et al. 2007) using a normalization approach to the base group.

4.2. Second Stage Estimation

Now, consider estimation of the regression coefficients γ_1^ν, γ_0^ν, and γ_C^ν:

$$\begin{aligned}\widehat{\gamma}_t^\nu &= \left(\sum_{i=1}^N \widehat{\omega}_t^*(T_i)X_i{X_i}'\right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_t^*(T_i)\widehat{\text{RIF}}(Y_i;\nu_t,F_t)X_i, \quad t=0,1\\ \widehat{\gamma}_C^\nu &= \left(\sum_{i=1}^N \widehat{\omega}_C^*(T_i,X_i)X_i{X_i}'\right)^{-1} \cdot \sum_{i=1}^N \widehat{\omega}_C^*(T_i,X_i)\widehat{\text{RIF}}(Y_i;\nu_C,F_C)X_i\end{aligned}$$

where for $t=0,1$

$$\widehat{\text{RIF}}(y;\nu_t,F_t) = \widehat{\nu}_t + \widehat{\text{IF}}(y;\nu_t,F_t) \qquad \text{and} \qquad \widehat{\text{RIF}}(y;\nu_C,F_C) = \widehat{\nu}_C + \widehat{\text{IF}}(y;\nu_C,F_C),$$

and $\widehat{\text{IF}}(\cdot;\nu,F)$ is a proper estimator of the influence function. We discuss how to estimate the influence function for a number of specific cases in Section 4.3.

We can thus decompose the effect of changes from $T=0$ to $T=1$ on the distributional statistic ν as:

$$\begin{aligned}\widehat{\Delta}_S^\nu &= \left(\sum_{i=1}^N \widehat{\omega}_1^*(T_i)X_i\right)'\left(\widehat{\gamma}_1^\nu - \widehat{\gamma}_C^\nu\right)\\ \widehat{\Delta}_X^\nu &= \left(\sum_{i=1}^N \widehat{\omega}_1^*(T_i)X_i\right)'\widehat{\gamma}_C^\nu - \left(\sum_{i=1}^N \widehat{\omega}_0^*(T_i)X_i\right)'\widehat{\gamma}_0^\nu\end{aligned}$$

It is also useful to rewrite the estimate of the composition effect as

$$\widehat{\Delta}_X^\nu = \left(\sum_{i=1}^N \left(\widehat{\omega}_1^*(T_i) - \widehat{\omega}_0^*(T_i)\right)X_i\right)'\widehat{\gamma}_0^\nu + \widehat{R}^\nu,$$

where $\widehat{R}^\nu = \left(\sum_{i=1}^N \widehat{\omega}_1^*(T_i)X_i\right)'\left(\widehat{\gamma}_C^\nu - \widehat{\gamma}_0^\nu\right)$ is an estimate of the approximation error previously discussed. This generalizes the OB decomposition to any distributional statistic, including quantiles, the variance or the Gini coefficient.

4.3. Examples

We now turn to popular statistics, (unconditional) quantiles, the variance, and the Gini coefficient to illustrate how the different elements of the decomposition can be computed in these specific cases.

4.3.1. Quantiles and Interquantile Ranges

Quantiles are a set of distributional measures that have been used extensively for the decomposition of wage distributions. Several methodologies (Machado and Mata 2005; Melly 2005) use conditional quantiles regressions as primary tools to infer entire distributions and counterfactual distributions even when the object of interest is the unconditional quantiles. For instance, in decompositions of the gender wage gap, they are used to address issues such as glass ceilings and sticky floors.

The τ-th quantile of the distribution F is defined as the functional, $Q(F,\tau) = \inf\{y|F(y) \geq \tau\}$, or as q_τ for short, and its influence function is:

$$\text{IF}(y;q_\tau,F) = \frac{\tau - \mathbb{1}\{y \leq q_\tau\}}{f_Y(q_\tau)}. \tag{14}$$

As shown in FFL, the recentered influence function of the τth quantile is

$$\text{RIF}(y;q_\tau,F) = q_\tau + \text{IF}(y;q_\tau,F) = q_\tau + \frac{\tau - \mathbb{1}\{y \leq q_\tau\}}{f_Y(q_\tau)} = c_{1,\tau}\cdot \mathbb{1}\{y > q_\tau\} + c_{2,\tau}.$$

where $c_{1,\tau} = 1/f_Y(q_\tau)$, $c_{2,\tau} = q_\tau - c_{1,\tau}\cdot(1-\tau)$, and $f_Y(q_\tau)$ is the density of Y evaluated at q_τ. Thus,

$$\mathbb{E}[\text{RIF}(Y;q_\tau,F)|X=x] = c_{1,\tau}\cdot \Pr[Y > q_\tau|X=x] + c_{2,\tau}.$$

and the estimation of conditional mean of the $\text{RIF}(Y;q_\tau,F)$ can be seen more intuitively as the estimation of a conditional probability model of being below or above the quantile of interest q_τ, rescaled by a factor $c_{1,\tau}$ to reflect the relative importance of the quantile to the distribution, and recentered by a constant $c_{2,\tau}$.

The decomposition of (unconditional) quantiles proceeds along the same steps as in the case of the mean. In the first stage, the estimates of $q_{\tau t}$, $t=0,1$ and $q_{\tau C}$ are obtained by reweighting as $\widehat{q}_{\tau t} = \arg\min_q \sum_{i=1}^N \widehat{\omega}_t(T_i)\cdot\rho_\tau(Y_i-q)$, $t=0,1$, and $\widehat{q}_{\tau C} = \arg\min_q \sum_{i=1}^N \widehat{\omega}_C(T_i,X_i)\cdot\rho_\tau(Y_i-q)$. The function $\rho_\tau(\cdot)$ is the well known check function, proposed by Koenker and Bassett (1978), where, for any u in $\mathbb{R}$, $\rho_\tau(u) = u\cdot(\tau - 1\{u\leq 0\})$. Note that $\widehat{q}_{\tau t}$ and $\widehat{q}_{\tau C}$ can simply be computed using standard software packages with the appropriate weighting factor.

The estimators for the gaps are computed as:

$$\widehat{\Delta}_O^{q_\tau} = \widehat{q}_{\tau 1} - \widehat{q}_{\tau 0}; \quad \widehat{\Delta}_S^{q_\tau} = \widehat{q}_{\tau 1} - \widehat{q}_{\tau C} \quad \text{and} \quad \widehat{\Delta}_X^{q_\tau} = \widehat{q}_{\tau C} - \widehat{q}_{\tau 0}. \tag{15}$$

In the second stage, we estimate the linear RIF-regressions. First, the recentered influence function is computed for each observation by plugging the sample estimate of the quantile, $\widehat{q}_\tau$, and estimating the density at the sample quantile, $\widehat{f}(\widehat{q}_\tau)$.

For the τ quantile of $Y_1|T=1$, we would use $\widehat{\text{RIF}}(y;q_{\tau 1},F) = \widehat{q_{\tau 1}} + \left(\widehat{f}_1(\widehat{q}_{\tau 1})\right)^{-1}\cdot(\tau - \mathbb{1}\{y\leq\widehat{q_{\tau,1}}\})$ where $\widehat{f}_1(\cdot)$ is a consistent estimator for the density of $Y_1|T=1$, $f_1(\cdot)$. For example, kernel methods can be used to estimate the density, but other simpler alternative methods are also available. For example, one may dispense with estimation of the density by kernel by noticing that $c_{1,\tau} = dq_\tau/d\tau$. By estimating sufficiently close quantiles, say q_τ and $q_{\tau+\lambda}$, where λ is a small positive real number, an estimate of $c_{1,\tau}$ is $\widehat{c}_{1,\tau} = (\widehat{q}_{\tau+\lambda} - \widehat{q}_\tau)/\lambda$, which is the inverse of the sparsity density estimator (Koenker 2005, p. 139). Another interesting alternative method is the recent one suggested by Cattaneo et al. (2017), which uses local polynomial regressions.

In the example of $Y_1|T=1$, the RIF-regressions are estimated by replacing the usual dependent variable, Y, by the estimated value of $\widehat{\text{RIF}}(y;q_{\tau 1},F)$. Standard software packages can be used to do so. The resulting regression coefficients are therefore

$$\widehat{\gamma}_t^{q_\tau} = \left(\sum_{i=1}^N \widehat{\omega}_t(T_i)X_iX_i'\right)^{-1}\cdot\sum_{i=1}^N \widehat{\omega}_t(T_i)X_i\widehat{\text{RIF}}(Y_i;q_{\tau t},F_t), \quad t=0,1, \tag{16}$$

$$\widehat{\gamma}_C^{q_\tau} = \left(\sum_{i=1}^N \widehat{\omega}_C(T_i,X_i)X_iX_i'\right)^{-1}\cdot\sum_{i=1}^N \widehat{\omega}_C(T_i,X_i)X_i\widehat{\text{RIF}}(Y_i;q_{\tau C},F_C). \tag{17}$$

Similar to the case of the mean, we get:

$$\widehat{\Delta}_S^{q_\tau} = \mathbb{E}[X,T=1]'\left(\widehat{\gamma}_1^{q_\tau} - \widehat{\gamma}_C^{q_\tau}\right), \tag{18}$$

$$\widehat{\Delta}_X^{q_\tau} = (\mathbb{E}[X|T=1] - \mathbb{E}[X|T=0])'\widehat{\gamma}_0^{q_\tau} + \widehat{R}^{q_\tau}, \tag{19}$$

where $\widehat{R}^{q_\tau} = \mathbb{E}[X|T=1]'\left(\widehat{\gamma}_C^{q_\tau} - \widehat{\gamma}_0^{q_\tau}\right)$.

Interquantile ranges, such as the difference between the 75th and the 25th percentiles, and the 90–10 gap (difference between 90th and the 10th percentiles) are also popular inequality measures that only depend on quantiles. Because they are simple differences between quantiles, their γ coefficients are the differences in the γ coefficients of their respective quantiles. For that reason, we omit the theoretical discussion about interquantile ranges, but present their estimates in the empirical section.

4.3.2. Variance

There are other applications where it is useful to decompose the impact of covariates on the variance of the distributions of log wages. Examples include the compression effect of unions and of public sector wage setting.

The estimators of these gaps can be computed as:

$$\widehat{\Delta}_O^{\sigma^2} = \widehat{\sigma}_1^2 - \widehat{\sigma}_0^2; \quad \widehat{\Delta}_S^{\sigma^2} = \widehat{\sigma}_1^2 - \widehat{\sigma}_C^2 \quad \text{and} \quad \widehat{\Delta}_X^{\sigma^2} = \widehat{\sigma}_C^2 - \widehat{\sigma}_0^2, \tag{20}$$

using the reweighting scheme $\widehat{\sigma}_t^2 = \sum_{i=1}^N \widehat{\omega}_t^*(T_i)\,(Y_i - \widehat{\mu}_t)^2$, $t = 0, 1$, and $\widehat{\sigma}_C^2 = \sum_{i=1}^N \widehat{\omega}_C^*(T_i, X_i) \cdot (Y_i - \widehat{\mu}_C)^2$. The influence function of the variance is well-known to be

$$\text{IF}(y; \sigma^2, F_Y) = \left(y - \int z \cdot dF_Y(z)\right)^2 - \sigma^2, \tag{21}$$

and the recentered influence function is the first term of this expression $\text{RIF}(y; \sigma^2, F_Y) = (y - \int z \cdot dF_Y(z))^2 = (Y - \mu)^2$.

The decomposition in terms of individual covariates, such as union coverage, follows by replacing $\text{RIF}(\cdot; q_\tau)$ by $\text{RIF}(\cdot; \sigma^2, F)$ in Equations (16)–(19).

4.3.3. The Gini coefficient

Finally, another popular measure of wage inequality is the Gini coefficient. There are a few papers (Choe and Van Kerm 2014; Gradín 2016) that have begun to use RIF-Gini regressions to investigate changes in income inequality. Recall that the Gini coefficient is defined as

$$\nu^G(F_Y) = 1 - 2\mu^{-1}R(F_Y) \tag{22}$$

where $R(F_Y) = \int_0^1 GL(p; F_Y)dp$ with $p(y) = F_Y(y)$ and where $GL(p; F_Y)$ is the generalized Lorenz ordinate of F_Y given by $GL(p; F_Y) = \int_{-\infty}^{F^{-1}(p)} z\, dF_Y(z)$. The generalized Lorenz curve tracks the cumulative total of y divided by total population size against the cumulative distribution function. The generalized Lorenz ordinate can be interpreted as the proportion of earnings going to the $100p$% lowest earners.

Monti (1991) derives the influence function of the Gini coefficient as

$$\text{IF}(y; \nu^G, F_Y) = A_2(F_Y) + B_2(F_Y)y + C_2(y; F_Y) \tag{23}$$

where $A_2(F_Y) = 2/\mu^{-1}R(F_Y)$, $B_2(F_Y) = 2\mu^{-2}R(F_Y)$, and $C_2(y; F_Y) = -2/\mu^{-1}[y\,[1 - p(y)] + GL\,(p(y); F_Y)$ with $R(F_Y)$ and $GL(p(y); F_Y)$ as defined underneath Equation (22). Recentering yields

$$\text{RIF}(y; \nu^G, F_Y) = 1 + B_2(F_Y)y + C_2(y; F_Y). \tag{24}$$

The recentered influence function of the Gini coefficient can also be written as

$$\text{RIF}(y; \nu^G, F_Y) = 2\frac{y}{\mu}\nu^G + \frac{(1-y)}{\mu} + \frac{2}{\mu}\int z F_Y(z) dz,$$

which gives a more intuitive expression after integrating by parts

$$\text{RIF}(y;\nu^G,F_Y) = 2\frac{y}{\mu}\left[F_Y(y) - \frac{(1+\nu^G)}{2}\right] + 2\left[\frac{(1-\nu^G)}{2} - GL(p;F_Y)\right] + \nu^G,$$

where $(1+\nu^G)/2$ and $(1-\nu^G)/2$ correspond, respectively, to the areas above and below the Lorenz curve. As pointed out by Monti (1991), the first term is unbounded because it increases by the factor y/μ, while the second is bounded between $\nu^G - 1$ and $1+\nu^G$. Thus, the $\text{RIF}(y;\nu^G,F_Y)$ is continuous and convex in y; its first derivative is equal to $2/\mu[F_Y(y) - (1+\nu^G)/2]$, and it reaches its minimum when $F_Y(y) = (1+\nu^G)/2$. The function is theoretically unbounded from above, but in practice it reaches its maximum at the upper bound of the empirical support of the distribution. This implies that the Gini coefficient is not robust to measurement error in high earnings, as pointed out by Cowell and Victoria-Feser (1996).

The GL coordinates are estimated using a series of discrete data points $y_1, \ldots y_N$, where observations have been ordered so that $y_1 \leq y_2 \leq \ldots \leq y_N$. Consider

$$\begin{aligned}
\widehat{p_t}(y_i) &= \frac{\sum_{j=1}^{i}\widehat{\omega}_t(T_j)}{\sum_{j=1}^{N}\widehat{\omega}_t(T_j)}, \qquad \widehat{GL_t}(p(y_i)) = \frac{\sum_{j=1}^{i}\widehat{\omega}_t(T_j)\cdot Y_j}{\sum_{j=1}^{N}\widehat{\omega}_t(T_j)} \quad t=0,1 \\
\widehat{p_C}(y_i) &= \frac{\sum_{j=1}^{i}\widehat{\omega}_C(T_j,X_j)}{\sum_{j=1}^{N}\widehat{\omega}_C(T_j,X_j)}, \qquad \widehat{GL_C}(p(y_i)) = \frac{\sum_{j=1}^{i}\widehat{\omega}_C(T_j,X_j)\cdot Y_j}{\sum_{j=1}^{N}\widehat{\omega}_C(T_j,X_j)}
\end{aligned}$$

where the numerators are the sum of the i ordered values of Y. The $\widehat{R}(F_t)$, $t=0,1$ and $\widehat{R}(F_C)$ are obtained by numerical integration of $\widehat{GL_t}(p(y_i))$ over $\widehat{p_t}(y_i)$, and of $\widehat{GL_C}(p(y_i))$ over $\widehat{p_C}(y_i)$.[18] The estimates of $\widehat{v}^G(F_t)$, $t=0,1$ and $\widehat{v}^G(F_C)$ are obtained by substituting $\widehat{R}(F_t)$ and $\widehat{R}(F_C)$, as well as $\widehat{\mu}_t$ and $\widehat{\mu}_C$, into Equation (22). We can then compute the gaps for the changes in the Gini coefficient as in Equation (20).

Similar substitutions into Equation (24) allows the estimation of $\widehat{\text{RIF}}(y;\nu_t^G,F_t)$, $t=0,1$ and $\widehat{\text{RIF}}(y;\nu_C^G,F_C)$. As before, the decomposition in terms of individual covariates, follows by replacing $\widehat{\text{RIF}}(\cdot;q_\tau,F)$ by $\widehat{\text{RIF}}(\cdot;\nu^G,F)$ in Equations (16)–(19).

5. Empirical Application: Changes in Male Wage Inequality between 1988 and 2016

Our empirical application focuses on changes in wage inequality over the past 30 years. It is well known that wage inequality increased sharply in the United States since the beginning of the 1980s. Using various distributional methods, Juhn et al. (1993) and DiNardo et al. (1996) showed that inequality expanded all through the wage distribution during the 1980s. In particular, both the "90–50 gap" (the difference between the 90th and the 50th quantile of log wages) and the "50–10 gap" increased during this period.

Since the late 1980s, however, changes in inequality have increasingly been concentrated at the top end of the wage distribution. In fact, Autor et al. (2006) showed that, while the 90–50 gap kept expanding after the late 1980s, the 50–10 gap declined during the same period. They refer to these changes as an increased polarization of the labor market. An obvious question is why wage dispersion has changed so differently at different points of the distribution. Autor et al. (2006) suggest that technological change is a possible answer, provided that computerization resulted in a decline in the

[18] In practice, we simply use the Stata `integ` command.

demand for skilled but " routine" tasks that used to be performed by workers around the middle of the wage distribution.[19]

Lemieux (2008) reviewed possible explanations for the increased polarization in the labor market, including the technological-based explanation of Autor, Katz, and Kearney. He suggested that, if this explanation is an important one, then changes in relative wages by occupation, i.e., the contribution of occupations to the wage structure effect, should play an important role in changes in the wage distribution. Furthermore, since it is well known that education wage differentials kept expanding after the late 1980s (e.g., Acemoglu and Autor 2011), the contribution of education to the wage structure effect is another leading explanation for inequality changes over this period. More recent studies have also implicated the role of offshorability and trade (Firpo et al. 2011; Autor et al. 2014) which may be more salient at the industry level, given that some "local" industries such as the construction, distribution (wholesale trade, transportation), and personal service sectors are likely less affected by these economic forces.

Previous studies also show that composition effects played an important role in increasing wage inequality. Lemieux (2006b) showed that all the growth in residual inequality over this period is due to composition effects linked to the fact that the workforce became older and more educated, two factors associated with more wage dispersion. Furthermore, Lemieux (2008) argued that de-unionization, defined as a composition effect in this paper, still contributed to the changes in the wage distribution over this period.

These various explanations can all be understood in terms of the respective contributions of a few broad sets of factors (unions, education, experience, occupations, industries, etc.) to either wage structure or composition effects. This makes the decomposition method proposed in this paper ideally suited for estimating the contribution of each of these possible explanations to changes in the wage distribution. Unlike other procedures, our method allows us to estimate the relative contribution of each of the factors mentioned above to recent changes in the U.S. wage distribution.[20]

Our empirical analysis is based on data for men from the 1988–1990 and 2014–2016 Outgoing Rotation Group (ORG) Supplements of the Current Population Survey, yielding about a quarter million observations for each time period. As in Fortin and Lemieux (2016), for conciseness, we focus exclusively on men. The extent of occupational gender segregation is such that we would have to perform the analysis and choose the base group separately by gender. Increasing inequality appears to have worked through different channels and time period for men and women. Autor et al. (2015) showed that men's employment was impacted by the automation of production activities in the manufacturing sector at the beginning of the period, while women suffered employment losses associated with the impact of computerization of information-processing tasks in non-manufacturing later in the period.

The data files were processed as in Lemieux (2006b) who provided detailed information on the relevant data issues. The wage measure used is an hourly wage measure computed by dividing earnings by hours of work for workers not paid by the hour. For workers paid by the hour, we use a direct measure of the hourly wage rate. In light of the above discussion, the key set of covariates on which we focus are education (six education groups), potential experience (nine groups), union coverage, occupation (17 categories), and industry (14 categories). We also include controls for marital

---

[19] This technological change explanation was first suggested by Autor et al. (2003). It also implies that the wages of both skilled (e.g., doctors) and unskilled (e.g., truck drivers) non-routine jobs, at the top and low end of the wage distribution, increased relative to those of "routine" workers in the middle of the wage distribution.

[20] Autor et al. (2005) used the Machado and Mata (2005) method to decompose changes at each quantile into a "price" (wage structure) and "quantity" (composition) effect. They did not further consider, however, the contribution of each individual covariate to the wage structure effect, except for separating the contribution of (all) covariates from the residual change in inequality. See also Lemieux (2002) for a similar decomposition based on a reweighting procedure.

status and race in all the estimated models. The sample means for all these variables are provided in Table A1.[21]

Before proceeding to the estimation of RIF-regressions, it is important to inspect the density of wages for unusual features that would challenge the estimation of the RIF at the quantiles of interest or the wage model that w use. Figure 1 presents kernel density estimates of male wages for 1988–1990 and 2014–2016 estimated using the Epanechnikov kernel and bandwidths of 0.06 and 0.08, respectively.[22] The figure also shows the 1988–1990 density reweighted to have the same distribution of characteristics as in 2014–2016. The typical issues to look for include cliffs associated with minimum wage effects at the bottom of the distribution, peaks associated with heaping (the fact that hourly wage workers, in particular, are more likely to round their wages at next dollar amount) in the middle of the distribution, and top-coding at the top of the distribution. The impact of minimum wages is clearly seen in Figure 1 when vertical lines corresponding to the minimum and maximum of federal and state minimum wages are displayed. Because we do not model minimum wages in the current paper, the 1988–1990 density and the reweighted density are superimposed in those wage ranges, showing the wage setting variables that we include are inadequate for modeling the distribution of wages when minimum wages matter.[23] Thus, we remain cautious with regards to the interpretation of any effect at the bottom of the distribution.

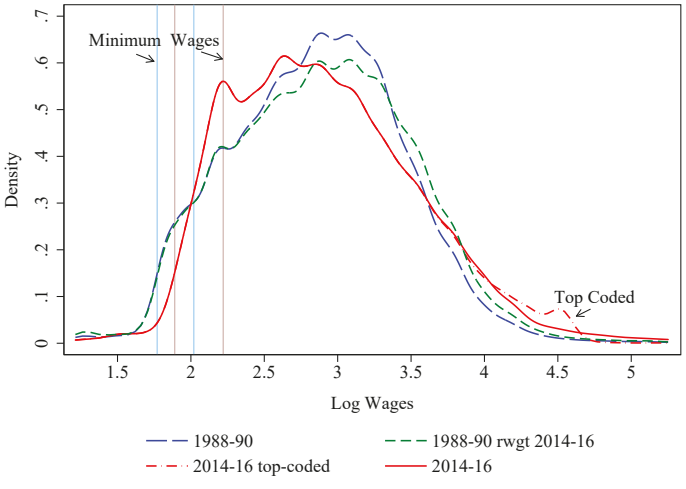


Figure 1. Density of Log Wages ($2010)—Men CPS. Note: The vertical lines show the minimum and maximum of state and federal minimum wages in each time period.

Heaping and top-coding can be problematic if they imply an unusually high value of the density at a particular quantile of interest that potentially biases the estimation of the denominator $\widehat{f_Y}(\widehat{q}_\tau)$ of the influence function (14). While only 0.7% of workers are top-coded in 1988–1990, this proportion

[21] Table A2 gives the details of the occupation and industry categories used.

[22] Several cross-validation tools suggested tuning parameters in that range, but the graphs were indistinguishable. In addition to the reweighting factors discussed in Sections 3 and 4, we also use CPS sample weights throughout the empirical analysis. In practice, this means that we multiply the relevant reweighting factor with CPS sample weight.

[23] See Brochu et al. (2017) for a more precise modeling of the effect of minimum wages on the distribution of wages.

increases to 3.6% in 2014–2016.[24] A standard adjustment for top-coding consists of multiplying top-coded wages by a fixed adjustment factor. In Figure 1, we use the adjustment factor of 1.4 suggested by Lemieux (2006b). While there is no visual evidence of an impact of top-coding in 1988–1990, there is a clear spike in the 2014–2016 distribution around the point (log wage of about 4.5) where most top-coded observations lie.[25] We deal with this issue using a more sophisticated stochastic imputation procedure (shown as the solid line) based on a Pareto distribution estimated using tax data from Alvaredo et al. (2013).

Given our large sample of hourly paid and salaried workers, heaping does not appear to be a serious issue in Figure 1.[26] However, heaping is more visible in Figure 2, which plots the 1988–1990 and 2014–2016 densities of wages for our base group. This group of about 400 workers in each period consists of non-unionized, white, married, high school educated men with 20 to 25 years of experience, working as construction workers in the construction industry, but not in the public sector.[27] The figure shows that the densities have changed very little over time, aside from different positioning of some local peaks associated with heaping.[28] This group was chosen because the economic forces that impact the overall wage distribution are less likely at play among this non-unionized group of low-educated workers in non-routine manual jobs with little exposure to international trade.[29]

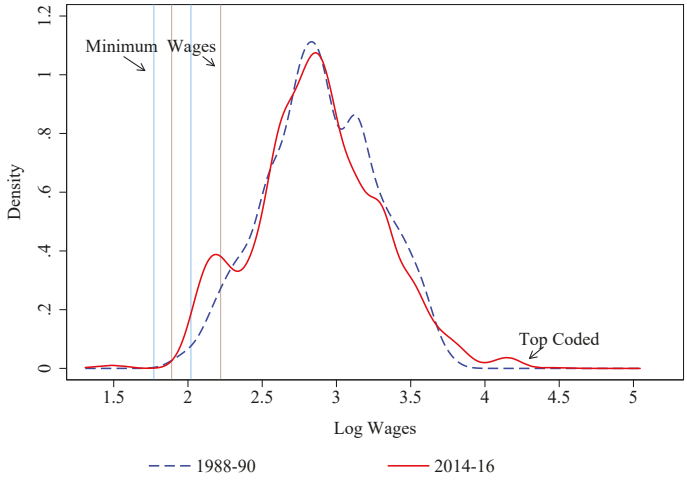


Figure 2. Density of Log Wages ($2010)—Base Group. Note: The vertical lines show the minimum and maximum of state and federal minimum wages in each time period.

[24] Weekly earnings are top-coded at $1923 in 1988–1990 and $2884 in 2014–2016. The latter is substantially lower in constant dollars. Furthermore, the top-code is even higher in relative terms because of the substantial growth in real wages at the top end of the distribution.

[25] A large fraction of workers top-coded at $2884 a week work 40 h a week, which yields an hourly wage rate of $72.1. Applying the 1.4 adjustment factor increases the wage to $100.9, or about $92.5 in dollars of 2010. This precisely matches the spike in Figure 1 since log(92.5) = 4.53.

[26] Deflating wages with monthly CPI while combining several years of data helps mitigate the issue of heaping.

[27] There are only 5–6 women in this category, which highlights the need of using different base groups for men and women.

[28] In nominal terms, the mode of the distributions is around $10.00/h in 1988–1990 and around $19.00/h in 2014–2016. In 1988–1990, there is a second local peak around $12.00/h, while, in 2014–2016, the second lower local peak is around $10.00/h.

[29] In Firpo et al. (2007), we used a mixed approach for the base group normalizing the coefficients of the occupation and industry dummies. That approach, although superficially attractive, has the important disadvantage of limiting the explanatory power of the variables whose coefficients are constrained. As a result, in this earlier version of the paper, very little of the changes in inequality were attributable to occupations and industries.

5.1. RIF-Regressions

Before showing the decomposition results, we first present some estimates from the RIF-regressions for different wage quantiles, the variance of log wages, and the Gini coefficient. From Equation (14), we compute $\text{IF}(y; q_\tau, F)$ for each observation using the sample estimate of q_τ, and the kernel density estimate of $f(q_\tau)$.

The RIF-regression coefficients for the 10th, 50th, and 90th quantiles in 1988–1990 and 2014–2016, along with bootstrapped standard errors, are reported in Table 1. The RIF-regression coefficients for the variance and the Gini are reported in Table 2. Detailed estimates for each of the 19 quantiles from the 5th to the 95th are also reported in Figures 3–5. For several covariates (for example, union status, non-white, married, clerical, production, and service occupations, transportation and utility, public administration sectors). Figure 3 illustrates highly non-monotonic effects across the different quantiles for some demographics. For instance, in Panel 1, the effect of union status first increases up to around the 40th quantile in 1988–1990, and up the 50th quantile in 2014–2016, and then declines, even turning negative for the 90th and 95th quantiles.

As shown by the RIF-regressions for the more global measures of inequality—the variance of log wages and the Gini coefficient of the wage distribution—displayed in Table 2, the effect of unions on these measures is negative, although the magnitude of that effect has decreased over time. This is consistent with the well-known result (e.g., Freeman 1980) that unions tend to reduce the variance of log wages for men. More importantly, as shown in Table 1, the results also indicate that unions increase inequality in the lower end of the distribution, but decrease inequality even more in the higher end of the distribution. As we will see later in the decomposition results, this means that the continuing decline in the rate of unionization can account for some of the "polarization" of the labor market (decrease in inequality at the low-end, but increase in inequality at the top end). The results for unions also illustrate an important feature of RIF regressions for quantiles, namely that they capture both the between-group effect (arising from union wage premia) and the within-group effect (arising from wage union compression) of unions on wage dispersion, which go in opposite direction in this case.[30]

The RIF-regression estimates in Table 1 for other covariates also illustrate this point. Consider, for instance, the case of college education. Table 1 and Figure 3 show that the effect of college increases monotonically as a function of percentiles. In other words, increasing the fraction of the workforce with a college degree has a larger impact on higher than lower quantiles. The reason why the effect is monotonic is that education increases both the level and the dispersion of wages (see, e.g., Lemieux 2006a). As a result, both the within- and the between-group effects go in the same direction of increasing inequality.

Another clear pattern that emerges in Figures 3 and 4 is that for most inequality enhancing covariates, i.e., those with a positively sloped curve, the inequality enhancing effect increases over time. In particular, the slopes for high levels of education (college graduates and post-graduates) and high-wage occupations (upper management, engineers and computer scientists, doctors, and lawyers) become steeper over time. This suggests that these covariates make a positive contribution to the wage structure effect.

There are some changes in the contribution of occupations and industries that are consistent with technological change and the routine-biased polarization of wages. For example, as shown in Figures 4 and 5, there are increases in the returns to high-tech service industries at the upper end of the wage distribution, but decreases in the returns to production and clerical occupations in the middle of the wage distribution. There are also decreases in the penalties to some low skilled non-routine

[30] As argued in FFL, the different relative strength of between and within effects at different quantiles explain the inverse U-shaped effect of unions. This is in sharp contrast with the effect of unions found estimated using conditional quantile regressions which captures only within-group effects and declines monotonically over the wage distribution (Chamberlain 1994).

occupations and associated industries, such as service occupations and truck driving and the retail industry, although some increases at the lower end appear to be driven by changes in minimum wages. On the other hand, there are some offsetting effects in industries that could have compensated the decline in manufacturing employment, such as the primary (e.g., mining), wholesale and retail trade, and personal services industries. In summary, the changes in the rewards and penalties associated with occupations and industries provide a descriptive account of factors potentially offsetting the wage effects of the polarization of employment. We turn next to the evaluation of the magnitude of these effects.

Table 1. Unconditional Quantile Regression Coefficients on Log Wages.

Years:	**1988/90**			**2014/16**		
Quantiles:	10	50	90	10	50	90
Explanatory Variables						
Union covered	0.146***	0.343***	−0.025***	0.058***	0.240***	−0.008
	(0.003)	(0.005)	(0.004)	(0.003)	(0.006)	(0.007)
Non-white	−0.063***	−0.137***	−0.072***	−0.053***	−0.106***	−0.041***
	(0.006)	(0.005)	(0.005)	(0.004)	(0.004)	(0.006)
Non-Married	−0.111***	−0.109***	−0.031***	−0.046***	−0.107***	−0.064***
	(0.004)	(0.003)	(0.004)	(0.003)	(0.004)	(0.005)
Education (High School omitted)						
Primary	−0.301***	−0.312***	−0.109***	−0.212***	−0.415***	−0.110***
	(0.011)	(0.006)	(0.005)	(0.01)	(0.009)	(0.006)
Some HS	−0.305***	−0.112***	0.005	−0.275***	−0.215***	0.002
	(0.007)	(0.005)	(0.003)	(0.008)	(0.007)	(0.004)
Some College	0.055***	0.135***	0.112***	0.036***	0.098***	0.023***
	(0.005)	(0.004)	(0.005)	(0.004)	(0.005)	(0.004)
College	0.143***	0.343***	0.410***	0.125***	0.409***	0.493***
	(0.005)	(0.005)	(0.008)	(0.004)	(0.006)	(0.009)
Post-grad	0.094***	0.418***	0.772***	0.099***	0.502***	0.962***
	(0.006)	(0.006)	(0.013)	(0.004)	(0.008)	(0.017)
Potential Experience ($20 \leq$ Experience < 25 omitted)						
Experience < 5	−0.486***	−0.448***	−0.312***	−0.335***	−0.425***	−0.301***
	(0.009)	(0.006)	(0.008)	(0.007)	(0.007)	(0.011)
$5 \leq$ Experience < 10	−0.056***	−0.270***	−0.278***	−0.067***	−0.285***	−0.306***
	(0.006)	(0.006)	(0.008)	(0.005)	(0.007)	(0.011)
$10 \leq$ Experience < 15	−0.005	−0.122***	−0.172***	−0.022***	−0.157***	−0.182***
	(0.005)	(0.006)	(0.008)	(0.004)	(0.006)	(0.011)
$15 \leq$ Experience < 20	0.002	−0.051***	−0.091***	−0.009*	−0.051***	−0.034***
	(0.005)	(0.005)	(0.008)	(0.004)	(0.006)	(0.012)
$25 \leq$ Experience < 30	0.010	0.033***	0.060***	−0.001	0.020***	0.036***
	(0.006)	(0.006)	(0.01)	(0.004)	(0.006)	(0.012)
$30 \leq$ Experience < 35	0.017*	0.048***	0.071***	0.008	0.037***	0.042***
	(0.006)	(0.006)	(0.011)	(0.004)	(0.007)	(0.012)
$35 \leq$ Experience < 40	0.022**	0.028***	0.061***	0.013**	0.054***	0.062***
	(0.007)	(0.008)	(0.012)	(0.004)	(0.007)	(0.013)
Experience ≥ 40	0.068***	0.020**	−0.010	0.030***	0.058***	−0.013
	(0.008)	(0.008)	(0.009)	(0.005)	(0.007)	(0.012)
R−square	0.253	0.359	0.206	0.182	0.353	0.202
No. of observations		268,494			236,296	

Note: Linear limited dependent variable model. Bootstrapped standard errors (500 repetitions) are in parentheses. Statistical signifiance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$. Also included in the regression are a public sector dummy, 16 occupation dummies, and 14 industry dummies. The base group is made up of individuals who are non-unionized (not covered), not in the public sector, white, married, have a high school degree, work as construction workers in the construction industry.

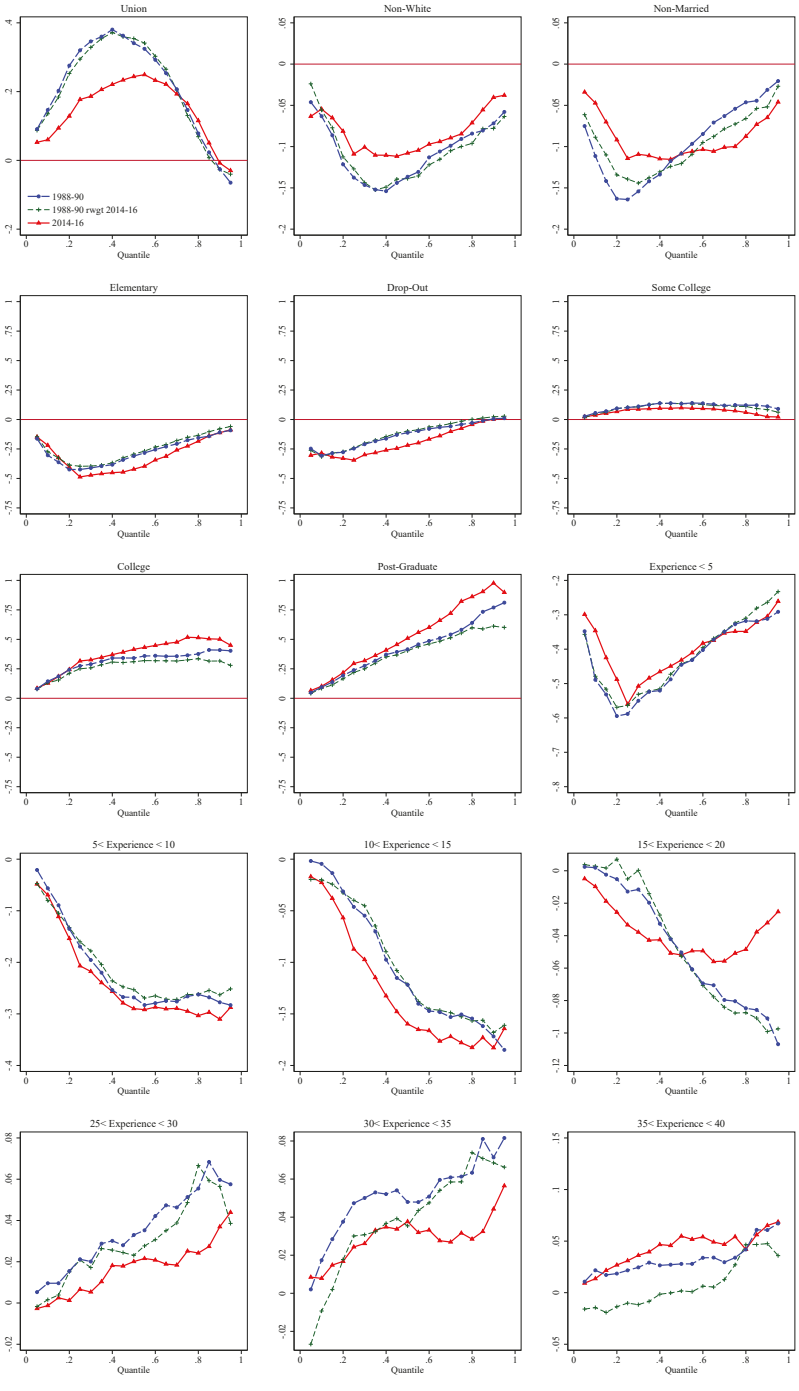


Figure 3. Unconditional Quantile Coefficients—Demographics and Human Capital.

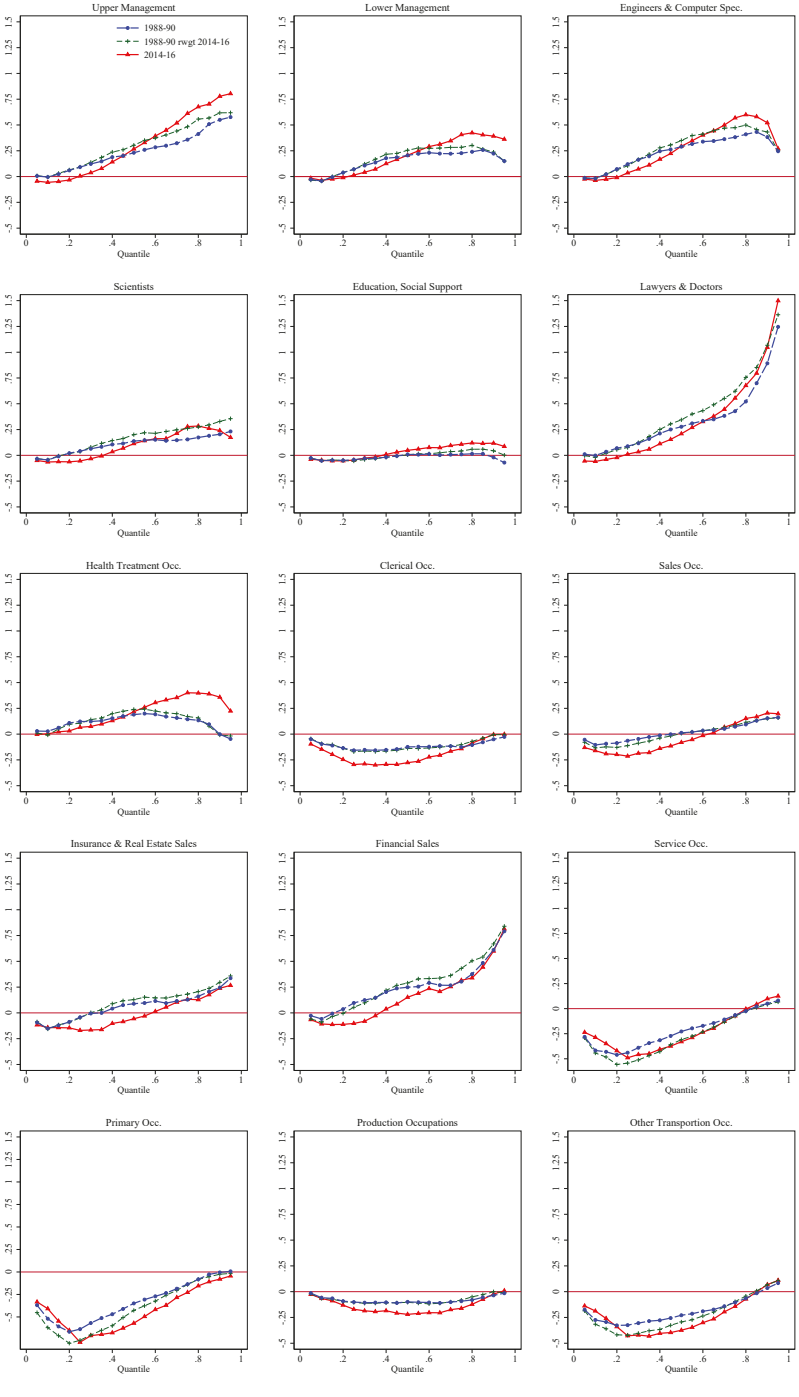


Figure 4. Unconditional Quantile Coefficients—Occupations.

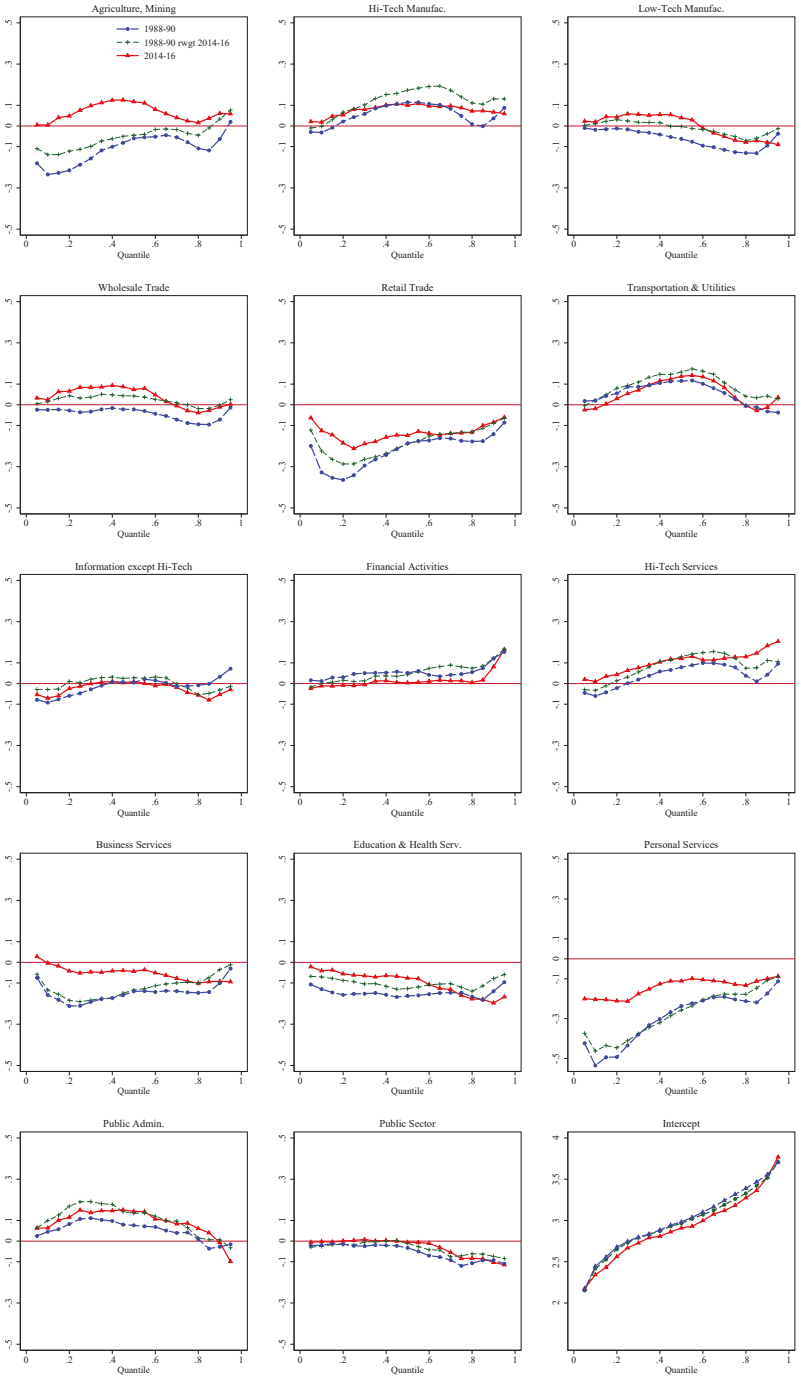


Figure 5. Unconditional Quantile Coefficients—Industries.

Table 2. RIF Regression of Inequality Measures.

Years:	**1988/90**	**2014/16**	**1988/90**	**2014/16**
Inequality Measures	**Variance of Log Wages**		**Gini**	
Estimated Values:	**0.341**	**0.418**	**0.330**	**0.396**
Explanatory Variables				
Constant	0.203***	0.205***	0.261***	0.290***
	(0.004)	(0.006)	(0.002)	(0.002)
Union covered	−0.075***	−0.040***	−0.067***	−0.039***
	(0.002)	(0.004)	(0.001)	(0.001)
Non-white	−0.002	0.005	0.006***	0.005***
	(0.003)	(0.004)	(0.001)	(0.001)
Non-Married	0.039***	0.001	0.022***	0.008**
	(0.002)	(0.004)	(0.001)	(0.001)
Education (High School omitted)				
Primary	0.074***	0.073***	0.051***	0.057***
	(0.004)	(0.006)	(0.002)	(0.002)
Some HS	0.104***	0.129***	0.048***	0.063***
	(0.003)	(0.005)	(0.001)	(0.001)
Some College	0.028***	−0.001	0.006***	−0.006***
	(0.003)	(0.003)	(0.002)	(0.003)
College	0.121***	0.166***	0.053***	0.061***
	(0.005)	(0.005)	(0.002)	(0.001)
Post-grad	0.301***	0.401***	0.157***	0.177***
	(0.007)	(0.01)	(0.003)	(0.002)
Potential Experience($20 \leq$ Experience <25 omitted)				
Experience < 5	0.047***	0.027***	0.031***	0.021***
	(0.004)	(0.007)	(0.002)	(0.002)
$5 \leq$ Experience < 10	−0.098***	−0.093***	−0.036***	−0.030***
	(0.005)	(0.007)	(0.002)	(0.002)
$10 \leq$ Experience < 15	−0.078***	−0.070***	−0.035***	−0.028***
	(0.004)	(0.007)	(0.002)	(0.002)
$15 \leq$ Experience < 20	−0.050***	−0.006	−0.026***	0.003**
	(0.005)	(0.008)	(0.002)	(0.002)
$25 \leq$ Experience < 30	0.023***	0.024***	0.012***	0.014***
	(0.006)	(0.008)	(0.002)	(0.002)
$30 \leq$ Experience < 35	0.022***	0.017**	0.008***	0.007***
	(0.006)	(0.008)	(0.002)	(0.002)
$35 \leq$ Experience < 40	0.015**	0.022***	0.008***	0.008***
	(0.007)	(0.008)	(0.003)	(0.002)
Experience≥ 40	−0.031***	−0.012	−0.015***	−0.005**
	(0.005)	(0.008)	(0.003)	(0.002)
Occupations (Construction & Repair Occ. omitted)				
Upper Management	0.235***	0.415***	0.132***	0.203***
	(0.007)	(0.011)	(0.003)	(0.002)
Lower Management	0.090***	0.200***	0.027***	0.080***
	(0.008)	(0.009)	(0.003)	(0.002)
Engineers & Computer Occ.	0.107***	0.202***	0.013**	0.054***
	(0.006)	(0.009)	(0.003)	(0.002)
Other Scientists	0.081***	0.134***	0.025**	0.068***
	(0.011)	(0.027)	(0.005)	(0.006)
Social Support Occ.	−0.001	0.065***	−0.012**	0.012***
	(0.007)	(0.009)	(0.003)	(0.003)
Lawyers & Doctors	0.524***	0.637***	0.337***	0.363***
	(0.027)	(0.032)	(0.010)	(0.008)
Health Treatment Occ.	−0.020	0.115***	−0.035***	0.011***
	(0.0101)	(0.012)	(0.005)	(0.005)
Clerical Occ.	0.013**	0.069***	0.017***	0.044***
	(0.004)	(0.005)	(0.002)	(0.002)

Table 2. *Cont.*

Years:	**1988/90**	**2014/16**	**1988/90**	**2014/16**
Inequality Measures	**Variance of Log Wages**		**Gini**	
Estimated Values:	**0.341**	**0.418**	**0.330**	**0.396**
Explanatory Variables				
Occupations (cnt.)				
Sales Occ.	0.088***	0.177***	0.043***	0.084***
	(0.005)	(0.008)	(0.002)	(0.002)
Insur. & Real Estate Sales	0.208***	0.197***	0.152***	0.105***
	(0.031)	(0.038)	(0.011)	(0.010)
Financial Sales	0.525***	0.409***	0.429***	0.219***
	(0.06)	(0.076)	(0.018)	(0.014)
Service Occ.	0.188***	0.208***	0.101***	0.107***
	(0.004)	(0.005)	(0.002)	(0.002)
Primary Occ.	0.226***	0.222***	0.114***	0.127***
	(0.008)	(0.015)	(0.004)	(0.004)
Production Occ.	0.004	0.020***	0.011***	0.028***
	(0.003)	(0.005)	(0.001)	(0.002)
Transportation Occ.	0.119***	0.145***	0.079***	0.094***
	(0.004)	(0.006)	(0.002)	(0.002)
Truckers	0.015***	0.042***	0.030***	0.040***
	(0.004)	(0.006)	(0.002)	(0.002)
Industries (Construction omitted)				
Agriculture, Mining	0.079***	0.013	0.036***	−0.001
	(0.008)	(0.012)	(0.003)	(0.003)
Hi-Tech Manufac	0.018***	0.014	−0.001	0.002
	(0.005)	(0.009)	(0.002)	(0.002)
Low-Tech Manufac	−0.037***	−0.053***	−0.011***	−0.019***
	(0.004)	(0.007)	(0.002)	(0.002)
Wholesale Trade	−0.012	−0.027**	0.001	−0.006*
	(0.006)	(0.012)	(0.002)	(0.003)
Retail Trade	0.060***	0.016*	0.038***	0.023***
	(0.005)	(0.007)	(0.002)	(0.002)
Transportation & Utilities	0.013***	−0.029***	−0.005*	−0.019***
	(0.005)	(0.007)	(0.002)	(0.002)
Information except Hi-Tech	−0.001	0.055***	−0.010***	0.041***
	(0.008)	(0.019)	(0.003)	(0.005)
Financial Activities	0.065***	0.064***	0.052***	0.053***
	(0.009)	(0.013)	(0.004)	(0.003)
Hi-Tech Services	0.048***	0.071***	0.018**	0.035***
	(0.008)	(0.01)	(0.004)	(0.003)
Business Services	0.018**	−0.042***	0.019***	−0.014***
	(0.005)	(0.008)	(0.002)	(0.002)
Education & Health Services	−0.008	−0.064***	−0.001	−0.018***
	(0.006)	(0.008)	(0.003)	(0.002)
Personal Services	0.136***	0.054***	0.051***	0.023***
	(0.006)	(0.006)	(0.002)	(0.002)
Public Admin	−0.038***	−0.071***	−0.036***	−0.029***
	(0.007)	(0.011)	(0.003)	(0.003)
Public Sector	−0.058***	−0.055***	−0.030***	−0.048***
	(0.005)	(0.007)	(0.002)	(0.002)
R-squared	0.115	0.087	0.048	0.025
No. of observations	268,492	236,287	268,492	236,287

Note: Bootstrapped standard errors (500 repetitions) are in parentheses. Statistical signifiance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$. The base group is made up of individuals who are non-unionized (not covered), not public sector, white, married, have a high school degree, work as construction workers in the construction industry. Trimmed sample drops 15 observations with hourly wages > \$1,636 (\$2010).

5.2. Decomposition Results

The results for the aggregate decomposition are presented in Figure 6. Tables 3 and 4 summarize the results for the standard measures of top-end (90–50 log wage differential) and low-end (50–10 log wage differential) wage inequality, as well as for the variance of log wages and the Gini coefficient. The covariates used in the RIF-regression models are those discussed above and listed in Table A1. A richer specification with additional interaction terms is used to estimate the logit models used compute the reweighting factor $\widehat{\omega}_C(T_i, X_i)$.[31]

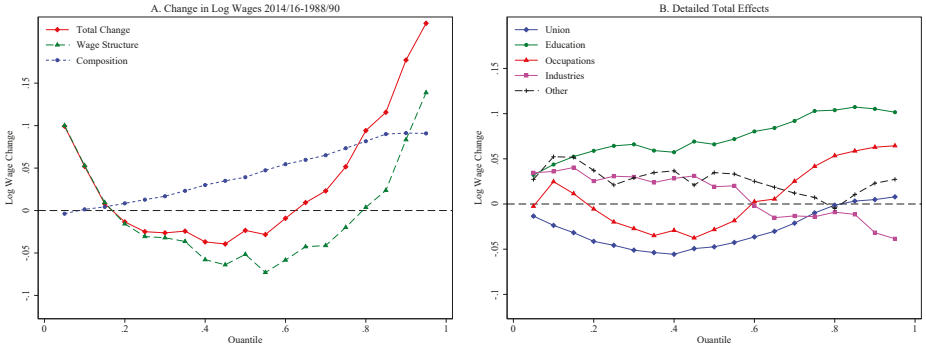


Figure 6. Decomposition of Total Change into Composition and Wage Structure Effects.

Figure 6a shows the overall change in (real log) wages at each percentile τ, $\Delta_O^{q_\tau}$, and decomposes this overall change into a composition ($\Delta_X^{q_\tau}$) and wage structure ($\Delta_S^{q_\tau}$) effect computed using the reweighting procedure of Result 1. Consistent with the pattern first documented in Autor et al. (2006), the overall change is U-shaped as wage dispersion increases in the top-end of the distribution, but declines in the lower end.[32] Most summary measures of inequality such as the 90–10 gap nonetheless increase over the 1988–1990 to 2014–2016 period as wage gains in the top-end of the distribution exceed those at the low-end. In other words, although the curve for overall wage changes is U-shaped, its slope is positive, on average, suggesting that inequality generally goes up. This overall increase shows up as positive total changes in the 90–10 gap, the variance of log wages, and the Gini, reported in Tables 3 and 4. In all cases, the aggregate decomposition of these overall measures attributes most (from 55% to 66%) of the changes to composition effects.

Figure 6a also shows that, consistent with Lemieux (2006b), composition effects have contributed to a substantial increase in inequality. In fact, once composition effects are accounted for, the remaining wage structure effects (estimated using reweighting) follow a "purer" U-shape than overall changes in wages. The wage declines are now right in the middle of the distribution (20th to 80th percentile), while wage gains at the top and low end are more similar. By the same token, however, composition effects cannot account at all for the U-shaped nature of wage changes.

Figure 7 moves to the next step of the decomposition using linear RIF-regressions to attribute the contribution of each set of covariates to the composition effect.[33] Figure 8, which we discuss below, does the same for the wage structure effect. Figure 6b summarizes the total of the composition and

---

[31] The logit specification also includes a full set of interaction between experience and education, union status and education, union status and experience, between education and occupations, and experience and industries.

[32] This stands in sharp contrast with the situation that prevailed in the 1980s when the corresponding curve was positively sloped as wage dispersion increased at all points of the distribution (Juhn et al. 1993).

[33] The effect of each set of factors is obtained by summing up the contribution of the relevant covariates. For example, the effect for "education" is the sum of the effect of each of the five education categories shown in Table 1. Showing the effect of each individual dummy separately would be cumbersome and harder to interpret.

wage structure effects by the sets of factors of interest. The combination of composition and wage structure effects shows the strong monotonic effect of education on wage changes, the mild U-shaped effect of union and occupations, and the offsetting hump-shaped effect of industries.

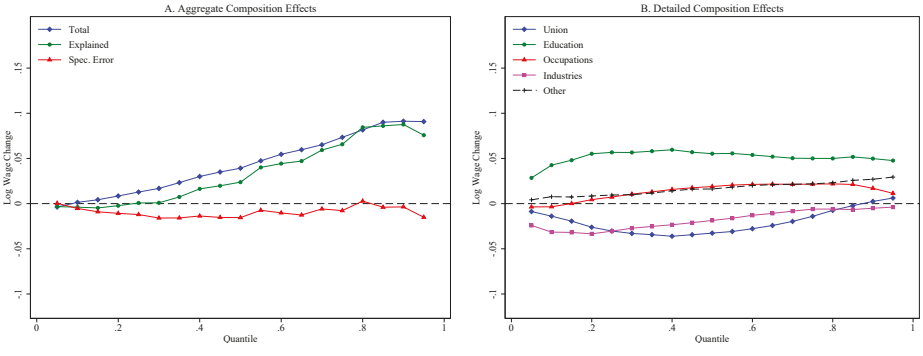


Figure 7. Decomposition of Composition Effects.

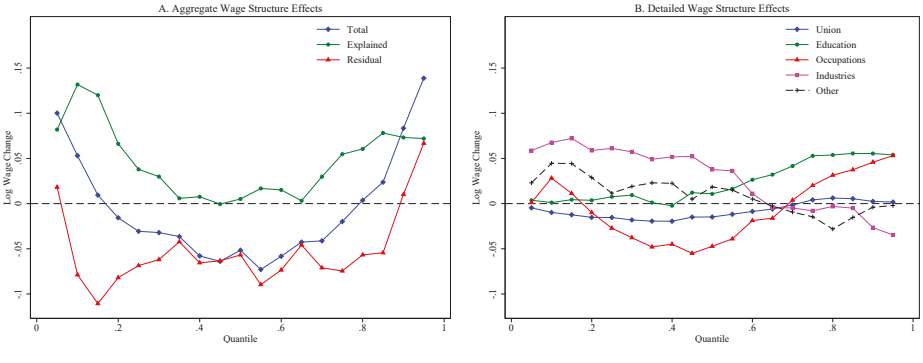


Figure 8. Decomposition of Wage Structure Effects.

Figure 7a compares the overall composition effect obtained by reweighting and displayed in Figure 6a, $\widehat{\Delta}_X^{q_\tau}$, to the composition effect explained using the RIF-regressions, $(\overline{X}_0^C - \overline{X}_0)'\widehat{\gamma}_0^{q_\tau}$. The difference between the two curves is the specification (approximation) error R^{q_τ}. The error term is relatively small and does not exhibit much of a systematic pattern. This means that the RIF-regression model does relatively well at tracking down the composition effect estimated consistently using the reweighting procedure; however, as we discuss below, in some cases, the specification error is significantly different from zero.

Figure 7b then divides the composition effect (explained by the RIF-regressions) into the contribution of five main sets of factors. To simplify the discussion, we focus on the impact of each factor on overall wage inequality summarized by the 90–10 log wage differential in comparison to the 50–10 and 90–50 log wage differentials that capture what happened in the lower and upper parts of the distribution, respectively. The decomposition of the log wage differentials, the log variance, and the Gini are reported in Tables 3 and 4. Table 3 presents the simple OB type decomposition computed from RIF-regressions of the five inequality measures, without reweighting. Table 4 applies the complete two-step procedure described above.

As discussed in Section 4.3, we compute the RIF of the difference between two (log) quantiles q_1 and q_2, where $q_2 > q_1$, as $\text{RIF}(y_i; q_2 - q_1) = \text{RIF}(y_i; q_2) - \text{RIF}(y_i; q_2)$, and use these differences

as dependent variables in the regressions. For the variance of log wages and the Gini, the RIF are as described above. Using the estimation results from these sets of regressions, we compute the components of the simple OB-type decomposition for the changes over time, $\widehat{\nu}_1 - \widehat{\nu}_0 = \widehat{\Delta}^{\nu}_{OB}$, from 1988–1990 ($T = 0$) to 2014–2016 ($T = 1$) as:

$$\widehat{\Delta}^{\nu}_{OB} = \underbrace{\left(\overline{X}_1 - \overline{X}_0\right)' \widehat{\gamma}^{\nu}_0}_{\widehat{\Delta}^{\nu}_{X,OB}} + \underbrace{\overline{X}_1{}' \left(\widehat{\gamma}^{\nu}_1 - \widehat{\gamma}^{\nu}_0\right)}_{\widehat{\Delta}^{\nu}_{S,OB}}.$$

Table 3. Decomposition Results without Reweighting.

Inequality Measures	90–10	50–10	90–50	Variance (× 100)	Gini (× 100)
Total Change	0.125***	−0.075***	0.201***	7.775***	6.599***
Composition	0.089***	0.037***	0.052***	4.163***	1.966***
Wage Structure	0.037***	−0.112***	0.149***	3.612***	4.633***
Composition Effects:					
Union	0.016***	−0.019***	0.035***	0.713***	0.639***
Other	0.019***	0.008***	0.011***	0.984***	0.473***
Education	0.009***	0.013***	−0.005***	0.665***	0.207**
Occupation	0.019***	0.022***	−0.002**	0.672***	0.112***
Industry	0.026***	0.013***	0.013***	1.128***	0.536***
Wage Structure Effects:					
Union	0.014***	−0.002*	0.015***	0.442***	0.360***
Other	−0.048***	−0.034***	−0.014	−0.983	−0.161
Education	0.015**	0.008***	0.007	1.444***	0.188*
Occupation	0.057***	−0.066***	0.123***	5.664***	2.423***
Industry	−0.079***	−0.048***	−0.031***	−3.212***	−1.044**
Constant	0.079***	0.030**	0.049***	0.257	0.287***
Total Effects:					
Union	0.030***	−0.021***	0.051***	1.156***	0.998***
Other	−0.029**	−0.026***	−0.003	0.001	0.312
Education	0.024***	0.022***	0.002	2.110***	0.395**
Occupation	0.076***	−0.045***	0.121***	6.336***	2.534***
Industry	−0.054***	−0.036***	−0.018	−2.084***	−0.508

Note: Other includes non-white, non-married, and five categories of experience. Statistical signifiance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$. Bootstrapped standard errors over the entire procedure (500 replications) were used to compute the p-value. Trimmed sample for the variance and Gini drops 15 observations with hourly wages > \$1,636 (\$2010).

These results are displayed in Table 3 by groups of variables.[34] In Table 4, we present the results of the decomposition that also applies the reweighting procedure

$$\widehat{\Delta}^{\nu}_{O} = \underbrace{\left(\overline{X}^{C}_0 - \overline{X}_0\right)' \cdot \widehat{\gamma}^{\nu}_0}_{\widehat{\Delta}^{\nu}_{X,p}} + \underbrace{\overline{X}^{C\prime}_0 \cdot \left(\widehat{\gamma}^{\nu}_C - \widehat{\gamma}^{\nu}_0\right)}_{\widehat{\Delta}^{\nu}_{X,e}} + \underbrace{\overline{X}_1{}' \cdot \left(\widehat{\gamma}^{\nu}_1 - \widehat{\gamma}^{\nu}_C\right)}_{\widehat{\Delta}^{\nu}_{S,p}} + \underbrace{\left(\overline{X}_1 - \overline{X}^{C}_0\right)' \cdot \widehat{\gamma}^{\nu}_C}_{\widehat{\Delta}^{\nu}_{S,e}}.$$

The four terms in this decomposition are easily obtained by running two OB decompositions using RIF regressions. First, we perform an OB decomposition using the $T = 0$ sample and the counterfactual sample ($T = 0$ sample reweighted to be as in $T = 1$) to get the pure composition effect,

[34] In practice, we use the popular Jann (2008) "oaxaca" Stata ado file and obtain bootstrapped standard errors over the entire procedure given the statistics and the RIF are estimated values. We opted for boostrapped instead of analytical standard errors by simplicity. Computation of analytical standard errors would involve estimation of different functionals, increasing the degree of complexity of the estimation step, whereas bootstrapped standard errors, although being potentially computationally more demanding are typically simpler to implement.

$\widehat{\Delta}^{v}_{X,p}$, using $T = 0$ as reference wage structure. The total unexplained effect in this decomposition corresponds to the specification error, $\widehat{\Delta}^{v}_{X,e}$, and allows one to assess the importance of departures from the linearity assumption. Second, we perform the decomposition using the $T = 1$ sample and the counterfactual sample, using the counterfactual wage structure as reference, and obtain the pure wage structure effect, $\widehat{\Delta}^{v}_{S,p}$, in the "unexplained" part of the decomposition. The total explained effect in this decomposition, $\widehat{\Delta}^{v}_{S,e}$, corresponds to the reweighting error which should go to zero in large samples. It provides an easy way of assessing the quality of the reweighting.[35]

Table 4. Decomposition Results with Reweighting.

Inequality Measures	90–10	50–10	90–50	Variance (×100)	Gini (×100)
Total Change	0.125***	−0.075***	0.201***	7.775***	6.599***
Composition	0.090***	0.038***	0.052***	4.193***	1.966***
Wage Structure	0.030***	−0.105***	0.135***	3.149***	4.402***
Composition Effects:					
Union	0.016***	−0.019***	0.035***	0.712***	0.638***
Other	0.019***	0.009***	0.011***	1.007***	0.481***
Education	0.007***	0.013***	−0.005***	0.600***	0.173
Occupation	0.020***	0.022***	−0.002*	0.719***	0.129***
Industry	0.026***	0.013***	0.014***	1.155***	0.546***
Specification Error	0.002	−0.010	0.012***	−0.308***	0.175***
Wage Structure Effects:					
Union	0.012***	−0.005**	0.017***	0.338***	0.220***
Other	−0.049***	−0.026***	−0.023	−0.871	−0.068
Education	0.054***	0.010	0.045***	2.303***	1.183***
Occupation	0.018	−0.075***	0.093***	2.872***	1.416***
Industry	−0.094***	−0.030**	−0.064***	−3.852***	−1.306***
Constant	0.089***	0.022	0.067***	2.359***	2.957***
Reweighting Error	0.003***	0.002***	0.001***	0.125***	0.057***
Total Effects:					
Union	0.029***	−0.024***	0.052***	1.050***	0.857***
Other	−0.029**	−0.018*	−0.012	0.135	0.413
Education	0.062***	0.022***	0.039***	2.903***	1.356***
Occupation	0.038***	−0.053***	0.091***	3.591***	1.545***
Industry	−0.068***	−0.017	−0.051***	−2.697***	−0.760*

Note: Other includes non-white, non-married, and five categories of experience. Statistical signifiance levels: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.1$. Bootstrapped standard errors over the entire procedure (500 replications) were used to compute the p-value. Trimmed sample for the variance and Gini drops 15 observations with hourly wages > \$1,636 (\$2010).

Consistent with Figure 7a, specification errors reported in Table 4 are generally small. As discussed in Section 3, the specification error reflects departures from non-linearity of the RIF-regressions and the fact that, except for the mean, the RIF depends on the distribution of Y (and X through its effect on Y). In Table 4, we formally test whether the specification error is significantly different from zero. The results are mixed. The specification error is not significantly different from zero for the 90–10 and the 50–10 gaps, but is statistically significant for the 90–50 gap, the variance, and the Gini. The specification error is nonetheless small relative to the overall changes in the distributional statistics, which indicates that RIF-regressions provide highly accurate estimates of the overall composition and wage structure effects in the empirical example being studied here. However, as we discuss below,

[35] Adding more terms in the specification of the reweighting function helps reducing the reweighting error. This has to be balanced with issues of common support, as more terms may lead to more perfect predictions, an undesirable outcome. As we discuss below, the specification we use yields a very small reweighting error.

although the specification error is small, using the two-step decomposition instead of a standard OB decomposition matters much more when looking at the contribution of individual covariates to the wage structure effect.

In both Table 3 and 4, the composition effects linked to factors other than unions go the "wrong way" in the sense that they account for rising inequality at the bottom end while inequality is rising at the top end, a point noted earlier by Autor et al. (2005). This applies in particular to education and occupations effects that are larger for the 50–10 than for the 90–50, while the effects of industry and other factors (race, marital status, and experience) on the 50–10 and 90–50 are similar. In contrast, composition effects linked to unions (the impact of de-unionization) reduce inequality at the low end (effect of −0.019 on the 50–10) but increases inequality at the top end (effect of 0.035 on the 90–50). Note that, just as in an OB decomposition, these effects on the 50–10 and the 90–50 gap can be obtained directly by multiplying the 9.5 percent decline in the unionization rate (Table A1) by the relevant union effects in 1988–1990 shown in Table 1. The effect of de-unionization accounts for about 25 percent of the total change in the 50–10 gap, which is remarkably similar to the relative contribution of de-unionization to the growth in inequality in the 1980s (see Freeman 1993; Card 1992; and DiNardo et al. 1996).

Figure 8a divides the wage structure effect, $\widehat{\Delta}_S^{q_\tau}$, into the part explained by the RIF-regression models, $\sum_{k=2}^{M}(\widehat{\gamma}_{1,k}^{v} - \widehat{\gamma}_{C,k}^{v})\overline{X}_1$, and the residual change $\widehat{\gamma}_{1,1}^{v} - \widehat{\gamma}_{C,1}^{v}$ (the change in for the base group captured by the intercepts). The contribution of each set of factors is then shown in Figure 8b. As in the case of the composition effects, it is easier to discuss the results by focusing on the 90–50 and 50–10 gaps shown in Tables 3 and 4.

Here, we note that the contribution of different covariates to the wage structure effect are quite different in Tables 3 and 4. This indicates that the OB decomposition of Table 3 is inaccurate because of differences between the estimated RIF-regression coefficients $\widehat{\gamma}_C^v$ and $\widehat{\gamma}_0^v$. As discussed in Section 3, the difference between $\widehat{\gamma}_1^v$ and $\widehat{\gamma}_C^v$ used to compute wage structure effects in Table 4 solely reflects changes in the wage structure. By contrast, the difference between $\widehat{\gamma}_1^v$ and $\widehat{\gamma}_0^v$ used in Table 3 is likely contaminated by changes in the distribution of X that are being adjusted for (by reweighting) when estimating $\widehat{\gamma}_C^v$. The difference is particularly striking in the case of education. As expected, Table 4 shows that wages structure effects linked to education play an important role in the growth of the 90–50 gap. By contrast, the effect is small and insignificant when using a conventional OB decomposition in Table 3. The case of education, a central variable in most studies on the sources of growing inequality, dramatically illustrates the importance of using the two-step decomposition with reweighting proposed in this paper.

The wage structure results of Table 4 first show that covariates overexplain −0.127 (sum of the five effects) of the −0.105 change (decline) in the 50–10 gap, the constant capturing the difference. Covariates do a less impressive job explaining changes in the 90–50 gap explaining only 0.068 (half) of the 0.136 change. Occupations are the set of the covariates that best capture the changes in the wage structure. They account for −0.075 of the −0.105 decline (73%) in the 50–10 gap and 0.088 of the 0.135 increase (68%) in the 90–50 gap. These results justify the increased attention given in the literature to the role of occupational tasks (Firpo et al. 2011; Fortin and Lemieux 2016). Changes in the returns to education continue to play an important role at the top of distribution accounting for 0.045 of the 0.135 increase (33%) in the 90–50. This supports Lemieux (2006a)'s conjecture that increases in the return to post-secondary education contribute to the convexification of the wage distribution.

Finally, the total effect of each covariate (wage structure plus composition effect) is reported in Figure 6b and the bottom panel of Table 4. Unions and occupations are the two factors that best account for the differential changes at the bottom and top of the distribution, capturing both a negative effect on the 50–10 and a positive effect on the 90–50. The total effect of the two factors on the 50–10 gap corresponds to −0.078 out of −0.105 (74%) of the change, while they account for 0.139 out of 0.136 change in the 90–50 (102%). This goes a substantial way towards explaining the polarization of the labor market.

6. Conclusions

We provide a detailed exposition of a two-stage method to decompose changes in the distribution of wages (or other outcome variables). In Stage 1, distributional changes are divided into a wage structure effect and a composition effect using a reweighting method. In Stage 2, these two components are further divided into the contribution of each individual covariate using the recentered influence function regression technique introduced by FFL. This two-stage procedure generalizes the popular OB decomposition method by extending the decomposition to any distributional measure (besides the mean), and allowing for a more flexible wage setting model. Other procedures (Machado and Mata 2005; Melly 2005; Rothe 2012; CFM) have been suggested for performing part of this decomposition for distributional parameters besides the means. One important advantage of our procedure is that it is easy to use in practice, as it simply involves estimating a logit model (first stage) and running least-square regressions (second stage). Another more distinctive advantage is that it can be used to divide the contribution of each covariate to the composition effect, something that most existing methods cannot do.

We illustrate the workings of our method by looking at changes in male wage inequality in the United States between 1988 and 2016. This is an interesting case to study as the wage distribution changed very differently at different points of the distribution, a phenomenon that cannot be captured by summary measures of inequality such as the variance of log wages. Our method is particularly well suited for looking in detail at the source of wage changes at each percentile of the wage distribution. Our findings indicate that unions, occupations, and education are the most important factors accounting for the observed changes in the wage distribution over this period.

Author Contributions: All authors contributed equally to the paper.

Funding: Fortin and Lemieux thank the Social Sciences and Humanities Research Council of Canada (grant# for financial support. Firpo thanks CNPq-Brazil for financial support.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Tables

Table A1. Sample Means.

Years:	1988/90	2014/16	Difference
Log wages	2.860	2.901	0.041
Std of log wages	0.579	0.622	0.043
Union covered	0.223	0.127	−0.095
Non-white	0.134	0.186	0.052
Non-Married	0.388	0.457	0.068
Age	36.204	39.882	3.677
Education			
Primary	0.059	0.034	−0.025
Some HS	0.118	0.054	−0.064
High School	0.381	0.307	−0.074
Some College	0.202	0.275	0.072
College	0.139	0.218	0.078
Post-grad	0.101	0.113	0.012
Occupations			
Upper Management	0.082	0.080	−0.002
Lower Management	0.040	0.068	0.028
Engineers & Computer Occ.	0.061	0.081	0.019
Other Scientists	0.014	0.010	−0.004
Social Support Occ.	0.052	0.061	0.009
Lawyers & Doctors	0.010	0.015	0.005
Health Treatment Occ.	0.010	0.019	0.009
Clerical Occ.	0.066	0.068	0.002

Table A1. *Cont.*

Years:	1988/90	2014/16	Difference
Sales Occ.	0.086	0.085	−0.001
Insur. & Real Estate Sales	0.007	0.006	−0.001
Financial Sales	0.003	0.002	−0.001
Service Occ.	0.107	0.149	0.042
Primary Occ.	0.026	0.011	−0.015
Construction & Repair Occ.	0.164	0.155	−0.009
Production Occ.	0.141	0.086	−0.055
Transportation Occ.	0.086	0.060	−0.026
Truckers	0.045	0.041	−0.004
Industries			
Agriculture, Mining	0.033	0.026	−0.007
Construction	0.097	0.101	0.005
Hi-Tech Manufac	0.102	0.066	−0.037
Low-Tech Manufac	0.137	0.087	−0.050
Wholesale Trade	0.051	0.033	−0.018
Retail Trade	0.105	0.113	0.008
Transportation & Utilities	0.086	0.079	−0.008
Information except Hi-Tech	0.018	0.012	−0.006
Financial Activities	0.047	0.058	0.011
Hi-Tech Services	0.035	0.064	0.029
Business Services	0.051	0.065	0.014
Education & Health Services	0.097	0.113	0.016
Personal Services	0.081	0.127	0.046
Public Admin	0.058	0.054	−0.005
Public Sector	0.149	0.126	−0.024

Note: Computed using sample weights. All differences over time are statistically significant at the $p = 0.001$ level.

Table A2. Occupation and Industry Definitions.

Code Sources:	2010 Census SOC	1980 SOC
Occupations		
Upper Management	10–200, 430	1–13, 19
Lower Management	200–950	14–18, 20–37, 473–476
Engineers & Computer Occ.	1000–1560	43–68, 213–218, 229
Other Scientists	1600–1960	69–83, 166–173, 223–225, 235
Social Support Occ.	2000–2060, 2140–2960	113–165, 174–177, 183–199, 228, 234
Lawyers & Doctors	2100–2110, 3010, 3060	84–85, 178–179
Health Treatment Occ.	3000, 3030–3050, 3110–3540	86–106, 203–208
Clerical Occ.	5000–5940	303–389
Sales Occ.	4700–4800, 4830–4900, 4930–4965	243–252, 256–285
Insur. & Real Estate Sales	4810,4920	253–254
Financial Sales	4820	255
Service Occ.	3600–4650	430–470
Primary Occ.	6000–6130	477–499
Construction & Repair Occ.	6200–7620	503–617, 863–869
Production Occ.	7700–8960	633–799, 873, 233
Transportation Occ.	9000–9120, 9140–9750	803, 808–859, 876–889, 226–227
Truck Drivers	9130	804–806
Industries		
Agriculture, Mining	170–490	10–50
Construction	770	60
Hi-Tech Manufac	2170–2390, 3180, 3360–3690, 3960	180–192, 210–212, 310, 321–322, 340–372
Low-Tech Manufac	1070–2090, 2470–3170, 3190–3290, 3770–3890, 3970–3990	100–162, 200–201,220–301, 311–320, 331–332, 380–392
Wholesale Trade	4070–4590	500–571
Retail Trade	4670–5790	580–640, 642–691
Transportation & Utilities	570–690, 6070–6390	400–432, 460–472
Information except Hi-Tech	6470–6480, 6570–6670, 6770–6780	171–172, 852
Financial Activities	6870–7190	700–712
Hi-Tech Services	6490, 6675–6695, 7290–7460	440–442, 732–740, 882
Business Services	7270–7280, 7470–7790	721–731, 741–791, 890, 892
Education & Health Services	7860–8470	812–851, 860–872, 891
Personal Services	8560–9290	641, 750–802, 880–881
Public Admin	9370–9590	900–932

Appendix B. Supplemental Material

Appendix B.1. Details of Weighting Functions Estimation

Appendix B.1.1. Estimating the Weights

We are interested in estimating weights ω that are generally functions of the distribution of (T, X). The three weighting functions under consideration are $\omega_1(T)$, $\omega_0(T)$, and $\omega_C(T, X)$. The first two weights are trivially estimated as:

$$\widehat{\omega}_1(T) = \frac{T}{\widehat{p}} \qquad \text{and} \qquad \widehat{\omega}_0(T) = \frac{1-T}{1-\widehat{p}}$$

where $\widehat{p} = N^{-1}\sum_{i=1}^{N} T_i$.

The weighting function $\omega_C(T, X)$ can be estimated as

$$\widehat{\omega}_C(T, X) = \frac{1-T}{\widehat{p}} \cdot \left(\frac{\widehat{p}(X)}{1-\widehat{p}(X)}\right),$$

where $\widehat{p}(\cdot)$ is an estimator of the true probability of being in Group 1 given X. We describe in detail below the two approaches that we consider, a parametric one and a non-parametric one. In addition, to have weights summing up to one, we use the following normalization procedures:

$$\begin{aligned}
\widehat{\omega}_1^*(T_i) &= \frac{\widehat{\omega}_1(T_i)}{\sum_{j=1}^{N}\widehat{\omega}_1(T_j)} = \frac{T_i}{N\cdot\widehat{p}}, \\
\widehat{\omega}_0^*(T_i) &= \frac{\widehat{\omega}_0(T_i)}{\sum_{j=1}^{N}\widehat{\omega}_0(T_j)} = \frac{1-T_i}{N\cdot(1-\widehat{p})}, \\
\widehat{\omega}_C^*(T_i, X_i) &= \frac{\widehat{\omega}_C(T_i)}{\sum_{j=1}^{N}\widehat{\omega}_C(T_j)} = \frac{(1-T_i)\cdot\left(\frac{\widehat{p}(X_i)}{1-\widehat{p}(X_i)}\right)}{\sum_{j=1}^{N}(1-T_j)\cdot\left(\frac{\widehat{p}(X_j)}{1-\widehat{p}(X_j)}\right)}.
\end{aligned}$$

Appendix B.1.2. Estimating the Distributional Statistics

We are interested in the estimation and inference of ν_1, ν_0, and ν_C. It can be shown that, under certain regularity conditions, estimators of these objects will be distributed asymptotically normal. We now show how to estimate those quantities, and derive their asymptotic distributions below.

The estimation follows a plug-in approach. Replacing the CDF by the empirical distribution function yields the estimators of interest:

$$\widehat{\nu}_t = \nu\big(\widehat{F}_t\big), \qquad t = 0, 1; \qquad \widehat{\nu}_C = \nu\big(\widehat{F}_C\big)$$

where

$$\widehat{F}_t(y) = \sum_{i=1}^{N}\widehat{\omega}_t^*(T_i)\cdot\mathbb{1}\{Y_i \le y\}, \quad t = 0, 1$$

$$\widehat{F}_C(y) = \sum_{i=1}^{N}\widehat{\omega}_C^*(T_i, X_i)\cdot\mathbb{1}\{Y_i \le y\}.$$

Note that, in practice, it is not usually necessary to compute these empirical distribution functions to get estimates of a distributional statistic, $\widehat{\nu}$. Standard software programs such as Stata can be used to compute distributional statistics directly from the observations on Y using the appropriate weighting factor.

The estimated distributional statistics can then be used to estimate the wage structure and composition effects as $\widehat{\Delta}_S^{\nu} = \widehat{\nu}_1 - \widehat{\nu}_C$ and $\widehat{\Delta}_X^{\nu} = \widehat{\nu}_C - \widehat{\nu}_0$.

Appendix B.1.3. Parametric Propensity Score Estimation

Suppose that $p(X)$ is correctly specified up to a finite vector of parameters δ_0. That is, $p(X) = p(X;\delta_0)$ or more formally:

Assumption A1. ***(Parametric p-score)*** $\Pr[T=1|X=x] = p(x;\delta_0)$*; where* $p(\cdot;\delta_0) : \mathcal{X} \to [0,1]$ *is a known function up to* $\delta_0 \in \mathbb{R}^d$, $d < +\infty$.

Estimation of δ_0 follows by maximum likelihood:

$$\widehat{\delta}_{MLE} = \arg\max_{\delta} \sum_{i=1}^{N} T_i \cdot \log(p(X_i;\delta)) + (1-T_i) \cdot \log(1-p(X_i;\delta))$$

Define the derivative of $p(X;\delta)$ with respect to δ as $\dot{p}(X;\delta) = \partial p(X;\delta)/\partial\delta$. The score function $s(T,X;\delta)$ is:

$$s(T,X;\delta) = \dot{p}(X;\delta) \cdot \frac{T - p(X;\delta)}{p(X;\delta) \cdot (1-p(X;\delta))}$$

Using a normalization argument, we suppress the entry for δ whenever a function of it is evaluated at the true δ. Therefore,

$$s(T,X;\delta_0) = s(T,X) = \dot{p}(X) \cdot \frac{T - p(X)}{p(X) \cdot (1-p(X))}$$

and finally

$$\widehat{\omega}_C(T,X) = \frac{1-T}{\widehat{p}} \cdot \left(\frac{p\left(X;\widehat{\delta}_{MLE}\right)}{1-p\left(X;\widehat{\delta}_{MLE}\right)} \right)$$

In particular, in this paper, we assume that the $p(x;\delta_0)$ can be modeled as a logit, that is,

$$p(x;\delta_0) = L(x'\delta_0)$$

where $L : \mathbb{R} \to \mathbb{R}$, $L(z) = (1 + exp(-z))^{-1}$.

Appendix B.1.4. Nonparametric Propensity Score Estimation

Suppose that $p(X)$ is completely unknown to the researcher. In that case, following Hirano et al. (2003), we approximate the log odds ratio by a polynomial series. In practice, this is done by finding a vector $\widehat{\pi}$ that is the solution of the following problem:

$$\hat{\pi} = \arg\max_{\pi} \sum_{i=1}^{N} T_i \cdot \log\left(L\left(H_J(X_i)'\pi\right)\right) + (1-T_i) \cdot \log\left(1 - L\left(H_J(X_i)'\pi\right)\right)$$

where $H_J(x) = [H_{J,j}(x)] \quad (j = 1, ..., J)$, a vector of length J of polynomial functions of $x \in \mathcal{X}$ satisfying the following properties: (*i*) $H_J : \mathcal{X} \to \mathbb{R}^J$; and (*ii*) $H_{J,1}(x) = 1$. More details on this estimation procedure can be found at Hirano et al. (2003) or in Firpo (2007). The non-parametric feature of this estimation procedure comes from the fact that such approximation is refined as the sample size increases, that is, J will be a function of the sample size N, $J = J(N) \to +\infty$ as $N \to +\infty$.

In this approach, $p(X)$ is estimated by $\hat{p}(X) = L(H_J(X)'\hat{\pi})$, thus:

$$\widehat{\omega}_C(T,X) = \frac{1-T}{\widehat{p}} \cdot \left(\frac{L(H_J(X)'\hat{\pi})}{1-L(H_J(X)'\hat{\pi})} \right)$$

Appendix B.2. Asymptotic Distribution

We first show that the plug-in estimators $\widehat{\nu}$ are asymptotically normal and compute their asymptotic variances. We then do the same for the density estimators.

Appendix B.2.1. The Asymptotic Distribution of Plug-In Estimators

We start by assuming that the estimators $\widehat{\nu}$ are asymptotically linear in the following sense:

Assumption A2 (Asymptotic Linearity). *$\widehat{\nu}_t$ and $\widehat{\nu}_C$ are asymptotically linear, that is,*

$$\begin{aligned} \nu\left(\widehat{F}_t\right) - \nu\left(F_t\right) &= \sum_{i=1}^{N} \widehat{\omega}_t\left(T_i, X_i\right) \cdot \mathrm{IF}(Y_i; F_t, \nu) + o_p(1/\sqrt{N}) \\ \nu\left(\widehat{F}_C\right) - \nu\left(F_C\right) &= \sum_{i=1}^{N} \widehat{\omega}_C\left(T_i, X_i\right) \cdot \mathrm{IF}(Y_i; F_C, \nu) + o_p(1/\sqrt{N}) \end{aligned}$$

Assumption A2 establishes that the estimators are either exactly linear, as those that are based on sample moments, or they can be linearized and the remainder term will approach zero as the sample size increases.

An additional technical assumption is that the influence function are square integrable and its conditional expectation given X is differentiable. To simplify notation, let us write $\mathrm{IF}(Y_t; \nu, F) = \psi_t^{\nu}(Y)$.

Assumption A3. [***Influence Function***] *For all weighting functions ω considered,*
(i) $E\left[\left(\psi_t^{\nu}\left(Y; F_t\right)\right)^2\right] < \infty$, $E\left[\left(\psi_C^{\nu}\left(Y; F_C\right)\right)^2\right] < \infty$ and
(ii) $E\left[\psi_t^{\nu}\left(Y; F_t\right) | X = x\right]$ $E\left[\psi_C^{\nu}\left(Y; F_C\right) | X = x\right]$ and are continuously differentiable for all x in $\mathcal{X}$.

Under ignorability, both types of estimators (parametric and non-parametric first step) for $\widehat{\nu}_1$, $\widehat{\nu}_0$, and $\widehat{\nu}_C$ proposed before will remain asymptotically linear. The theorem below considers both the parametric and non-parametric cases.

Theorem A1. [***Asymptotic Normality of the $\widehat{\nu}$ Estimators***]**:**
Under Assumptions 1, 2, A2 and A3:
(i-ii) $\sqrt{N} \cdot \left(\widehat{\nu}_t - \nu_t\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \omega_t(T_i) \cdot \psi^{\nu}\left(Y_i; F_t\right) + o_p(1) \xrightarrow{D} N\left(0, V_t\right)$, $t = 0, 1$
(iii) (a) if in addition, Assumption A1 holds, then:

$$\begin{aligned} \sqrt{N} \cdot \left(\widehat{\nu}_C - \nu_C\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} & \omega_C(T_i, X_i) \cdot \psi^{\nu}\left(Y_i; F_C\right) \\ & + \left(\omega_1(T_i) - \omega_C(T_i, X_i)\right) \cdot \frac{\dot{p}\left(X_i\right)'}{p\left(X_i\right)} \cdot \left(\mathbb{E}\left[s\left(T, X\right) \cdot s\left(T, X\right)'\right]\right)^{-1} \\ & \cdot \mathbb{E}\left[\frac{\dot{p}\left(X\right)}{1-p\left(X\right)} \cdot \mathbb{E}\left[\psi_C^{\nu}\left(Y; F_C\right) \mid X, T=0\right]\right] + o_p(1) \xrightarrow{D} N\left(0, V_{C,P}\right) \end{aligned}$$

(iii) (b) otherwise, if in addition we assume [non-parametric], then:

$$\begin{aligned} \sqrt{N} \cdot \left(\widehat{\nu}_C - \nu_C\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} & \omega_C(T_i, X_i) \cdot \psi^{\nu}\left(Y_i; F_C\right) \\ & + \left(\omega_1(T_i) - \omega_C(T_i, X_i)\right) \cdot \mathbb{E}\left[\psi_C^{\nu}\left(Y; F_C\right) \mid X_i, T=0\right] + o_p(1) \xrightarrow{D} N\left(0, V_{C,NP}\right) \end{aligned}$$

where

$$V_t = \mathbb{E}\left[\left(\omega_t(T)\cdot\psi_t^{\nu}\left(Y;F_t\right)\right)^2\right], \quad t=0,1$$

$$V_{C,P} = \mathbb{E}\left[\left(\omega_C(T,X)\cdot\psi^{\nu}\left(Y;F_C\right) + \left(\omega_1(T)-\omega_C(T,X)\right)\cdot\frac{\dot{p}\left(X\right)'}{p\left(X\right)}\cdot\left(\mathbb{E}\left[s\left(T,X\right)\cdot s\left(T,X\right)'\right]\right)^{-1} \cdot\mathbb{E}\left[\frac{\dot{p}\left(X\right)}{1-p\left(X\right)}\cdot\mathbb{E}\left[\psi_C^{\nu}\left(Y;F_C\right)\mid X,T=0\right]\right]\right)^2\right]$$

$$V_{C,NP} = \mathbb{E}\left[\left(\omega_C(T,X)\cdot\psi^{\nu}\left(Y,X;F_C\right) + \left(\omega_1(T)-\omega_C(T,X)\right)\cdot\mathbb{E}\left[\psi_C^{\nu}\left(Y,X;F_C\right)\mid X,T=0\right]\right)^2\right]$$

Appendix B.3. Proofs

Proof of Result 1. A proof can be found in Firpo and Pinto (2016). □

Proof of Result 2. Part (*i*) is straightforward and follows from identification of the functionals ν_1, ν_0 and ν_C, a direct consequence of identification of F_1, F_0 and F_C. Part (*ii*) follows from the fact that

$$\begin{aligned} F_1\left(y\right) &= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{g_1\left(X,\varepsilon\right)\le y\}\mid T=1,X\right]\right] \\ &= \begin{array}{r}\mathbb{E}[\mathbb{E}\left[\mathbb{1}\{g_0\left(X,\varepsilon\right)\le y\}\mid T=1,X\right] \\ +\mathbb{E}\left[\mathbb{1}\{g_1\left(X,\varepsilon\right)\le y\}-\mathbb{1}\{g_0\left(X,\varepsilon\right)\le y\}\mid T=1,X\right]\end{array} \\ &= F_C\left(y\right)+F_{1-0}\left(y\right)\end{aligned}$$

where

$$F_{1-0}\left(y\right) = E\left[\mathbb{E}\left[\mathbb{1}\{g_1\left(X,\varepsilon\right)\le y\}-\mathbb{1}\{g_0\left(X,\varepsilon\right)\le y\}\mid T=1,X\right]\right]$$

thus, if $g_1\left(\cdot,\cdot\right)=g_0\left(\cdot,\cdot\right)$, then for all y, $F_{1-0}\left(y\right)=0$ and

$$\nu_1=\nu\left(F_1\right)=\nu\left(F_C+F_{1-0}\right)=\nu\left(F_C\right)=\nu_C.$$

Part (*iii*) follows from a similar argument:

$$\begin{aligned} F_0\left(y\right) &= \int\Pr\left[Y_0\le y\mid T=0,X=x\right]\cdot dF_{X|T}\left(x|0\right)\cdot dx \\ &= \int\Pr\left[Y_0\le y\mid T=0,X=x\right]\cdot dF_{X|T}\left(x|1\right)\cdot dx \\ &\quad+\int\Pr\left[Y_0\le y\mid T=0,X=x\right]\cdot\left(dF_{X|T}\left(x|0\right)-dF_{X|T}\left(x|1\right)\right)\cdot dx \\ &= F_C\left(y\right)+F_\Delta\left(y\right)\end{aligned}$$

where

$$F_\Delta\left(y\right)=\int\Pr\left[Y_0\le y\mid T=0,X=x\right]\cdot d\left(F_{X|T}\left(x|0\right)-F_{X|T}\left(x|1\right)\right)\cdot dx$$

thus if $F_{X|T}\left(\cdot|1\right)=F_{X|T}\left(\cdot|0\right)$, then for all x, $F_{X|T}\left(x|1\right)-F_{X|T}\left(x|0\right)=0$ and therefore, for all y, $F_\Delta\left(y\right)=0$ and

$$\nu_0=\nu\left(F_0\right)=\nu\left(F_C+F_\Delta\right)=\nu\left(F_C\right)=\nu_C.$$

□

Proof of Theorem A1. A proof of parts (*i*), (*ii*) and (*iii*) (*b*) can be found in Firpo and Pinto (2016). A proof of part (*iii*) (*a*) can be found in Chen et al. (2008). □

References

Acemoglu, Daron, and David H. Autor. 2011. Skills, Tasks, and Technologies: Implications for Employment and Earnings. In *Handbook of Labor Economics*. Edited by Orley Ashenfelter and David Card. Amsterdam: North-Holland, vol. IV.B, pp. 1043–172.

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. 2013. The Top 1 Percent in International and Historical Perspective. *Journal of Economic Perspectives* 27: 3–20. [CrossRef]

Autor, David H., and David Dorn. 2013. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review* 103: 1553–97. [CrossRef]

Autor, David H., David Dorn, Gordon H. Hanson, and Jae Song. 2014. Trade Adjustment: Worker-level Evidence. *Quarterly Journal of Economics* 129: 1799–860. [CrossRef]

Autor, David H., David Dorn, and Gordon H. Hanson. 2015. Untangling Trade and Technology: Evidence from Local Labour Markets. *Economic Journal* 125: 621–46. [CrossRef]

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2005. Rising Wage Inequality: The Role of Composition and Prices. NBER Working paper No. 11628, National Bureau of Economic Research, Cambridge, MA, USA.

Autor, David H., Lawrence F. Katz, and Melissa S. Kearney. 2006. The Polarization of the U.S. Labor Market. *American Economic Review* 96: 189–94. [CrossRef]

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. The Skill Content Of Recent Technological Change: An Empirical Exploration. *Quarterly Journal of Economics* 118: 1279–333. [CrossRef]

Barsky, Robert, John Bound, Kerwin Kofi Charles, and Joseph P. Lupton. 2002. Accounting for the Black-White Wealth Gap: A Nonparametric Approach. *Journal of the American Statistical Association* 97: 663–73. [CrossRef]

Bento, Antonio, Kenneth Gillingham, and Kevin Roth. 2017. The Effect of Fuel Economy Standards on Vehicle Weight Dispersion and Accident Fatalities. NBER Working paper No. w23340, National Bureau of Economic Research, Cambridge, MA, USA.

Blinder, Alan. 1973. Wage Discrimination: Reduced Form and Structural Estimates. *Journal of Human Resources* 8: 436–55. [CrossRef]

Brochu, Pierre, David A. Green, Thomas Lemieux, and James Townsend. 2017. The Minimum Wage, Turnover, and the Shape Effects of Wage Distribution. In *Mimeo*. Vancouver: University of British Columbia.

Card, David. 1992. The Effects of Unions on the Distribution of Wages: Redistribution or Relabelling? NBER Working paper No. 4195, National Bureau of Economic Research, Cambridge, MA, USA.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2017. Simple Local Polynomial Density Estimators. In *Mimeo*. Berkeley: UC Berkeley.

Chamberlain, Gary. 1994. Quantile Regression Censoring and the Structure of Wages. In *Advances in Econometrics*. Edited by Christopher Sims. New York: Elsevier.

Chernozhukov, Victor, Ivan Fernandez-Val, and Blaise Melly. 2013. Inference on Counterfactual Distributions. *Econometrica* 81: 2205–68.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. 2008. Semiparametric Efficiency in GMM Models with Auxiliary Data. *The Annals of Statistics* 36: 808–43. [CrossRef]

Choe, Chung, and Philippe Van Kerm. 2014. *Foreign Workers and the Wage Distribution: Where Do They Fit in*? Technical Report 2014-02. Esch-sur-Alzette: Luxembourg Institute of Socio-Economic Research.

Cowell, Frank, and Maria-Pia Victoria-Feser 1996. Robustness Properties of Inequality Measures. *Econometrica* 64: 77–101. [CrossRef]

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach. *Econometrica* 64: 1001–44. [CrossRef]

Eeckhout, Jan, Roberto Pinheiro, and Kurt Schmidheiny. 2014. Spatial sorting. *Journal of Political Economy* 122: 554–620. [CrossRef]

Essama-Nssah, Boniface, and Peter J. Lambert. 2012. Influence functions for policy impact analysis. In *Inequality, Mobility and Segregation: Essays in Honor of Jacques Silber.* Edited by John A. Bishop and Rafael Salas. Cheltenham: Emerald Group Publishing Limited, chp. 6, pp. 135–59.

Firpo, Sergio. 2007. Efficient Semiparametric Estimation of Quantile Treatment Effects. *Econometrica* 75: 259–76. [CrossRef]

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2007. Decomposing Wage Distributions using Recentered Influence Functions Regressions. In *Mimeo.* Vancouver: University of British Columbia.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. Unconditional Quantile Regressions. *Econometrica* 77: 953–973.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2011. Occupational Tasks and Changes in the Wage Structure. In *Mimeo.* Vancouver: University of British Columbia.

Firpo, Sergio, and Cristine Pinto. 2016. Identification and Estimation of Distributional Impacts of Interventions Using Changes in Inequality Measures. *Journal of Applied Econometrics* 31: 457–86. [CrossRef]

Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. Decomposition Methods in Economics. In *Handbook of Labor Economics.* Edited by Orley Ashenfelter and David Card. Amsterdam: North-Holland, vol. IV.A, pp. 1–102.

Fortin, Nicole, and Thomas Lemieux. 2016. Inequality and Changes in Task Prices: Within and between Occupation Effects?. In *Income Inequality, Causes and Consequences (Research in Labor Economics, Vol. 43).* Edited by Lorenzo Cappellari, Solomon W. Polachek, and Konstantinos Tatsiramos. Cheltenham: Emerald Group Publishing Limited, pp. 195–226.

Freeman, Richard B. 1980. Unionism and the Dispersion of Wages. *Industrial and Labor Relations Review* 34: 3–23. [CrossRef]

Freeman, Richard B. 1993. How Much has Deunionization Contributed to the Rise of Male Earnings Inequality? In *Uneven Tides: Rising Income Inequality in America.* Edited by Sheldon Danziger and Peter Gottschalk. New York: Russell Sage Foundation, pp. 133–63.

Gâteaux, René. 1913. Sur les fonctionnelles continues et les fonctionnelles analytiques. *Comptes Rendus de l'Académie des Sciences-Series I—Mathematics* 157: 325–27.

Gardeazabal, Javier, and Arantza Ugidos. 2004. More on the Identification in Detailed Wage Decompositions. *Review of Economics and Statistics* 86: 1034–57. [CrossRef]

Gradín, Carlos. 2016. Why Is Income inequality so High in Spain? In *Income Inequality Around the World (Research in Labor Economics, Vol. 44).* Edited by Lorenzo Cappellari, Solomon W. Polachek and Konstantinos Tatsiramos. Cheltenham: Emerald Group Publishing Limited, pp. 109–77.

Hampel, Frank R. 1974. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association* 60: 383–93. [CrossRef]

Heckman, James J. 1990. Varieties of Selection Bias. *American Economic Review* 80: 313–18.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. Matching as an Econometric Evaluation Estimator. *Review of Economic Studies* 65: 261–94. [CrossRef]

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra Todd. 1998. Characterizing Selection Bias Using Experimental Data. *Econometrica* 66: 1017–98. [CrossRef]

Heckman, James J., and Richard Robb. 1985. Alternative Methods for Evaluating the Impact of Interventions: An Overview. *Journal of Econometrics* 30: 239–67. [CrossRef]

Heckman, James J., and Richard Robb. 1986. Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes. In *Drawing Inference from Self-Selected Samples.* Edited by Howard Wainer. New York: Springer, pp. 63–107.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71: 1161–89. [CrossRef]

Jann, Ben. 2008. The Oaxaca-Blinder Decomposition for Linear Regression Models. *Stata Journal* 8: 435–79.

Juhn, Chinhui, Kevin Murphy, and Brooks Pierce. 1993. Wage Inequality and the Rise in Returns to Skill. *Journal of Political Economy* 101: 410–42. [CrossRef]

Kline, Patrick. 2011. Oaxaca-Blinder as a Reweighting Estimator. *American Economic Review* 101: 532–37. [CrossRef]

Koenker, Roger. 2005. *Quantile Regression.* Cambridge: Cambridge University Press.

Koenker, Roger, and Gilbert Bassett, Jr. 1978. Regression Quantiles. *Econometrica* 46: 33–50. [CrossRef]

Lemieux, Thomas. 2002. Decomposing Changes in Wage Distributions: A Unified Approach. *Canadian Journal of Economics* 35: 646–88. [CrossRef]

Lemieux, Thomas. 2006a. Post-secondary Education and Increasing Wage Inequality. *American Economic Review* 96: 195–99. [CrossRef]

Lemieux, Thomas. 2006b. Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill? *American Economic Review* 96: 461–98. [CrossRef]

Lemieux, Thomas. 2008. The Changing Nature of Wage Inequality. *Journal of Population Economics* 21: 21–48. [CrossRef]

Machado, José A. F., and José Mata. 2005. Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression. *Journal of Applied Econometrics* 20: 445–65. [CrossRef]

Melly, Blaise. 2005. Decomposition of Differences in Distribution Using Quantile Regression. *Labour Economics* 12: 577–1990. [CrossRef]

Monti, Anna Clara. 1991. The Study of the Gini Concentration Ratio by Means of the Influence Function. *Statistica* 51: 561–77.

Oaxaca, Ronald. 1973. Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review* 14: 693–709. [CrossRef]

Oaxaca, Ronald, and Michael R. Ransom. 1999. Identification in Detailed Wage Decompositions. *Review of Economics and Statistics* 81: 154–57. [CrossRef]

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70: 41–55. [CrossRef]

Rosenbaum, Paul R., and Donald B. Rubin. 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 79: 516–24. [CrossRef]

Rothe, Christoph. 2010. Nonparametric Estimation of Distributional Policy Effects. *Journal of Econometrics* 155: 56–70. [CrossRef]

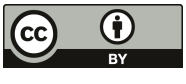
Rothe, Christoph. 2012. Partial Distributional Policy Effects. *Econometrica* 80: 2269–301.

Rothe, Christoph. 2015. Decomposing the Composition Effect. *Journal of Business Economics and Statistics* 33: 323–37. [CrossRef]

Von Mises, Richard. 1947. On the Asymptotic Distribution of Differentiable Statistical Functions. *The Annals of Mathematical Statistics* 18: 309–48. [CrossRef]

White, Halbert. 1980. Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review* 21: 149–70. [CrossRef]

Yun, Myeong-Su. 2005. A simple Solution to the Identification Problem in Detailed Wage Decompositions. *Economic Inquiry* 43: 766–72. [CrossRef]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Statistical Inference on the Canadian Middle Class

Russell Davidson [1,2]

[1] Department of Economics and CIREQ, McGill University, Montréal, QC H3A 2T7, Canada; russell.davidson@mcgill.ca

[2] AMSE-GREQAM, 5 Boulevard Maurice Bourdet, CS 50498, CEDEX 01, 13205 Marseille, France

Received: 25 December 2017; Accepted: 8 March 2018; Published: 13 March 2018

Abstract: Conventional wisdom says that the middle classes in many developed countries have recently suffered losses, in terms of both the share of the total population belonging to the middle class, and also their share in total income. Here, distribution-free methods are developed for inference on these shares, by means of deriving expressions for their asymptotic variances of sample estimates, and the covariance of the estimates. Asymptotic inference can be undertaken based on asymptotic normality. Bootstrap inference can be expected to be more reliable, and appropriate bootstrap procedures are proposed. As an illustration, samples of individual earnings drawn from Canadian census data are used to test various hypotheses about the middle-class shares, and confidence intervals for them are computed. It is found that, for the earlier censuses, sample sizes are large enough for asymptotic and bootstrap inference to be almost identical, but that, in the twenty-first century, the bootstrap fails on account of a strange phenomenon whereby many presumably different incomes in the data are rounded to one and the same value. Another difference between the centuries is the appearance of heavy right-hand tails in the income distributions of both men and women.

Keywords: middle class; Canada; bootstrap

JEL Classification: C10; C12; C15

1. Introduction

There has been much discussion in many countries about the fate of the middle class, variously defined. It appears clearly that middle classes in different developed countries have had rather different experiences; in particular, the case of the USA, about which a lot has been written, for instance, Heathcote et al. (2010), is in no way typical or representative. Canada shares a long border with the USA, and has a culture more similar to the American one than any other country, but it maintains a separate identity, and differs from the US markedly on matters of social security and immigration. Nevertheless, a couple of decades ago, it was pointed out by Foster and Wolfson (2010) that, in both countries, a decline of the middle class had led to a polarisation of the income distribution. In Canada specifically, the situation is reviewed by Brzozowski et al. (2010), for inequality not only of income, but also of wealth and consumption. For the USA, an early article by Wolfson (1994) discusses polarisation, while Wolff (2013) describes the fate of the wealth of the middle class following the crisis of 2008. Some recent trends in income inequality in different European regions have been analysed by Castells-Quintana et al. (2015).

The study of income inequality, and its effects on growth, social stability, and many other features of society, started more than half a century ago, with Kuznets (1955). A landmark contribution to the measurement of income inequality was Atkinson (1970). A useful article is Cowell (1999), which appears in the Handbook of Income Inequality Measurement, and contains many chapters on different aspects of the topic, some purely theoretical, such as the seminal contributions of Blackorby et al. (1999). An interesting recent paper, Ryu (2013), develops a sort of inverted Gini index

that emphasises the distribution of the poor, and describes ways of estimating income distributions based on the principle of maximum entropy.

The Canadian Liberal federal government elected in late 2015 has made a point of trying to improve the lot of the Canadian middle class, claiming, no doubt with some justice, that the share of the middle class, however defined, has declined over the last several decades, in terms of both the share of the population belonging to the middle class, and also its share in total national income.

Beach (2016), in his presidential address to the Canadian Economics Association, drew a wide-ranging portrait of the evolution of Canadian middle-class fortunes since the 1970s. His analysis tries to understand the different mechanisms that have shaped the economic environment in which this evolution has taken place. He provides abundant statistical information on earnings in Canada, duly separating the two sexes in his analysis, given that their position in the labour market has changed very considerably in the last fifty years.

The aim of this paper is to bring some formal statistical analysis to bear on the Canadian census data. The work of Davidson and Duclos[1], found in Davidson and Duclos (1997) and Davidson and Duclos (2000), introduced a set of statistical procedures that permit distribution-free inference on income data, many of which can be used directly for the analysis in this paper. Some extensions of their methodology are developed here to deal with the specific problems addressed.

Formal analysis requires a formal definition of the middle class. An ideal definition would have to be based on all sorts of socioeconomic characteristics of individuals and households, but such a thing is well outside the scope of this paper. Instead, we consider definitions based solely on individual income. Usually different segments of the income distribution are defined by use of quantiles, and income data are sometimes grouped by deciles or vigintiles. Thus, a possible definition of the middle class could be those households or individuals whose incomes lie between the second decile and the eighth. Another approach would be to define the upper and lower bounds of middle-class incomes as multiples of the mean or median income. However, given the stylised fact that the recent changes in income inequality in most developed countries have favoured the rich and the super-rich, use of the mean as a criterion for defining income classes is likely to distort inference. It is easy to see that a substantial increase in the income of the upper 10% of the distribution, with no changes for the lower 90%, leads to an increase in mean income and no change in the median. Similarly, quantile-based definitions of the middle class are unaffected by an increase in the income of the rich and only the rich.

If the middle class is defined as the set of individuals with incomes between the p_{lo} quantile of the income distribution and the p_{hi} quantile, where a possible choice might be $p_{\mathrm{lo}} = 0.2$ and $p_{\mathrm{hi}} = 0.8$, it is not possible to measure changes in the population share of the middle class, because this share is always just $p_{\mathrm{hi}} - p_{\mathrm{lo}}$. It remains possible to measure changes in the income share.

In the next section, distribution-free plug-in estimators are presented for the population and income shares of the middle class, according to three different sorts of definition of the middle class—based on the median income, based on the mean income, and based on quantiles of the income distribution. These estimators are shown to be consistent and asymptotically normal, and feasible estimators are given for the asymptotic variance. Then, in Section 3, the evolution over time of the middle-class shares in Canada is analysed using census data from the 1971 census to that in 2006. Section 4 concludes.

2. Asymptotic Analysis

We begin with a definition of the middle class as the section of the population with incomes between a fraction a of median income and a multiple b of it. Typically, we might have $a = 0.5$ and $b = 1.5$. It is desired to estimate the size of this section of the population, and also to estimate its share in total income. Other definitions will be considered later.

---

[1] Currently (December 2017) Minister of Families, Children and Social Development in the Canadian federal government.

2.1. Definition in Terms of the Median

Let m denote median income. Then, the proportion of the population considered to be middle class is $F(bm) - F(am)$, where F is the cumulative distribution function (CDF) of income in the population. To estimate this quantity based on a random sample of size N, it is necessary to have an estimate of F, i.e., $\hat{F}$, from which an estimate of m may be deduced, or else obtained directly using order statistics, by use of the formula

$$\hat{m} = \begin{cases} y_{(n+1)} & \text{if } N = 2n+1 \ (N \text{ odd}) \\ (y_{(n)} + y_{(n+1)})/2 & \text{if } N = 2n \ (N \text{ even}) \end{cases}$$

The natural choice for $\hat{F}$ is the empirical distribution function (EDF):

$$\hat{F}(y) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}(y_i \leq y), \tag{1}$$

where the y_i are the incomes observed in the sample, and I is the indicator function, equal to 1 when its argument is true, to 0 otherwise. If PS denotes the share of the middle class in the whole population, then it can be estimated by

$$\widehat{PS} = \hat{F}(b\hat{m}) - \hat{F}(a\hat{m}) \tag{2}$$

The income share, i.e., IS, that accrues to the middle class is by definition given by

$$\int_{am}^{bm} y \, \mathrm{d}F(y)$$

divided by the mean income, denoted by μ, and equal to $\int_0^\infty y \, \mathrm{d}F(y)$. The plug-in estimator of μ is

$$\hat{\mu} = \int_0^\infty y \, \mathrm{d}\hat{F}(y) = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

Consequently, a suitable estimate of IS is

$$\widehat{IS} \equiv \frac{1}{\hat{\mu}} \int_{a\hat{m}}^{b\hat{m}} y \, \mathrm{d}\hat{F}(y). \tag{3}$$

For asymptotic statistical inference, we need estimates of the asymptotic covariance matrix of $(\widehat{PS}, \widehat{IS})$. Consider first the asymptotic variance of $\widehat{PS}$, which is by definition the variance of the limit in distribution as $N \to \infty$ of $N^{1/2}(\widehat{PS} - PS)$. We have

$$\widehat{PS} - PS = \hat{F}(b\hat{m}) - F(bm) - \big(\hat{F}(a\hat{m}) - F(am)\big). \tag{4}$$

Now

$$\hat{F}(b\hat{m}) - F(bm) = \int_0^{bm} \mathrm{d}(\hat{F} - F)(y) + \int_{bm}^{b\hat{m}} \mathrm{d}F(y) + \int_{bm}^{b\hat{m}} \mathrm{d}(\hat{F} - F)(y).$$

The first two terms on the right-hand side are of order $N^{-1/2}$ if, as we can reasonably assume, things are regular enough for both $(\hat{F} - F)(y)$ and $\hat{m} - m$ to be of that order. The last term, on the other hand, is of order N^{-1}, and so can be dropped for the purposes of asymptotic analysis. The first term is

$$\frac{1}{N} \sum_{i=1}^{N} \big[\mathrm{I}(y_i \leq bm) - F(bm)\big], \tag{5}$$

and the second is

$$bf(bm)(\hat{m}-m)+O(N^{-1}), \tag{6}$$

where $f=F'$ is the density function. By the Bahadur (1966) almost-sure representation of quantiles, we have

$$\hat{m}-m=-\frac{1}{Nf(m)}\sum_{i=1}^{N}\left[\mathrm{I}(y_i<m)-\frac{1}{2}\right]+O\big(N^{-3/4}(\log N)^{1/2}(\log\log N)^{1/4}\big). \tag{7}$$

From (4), (5), (6), and (7), we conclude that

$$N^{1/2}(\widehat{PS}-PS)=N^{-1/2}\sum_{i=1}^{N}\left\{\Big[\mathrm{I}(am\le y_i\le bm)-\big(F(bm)-F(am)\big)\Big]\right.$$
$$\left.-\frac{bf(bm)-af(am)}{f(m)}\left[\mathrm{I}(y_i<m)-\frac{1}{2}\right]\right\}+o_p(1).$$

It is convenient to make the following definition:

$$u_i=\mathrm{I}(am<y_i<bm)-\frac{bf(bm)-af(am)}{f(m)}\mathrm{I}(y_i<m). \tag{8}$$

Since the y_i are IID, as elements of a random sample, so are the u_i, so that, to leading order asymptotically,

$$N^{1/2}(\widehat{PS}-PS)=N^{-1/2}\sum_{i=1}^{N}\big(u_i-\mathrm{E}(U)\big), \tag{9}$$

where U denotes a random variable that has the distribution of which the u_i are IID realisations. We may therefore apply the central-limit theorem to show that $N^{1/2}(\widehat{PS}-PS)$ is asymptotically normal, with expectation zero and variance equal to that of U. If we make the definition

$$\hat{u}_i=\mathrm{I}(a\hat{m}<y_i<b\hat{m})-\frac{b\hat{f}(b\hat{m})-a\hat{f}(a\hat{m})}{\hat{f}(\hat{m})}\mathrm{I}(y_i<\hat{m}),$$

where the density estimate $\hat{f}$ could be a kernel density estimate, we can estimate $\mathrm{var}(U)$ by

$$N^{-1}\sum_{i=1}^{N}\hat{u}_i^2-\Big[N^{-1}\sum_{i=1}^{N}\hat{u}_i\Big]^2.$$

We now turn to $N^{1/2}(\widehat{IS}-IS)$. From (3), we see that

$$\widehat{IS}-IS=\frac{\mu\int_{a\hat{m}}^{b\hat{m}}y\,\mathrm{d}\hat{F}(y)-\hat{\mu}\int_{am}^{bm}y\,\mathrm{d}F(y)}{\mu\hat{\mu}}. \tag{10}$$

The numerator is clearly of order $N^{-1/2}$, while the denominator is $O_p(1)$, being equal to $\mu^2+O_p(N^{-1/2})$ To leading order, therefore, we can replace the denominator by its leading term, namely μ^2. Make the definition

$$\mu_{ab}=\int_{am}^{bm}y\,\mathrm{d}F(y).$$

Now, by arguments like those used above for $\widehat{PS}$, we have to leading order that

$$\begin{aligned}\int_{a\hat{m}}^{b\hat{m}} y\,\mathrm{d}\hat{F}(y) &= \int_{am}^{bm} y\,\mathrm{d}\hat{F}(y) + \int_{a\hat{m}}^{am} y\,\mathrm{d}F(y) + \int_{bm}^{b\hat{m}} y\,\mathrm{d}F(y)\\ &= \int_{am}^{bm} y\,\mathrm{d}\hat{F}(y) + m\big(b^2 f(bm) - a^2 f(am)\big)(\hat{m} - m)\end{aligned} \tag{11}$$

and

$$\int_{am}^{bm} y\,\mathrm{d}\hat{F}(y) = \frac{1}{N}\sum_{i=1}^{N}\big[y_i\,\mathrm{I}(am < y_i < bm)\big]. \tag{12}$$

Note that

$$\hat{\mu} = \mu + \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu). \tag{13}$$

If we make the definition

$$v_i = \frac{1}{\mu^2}\Big[\mu y_i\,\mathrm{I}(am < y_i < bm) - \mu_{ab} y_i - \frac{\mu m}{f(m)}\big(b^2 f(bm) - a^2 f(am)\big)\mathrm{I}(y_i < m)\Big],$$

we see that, to leading order,

$$N^{1/2}(\widehat{IS} - IS) = N^{-1/2}\sum_{i=1}^{N}\big(v_i - \mathrm{E}(V)\big), \tag{14}$$

with V a random variable whose distribution is that of which the v_i are IID realisations. We may once more apply the central-limit theorem to conclude that $N^{1/2}(\widehat{IS} - IS)$ is asymptotically normal with variance equal to that of V.

Define

$$\hat{v}_i = \frac{1}{\hat{\mu}^2}\Big[\hat{\mu}\,\mathrm{I}(a\hat{m} < y_i < b\hat{m}) - \hat{\mu}_{ab} y_i - \frac{\hat{\mu}\hat{m}}{\hat{f}(\hat{m})}\big(b^2\hat{f}(b\hat{m}) - a^2\hat{f}(a\hat{m})\big)\mathrm{I}(y_i < \hat{m})\Big]$$

where

$$\hat{\mu}_{ab} = N^{-1}\sum_{i=1}^{N} y_i\,\mathrm{I}(a\hat{m} < y_i < b\hat{m}).$$

It is then clear that we can estimate $\mathrm{var}(V)$ by

$$N^{-1}\sum_{i=1}^{N}\hat{v}_i^2 - \Big[N^{-1}\sum_{i=1}^{N}\hat{v}_i\Big]^2, \tag{15}$$

and the covariance of U and V by

$$N^{-1}\sum_{i=1}^{N}\hat{u}_i\hat{v}_i - \Big[N^{-1}\sum_{i=1}^{N}\hat{u}_i\Big]\Big[N^{-1}\sum_{i=1}^{N}\hat{v}_i\Big]. \tag{16}$$

Remark 1. *In some cases, the sample is not supposed to be completely random. Rather, observation i is associated with a weight p_i, defined such that $\sum_{i=1}^{N} p_i = N$. In that case, the empirical distribution function in Equation (1) should be replaced by*

$$\hat{F}(y) = \frac{1}{N}\sum_{i=1}^{N} p_i\mathrm{I}(y_i \le y). \tag{17}$$

Similarly, the mean income should be defined as $\hat{\mu} = N^{-1}\sum_{i=1}^{N} p_i y_i$, the expectation of the EDF in Equation (17), and term i in the sums (9) and (14) should be weighted by p_i.

The use of non-uniform weights also has consequences for the bootstrap, as discussed later.

2.2. Definition in Terms of the Mean

Although, for the current study, it is not very sensible to define the range of middle-class incomes as delimited by multiples of the mean income, it may be useful in other circumstances to be able to perform inference on shares thus defined. Let a and b, $a < b$ define the middle class as those individuals that have incomes between $a\mu$ and $b\mu$. The population share is then

$$PS = F(b\mu) - F(a\mu), \quad \text{with} \quad \widehat{PS} = \hat{F}(b\hat{\mu}) - \hat{F}(a\hat{\mu}) = N^{-1}\sum_{i=1}^{N} \mathrm{I}(a\hat{\mu} < y_i < b\hat{\mu}).$$

From this, we see that

$$\widehat{PS} - PS = \hat{F}(b\hat{\mu}) - F(b\mu) - \big(\hat{F}(a\hat{\mu}) - F(a\mu)\big).$$

Now, as usual neglecting terms of order N^{-1}, we see that

$$\begin{array}{ll} \hat{F}(b\hat{\mu}) - F(b\mu) & = \int_0^{b\mu} \mathrm{d}(\hat{F} - F)(y) + \int_{b\mu}^{b\hat{\mu}} \mathrm{d}F(y) \\ & = N^{-1}\sum_{i=1}^{N}\big[\mathrm{I}(y_i < b\mu) - F(b\mu)\big] + bf(b\mu)(\hat{\mu} - \mu) \\ & = N^{-1}\sum_{i=1}^{N}\big[\mathrm{I}(y_i < b\mu) + bf(b\mu)y_i - \big(F(b\mu) + bf(b\mu)\mu\big)\big], \end{array} \tag{18}$$

where $f = F'$ is the density, and the last equality makes use of (13). The terms in (18) clearly have expectation zero.

It is straightforward now to see that, to leading order,

$$N^{1/2}(\widehat{PS} - PS) = N^{-1/2}\sum_{i=1}^{N}\big(u_i - \mathrm{E}(U)\big),$$

with $u_i = \mathrm{I}(a\mu < y_i < b\mu) + y_i\big(bf(b\mu) - af(a\mu)\big)$ and U a random variable with the distribution of which the u_i are realisations. The asymptotic variance of $N^{1/2}(\widehat{PS} - PS)$ can therefore be estimated by

$$N^{-1}\sum_{i=1}^{N}\hat{u}_i^2 - \Big[N^{-1}\sum_{i=1}^{N}\hat{u}_i\Big]^2,$$

where $\hat{u}_i = \mathrm{I}(a\hat{\mu} < y_i < b\hat{\mu}) + y_i\big(b\hat{f}(b\hat{\mu}) - a\hat{f}(a\hat{\mu})\big)$, with $\hat{f}$ a kernel density estimator.

For the income share, we have

$$IS = \frac{1}{\mu}\int_{a\mu}^{b\mu} y\,\mathrm{d}F(y) \quad \text{with} \quad \widehat{IS} = \frac{1}{\hat{\mu}}\int_{a\hat{\mu}}^{b\hat{\mu}} y\,\mathrm{d}\hat{F}(y).$$

Analogously to (10), we have

$$\widehat{IS} - IS = \frac{\mu\int_{a\hat{\mu}}^{b\hat{\mu}} y\,\mathrm{d}\hat{F}(y) - \hat{\mu}\int_{a\mu}^{b\mu} y\,\mathrm{d}F(y)}{\mu\hat{\mu}}.$$

Now, as in (11) and (12), to leading order, we have

$$\begin{array}{ll} \int_{a\hat{\mu}}^{b\hat{\mu}} y\,\mathrm{d}\hat{F}(y) & = \Big[\int_{a\mu}^{b\mu} + \int_{a\hat{\mu}}^{a\mu} + \int_{b\mu}^{b\hat{\mu}}\Big] y\,\mathrm{d}\hat{F}(y) \\ & = N^{-1}\sum_{i=1}^{N} y_i\,\mathrm{I}(a\mu < y_i < b\mu) + \mu\big(b^2 f(b\mu) - a^2 f(a\mu)\big)(\hat{\mu} - \mu) \\ & = N^{-1}\sum_{i=1}^{N}\Big[y_i\,\mathrm{I}(a\mu < y_i < b\mu) + \mu\big(b^2 f(b\mu) - a^2 f(a\mu)\big)(y_i - \mu)\Big] \end{array}$$

Here, let us redefine μ_{ab} as:

$$\mu_{ab} = \int_{a\mu}^{b\mu} y \, \mathrm{d}F(y).$$

Then,

$$N^{1/2}(\widehat{IS} - IS) = N^{-1/2} \sum_{i=1}^{N} \big(v_i - \mathrm{E}(V)\big),$$

where

$$v_i = \frac{y_i}{\mu^2}\Big[\mu \, \mathrm{I}(a\mu < y_i < b\mu) + \mu^2\big(b^2 f(b\mu) - a^2 f(a\mu)\big) - \mu_{ab}\Big] \quad \text{and}$$
$$\hat{v}_i = \frac{y_i}{\hat{\mu}^2}\Big[\hat{\mu} \, \mathrm{I}(a\hat{\mu} < y_i < b\hat{\mu}) + \hat{\mu}^2\big(b^2 \hat{f}(b\hat{\mu}) - a^2 \hat{f}(a\hat{\mu})\big) - \hat{\mu}_{ab}\Big]$$

with obvious definitions of $\hat{f}$ and $\hat{\mu}_{ab}$. Except for notational changes, the estimates (15) and (16) hold for this case as well.

2.3. Definition by Quantiles

Let the two proportions, p_{lo} and p_{hi}, with $p_{\mathrm{lo}} < p_{\mathrm{hi}}$, define the middle class as the set of individuals whose incomes lie between the quantiles q_{lo} and q_{hi}, where $F(q_{\mathrm{lo}}) = p_{\mathrm{lo}}$ and $F(p_{\mathrm{hi}}) = q_{\mathrm{hi}}$. Then the share of the population that belongs to the middle class is fixed at $p_{\mathrm{hi}} - p_{\mathrm{lo}}$. The income share is

$$IS = \frac{1}{\mu} \int_{q_{\mathrm{lo}}}^{q_{\mathrm{hi}}} y \, \mathrm{d}F(y),$$

and it can be estimated by

$$\widehat{IS} = \frac{1}{\hat{\mu}} \int_{\hat{q_{\mathrm{lo}}}}^{\hat{q_{\mathrm{hi}}}} y \, \mathrm{d}\hat{F}(y),$$

where $\hat{q_{\mathrm{lo}}}$ and $\hat{q_{\mathrm{hi}}}$ are the p_{lo} and p_{hi} quantiles of the EDF $\hat{F}$.

By an asymptotic argument such as those used in the preceding subsection, it can be seen that

$$\widehat{IS} - IS = \frac{1}{\mu^2}\Big[\mu \int_{\hat{q_{\mathrm{lo}}}}^{\hat{q_{\mathrm{hi}}}} y \, \mathrm{d}\hat{F}(y) - \hat{\mu} \int_{q_{\mathrm{lo}}}^{q_{\mathrm{hi}}} y \, \mathrm{d}F(y)\Big] + O_p(N^{-1}). \tag{19}$$

Neglecting terms of order N^{-1}, we have

$$\begin{aligned}
&\int_{\hat{q_{\mathrm{lo}}}}^{\hat{q_{\mathrm{hi}}}} y \, \mathrm{d}\hat{F}(y) = \int_{q_{\mathrm{lo}}}^{q_{\mathrm{hi}}} y \, \mathrm{d}\hat{F}(y) + \int_{\hat{q_{\mathrm{lo}}}}^{q_{\mathrm{lo}}} y \, \mathrm{d}\hat{F}(y) + \int_{q_{\mathrm{hi}}}^{\hat{q_{\mathrm{hi}}}} y \, \mathrm{d}\hat{F}(y) \\
&= N^{-1} \sum_{i=1}^{N} y_i \, \mathrm{I}(q_{\mathrm{lo}} < y_i < q_{\mathrm{hi}}) - q_{\mathrm{lo}}\big(p_{\mathrm{lo}} - \hat{F}(q_{\mathrm{lo}})\big) + q_{\mathrm{hi}}\big(p_{\mathrm{hi}} - \hat{F}(q_{\mathrm{hi}})\big) \\
&= p_{\mathrm{hi}} q_{\mathrm{hi}} - p_{\mathrm{lo}} q_{\mathrm{lo}} + N^{-1} \sum_{i=1}^{N} \Big[y_i \, \mathrm{I}(q_{\mathrm{lo}} < y_i < q_{\mathrm{hi}}) - q_{\mathrm{hi}} \mathrm{I}(y_i < q_{\mathrm{hi}}) + q_{\mathrm{lo}} \mathrm{I}(y_i < q_{\mathrm{lo}})\Big].
\end{aligned}$$

Define

$$\mu_{lh} = \int_{q_{\mathrm{lo}}}^{q_{\mathrm{hi}}} y \, \mathrm{d}F(y).$$

Since

$$\mathrm{E}\big(Y \, \mathrm{I}(q_{\mathrm{lo}} < Y < q_{\mathrm{hi}})\big) = \mu_{lh}, \quad \mathrm{E}\big(\mathrm{I}(Y < q_{\mathrm{lo}})\big) = p_{\mathrm{lo}}, \quad \text{and} \quad \mathrm{E}\big(\mathrm{I}(Y < q_{\mathrm{hi}})\big) = p_{\mathrm{hi}},$$

where Y is a random variable that has the distribution of which the y_i are realisations, it follows that

$$\int_{\hat{q_{\mathrm{lo}}}}^{\hat{q_{\mathrm{hi}}}} y \, \mathrm{d}\hat{F}(y) = \mu_{lh} + N^{-1} \sum_{i=1}^{N} \big(w_i - \mathrm{E}(W)\big),$$

where

$$w_i = y_i \, \mathrm{I}(q_{\mathrm{lo}} < y_i < q_{\mathrm{hi}}) - q_{\mathrm{hi}} \mathrm{I}(y_i < q_{\mathrm{hi}}) + q_{\mathrm{lo}} \mathrm{I}(y_i < q_{\mathrm{lo}}),$$

and W is a random variable that has the distribution of which the w_i are realisations. From (19), it can now be seen that

$$N^{1/2}(\widehat{IS} - IS) = N^{-1/2} \sum_{i=1}^{N} \big(v_i - \mathrm{E}(V)\big),$$

where

$$v_i = \frac{w_i}{\mu} - \frac{y_i \mu_{Ih}}{\mu^2},$$

the v_i being realisations of the distribution of V.

The asymptotic variance of the asymptotically normal random variable $N^{1/2}(\widehat{IS} - IS)$ is therefore equal to the variance of V. This variance can be estimated in a distribution-free manner by

$$N^{-1} \sum_{i=1}^{N} \hat{v}_i^2 - \Big[N^{-1} \sum_{i=1}^{N} \hat{v}_i\Big]^2,$$

with

$$\hat{v}_i = \frac{1}{\hat{\mu}} \big\{ y_i \, \mathrm{I}(\hat{q_{\mathrm{lo}}} < y_i < \hat{q_{\mathrm{hi}}}) - \hat{q_{\mathrm{hi}}} \mathrm{I}(y_i < \hat{q_{\mathrm{hi}}}) + \hat{q_{\mathrm{lo}}} \mathrm{I}(y_i < \hat{q_{\mathrm{lo}}}) \big\} - \frac{y_i \hat{\mu}_{Ih}}{\hat{\mu}^2}.$$

2.4. Accuracy Measured by Simulation

Since everything in this section is asymptotic, it may be helpful to look briefly at evidence for finite-sample behaviour as revealed by simulation. For the case in which middle class incomes are defined as lying between specified multiples of the median income, random samples of different numbers of observations were drawn from the lognormal distribution, with an underlying standard normal distribution. The proportions a and b were set equal to 0.5 and 1.5, respectively. The values of the mean, median, and the population and income shares can be computed analytically, and are:

$$m = 1, \quad \mu = 1.648721, \quad PS = 0.413324, \quad IS = 0.230863.$$

For each of 9999 samples, and for each sample size, $n = 1,012,015,011,001$, the estimates of these four quantities were obtained. The variances of the estimates of the shares, and their covariance, were estimated by the sample variances and covariance from the 9999 samples. These were compared with the estimates of the asymptotic variances and covariances, averaged over the samples. For the purposes of the comparison, the variances were multiplied by the sample size. Results are in Table 1.

With the middle class defined using the mean income, the proportions a and b were set to 0.4 and 1.6. The mean and median are as before, and the exact shares are

$$PS = 0.495379 \quad \text{and} \quad IS = 0.409690.$$

The results are in Table 2.

Finally, using quantiles, the results in Table 3 are for the middle class contained between the 0.2 quantile and the 0.8 quantile. (Recall that the population share is by definition always $0.8 - 0.2 = 0.6$.)

The variances and covariance estimates derived in this section are clearly asymptotically correct, but are naturally not exact for finite n.

Table 1. Comparison of finite-sample and asymptotic variance: median definition.

	n	$\mathbf{var}(\widehat{PS})$	$\mathbf{var}(\widehat{IS})$	$\mathbf{cov}(\widehat{PS},\widehat{IS})$
Sample variances	101	0.239325	0.224096	0.176514
Averaged estimates	101	0.261119	0.218908	0.202878
Sample variances	201	0.244931	0.222913	0.180768
Averaged estimates	201	0.249148	0.207283	0.189229
Sample variances	501	0.245171	0.219862	0.180843
Averaged estimates	501	0.240752	0.200225	0.180011
Sample variances	1001	0.246202	0.218693	0.179762
Averaged estimates	1001	0.236738	0.197485	0.175393

Table 2. Comparison of finite-sample and asymptotic variance: mean definition.

	n	$\mathbf{var}(\widehat{PS})$	$\mathbf{var}(\widehat{IS})$	$\mathbf{cov}(\widehat{PS},\widehat{IS})$
Sample variances	101	0.289240	0.270821	0.251248
Averaged estimates	101	0.269630	0.262705	0.236283
Sample variances	201	0.295019	0.270204	0.254169
Averaged estimates	201	0.268601	0.259170	0.234529
Sample variances	501	0.290917	0.268718	0.237937
Averaged estimates	501	0.273562	0.259882	0.251659
Sample variances	1001	0.292915	0.268624	0.251931
Averaged estimates	1001	0.279508	0.262628	0.242509

Table 3. Comparison of finite-sample and asymptotic variance: quantile definition.

	n	$\mathbf{var}(\widehat{IS})$
Sample variances	101	0.137487
Averaged estimates	101	0.124903
Sample variances	201	0.145837
Averaged estimates	201	0.137819
Sample variances	501	0.147931
Averaged estimates	501	0.149558
Sample variances	1001	0.149601
Averaged estimates	1001	0.154112

3. Inference

The results of the previous section allow us to construct asymptotic confidence intervals for the population and income shares of the middle class, according to the different definitions considered. However, because we can also construct asymptotically pivotal functions, it is possible to construct bootstrap confidence intervals, and to perform bootstrap tests of specific hypotheses about these shares.

3.1. Data

The data used for the empirical analysis in this paper come from Canadian Census Public Use Microdata Files (PUMF) for Individuals for 1971, 1981, 1991, 2001, and 2006. Beach (2016) used these data, along with data from other sources, for his comprehensive account of the evolving fate of the Canadian middle class. In the Census files, the term earnings refers to annual earnings in the full year before the Census. Although the individuals of the samples provided for each of the census years are not identified by name, for obvious reasons, they are characterised by age (or age group), sex, and the number of weeks worked in the year. Income is split into wage income and income from self-employment. In the census data from 1991 onwards, individuals are assigned weights in order that the weighted sample should be more representative of the population than the unweighted one.

However, the weights vary little in the samples, and, indeed, they are all identical in the 2006 data. They are therefore not taken into account in the subsequent analysis.

It is of interest to compare formally the fates of men and women. Accordingly, for each census year, two samples are treated separately, one with data on men, the other on women, only. In both cases, individuals younger than 15 years of age are dropped from the sample, as well as individuals who did not work in that year, or for whom the information on weeks worked is missing. In addition, income from wages and salaries and income from self-employment are simply combined to yield the income variable.

3.2. Confidence Intervals

The confidence intervals given in this section are either asymptotic, using the estimates of asymptotic variances derived in the previous section, or bootstrap intervals, of the sort usually called percentile-*t*, or bootstrap-*t*; see for instance DiCiccio and Efron (1996), Davison and Hinkley (1997), and Hall (1992) for a discussion of the relative merits of different types of bootstrap confidence interval.

A bootstrap-*t* interval is constructed as follows using a resampling bootstrap. For a suitable number B of bootstrap repetitions, a bootstrap sample is created by resampling from the original sample. Let the parameter of interest be denoted by θ, its estimate from the original sample by $\hat{\theta}$, and its standard error by $\hat{\sigma}_\theta$. If the true, or population, value is θ_0, an asymptotically pivotal quantity is $\tau \equiv (\hat{\theta} - \theta_0)/\hat{\sigma}_\theta$. A bootstrap sample yields a parameter estimate θ^* and a standard error σ^*_θ. Then, the bootstrap counterpart of τ is $\tau^* \equiv (\theta^* - \hat{\theta})/\sigma^*_\theta$, since $\hat{\theta}$ is the "true" parameter value for the resampling bootstrap data-generating process (DGP).

If non-uniform weights are associated with the sample observations, then the reampling should also be non-uniform, whereby observation i is resampled with probability p_i/N, where p_i is the weight associated with the observation. This amounts to generating bootstrap samples from the weighted EDF (17). Then, each bootstrap sample is to treated as though it were a genuinely random sample, so that the weights do not appear in the estimation of the shares or in their standard errors. However, since, in some of the samples analysed here, there are no weights, and, even if they are present, they are very nearly, if not exactly, uniform, all of the empirical results are computed without use of weighting.

The distribution of τ^* is estimated by the empirical distribution of its B realisations. For an equal-tailed confidence interval of confidence level $1 - \alpha$, the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution are estimated by the order statistics $\alpha(B+1)/2$ and $(1 - \alpha/2)(B+1)$ of the realisations of τ^*. Let these estimated quantiles be $q^*_{\alpha/2}$ and $q^*_{1-\alpha/2}$. The bootstrap-*t* confidence interval is then

$$[\hat{\theta} - \hat{\sigma}_\theta q^*_{1-\alpha/2},\ \hat{\theta} - \hat{\sigma}_\theta q^*_{\alpha/2}].$$

This approach requires $\alpha(B+1)/2$ to be an integer; see, among many other references, Davidson and MacKinnon (2006).

Tables 4–8 present point estimates as well as asymptotic and bootstrap confidence intervals, at nominal confidence level of 95%, of the population and income shares, for the median-based definition of the middle class in 1971, 1981, 1991, 2001, and 2006.

Table 4. Estimates and confidence intervals: 1971.

		$\widehat{PS}$	$\widehat{IS}$
Male	point estimate	0.544	0.492
59,123 obs	asymptotic interval	[0.539, 0.549]	[0.488, 0.496]
median \$6000	bootstrap interval	[0.540, 0.554]	[0.487, 0.497]
Female	point estimate	0.399	0.362
32,164 obs	asymptotic interval	[0.392, 0.407]	[0.355, 0.369]
median \$2900	bootstrap interval	[0.392, 0.410]	[0.353, 0.377]

Table 5. Estimates and confidence intervals: 1981.

		$\widehat{PS}$	$\widehat{IS}$
Male	point estimate	0.519	0.481
143,248 obs	asymptotic interval	[0.515, 0.522]	[0.478, 0.484]
median $15,715	bootstrap interval	[0.515, 0.522]	[0.477, 0.485]
Female	point estimate	0.390	0.335
101,619 obs	asymptotic interval	[0.386, 0.394]	[0.331, 0.339]
median $7800	bootstrap interval	[0.387, 0.393]	[0.331, 0.339]

Table 6. Estimates and confidence intervals: 1991.

		$\widehat{PS}$	$\widehat{IS}$
Male	point estimate	0.483	0.436
234,636 obs	asymptotic interval	[0.481, 0.486]	[0.434, 0.438]
median $27,000	bootstrap interval	[0.481, 0.486]	[0.434, 0.439]
Female	point estimate	0.390	0.318
196,143 obs	asymptotic interval	[0.386, 0.392]	[0.316, 0.321]
median $15,139	bootstrap interval	[0.385, 0.391]	[0.314, 0.321]

Table 7. Estimates and confidence intervals: 2001.

		$\widehat{PS}$	$\widehat{IS}$
Male	point estimate	0.437	0.364
227,828 obs	asymptotic interval	[0.435, 0.440]	[0.363, 0.366]
median $31,700	bootstrap interval	[0.429, 0.440]	[0.354, 0.368]
Female	point estimate	0.414	0.333
20,2491 obs	asymptotic interval	[0.411, 0.416]	[0.330, 0.335]
median $20,000	bootstrap interval	[0.411, 0.416]	[0.330, 0.335]

Table 8. Estimates and confidence intervals: 2006.

		$\widehat{PS}$	$\widehat{IS}$
Male	point estimate	0.418	0.302
238,356 obs	asymptotic interval	[0.416, 0.420]	[0.300, 0.304]
median $35,000	bootstrap interval	[0.400, 0.420]	[0.282, 0.305]
Female	point estimate	0.415	0.320
202,491 obs	asymptotic interval	[0.413, 0.417]	[0.318, 0.322]
median $24,000	bootstrap interval	[0.413, 0.445]	[0.318, 0.355]

Remark 2. *In many cases, the asymptotic and bootstrap intervals very nearly coincide. The bootstrap intervals are a bit wider for 1971. For 2001 and 2006, however, the bootstrap population-share and income-share intervals for males extend far to the left of the asymptotic ones. For females, the pattern is different. In 2001, the asymptotic and bootstrap intervals are very close, but, in 2006, the bootstrap intervals extend far to the right of the asymptotic ones.*

The reason for these phenomena with the 2001 and 2006 data emerges from looking at the distributions of the bootstrap statistics, of which kernel density plots in 2006 for males and for females are shown in Figures 1 and 2 respectively.

One might expect the plots to resemble roughly a plot of the standard normal density. This would be the case if the long right-hand tail for men, and the long left-hand tail for women, each with a second mode, are neglected. It is well known that the resampling bootstrap can give highly misleading results with heavy-tailed data; see for instance Davidson (2012).

By looking at kernel density plots in Figure 3 of the sample income distributions for men and women in 2006, one can see evidence of the heavy right-hand tails for both sexes.

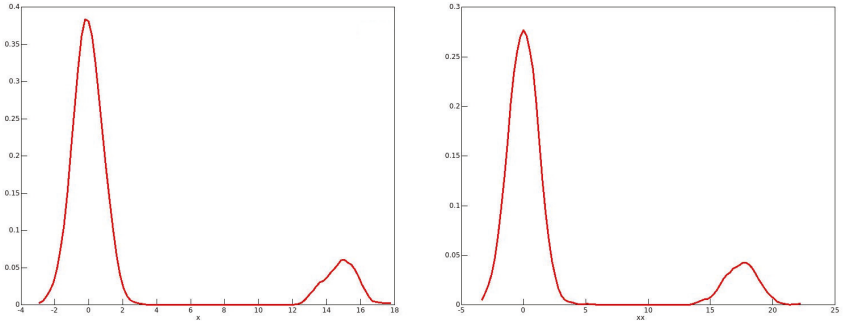


Figure 1. Kernel density plots of bootstrap statistics: 2006 males.

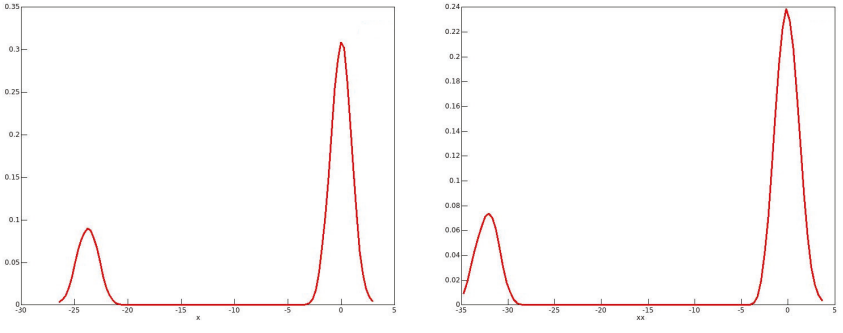


Figure 2. Kernel density plots of bootstrap statistics: 2006 females.

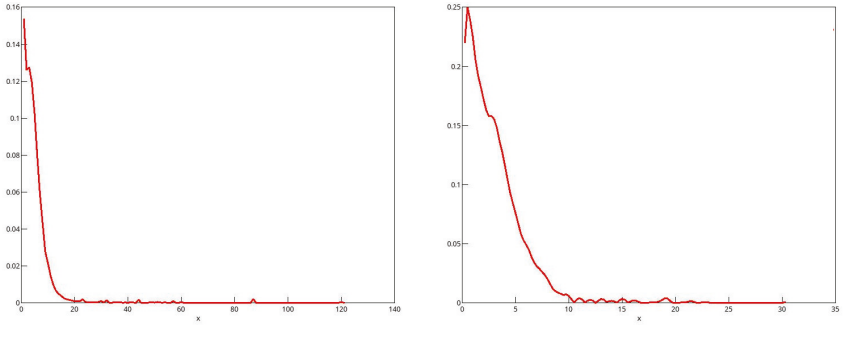


Figure 3. Kernel density plots of income distributions in 2006.

In addition, for all of the twenty-first century data, there is clear evidence of top-coding, since, in all cases, there are several observations equal to the largest income in the sample, while the next highest income is much lower. For instance, in the 2006 male sample, out of the 238,356 observations,

there are 121 equal to the highest income of $1,202,480, while the next highest income in the sample is $872,522.

However, there is no reason to think that top-coding would have any effect on the estimated population shares, since their exact values do not matter. They do, of course, for the income shares, and so these are overestimated with top-coding. It turns out that the reason for the bimodal distributions of the bootstrap statistics is quite unrelated to top-coding. A closer look at the data for 2006 shows that a phenomenon that we may call "heaping" occurs in the data. What this means is that, for each recorded income, there are multiple instances, with comparatively large gaps between the distinct recorded incomes. While there is some measure of a similar heaping in the twentieth-century data, the phenomenon is much less marked. As an example, there is only one observation in the 1971 male data equal to the maximum value.

The consequences of this heaping are most salient with the 2006 data. For men, the median income is $35,000, and there are no fewer than 3228 observations of incomes apparently exactly equal to $35,000. The upper and lower limits for middle-class incomes that have been used in this study are $52,500 and $17,500, respectively. There are no observations of incomes equal to either of these limits, and this follows inevitably from the fact that *all* incomes no greater than $200,000 are recorded as exact integer multiples of $1000.

The data for women present a different picture, because the limits of $12,000 and $36,000 are integer multiples of $1000, and all incomes no greater than $100,000 are recorded as integer multiples of $1000. The maximum income of $310,136 is assigned to 99 observations; the median of $24,000 to 3316 observations; the lower limit of $12,000 to 4282 observations; and the upper limit of $36,000 to 2694 observations. The second highest recorded income is $306,763.

What this has meant for the bootstrap is that, of the 999 bootstrap repetitions with the data for men, all but 146 had a median of $35,000, the others having a median of $36,000. For the latter, the limits for middle-class income were $18,000 and $54,000, and including the 2052 observations of $54,000 in the numbers of the middle class greatly increases the population and income shares in those bootstrap samples relative to the shares of the 853 samples with a median of $35,000. At the other end, increasing the limit from $17,500 to $18,000 made no difference to the numbers, since there are no observations recorded in the interior of the range of the increase.

A similar analysis can be conducted with the data for women, but the reason for the bimodal distributions of the bootstrap statistics is clear: it arises on account of the data heaping. With the 2001 data, a bimodal distribution might have been expected, but all but five out of 999 bootstrap samples had a median equal to that of the original data, and, as expected, the distribution of the bootstrap statistics is unimodal in that case.

The data for years before 2001 have a much lesser amount of heaping and have unimodal bootstrap distributions. This no doubt implies that the bootstrap results are credible, although this conclusion is not of much worth since the bootstrap and asymptotic confidence intervals are nearly coincident.

3.3. Smoothing

An obvious remedy for the heaping in the later datasets is to smooth them. The smoothed sample distribution may well be a better estimate of the population distribution than the heaped estimate, since the heaping is manifestly an artefact of the way in which the datasets were constructed. As always with smoothing, a troublesome question is the choice of bandwidth. Since the heaping occurs at integer multiples of $1000, the bandwidth h should be of a comparable magnitude in order to avoid an excessively discrete distribution. For $h = 1000$, the raw EDFs of the 2006 data for men and women are plotted in Figure 4 along with the smoothed EDFs, for the range of incomes from half the median to 1.5 times the median. The heaped nature of the data for both sexes is quite evident in the green, unsmoothed, plots.

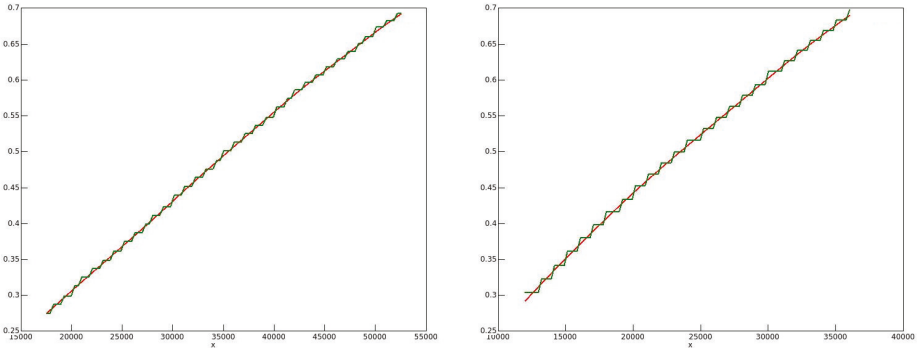
The (cumulative) kernel used for smoothing was the integrated Epanechnikov kernel. The smoothed estimate of the distribution is

$$F_{\mathrm{sm}}(y) = \frac{1}{N}\sum_{i=1}^{N} K\big(h^{-1}(y_i - y)\big), \tag{20}$$

where h is the bandwidth, and the cumulative kernel K is defined as

$$K(z) = \mathrm{I}\big(|z| \le \sqrt{5}\big)\Big(\frac{3}{4\sqrt{5}}(z - z^3/15) + \frac{1}{2}\Big) + \mathrm{I}\big(z > \sqrt{5}\big). \tag{21}$$

where h is the bandwidth. Other choices of h greater than around 500 give qualitatively similar results.



(**a**) Males (**b**) Females

Figure 4. Smoothed (red) and unsmoothed (green) EDFs for 2006 data.

For bootstrapping, resampling from the unsmoothed EDF is replaced by resampling from the smoothed EDF. Since the heaping phenomenon is banished by the smoothing, we can expect dramatically different results, in particular, a unimodal distribution of the bootstrap statistics. The CDF (20) describes a mixture distribution which assigns a weight of $1/N$ to the each of the distributions characterised by the terms in the sum. It is easily checked that K in (21) is a valid CDF, with support $[-\sqrt{5}, \sqrt{5}]$. The term indexed by i in (20) has support $[y_i - h\sqrt{5}, y_i + h\sqrt{5}]$.

To draw from the distribution (21), one starts from a uniform variate p from the U(0,1) distribution, and the draw is then $K^{-1}(p)$. The analytic form of K^{-1} is not, I think, well known, and so I give it here for reference. It is[2]

$$K^{-1}(p) = 2\sqrt{5}\cos\Big(\frac{1}{3}\big(2\pi - \cos^{-1}(1-2p)\big)\Big).$$

Thus, to draw from distribution (20), one may first draw the index i from the uniform distribution on $\{1, 2, \ldots, N\}$, then draw p from U(0,1), and get the draw

$$y^* = y_i + hK^{-1}(p).$$

The effect is to resample from the unsmoothed distribution and then add some smoothing "noise" from the Epanechnikov distribution.

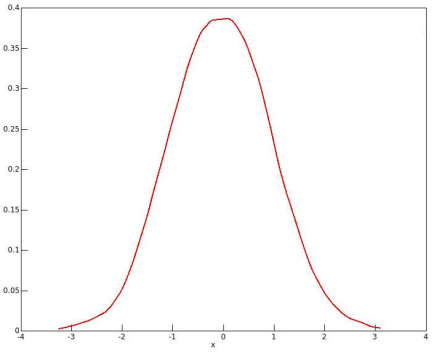
[2] It can be found, in a somewhat different version, in the documentation of the epandist package for R.

Although the smoothing preserves the mean of the distribution, it does not preserve the median, nor the population or income shares. If we accept the argument that the smoothed CDF is a better estimate of the true distribution than the unsmoothed one, then the smoothed median, and the shares in the smoothed distribution are also better estimators. In addition, the smoothed shares are the "true" values for the bootstrap DGP, and so the bootstrap statistics should test the hypothesis that they are true, not the hypothesis that the unsmoothed shares are true.

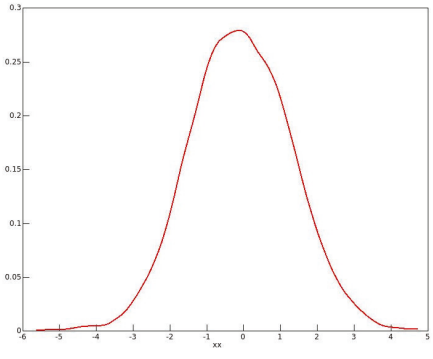
With the 2006 data for men, the new estimates of the shares are 0.421 for the population and 0.307 for income, slightly higher than the estimates from the raw data. The bootstrap confidence intervals are, for the population share, $[0.419, 0.423]$ and, for the income share, $[0.305, 0.310]$. They are of roughly the same width as the asymptotic intervals.

With the data for women, the new share estimates are 0.393 and 0.298, substantially lower than the unsmoothed estimates, and the confidence interval for the population share is $[0.390, 0.395]$, and, for the income share $[0.295, 0.301]$. Unsurprisingly, the smoothed share estimates are roughly in the middle of the respective intervals.

In Figures 5 (men) and 6 (women), kernel density plots are shown for the distribution of the bootstrap statistics. There is no trace of bimodality, and so it seems that smoothing has indeed corrected the heaping problem.

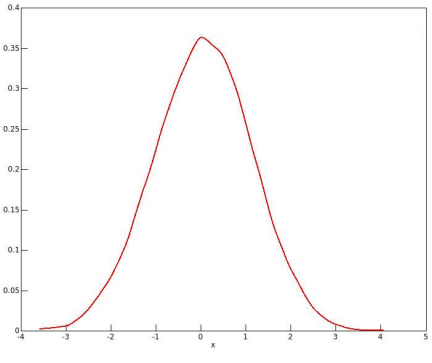


(**a**) Population share

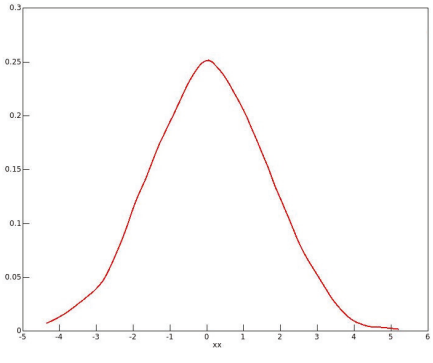


(**b**) Income share

Figure 5. Kernel density plots of smoothed bootstrap statistics: 2006 males.



(**a**) Population share



(**b**) Income share

Figure 6. Kernel density plots of smoothed bootstrap statistics: 2006 females.

3.4. Hypothesis Tests

In this section, the results of testing various hypotheses are found. All of the test statistics are asymptotic, as we have seen that when bootstrap inference differs greatly from asymptotic, the unsmoothed bootstrap, at least, is likely to be unreliable.

First are tests of hypotheses that the population and income shares for each sex did not change from one census until the next one. For instance, can one reject the hypothesis that the population share of the male middle class did not change from 1981 to 1991? Next are tests of hypotheses that the shares of men and women are the same in each census. For instance, can one reject the hypothesis that the income shares of men and women were the same in 2001?

The test results are expressed as asymptotic t statistics, rather than asymptotic p values, since in most cases the hypothesis is rejected strongly, and a p value very close to zero does not let one judge just how strong the rejection is. However, in some cases, the hypotheses are not rejected, and in some other cases, the sign of the statistic differs from the signs of the other statistics for the same sort of hypothesis.

For the first group of tests, the results of which are found in Table 9, the sign of the statistic is positive if the decline in a share from the earlier to the later census is positive. A negative statistic indicates that the estimated share rose between the two censuses.

Table 9. t statistics for hypothesis of no change in share between consecutive censuses.

Period	PS (Men)	PS (Women)	IS (Men)	IS (Women)
1971–1981	8.571726	2.299586	4.740735	6.571228
1981–1991	16.702812	0.311744	26.933620	6.875789
1991–2001	26.128047	−12.835861	53.350095	−7.954860
2001–2006	11.322294	−0.752581	43.943449	7.824492

Remark 3. *All but two hypotheses of no change between two censuses are strongly rejected. The two exceptions concern the female population share, which did not change significantly either between 1981 and 1991 or between 2001 and 2006. There are two significantly positive increases, for the female population and income shares from 1991 to 2001.*

In Table 10 are found the statistics for testing the hypothesis that the share of men and women is the same for a given census. A positive statistic means that the estimated male share is greater than the female.

Table 10. t statistics for hypothesis of equal shares for men and women.

Census	PS	IS
1971	32.526094	32.306558
1981	49.137099	60.112426
1991	50.265363	69.768414
2001	12.902812	20.345573
2006	7.824492	−12.143588

4. Conclusions

The main contribution of this paper is probably the theoretical part. The empirical results are not really surprising, although they do document clearly how the population and income shares of the male middle class have fallen over the period since 1970. In addition, one sees the results of the considerable increase in female labour market participation. Although the bootstrap has not shown itself especially useful for formal inference, the evolution over time of the distribution of the bootstrap statistics shows very clearly the increasing polarisation of Canadian society, with the growth of a heavy right-hand tail in the income distributions of both men and women.

The main obstacle to inference, whether asymptotic or bootstrap, with the twenty-first century data has been seen to be the problem of heaping, or excessively rounding, the data. The smoothing technique proposed here appears to lead to more reliable inference, but truly reliable inference would need better data.

Acknowledgments: This research was supported by the Canada Research Chair program (Chair in Economics, McGill University), and by grants from the Fonds de Recherche du Québec: Société et Culture. I am grateful for discussions with Charles Beach, and I thank his research assistant, Aidan Worswick, for providing me with data in a manageable form. The paper has benefited from comments from participants at the Econometric Study Group (Bristol 2017) and the second Lebanese Econometric Study Group (Beirut 2017).

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CDF cumulative distribution function
EDF empirical distribution function
DGP Data-generating process

References

Atkinson, Anthony B. 1970. On the measurement of inequality. *Journal of Economic Theory* 2: 244–63.

Bahadur, R. Raj. 1966. A Note on Quantiles in Large Samples. *Annals of Mathematical Statistics* 37: 577–80.

Beach, Charles M. 2016. Changing income inequality: A distributional paradigm for Canada. *Canadian Journal of Economics* 49: 1229–92.

Blackorby, Charles, Walter Bossert, and David Donaldson. 1999. Income Inequality Measurement: The Normative Approach. In *Handbook of Income Inequality Measurement*. Edited by Jacques Silber. New York: Springer, pp. 133–62.

Brzozowski, Matthew, Martin Gervais, Paul Klein, and Michio Suzuki. 2010. Consumption, income, and wealth inequality in Canada. *Review of Economic Dynamics* 13: 52–75.

Castells-Quintana, David, Raul Ramos, and Vicente Royuela. 2015. Income inequality in European Regions: Recent trends and determinants. *Review of Regional Research* 35: 123–46.

Cowell, Frank A. 1999. Estimation of Inequality Indices. In *Handbook of Income Inequality Measurement*. Editor by Jacques Silber. New York: Springer, pp. 269–90.

Davidson, Russell. 2012. Statistical Inference in the Presence of Heavy Tails. *Econometrics Journal* 15: C31–C53.

Davidson, Russell, and Jean-Yves Duclos. 1997. Statistical Inference for Measurement of the Incidence of Taxes and Transfers. *Econometrica* 65: 1453–65.

Davidson, Russell, and Jean-Yves Duclos. 2000. Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality. *Econometrica* 68: 1435–64.

Davidson, Russell, and James G. MacKinnon. 2006. Bootstrap Methods in Econometrics. In *Palgrave Handbook of Econometrics*. Edited by Terence C. Mills and Kerry Patterson. London: Palgrave-Macmillan, vol. 1. Econometric Theory.

Davison, Anthony Christopher, and David Victor Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

DiCiccio, Thomas J., and Bradley Efron. 1996. Bootstrap confidence intervals (with discussion). *Statistical Science* 11: 189–228.

Foster, James E., and Michael C. Wolfson. 2010. Polarization and the decline of the middle class: Canada and the U.S. *Journal of Economic Inequality*. 8: 247–73. Reprint of 1992 original article.

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Heathcote, Jonathan, Fabrizio Perri, and Giovanni L. Violante. 2010. Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006. *Review of Economic Dynamics* 13: 15–51.

Kuznets, Simon. 1955. Economic Growth and Income Inequality. *American Economic Review* 45: 1–28.

Ryu, Hang Keun. 2013. A bottom poor sensitive Gini coefficient and maximum entropy estimation of income distributions. *Economics Letters* 118: 370–74.

Wolff, Edward N. 2013. The Asset Price Meltdown, Rising Leverage, and the Wealth of the Middle Class. *Journal of Economic Issues* 47: 333–42.
Wolfson, Michael C. 1994. When Inequalities Diverge. *American Economic Review* 84: 353–58.

© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Parametric Inference for Index Functionals

Stéphane Guerrier [1], Samuel Orso [2] and Maria-Pia Victoria-Feser [2,*]

[1] Department of Statistics & Institute for CyberScience, Eberly College of Science, Pennsylvania State University, University Park, 16802 PA, USA; stephane@psu.edu

[2] Research Center for Statistics, Geneva School of Economics and Management, University of Geneva, 1202 Geneva, Switzerland; Samuel.Orso@unige.ch

* Correspondence: maria-pia.victoriafeser@unige.ch

Received: 13 December 2017; Accepted: 13 April 2018; Published: 20 April 2018

Abstract: In this paper, we study the finite sample accuracy of confidence intervals for index functional built via parametric bootstrap, in the case of inequality indices. To estimate the parameters of the assumed parametric data generating distribution, we propose a Generalized Method of Moment estimator that targets the quantity of interest, namely the considered inequality index. Its primary advantage is that the scale parameter does not need to be estimated to perform parametric bootstrap, since inequality measures are scale invariant. The very good finite sample coverages that are found in a simulation study suggest that this feature provides an advantage over the parametric bootstrap using the maximum likelihood estimator. We also find that overall, a parametric bootstrap provides more accurate inference than its non or semi-parametric counterparts, especially for heavy tailed income distributions.

Keywords: parametric bootstrap; generalized method of moments; income distribution; inequality measurement; heavy tail

JEL Classification: C10; C13; C15; C43; C46; D31

1. Introduction

In this paper, we consider the problem of inference for an index functional T, i.e., quantities of interest that can be written as a function of the data generating model. Given a sample $x_i, i = 1, \ldots, n$ and an associated distribution F such that one can assume that $X_i \sim F$, $i = 1, \ldots, n$, we are interested in computing confidence intervals or proceeding with hypothesis testing for $T(F)$. For that, there exists many different approaches that are based on either $T(F^{(n)})$ or $T(F_\theta)$, where $F^{(n)}$ is the empirical distribution (hence leading to a nonparametric approach) and F_θ, $\theta \in \Theta \subset \mathbb{R}^p$ is a parametric model for which θ needs to be estimated from the sample (hence leading to a parametric approach).

As a leading example, we consider T to be an inequality index and F an income distribution. Inequality indices are welfare indices which can be very generally written in the following quasi-additively decomposable form (see Cowell and Victoria-Feser (2002, 2003) for the original formal setting)

$$W_{\mathrm{QAD}}(F) = \int \varphi\left(x, \mu(F)\right) dF(x), \tag{1}$$

where φ is piecewise differentiable in $\mathbb{R}$. The generalized entropy family of inequality indices given by

$$I^{\xi}_{\mathrm{GE}}(F) = \frac{1}{\xi^2 - \xi}\left[\int \left[\frac{x}{\mu(F)}\right]^{\xi} dF(x) - 1\right], \tag{2}$$

is obviously obtained by setting

$$\varphi\left(x,\mu(F)\right)=\frac{1}{\xi^2-\xi}\left[\left[\frac{x}{\mu(F)}\right]^{\xi}-1\right]. \tag{3}$$

For example, the cases $\xi=0$ and $\xi=1$ are given by

$$\begin{aligned} I^0_{\mathrm{GE}}(F) &= -\int \log\left(\frac{x}{\mu(F)}\right)dF(x), \\ I^1_{\mathrm{GE}}(F) &= \int \frac{x}{\mu(F)}\log\left(\frac{x}{\mu(F)}\right)dF(x), \end{aligned} \tag{4}$$

with $I^0_{\mathrm{GE}}(F)$ being the Mean Logarithmic Deviation (see Cowell and Flachaire 2015) and $I^1_{\mathrm{GE}}(F)$ being the Theil index. A notable exception to the class in (1) is the Gini coefficient which can be expressed in several forms, such as

$$I_{\mathrm{Gini}}(F)=1-2\int_0^1 \frac{C(F;q)}{\mu(F)}dq, \tag{5}$$

with $C(F;q)=\int^{F^{-1}(q)} xdF(x)$, the cumulative income functional. Inference on $T(F)$ can be done in several manners:

1. The (nonparametric) bootstrap is a distribution-free approach that allows to derive the sample distribution of $T(F^{(n)})$ from which quantiles (for confidence intervals) and variance (for testing) can be estimated; for application to inequality indices, see e.g., Mills and Zandvakili (1997) and Biewen (2002).
2. Another distribution-free approach consists in deriving the asymptotic variance of the index using the Influence Function (IF) of Hampel (1974) (see also Hampel et al. 1986) as is done in Cowell and Victoria-Feser (2003) (for different types of data features such as censoring and truncating) and estimate it directly from the sample (see also Victoria-Feser 1999; Cowell and Flachaire 2015).
3. A parametric (and asymptotic) approach, given a chosen parametric model F_θ for the data generating model, consists in first consistently estimating θ, say $\hat{\theta}$, then considering its asymptotic properties such as its variance $\mathrm{var}(\hat{\theta})$ and derive the corresponding asymptotic variance of $T(F_{\hat{\theta}})$ using e.g., the delta method (based on a first order Taylor series expansion).
4. A parametric (finite sample) approach, given a chosen parametric model F_θ for the data generating model, consists in first consistently estimating θ, say $\hat{\theta}$, then using parametric bootstrap to derive the sample distribution of $T(F_{\hat{\theta}})$ from which quantiles (for confidence intervals) and variance (for testing) can be estimated.
5. Refinements and combinations of these approaches.

While most would agree that the fully parametric and asymptotic approach based on the delta method cannot provide as accurate inference as the other methods, it is not clear that avoiding the specification of a parametric model is the way to go. Indeed, for example, Cowell and Flachaire (2015) notice that nonparametric bootstrap inference on inequality indices is sensitive to the exact nature of the upper tail of the income distribution, in that bootstrap inference is expected to perform reasonably well in moderate and large samples, unless the tails are quite heavy. Similar conclusions are also drawn in Davidson and Flachaire (2007); Cowell and Flachaire (2007); Davidson (2009); Davidson (2010) and Davidson (2012). This has for example motivated Schluter and van Garderen (2009) and Schluter (2012), using the results of Hall (1992), to propose normalizing transformations of inequality measures using Edgeworth expansions, to adjust asymptotic Gaussian approximations.

Alternatively, Davidson and Flachaire (2007) and Cowell and Flachaire (2007) consider a semi-parametric bootstrap, where bootstrap samples are generated from a distribution which

combines a parametric estimate of the upper tail, namely the Pareto distribution, with a nonparametric estimate the other part of the distribution. We note that modelling the upper tail with a parametric model is common in instances were not only the interest lies in the upper tail itself but also where the data are sparse. For example, in finance, determination of the value at risk or expected shortfall is central to portfolio management, and in insurance, it is important to estimate probabilities associated with given levels of losses. A critical challenge is then to select the threshold from which the upper tail is modelled parametrically (see for example Danielsson et al. 2001; Guillou and Hall 2001; Beirlant et al. 2002; Dupuis and Victoria-Feser 2006 and the references therein).

Cowell and Flachaire (2015) propose to use a another type of semi-parametric approach by which a mixture of lognormal distributions is first considered and then data are generated from the estimated mixture. A mixture of lognormal distributions to model the data can be thought of as a compromise between fully parametric and nonparametric estimation. The use of mixtures for income distribution estimation can be found for example in Flachaire and Nuñez (2007) and the references in Cowell and Flachaire (2015).

Through a simulation study, Cowell and Flachaire (2015), Table 7, compare the actual coverage probabilities of 95% confidence intervals for the Theil index, using, as data generating models, the lognormal distribution and the Singh-Maddala (SM) distribution (Singh and Maddala 1976), with varying parameters to increase the heaviness of the tail. The different methods cited above are compared. Cowell and Flachaire (2015) conclude that, in the presence of very heavy-tailed distributions, even if significant improvements can be obtained on the fully asymptotic and the standard bootstrap methods, none of the alternative methods provides very good results overall.

Moreover, Cowell and Flachaire (2015) do not consider a parametric bootstrap and this has motivated the present paper. Namely, we study the behaviour of coverage probabilities associate to the index functional $T(F)$ using a parametric bootstrap based on samples generated from $F_{\hat{\theta}}$ (i.e., Approach 4). A parametric model introduces a form of smoothness into the inferential procedure which can lead to more accurate inference. This is for example a fundamental argument for modelling the upper tail with a Pareto distribution. Specifying a parametric distribution for the data generating process can be considered as an additional risk of introducing "error" in the inferential procedure. With income distributions, common wisdom however suggests that some parametric models are sufficiently flexible to encompass most of the data generating processes observed with real data. For example, the four parameters generalized beta distribution of second kind (GB2) proposed by (McDonald 1984), which encompasses the generalized gamma, the Singh-Maddala and Dagum distribution (Dagum 1977) (see also McDonald and Xu 1995), can be considered as sufficiently general to model income data. If this is not the case, then one would wonder if the lack of flexibility of a general four parameter model is not due to a spurious amount of observations, and hence consider a robust estimation approach as proposed and motivated in Cowell and Victoria-Feser (1996), see also (Cowell and Victoria-Feser 2000).

In this paper, as an alternative to the classical Maximum Likelihood Estimator (MLE), we propose a *Target Matching Estimator* (TME), a member of the class of Generalized Method of Moments (GMM) estimators (Hansen 1982), where one of the "moments" is the targeted inequality index T. It has the advantage that for inference on T, the scale parameter does not need to be estimated (and hence can be set to an arbitrary value), so that the estimation exercise is simpler in that the optimization is performed in a smaller dimension. We derive its asymptotic properties and compare them to the MLE when targeting $T(F_\theta)$. As illustrated in a simulation study, it turns out that the finite sample coverage probabilities obtained from a parametric bootstrap based on this alternative estimator are far more accurate than the ones computed with other methods, especially with heavy tailed income distributions.

2. A Target Matching Estimator

Recall that we are interested in making inference on an inequality index T and we assume that the sample data are generated from a (sufficiently general) parametric mode F_θ, $\theta \in \Theta \subset \mathbb{R}^p$. We let $\nu = (T, S_1, \ldots, S_{q-1})'$ be a vector of statistics of length q, where the first element is the statistic of interest and the remaining $q-1$ elements are additional statistics. We denote by $\hat{\nu}$ the sample vector of statistics and by $\nu_n(\theta)$ its expectation at the model F_θ, for a fixed sample size n. Assuming that the mapping $\theta \mapsto \nu_n(\theta)$ is bijective, a GMM estimator can be defined as

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\| \hat{\nu} - \nu_n(\theta) \right\|_\Omega^2 \tag{6}$$

where Ω is positive definite $q \times q$ matrix of weights, possibly estimated from the sample (in that case one assumes that it converges to a non-stochastic quantity), used to adjust the statistical efficiency of $\hat{\theta}$. If $\nu_n(\theta)$ cannot be obtained in an analytically tractable form, one can use instead $\nu(\theta) = \lim_{n\to\infty} \nu_n(\theta)$, or alternatively, use Monte Carlo simulations to approximate $\nu_n(\theta)$, leading to a Simulated Method of Moments (SMM) estimator (McFadden 1989) given by

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \left\| \hat{\nu} - \bar{\nu}_n(\theta) \right\|_\Omega^2 \tag{7}$$

where $\bar{\nu}_n(\theta) = \frac{1}{B}\sum_{b=1}^{B} \hat{\nu}_b$ and $\hat{\nu}_b = \hat{\nu}_b(F_\theta)$ is the b-th vector of statistics obtained on pseudo-values simulated from F_θ. If the number of simulation B is infinite, then the estimators in (6) and (7) are equivalent, otherwise the latter is (asymptotically) less efficient.

It is computationally advantageous to have an analytic expression for $\nu(\theta)$ and thus prefer this approximation over $\bar{\nu}_n(\theta)$. However, in finite samples, the bias on $\hat{\theta}$ using $\nu(\theta)$ may be more important than the one resulting from using $\bar{\nu}_n(\theta)$ (see Guerrier et al. 2018). An other approach, considered for example by Arvanitis and Demos (2015), is to directly approximate $\nu_n(\theta)$ with expansions on analytical functions.

Given that the interest here is to make inference about a functional T, one also needs to consider a suitable choice for the (additional) statistics in ν. Obviously one needs to choose a number of statistics at least as large as the number of parameter in the assumed model, i.e., $q \geq p$. If these statistics are sufficient, then $q = p$. Moreover, T may depend only on $q_s < p$ of the elements of θ, and for this purpose, the whole estimation of θ maybe an unnecessary burden. Let $\theta = (\theta^{s\prime}, \theta^{c\prime})'$ where θ^s, of dimension $q_s \geq 1$ is the vector of parameters that (uniquely) determines T whereas θ^c, of dimension q_c, is the vector of "nuisance parameters" that do not influence T. Then, instead of solving (6) or (7), we propose to consider a *Target Matching Estimator* (TME) defined as

$$\hat{\theta}^s = \underset{\theta^s \in \Theta^s \subset \mathbb{R}^{q_s}}{\operatorname{argmin}} \left\| \hat{\nu}^s - \nu(\theta^s) \right\|_\Sigma^2 \quad . \tag{8}$$

It is known that in an homogeneous system the asymptotic covariance of $\hat{\theta}^s$ is not influenced by the weighting matrix Σ (supposedly independent from θ) as long as Σ is a positive-definite matrix. Since we consider the case when the dimension of the statistics and the parameters of interest are the same, i.e., $\dim(\nu) = \dim(\theta^s) = q^s$, taking the identity matrix for Σ, and assuming that the minimum of the quadratic function is attained in the interior of the parameter space Θ^s, we then have that (8) can be equivalently written as

$$\hat{\theta}^s = \underset{\theta^s \in \Theta^s \subset \mathbb{R}^{q_s}}{\operatorname{argzero}} \left[\hat{\nu}^s - \nu(\theta^s) \right].$$

The generalized entropy family of measures and the Gini index are scale invariant whereas the models F_θ usually suggested in the literature (Kleiber and Kotz 2003) are parametrised with a scale component. Indeed, let δ, an element of θ, denote the scale parameter, then with the linear property of the expectation, $I_{\mathrm{GE}}^{\xi}(F)$ in (2) is invariant to any transformation δX. The same statement is

true for the Gini coefficient. This is not surprising as scale-invariance is indeed one of the required property of inequality indices. We hence have $(\partial/\partial\delta)T(F_\theta) = 0$, so that θ^s is θ without the scale parameter δ. Note that $(\partial/\partial\delta)T(F_\theta) = 0$ may be useful in situations where the analytical form of $T(F_\theta)$ is not available.

More generally, suppose we are in the situation where T is such that $(\partial/\partial\theta^c)T(F_\theta) = 0$ and $(\partial/\partial\theta^s)T(F_\theta) \neq 0$. Also suppose that the statistics $S_1, \ldots, S_{q-1}$ are chosen such that $(\partial/\partial\theta^c)S_j(F_\theta) = 0$ and $(\partial/\partial\theta^s)S_j(F_\theta) \neq 0$, $j = 1, \ldots, q-1$, $q = p$, then (8) provides a suitable estimator for inference on T. For scale invariant inequality measures T, any statistics of the form

$$S_k(x) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i}{\hat{\mu}}\right)^k, \quad k \in \mathbb{R}, \quad \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{9}$$

is also scale-invariant. This is also true with a logarithmic transformation as

$$U_l(x) = \frac{1}{n}\sum_{i=1}^{n}(\log(x_i) - \hat{m})^l, \quad l \in \mathbb{R}, \quad \hat{m} = \frac{1}{n}\sum_{i=1}^{n}\log(x_i). \tag{10}$$

Finally, for the choice of F_θ, one can consider the GB2 (see Section 4) which is sufficiently general to encompass real data situations with income data (Bandourian et al. 2002). Alternatively, as suggested for example in Cowell and Flachaire (2015), one can also consider the SM distribution.

In the simulation Section 4 we propose suitable statistics ν that are used in (8). Given these statistics ν and an assumed data generating model F_θ, inference about T, using the parametric bootstrap, is obtained using Algorithm 1.

Algorithm 1: TME-percentile confidence interval

Input : A given function ν^s; its sample version $\hat{\nu}^s$; number of iteration B; a confidence level $1-\alpha$.

Output: An interval: $[H_T^{(n)}(\alpha/2),\ H_T^{(n)}(1-\alpha/2)]$, where $H_T^{(n)}(\alpha) = \inf\{t : F_T^{(n)}(t) \geq \alpha\}$, $F_T^{(n)}$ is the empirical distribution function of T, with realizations $T_1, \ldots, T_B$.

Compute $\hat{\theta}^s = \text{argmin}_{\theta^s} \|\hat{\nu}^s - \nu(\theta^s)\|^2$.
Fix θ^c to an arbitrary value in Θ^c.
for $b \leftarrow 1$ **to** B **do**
  Draw a sample $X^{(b)} \sim F_{\theta=(\hat{\theta}^s,\theta^c)}$.
  Compute T_b on $X^{(b)}$.
end
Compute the percentiles $H_T^{(n)}(\alpha/2),\ H_T^{(n)}(1-\alpha/2)$ on the values $T_1, \ldots, T_B$.

Note that if $\tilde{\nu}(\theta^s)$ is used instead of $\nu(\theta^s)$ in (8), the last step of the optimization leading to $\hat{\theta}^s$ readily delivers $(T_1, \ldots, T_B)$.

3. Asymptotic Properties

We now look at the asymptotic distribution of the TME in (8). Since θ^c is fixed but θ^s is estimated by matching some statistics ν, a crucial question is on whether $\hat{\theta}^s$ is more efficient than say $\hat{\theta}^s_{\text{MLE}}$, the estimator that we would have obtained by the MLE on the whole vector θ. In order to answer this question consider a setting in which the regular conditions for the MLE $\hat{\theta}_{\text{MLE}}$ to be square root-n consistent are met. In this case, we let $\mathcal{I}$ denotes the Fisher information matrix evaluated at the point $\theta_0 \in \Theta$, we have

$$n^{1/2}\left(\hat{\theta}_{\text{MLE}} - \theta_0\right) \rightsquigarrow \mathcal{N}\left(0, \mathcal{I}^{-1}\right).$$

This setting is clearly not the weakest possible in theory for our analysis and may be further relaxed. We do not attempt to pursue the weakest possible conditions to avoid overly technical treatments in establishing the theoretical result given in this section.

Theorem 1. *Let $\Theta^s \subset \mathbb{R}^{q_s}$ be compact. Suppose that the point θ_0^s is in the interior of Θ^s. Suppose that $\nu(\theta_0^s)$ is the expectation of $\hat{\nu}^s$ when n is large. If $n^{1/2}\left(\hat{\nu}^s - \nu(\theta_0^s)\right)$ satisfies a central limit theorem with covariance matrix Ξ, the mapping $\theta \mapsto \nu$ is bijective, continuously once differentiable in an open neighborhood of the point $\theta_0^s \in \Theta^s$ and the derivative $\dot{\nu}$ is nonsingular at the point θ_0^s, then*

$$n^{1/2}\dot{\nu}(\theta_0^s)\left(\hat{\theta}^s - \theta_0^s\right) \rightsquigarrow \mathcal{N}\left(0, \Xi\right).$$

The proof is provided in the Appendix A.

Compared to the MLE, the additional condition that the statistics $\hat{\nu}^s$ satisfy a central limit theorem is mild and generally met in practice for sample moments and the inequality indices considered here. The results on the delta method and the continuous mapping theorem of Phillips (2012) may be employed to refine Theorem 1 to the case where the known function ν is replaced by the function evaluated by simulation $\bar{v}_n$.

The asymptotic covariance matrix of $\hat{\theta}^s$, given in Theorem 1 by $[\dot{\nu}(\theta_0^s)']^{-1}\Xi\dot{\nu}(\theta_0^s)^{-1}$, is proportional to the inverse of the derivative of the expectation of the statistics with respect to θ and the asymptotic covariance matrix of the statistics. The choice of statistics should then be guided by their sensitivity to θ and their variability at the model. The same argument is found in Heggland and Frigessi (2004).

If the statistics ν are sufficient, then the asymptotic covariance matrix of $\hat{\theta}^s$ is equivalent to the asymptotic covariance matrix of the MLE conditionally on $\hat{\theta}^c_{\text{MLE}}$ fixed. From the properties of the normal distribution, we have asymptotically that

$$n^{1/2}\left(\hat{\theta}^s_{\text{MLE}} - \theta_0^s\right) \Big| \left(\hat{\theta}^c_{\text{MLE}} = \theta_0^c\right) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, V_{ss}\right),$$

where $V_{ss} = \mathcal{I}_{ss}^{-1} - [\mathcal{I}^{-1}]_{sc}\mathcal{I}_{cc}[\mathcal{I}^{-1}]_{cs}$, $\mathcal{I}_{ss}$ denotes the partition of $\mathcal{I}$ corresponding to θ^s, $\mathcal{I}_{cc}$ for θ^c and $[\mathcal{I}^{-1}]_{sc}$ for the covariances between $\hat{\theta}^s_{\text{MLE}}$ and $\hat{\theta}^c_{\text{MLE}}$. Thus, the estimator $\hat{\theta}^s$ obtained from (8) has a smaller variance than the unconditional MLE by a factor $[\mathcal{I}^{-1}]_{sc}\mathcal{I}_{cc}[\mathcal{I}^{-1}]_{cs} \geq 0$. In particular, this gain could by substantial if $\hat{\theta}^c$ has a large variance. On the other hand, the gain would be null if $\hat{\theta}^s$ and $\hat{\theta}^c$ are independent as their covariances $[\mathcal{I}^{-1}]_{sc} = [\mathcal{I}^{-1}]'_{cs} = 0$.

Choosing "good" statistics $\hat{\nu}^s$ remains a difficult task: sufficient statistics with appropriate data reduction and with the property of being independent (asymptotically) from θ^c may be hard to find. Heggland and Frigessi (2004) suggest a graphical procedure based on simulation to find statistics "sensitive enough" to the parameter of interest. In a similar context, Gallant and Tauchen (1996) propose to use the likelihood score function of a model "close" to the one of interest as statistics. In the present context, it could be a probability model parametrised by θ^s only. There are however no guarantee that such a model exists, and if it does, it might be not unique.

4. Simulation Study

We consider here two parametric distributions, namely the four parameters GB2 and the three parameters SM distributions. We compare the coverage probabilities provided by the parametric bootstrap using on the one hand the MLE and on the other hand the TME approach presented in Section 2 (using Algorithm 1) to the nonparametric bootstrap for the GB2. We also compare the coverage probabilities assuming a SM data generating process, to a variance stabilizing transform of the index proposed by Schluter (2012) (Varstab), the semi-parametric approach of Davidson and Flachaire (2007) and Cowell and Flachaire (2007) (Semip) and when mixtures of lognormal distributions are used to fit the density as proposed in Cowell and Flachaire (2015).

The GB2 has density function

$$f_\theta(x) = \frac{a x^{ap-1}}{b^{ap}\mathcal{B}(p,q)(1+(x/b)^a)^{p+q}}, \quad x,a,b,p,q > 0, \tag{11}$$

where $\mathcal{B}$ is the beta function, b is the scale parameter, a, p and q are shape parameters. Note that here we consider a to be positive, yet, the distribution of the inverse may be obtained by allowing a to be negative (McDonald and Xu 1995). Suppose we are interested in the Theil index defined in (4), the population index, with $\theta = (a,b,p,q)'$, is given by

$$\begin{aligned} T(F_\theta) = \log(\Gamma(p)) + \log(\Gamma(q)) - \log\left(\Gamma\left(\frac{aq-1}{a}\right)\right) - \log\left(\Gamma\left(\frac{ap+1}{a}\right)\right) \\ + \frac{1}{a}\left[\psi\left(\frac{ap+1}{a}\right) - \psi\left(\frac{aq-1}{a}\right)\right], \end{aligned}$$

where Γ is the gamma function and ψ is the digamma function. Clearly the Theil index is scale invariant, so that we set $\theta^s = (a,p,q)'$ and $\theta^c = b$.

The population values of the statistics S_k in (9) are given by

$$S_k(F_\theta) = \frac{[\Gamma(p)\Gamma(q)]^{k-1}\,\Gamma\left(\frac{aq-k}{a}\right)\Gamma\left(\frac{ap+k}{a}\right)}{\left[\Gamma\left(\frac{aq-1}{a}\right)\Gamma\left(\frac{ap+1}{a}\right)\right]^k}, \; k \in \mathbb{R},$$

and the ones for U_l in (10), for $l = 2,3$, are given by

$$\begin{aligned} U_2(F_\theta) &= \frac{\psi^{(1)}(p) + \psi^{(1)}(q)}{a^2}, \\ U_3(F_\theta) &= \frac{\psi^{(2)}(p) - \psi^{(2)}(q)}{a^3}, \end{aligned}$$

where $\psi^{(m)}$ is the polygamma function, i.e., the m-th derivative of the digamma function ψ.

As is done in Cowell and Flachaire (2015), we consider the SM distribution with density

$$f_\theta(x) = \frac{aq x^{a-1}}{b^a(1+(x/b)^a)^{1+q}}, \quad x,a,b,q > 0, \tag{12}$$

and corresponding population statistics T, S_k and U_l, $l = 2,3$, given by

$$\begin{aligned} T(F_\theta) &= 1 + \log(\Gamma(q)) - \log\left(\Gamma\left(\frac{aq-1}{a}\right)\right) - \log\left(\Gamma\left(\frac{a+1}{a}\right)\right) \\ &\quad + \frac{1}{a}\left[\psi\left(\frac{1}{a}\right) - \psi\left(\frac{aq-1}{a}\right)\right], \\ S_k(F_\theta) &= \frac{a^k\Gamma(q)^{k-1}\Gamma\left(\frac{aq-k}{a}\right)\Gamma\left(\frac{k+a}{a}\right)}{\Gamma\left(\frac{1}{a}\right)\Gamma\left(\frac{aq-1}{a}\right)}, k \in \mathbb{R}, \\ U_2(F_\theta) &= \frac{\pi^2 + 6\psi^{(1)}(q)}{6a^2}, \\ U_3(F_\theta) &= \frac{-\psi^{(2)}(q) - 2\zeta(3)}{a^3}, \end{aligned}$$

where $\zeta(3)$ is the Apéry's constant.

Under the GB2, for generating the data, we set $\theta^s = (a = 3, p = 3.5, q = 0.8)'$, $\theta^c = (b = 10)$ and $n = 250, 500, 1000$. For the TME, we choose the vector of statistics to be $\nu = [T(x), U_2(x), U_3(x)]'$ with $T(x)$ the Theil index and $U_j(x)$, $j = 2, 3$ given in (10). We fix the value of the scale parameter to the arbitrary value of one ($b = 1$) in Algorithm 1. We repeat the experiment 10^4 times and set the number of bootstrap replicates to $B = 10^3$.

To solve for $\hat{\theta}^s$ in (8) or for the MLE, we use the classical quasi-Newton optimization algorithm with starting values obtained from the differential evolution heuristic (Storn and Price 1997), in order to mimic a real situation in which the true parameter's values are unknown.

In Table 1, we report the performances of the three approaches with respect to a nominal confidence level of 95% for the three sample sizes. As already shown in the literature (see e.g., Cowell and Flachaire 2015), we find poor performance for the nonparametric bootstrap (Boot), far from the nominal confidence level. The parametric bootstrap using the MLE provides reasonable finite sample coverage that are nevertheless conservatives. On the other hand, the performance of parametric bootstrap using the TME is overall satisfactory, with enhanced performance when sample size increases.

Table 1. Finite sample coverage probability with respect to a nominal confidence level (two-sided) of 95% for the Theil Index. Data are simulated under the GB2 with $\theta^s = (a = 3, p = 3.5, q = 0.8)'$, $\theta^c = (b = 10)$. $\nu = [T(x), U_2(x), U_3(x)]'$ with $T(x)$ the Theil index. In Algorithm 1, $b = 1$. The experiment is repeated 10^4 times and $B = 10^3$.

Sample Size	Boot	MLE	TME
$n = 250$	0.708	0.962	0.927
$n = 500$	0.753	0.978	0.942
$n = 1000$	0.790	0.990	0.949

In Table 2, we replicate the simulation study in (Cowell and Flachaire 2015, Table 6.6), and report the values for *Varstab*, *Semip* and *Mixture*. We have $\theta^s = (a = 2.8, q)'$, $\theta^c = (b = 0.193)$ and set $\nu = [T(x), U_2(x)]'$ with $T(x)$ the Theil index and $U_2(x)$ given in (10). We fix the value of the scale parameter to the arbitrary value of one ($b = 1$) in Algorithm 1. We repeat the experiment 10^4 times and set the number of bootstrap replicates to $B = 10^3$. The results reported in Table 2 are also presented graphically in Figure 1. Both parametric approaches present finite sample coverage probabilities that are far more accurate than the other approaches, especially in the heavy tail case. As with the GB2, the parametric bootstrap based on the MLE tends to provide conservative coverage probabilities.

Table 2. Finite sample coverage probability with respect to a nominal confidence level (two-sided) of 95% for the Theil Index. The values for *Varstab*, *Semip* and *Mixture* are directly reported from (Cowell and Flachaire 2015, Table 6.6). Data are simulated under the Singh-Madalla with $n = 500$, $\theta^s = (a = 2.8, q)$, $\theta^c = (b = 0.193)$. The parameter q accounts for the shape of the upper tail of the distribution, the smaller the heavier the tail. $\nu = [T(x), U_2(x)]'$ with $T(x)$ the Theil index. In Algorithm 1, $b = 1$. The experiment is repeated 10^4 times and $B = 10^3$.

Singh-Madalla	Varstab	Semip	Mixture	Boot	MLE	TME
$q = 1.7$	0.933	0.926	0.928	0.912	0.962	0.952
$q = 1.2$	0.899	0.905	0.912	0.859	0.979	0.957
$q = 0.7$	0.796	0.871	0.789	0.637	0.994	0.939

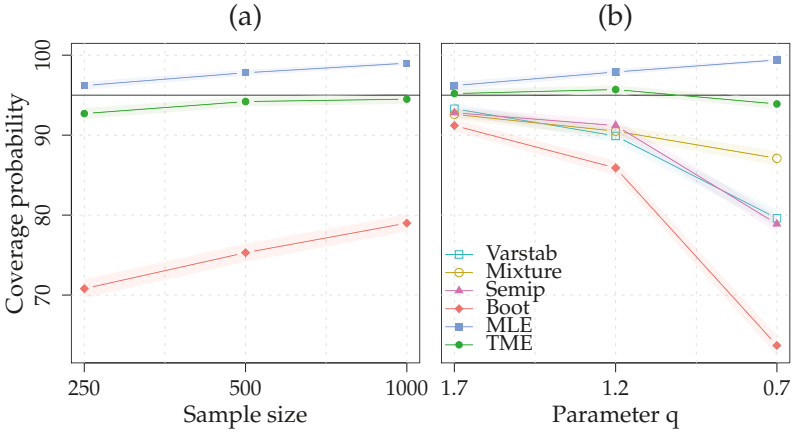


Figure 1. Illustration of the coverage probabilities obtained over 10,000 Monte Carlo experiments for the GB2 (**a**) (see Table 1) and the Singh-Madalla (**b**) (see Table 2). Each color represents a different method. The shade area around each line is the 99.9% asymptotic confidence interval for proportion. The black line is the nominal confidence level of 95%.

5. Conclusions

In this paper, we study the finite sample accuracy of confidence intervals built via parametric bootstrap. We also propose a GMM estimator, the TME, that targets the quantity of interest, namely the considered inequality index. Its primary advantage is that the scale parameter of the assumed parametric model does not need to be estimated to perform parametric bootstrap, since inequality measures are scale invariant. The theoretical result and the simulation study suggest that this feature provides an advantage over the parametric bootstrap using the MLE and also over other established simulation-based inferential methods.

As noted by an anonymous referee, an important point that has not been directly assessed is the specification robustness, i.e., the properties of the proposed method when the assumed general model is not the exact one. This point deserves more (formal) investigation that we leave for further research.

On the more practical side, although this study is limited to two income distributions and one inequality index, the methodology presented here can be extended to other settings in a relative straightforward manner. For example, it is possible to extend the TME to include trimmed inequality indices since it suffices to use the trimmed version of T in ν. If trimming is done for robustness purposes as proposed in Cowell and Victoria-Feser (2003), then the other statistics in $\hat{\nu}$ should also be robust (see also Victoria-Feser 2000). This is the case, for example, with trimmed moments.

Author Contributions: All authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proof of Theorem 1

Proof. Fix θ_0^s in the interior of Θ^s. Since Θ^s is compact, $\sup_{\theta^s \in \Theta^s} \nu(\theta^s)$ is bounded (see Theorem 4.15 in Rudin 1976). Since the mapping $\theta \mapsto \nu$ is bijective, $\nu(\theta^s) = 0$ only if $\theta^s = \theta_0^s$. The conditions for the consistency theorem of a GMM are satisfied (Theorem 2.6 in Newey and McFadden 1994) and $\hat{\theta}^s$ converges in probability to θ_0^s.

Now take an open neighborhood around θ_0^s, say B. Instead of solving the quadratic form in (8), it is equivalent to solve its derivative:

$$\hat{\theta}^s = \underset{\theta^s \in B}{\text{argzero}}\, \dot{\nu}(\theta^s)' g(\theta^s), \quad g(\theta^s) = \hat{\nu}^s - \nu(\theta^s).$$

By the delta method (see Van der Vaart (1998)), we have

$$g(\hat{\theta}^s) - g(\theta_0^s) = \dot{\nu}(\theta_0^s) \cdot \left(\hat{\theta}^s - \theta_0^s\right) + o_p\left(\left\|\hat{\theta}^s - \theta_0^s\right\|\right). \tag{A1}$$

Since $\hat{\theta}^s$ is consistent, the right-hand side element of (A1) is $o_p(1)$. Now multiplying (A1) by $\dot{\nu}(\hat{\theta}^s)'$ yields

$$\dot{\nu}(\hat{\theta}^s)' g(\hat{\theta}^s) - \dot{\nu}(\hat{\theta}^s)' g(\theta_0^s) = \dot{\nu}(\hat{\theta}^s)' \dot{\nu}(\theta_0^s) \cdot \left(\hat{\theta}^s - \theta_0^s\right) + \dot{\nu}(\hat{\theta}^s)' o_p(1).$$

By construction, $\dot{\nu}(\hat{\theta}^s)' g(\hat{\theta}^s) = 0$. By the continuity assumption on the mapping $\theta \mapsto \dot{\nu}$, the continuous mapping theorem applies (see Van der Vaart (1998)) so $\dot{\nu}(\hat{\theta}^s) = \dot{\nu}(\theta_0^s) + o_p(1)$. Next, multiplying by square-root n gives

$$-\dot{\nu}(\theta_0^s)' n^{1/2} g(\theta_0^s) + o_p(1) = \dot{\nu}(\theta_0^s)' \dot{\nu}(\theta_0^s) \cdot n^{1/2} \left(\hat{\theta}^s - \theta_0^s\right) + o_p(1).$$

The proof results from the central limit theorem on $n^{1/2} g(\theta_0^s)$, the invertibility of the derivative $\dot{\nu}(\theta_0^s)$ and the Slutsky's lemma. □

References

Arvanitis, Stelios, and Antonis Demos. 2015. A class of indirect inference estimators: Higher-order asymptotics and approximate bias correction. *The Econometrics Journal* 18: 200–41.

Bandourian, Ripsy, James McDonald, and Robert S. Turley. 2002. A Comparison of Parametric Models of Income Distribution Across Countries and over Time. Available online: http://www.lisdatacenter.org/wps/liswps/305.pdf (accessed on 28 November 2017).

Beirlant, Jan, Goedele Dierckx, A. Guillou, and Catalin Stărică. 2002. On exponential representations of log-spacings of extreme order statistics. *Extremes* 5: 157–80.

Biewen, Martin. 2002. Bootstrap inference for inequality, mobility and poverty measurement. *Journal of Econometrics* 108: 317–42.

Cowell, Frank A., and Emmanuel Flachaire. 2007. Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics* 141: 1044–72.

Cowell, Frank A., and Emmanuel Flachaire. 2015. Statistical Methods for Distributional Analysis. In *Handbook of Income Distribution*. Edited by François Bourguignon and Anthony B. Atkinson. Amsterdam: Elsevier, vol. 2, pp. 359–465.

Cowell, Frank A., and Maria-Pia Victoria-Feser. 1996. Robustness properties of inequality measures. *Econometrica* 64: 77–101.

Cowell, Frank A., and Maria-Pia Victoria-Feser. 2000. Distributional analysis: A robust approach. In *Putting Economics to Work, Volume in Honour of Michio Morishima*. Edited by Anthony Atkinson, Howard Glennerster and Nicholas Stern. London: STICERD.

Cowell, Frank A., and Maria-Pia Victoria-Feser. 2002. Welfare rankings in the presence of contaminated data. *Econometrica* 70: 1221–33.

Cowell, Frank A., and Maria-Pia Victoria-Feser. 2003. Distribution-free inference for welfare indices under complete and incomplete information. *Journal of Economic Inequality* 1: 191–219.

Dagum, Camilo. 1977. A new model of personal income distribution: Specification and estimation. *Economie Appliquée* 30: 413–36.

Danielsson, Jon, Lauredns de Haan, Liang Peng, and Casper G. de Vries. 2001. Using a bootstrap method to choose sample fraction in Tail index estimation. *Journal of Multivariate Analysis* 76: 226–48.

Davidson, Russell. 2009. Reliable inference for the Gini index. *Journal of Econometrics* 150: 30–40.

Davidson, Russell. 2010. Innis lecture: Inference on income distributions. *Canadian Journal of Economics* 43: 1122–48.

Davidson, Russell. 2012. Statistical inference in the presence of heavy tails. *Econometrics Journal* 15: 31–53.

Davidson, Russell, and Emmanuel Flachaire. 2007. Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics* 141: 141–66.

Dupuis, Debbie J., and Maria-Pia Victoria-Feser. 2006. A robust prediction error criterion for Pareto modeling of upper tails. *Canadian Journal of Statistics* 34: 639–58.

Flachaire, Emmanuel, and Olivier G. Nuñez. 2007. Estimation of income distribution and detection of subpopulations: An explanatory model. *Computational Statistics & Data Analysis* 51: 3368–80.

Gallant, A. Ronald, and George Tauchen. 1996. Which moments to match? *Econometric Theory* 12: 657–81.

Guerrier, Stephane, Elise Dupuis, Yanyuan Ma, and Maria-Pia Victoria-Feser. 2018. Simulation based bias correction methods for complex models. *Journal of the American Statistical Association (Theory & Methods)*, in press. doi:10.1080/01621459.2017.1380031.

Guillou, Armelle, and Peter Hall. 2001. A diagnostic for selecting the threshold in extreme-value analysis. *Journal of the Royal Statistical Society, Series B* 63: 293–305.

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansions*. New York: Springer Verlag.

Hampel, Frank R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69: 383–93.

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

Hansen, Lars Peter. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–54.

Heggland, Knut, and Arnoldo Frigessi. 2004. Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66: 447–62.

Kleiber, Christian, and Samuel Kotz. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. New York: John Wiley & Sons, vol. 470.

McDonald, James B. 1984. Some generalized functions for the size distribution of income. *Econometrica* 52: 647–64.

McDonald, James B., and Yexiao J. Xu. 1995. A generalization of the beta distribution with applications. *Journal of Econometrics* 66: 133–52.

McFadden, Daniel. 1989. Method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57: 995–1026.

Mills, Jeffrey A., and Sourushe Zandvakili. 1997. Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics* 12: 133–50.

Newey, Whitney K., and Daniel McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*. Amsterdam: Elsevier, vol. 4, pp. 2111–245.

Phillips, Peter C. B. 2012. Folklore theorems, implicit maps, and indirect inference. *Econometrica* 80: 425–54.

Rudin, Walter. 1976. *Principles of Mathematical Analysis (International Series in Pure & Applied Mathematics)*. New York: McGraw-Hill Education.

Schluter, Christian. 2012. On the problem of inference for inequality measures for heavy-tailed distributions. *The Econometrics Journal* 15: 125–53.

Schluter, Christian, and Kees Jan van Garderen. 2009. Edgeworth expansions and normalizing transforms for inequality measures. *Journal of Econometrics* 150: 16–29.

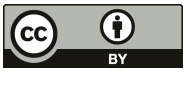
Singh, S. K., and G. S. Maddala. 1976. A function for the size distribution of income. *Econometrica* 44: 963–70.

Storn, Rainer, and Kenneth Price. 1997. Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* 11: 341–59.

Van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge: Cambridge University Press, vol. 3.

Victoria-Feser, Maria-Pia. 1999. Comment on Giorgi's chapter: The sampling properties of inequality indices. In *Income Inequality Measurement: From Theory to Practice*. Edited by J. Silber. Boston: Kluwer Academic Publisher, pp. 260–67.

Victoria-Feser, Maria-Pia. 2000. A general robust approach to the analysis of income distribution, inequality and poverty. *International Statistical Review* 68: 277–93.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

A Hybrid MCMC Sampler for Unconditional Quantile Based on Influence Function

El Moctar Laghlal and Abdoul Aziz Junior Ndoye *

Laboratoire d'Économie d'Orléans (LEO), Faculté de Droit, D'économie et de Gestion, University of Orleans, LEO (FRE CNRS 2014), Rue de Blois, F-45067 Orleans, France; el-moctar.laghlal@univ-orleans.fr

* Correspondence: abdoul-aziz.ndoye@univ-orleans.fr; Tel.: +33-2-3849-2410

Received: 29 December 2017; Accepted: 26 April 2018; Published: 4 May 2018

Abstract: In this study, we provide a Bayesian estimation method for the unconditional quantile regression model based on the Re-centered Influence Function (RIF). The method makes use of the dichotomous structure of the RIF and estimates a non-linear probability model by a logistic regression using a Gibbs within a Metropolis-Hastings sampler. This approach performs better in the presence of heavy-tailed distributions. Applied to a nationally-representative household survey, the Senegal Poverty Monitoring Report (2005), the results show that the change in the rate of returns to education across quantiles is substantially lower at the primary level.

Keywords: hybrid MCMC sampler; quantile regression; influence function; return to education

JEL Classification: C11; C14; C52

1. Introduction

Introduced by Koenker and Bassett (1978), quantile regression models have been increasingly used in empirical labor market studies[1] to parsimoniously describe the entire distribution of an outcome variable. To overcome some limitations[2] of conditional quantile regression models, Firpo et al. (2009) propose the Re-centered Influence Function (RIF)-regression. This regression evaluates the impact of changes in the distribution of covariates on the quantiles of the marginal distribution of the dependent variable. The two-step estimation of the RIF-regression requires first an estimation of the density of the RIF function. A "classical" approach consists of estimating independently the RIF and the regression coefficients (see Firpo et al. 2009). This approach does not take into account the uncertainty related to the first step of estimation. Lubrano and Ndoye (2014) provide a Bayesian estimation of the RIF-regression where they consider sequentially the two-steps of estimation by estimating the density function of the outcome variable by a mixture of normal distributions. While being consistent[3] in the presence of heavy tails, their approach makes the underlying restrictive hypothesis of linearity. However, the estimated RIF function is a binary dependent variable; the linearity and the normality assumptions are strong and may lead sometimes to predicted probabilities that are negative or greater than one. In this study, we implement a Bayesian estimation method for the RIF-regression by considering the dichotomous structure of the RIF function. The method consists of running a logistic-regression where coefficients are estimated by the Metropolis-Hastings sampler using Gibbs output in the first step of estimation.

1 (Buchinsky 1994; Chamberlain 1994; Machado and Mata 2001).
2 Unlike conditional means, conditional quantiles do not average up to their unconditional population counterparts.
3 Mixture models provide flexible extensions of parametric models, and the Bayesian approach takes into account the uncertainty related to the first step of the estimation.

Since the collective agreement in April 2000 to place education at the heart of the development priorities for eradicating extreme poverty, the last two decades have seen a large increase in the enrollment rate of primary education in most developing countries, responding also to the second priority of the Millennium Development Goals (MDGs), "primary education for all". While education is increasingly acknowledged as an important dimension of poverty reduction, there remains some challenges in measuring its return, for example on a household's welfare. Studies emphasizing the role of education on poverty reduction have recently exploded, and regression analysis relying on both household surveys and cross-country data has been widely used in this literature. These regressions, using reduced-form equations, generally provide a simple, but partial framework for examining the marginal effect of education on a household's income[4]. Since the distribution of income is generally skewed to the right, the mean regression models do not provide complete and meaningful information, and then, the analysis of each point of the distribution is of particular interest to assess changes at these different points.

The proposed approach is employed in the empirical analysis to measure the return to education and to address the extent to which the rate of the marginal effect of primary education on a household's income changes across quantiles compared with those of higher education.

The investment in primary education devotes the largest budget allocation in developing countries to fulfill development priorities (Psacharopoulos 1994; Psacharopoulos and Patrinos 2002). In Senegal, the enrollment rate in primary school has climbed from 54 percent in 1994 to 70 percent in 2001 and 82.5 percent in 2005, accompanied by an increase in the female enrollment rate and the rural sectors enrollment rate[5]. However, the IMF 2007's report reveals that 78.51% of Senegalese youth aged 15–19 dropped out before finishing lower secondary school.

The empirical analysis of this paper uses the data from a nationally-representative survey: the Senegal Poverty Monitoring Report (ESPS, 2005) conducted by the National Agency of Statistics and Demography (ANSD)[6]. This survey is largely used by empirical studies, government monitoring reports, institutional strategic documents and in poverty reduction strategies papers (PRSPs) in Senegal[7].

This study applies the RIF-regression method in a Mincer[8] equation type, to primarily investigate the changes in the return to education across quantiles.

The empirical results primarily demonstrate evidence from the heterogeneous pattern of changes in the rate of return to education across quantiles. The rate of change in the return to primary education does not vary much between the lower and the upper quantiles (0.50, 0.75, 0.90) compared to those to secondary and tertiary education. This result supports findings showing that in countries that rapidly expand access to primary education, the returns to primary education fall, while returns to higher education rise (Psacharopoulos 1994; Psacharopoulos and Patrinos 2002).

The paper is organized as follows: Section 2 presents the RIF-regression and the different estimation methods employed. It implements a Bayesian RIF-logit estimation by a Gibbs-Metropolis-Hastings sampler. Section 3 describes the data. Section 4 discusses the empirical results. Section 5 concludes and discusses some policy implications.

---

4 The consumption expenditure is considered as an indicator of a household's income.

5 Source: published reports and papers; see for instance (IMF 2007; Delaunay 2012). These ratios correspond to the number of students formally registered in primary school.

6 ESPS, "Enquête Suivie de la Pauvreté au Sénégal", 2005–2006; ANSD, "Agence National de la Statistique et de la Démographie".

7 Among the studies using the ESPS datasets, we can cite Boccanfuso et al. (2008); Boccanfuso et al. (2009); Diawara (2012), among others, and the national and institutional reports: DSRP 2005; IMF 2007; ANSD 2007.

8 The standard (Mincer 1974) earnings equation linearly regresses the log of wage on the year of education and the quadratic function of labor market experience.

2. Unconditional Quantile Regression Models

We consider the following quantile regression model:

$$y_i = x_i\beta_\tau + u_{i\tau}, \tag{1}$$

where (y_i, x_i), $i = 1, 2, \ldots, n$ are independent observations, y_i being the single-response variable and $x_i = (1, x_{i1}, \cdots, x_{ik})$ being the $(k+1)$ known covariates. $\beta_\tau = (\beta_{\tau 0}, \cdots, \beta_{\tau k})'$ represents the $(k+1)$ unknown regression parameters, and $u_{i\tau}$, $i = 1, \ldots, n$ are the error terms, which are supposed to be independent and identically distributed. The τ-th quantile of $u_{i\tau}$ is assumed equal to zero, $q_\tau(u_{i\tau}|X) = 0$.

2.1. RIF-Regression Models

Firpo et al. (2009) developed an unconditional quantile regression method based on the Re-centered Influence Function (RIF) to evaluate the marginal impact of changes in the distribution of the explanatory variables on the quantiles of the marginal distribution of the dependent variable.

The Influence Function (IF) studies how a change in the distribution of covariates affects a distributional statistic $\nu(F)$, where F is a class of distribution functions. It is defined as:

$$IF(y, \nu, F) = \lim_{\epsilon \to 0} \frac{\nu(F_{\epsilon,\Delta_y}) - \nu(F)}{\epsilon} = \frac{\partial \nu(F_{\epsilon,\Delta_y})}{\partial \epsilon}|_{\epsilon=0}, \tag{2}$$

where Δ_y is a perturbation distribution, which puts a mass of one at any point y and $F_{\epsilon,\Delta_y} = (1-\epsilon)F + \epsilon\Delta_y$ is a mixture model. Firpo et al. (2009) consider the τ-th quantile, q_τ as the distributional statistics $\nu(F)$, and show that the IFcan be expressed as:

$$IF(y_i, q_\tau) = \frac{\tau - \mathbb{1}(y_i \leq q_\tau)}{f_Y(q_\tau)},$$

where $f_Y(.)$ is the density of the variable of interest, Y. A convenient property of IF is that $\mathrm{E}_Y(IF(Y, \nu, F)) = 0$. Firpo et al. (2009) define the Re-centered Influence Function (RIF) as $RIF(y_i, \nu, F) = IF(y_i, \nu, F) + \nu(F)$. For quantiles, the RIF can be expressed in the following convenient way:

$$\begin{aligned} RIF(y_i, q_\tau) &= q_\tau + IF(y_i, q_\tau) \\ &= q_\tau + \frac{\mathbb{1}(y_i > q_\tau)}{f_Y(q_\tau)} - \frac{1-\tau}{f_Y(q_\tau)} \\ &= c_{1,\tau}\mathbb{1}(y_i > q_\tau) + c_{2,\tau}, \end{aligned} \tag{3}$$

where $c_{1\tau} = 1/f_Y(q_\tau)$ and $c_{2\tau} = q_\tau - (1-\tau)c_{1\tau}$.

The RIF-regression model consists of regressing the function RIF given in (3) on a set of covariates X.

2.2. Bayesian Estimation of the RIF-Regression

Running the two-step estimation of the RIF-regression remains a challenging problem. The "classical" approach consists of estimating independently the influence function by kernel estimation and the regression coefficients (see Firpo et al. 2009). However, the kernel density estimation in the first step may lead to unreliable inference in the presence of heavy-tailed distributions as theoretically shown by Bahadur and Savage (1956) and empirically evidenced by Davidson (2012). The Bayesian estimation method of the RIF consists of choosing a mixture representation for the density function by solving a data augmentation problem by a Gibbs sampler and then estimating the regression coefficients. A first MCMC algorithm, which combines the two steps of estimation in a sequential process in linear RIF-regression, was suggested by Lubrano and Ndoye (2014). However, the estimated RIF function is a binary dependent variable; the linearity and the normality assumptions are strong

and may lead sometimes to predicted probabilities that are negative or greater than one. Following the dichotomous structure of the RIF in (3), a non-linear model can be estimated using a logistic (probit) regression. We take the opportunity of this requirement to introduce a hybrid MCMC method, which is called a Gibbs within a Metropolis-Hastings algorithm.

The conditional expectation of the RIF is expressed as:

$$\begin{array}{ll} E\left[RIF\left(Y,q_\tau\right)|X=x\right] & = c_{1,\tau}E\left[\mathbb{1}(Y>q_\tau)|X=x\right]+c_{2,\tau} \\ & = c_{1,\tau}Pr[Y>q_\tau|X=x]+c_{2,\tau}. \end{array} \tag{4}$$

Since $E\left[RIF(Y,q_\tau)|X=x\right]$ in (4) is linear on $Pr[\mathbb{1}(y>q_\tau)|X=x]$, the average marginal effect of covariates is given by:

$$\beta_\tau = \hat{c}_{1\tau}\frac{\partial Pr[Y>q_\tau|X=x]}{\partial x},$$

where $\hat{c}_{1\tau}=1/\hat{f}(q_\tau|\theta)$ with θ are the mixture parameters estimated by the Gibbs sampler. The average marginal effect $\gamma_\tau = \frac{\partial Pr[Y>q_\tau|X=x]}{\partial x}$ can be consistently estimated by a logit regression considering the dummy variable $y_{i\tau}=\mathbb{1}(y_i>q_\tau)$ that is regressed on x_i to derive the RIF-regression coefficients, γ_τ. A Bayesian estimation of a logit regression can be done by a Metropolis-Hastings sampler where the starting values are derived from the estimation of the regression coefficients in a linear probability model.

The average marginal effect from a logit model will be consistent only if:

$$Pr\left(y>q_\tau|X=x\right)=\Lambda(x_i\gamma_\tau)^{1-y_{i\tau}}\left(1-\Lambda(x_i\gamma_\tau)\right)^{y_{i\tau}}, \tag{5}$$

where $\Lambda(.)$ is the cumulative distribution function of a logistic distribution.

The likelihood of the sample is then given by:

$$L(\gamma_\tau|y,x)\propto\prod_{i=1}^{n}\Lambda(x_i\gamma_\tau)^{1-y_{i\tau}}\left(1-\Lambda(x_i\gamma_\tau)\right)^{y_{i\tau}}.$$

For a given prior $\pi(\gamma_\tau)$, the posterior distribution $\pi(\gamma_\tau|y,x)$ is:

$$\pi(\gamma_\tau|y,x)\propto\pi(\gamma_\tau)\times\prod_{i=1}^{n}\Lambda\left(x_i\gamma_\tau\right)^{1-y_i}\left(1-\Lambda(x_i\gamma_\tau)\right)^{y_i}. \tag{6}$$

The Gibbs sampler is difficult to implement since conjugate priors do not exist because the logistic likelihood function does not belong to the exponential family. Therefore, we consider a Metropolis-Hastings sampler, which can be tuned only with the likelihood function under a flat prior on γ_τ.

The proposed approach for the RIF-logit developed is a Gibbs within a Metropolis-Hastings sampler algorithm, as it first requires the use of the Gibbs sampler to estimate the mixture of lognormal densities[9] for $\hat{c}_{1\tau}=1/\widehat{f(q_\tau|\theta)}$.

Gibbs within a Metropolis-Hastings sampler algorithm.

- Estimate the density function of y by Gibbs sampling to obtain $\hat{c}_{1\tau}=1/\widehat{f(q_\tau|\theta)}$
- Initialization: run a linear probability model to set $\gamma_\tau^{(0)}$, and compute $\hat{\Sigma}$.
- Iteration: for $t=1,\cdots,m$

  1. Generate $\tilde{\gamma}_\tau\sim N(\gamma_\tau^{(t-1)},\hat{\Sigma})$

[9] The Gibbs sampler for the mixture of lognormal densities was developed in Lubrano and Ndoye (2016); see also Marin and Robert (2007) for the mixture of normal distributions.

2. Compute the acceptance probability $\rho(\gamma_\tau^{(t-1)}, \tilde{\gamma}_\tau) = \min\left(1, \frac{\pi(\tilde{\gamma}_\tau|y)}{\pi(\gamma^{(t-1)}|y)}\right)$
3. With probability $\rho(\gamma_\tau^{(t-1)}, \tilde{\gamma}_\tau)$, set $\gamma_\tau^{(t)} = \tilde{\gamma}_\tau$ otherwise $\gamma_\tau^{(t)} = \gamma_\tau^{(t-1)}$
4. Compute $\hat{\beta}_\tau^{(t)} = \hat{c}_{1\tau} * \gamma_\tau^{(t)}$

- Average $\hat{\beta}_\tau^{(t)}$ to obtain the estimates of the RIF-regression coefficient, $\hat{\beta}_\tau$.

Without any prior information, the flat prior on γ_τ can be considered, $\pi(\gamma_\tau) \propto 1$. For comparison purposes, we will consider Zellner's non-informative G-prior:

$$\pi(\beta_\tau) \propto det\left((x'x)^{1/2}\right)\Gamma\left[(2k-1)/4\right]\left(\beta'(x'x)\beta\right)^{-(2k-1)/4}\pi^{-k/2}.$$

We can notice that the RIF-logit estimation approach makes assumptions about the functional forms of the $P(Y > q_\tau|X = x)$ in (4). Firpo et al. (2009) suggest the nonparametric-RIF (NP-RIF) regression method based on polynomial series approximations and show that RIF-logit regression yields estimates very close to the fully-nonparametric estimator. However, the choice of the nonparametric estimator is not crucial in large samples as discussed by Newey (1994); if the domain is unbounded, the polynomial series would also poorly approximate the tails.

3. Empirical Analysis

3.1. Data and Descriptive Statistics

The Senegal Poverty Monitoring Report (ESPS, 2005) is a nationally-representative survey conducted by the National Agency of Statistics and Demography. The survey is constructed to provide information related to the evaluation of poverty and to the assessment of the impact of public policies. The ESPS sample covers 13,500 of households of all social classes and from all geographical areas of residence.

Table 1 reports descriptive statistics concerning the characteristics of households and information on the head of the household. It shows that two-thirds of household-heads are illiterate, around 13 percent have reached primary education, 9 percent a secondary education level and less than 5 percent a tertiary level and equivalent. Senegalese families are often extended, nine persons per household on average, and more than half are between 40 and 65 years old. About 80 percent of household-heads are employed (self-employed or salaried). More details on the descriptive statistics of these data are given in the summary reports of the two surveys published by the National Agency of Demography (ANSD 2007).

The estimation of a given equivalence scale relies on a particular consumption model, which is rather restrictive and therefore may lead to identification problems. The usual practice consists of using the per capita income, dividing the household income by the household size. That is what we use in this study referring to Deaton and Muellbauer (1980) and Deaton (1997) and empirical work by the World Bank with Ravallion (2001).

Table 1. Characteristics of heads of households.

Education Level of the Head		**Age**	
Illiterate	71.22	Mean	50.62
Primary	12.63	less 40	21.97
Secondary	11.58	40–65	57.92
Tertiary	4.57	65 and plus	30.11
Gender		**Occupation of the head**	
Female	22.55	Employed	70.6
Marital status of the head		**Size of the household**	
Monogamy	57.03	Mean	9.01
Polygamy	25.39	1–4	20.13
Single	3.40	5–9	49.25
Widower	11.71	10–14	18.33
Divorced	2.39	15, +	12.29

Computations are based on ESPS 2005–2006 after dropping households without any information on educational attainment of the head or on the total consumption expenditures.

3.2. Real Consumption Expenditure Per Capita Distribution

We consider the annual real consumption expenditure as an indicator of permanent income. The consumption expenditures are expressed in CFA francs.[10] The WAEMU[11] Harmonized Consumer Prices Index (HCPI) was respectively 10.94 in 2001 and 11.3 in 2005, revealing a small inflation rate of 0.036 points. The total consumption expenditures in the survey are already deflated by sectors using the national Consumer Prices Index (CPI). The differences in weight in CPI between urban and rural sectors nicely reflect the consumption expenditure structure. In fact, foods are typically less expensive in the rural sectors, and urban households are more likely to consume higher quality goods, which increases their consumption expenditures. The total consumption expenditure in the sample is the sum of food and non-food expenditures, with self-consumption added.

Table 2 presents the distribution of the real annual consumption expenditure per capita.

Table 2. Real annual consumption expenditure per capita.

$q_{0.10}$	8.89
$q_{0.25}$	13.54
Median	20.71
Mean	27.11
$q_{0.75}$	32.40
$q_{0.90}$	50.07
N	13,326
Gini	0.388

The sample reveals that the largest part of the Senegalese household's consumption expenditure is on food (45.6%) and housing (20%); the remainder of the budget is mostly used to cover the clothing expenditure, health and items expenditure.

Since the distribution of the consumption expenditure is often skewed to the left, we impose a restriction on the form of the distribution. We estimate the density function by a mixture of normals

[10] CFA (Communauté Financière Africaine (African Financial Community)). CFA franc had a fixed exchange rate with the Euro (1 euro = 656 CFA) in 2013.

[11] West African Economic and Monetary Union.

using a Gibbs sampler. Figure 1 presents the estimation of the real consumption expenditure per capita 10^{-6} by a mixture of two lognormal distributions.

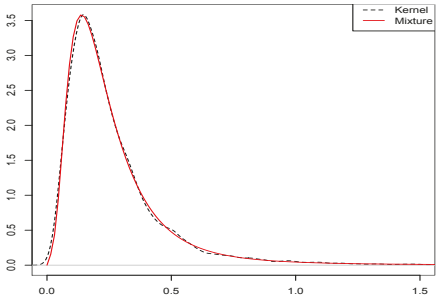


Figure 1. Mixture of two lognormal densities.

4. Empirical Application

In the RIF-regression models, we consider a Mincer type model where the logarithm of the consumption expenditure per capita is the dependent variable. We estimate returns to education at different levels by converting the continuous years of the schooling variable into three dummy variables referring to the completion of the main schooling cycles[12]. This return to education refers to the marginal effect of the level of education on the household's consumption expenditure per capita.

We consider the following set of covariates: *primary*, *secondary* and *tertiary* as dummies, which refer to the level of education of the head of household; *age* and its square[13] refer to the age of the heads of household; the dummy *female* refers to a female headed-household; the dummy *married* refers to a married household's head; the dummy *rural* is the rural geographical area of residence. We restrict the estimations to five quantiles (0.10, 0.25, 0.50, 0.75, 0.90).

In this case, the RIF-regression allows us to evaluate the marginal effect of the changes in the distribution of covariates on the quantiles of the marginal distribution of the total consumption expenditure per capita.

Tables 3 and 4 report the RIF-regression estimates. They show the marginal effects of different covariates on the household's expenditure consumption per capita and their changes across the five quantiles. The regression coefficients are estimated by the hybrid MCMC RIF-estimation methods developed in this paper. The density function of the dependent variable (log of the expenditure consumption per capita) is estimated by a mixture of normal distributions.

---

[12] Primary education corresponds to 6 years or less, secondary between 7 and 13 years and tertiary more than 13 years.

[13] We consider the quadratic function of age to capture the fact that on-the-job training investments decline over time in a standard life-cycle human capital model. This quadratic form of age is implied by a model in which investments decline linearly over time.

Table 3. Bayesian RIF estimates on the log-income without using prior β. RIF, Re-centered Influence Function.

	Lowest 0.10	**Lower Middle 0.25**	**Median 0.50**	**Upper Middle 0.75**	**Highest 0.90**
RIF-Logit Regression Using Flat Prior					
Intercept	18.321	6.497	2.992	1.250	−4.175
	(1.669)	(0.571)	(0.378)	(0.521)	(1.421)
primary	**0.482**	0.465	0.541	0.829	2.175
	(0.449)	(0.145)	(0.093)	(0.133)	(0.405)
secondary	1.421	1.564	1.391	2.322	6.060
	(0.555)	(0.182)	(0.103)	(0.129)	(0.346)
tertiary	5.905	4.145	3.653	4.712	11.332
	(1.651)	(0.554)	(0.271)	(0.238)	(0.490)
age	−0.697	−0.412	−0.308	−0.256	**0.089**
	(0.290)	(0.099)	(0.067)	(0.095)	(0.273)
age^2	0.030	0.017	0.014	0.013	**0.001**
	(0.012)	(0.004)	(0.003)	(0.004)	(0.012)
size	−0.222	−0.148	−0.167	−0.376	−1.468
	(0.020)	(0.008)	(0.007)	(0.013)	(0.053)
female	1.460	0.927	0.609	0.735	1.641
	(0.469)	(0.152)	(0.093)	(0.126)	(0.347)
rural	−8.137	−3.251	−2.412	−3.128	−6.341
	(0.318)	(0.098)	(0.071)	(0.130)	(0.473)
married	1.222	0.688	0.465	0.504	2.183
	(0.503)	(0.165)	(0.103)	(0.139)	(0.378)

The age variable was divided by 100. age^2 represents the square of age. Standard errors are indicated in parentheses. Bold figures correspond to posterior means for which 0 is contained in a 95% HPDinterval.

Table 4. Bayesian RIF estimates on the log-income.

	Lowest 0.10	**Lower Middle 0.25**	**Median 0.50**	**Upper Middle 0.75**	**Highest 0.90**
RIF-Logit Regression Using Zellner's Non-Informative Prior					
Intercept	18.272	6.492	3.001	1.204	−4.075
	(1.669)	(0.571)	(0.378)	(0.521)	(1.421)
primary	**0.487**	0.470	0.534	0.842	2.117
	(0.449)	(0.145)	(0.093)	(0.133)	(0.405)
secondary	1.391	1.558	1.392	2.317	6.013
	(0.555)	(0.182)	(0.103)	(0.129)	(0.346)
tertiary	5.984	4.065	3.621	4.686	11.266
	(1.651)	(0.554)	(0.271)	(0.238)	(0.490)
age	−0.701	−0.414	−0.309	−0.251	**0.066**
	(0.290)	(0.099)	(0.067)	(0.095)	(0.273)
age^2	0.030	0.017	0.014	0.013	**0.002**
	(0.012)	(0.004)	(0.003)	(0.004)	(0.012)
size	−0.220	−0.148	−0.167	−0.372	−1.455
	(0.020)	(0.008)	(0.007)	(0.013)	(0.053)
female	1.444	0.915	0.613	0.735	1.606
	(0.469)	(0.152)	(0.093)	(0.126)	(0.347)
rural	−8.127	−3.245	−2.409	−3.104	−6.341
	(0.318)	(0.098)	(0.071)	(0.130)	(0.473)
married	1.239	0.680	0.476	0.494	2.174
	(0.503)	(0.165)	(0.103)	(0.139)	(0.378)

The age variable was divided by 100. age^2 represents the square of age. Standard errors are indicated in parentheses. Bold figures correspond to posterior means for which 0 is contained in a 95% HPD interval.

Returns to education: For both estimations, the marginal effect of education monotonically increases with the level of education and with quantiles. The rate of change in the returns to education across quantiles provides evidence of significant differences between the bottom and the top of the distribution. For all educational attainment levels, the marginal effects and their rate of change are significantly larger for upper quantiles (0.5, 0.75, 0.90), especially the secondary and the tertiary levels. The marginal effects of the secondary and tertiary education largely dominate the upper part of the distribution. The primary education is significant for all quantiles except the lowest 10 percent; its return increases from the first quartile to the third quartile and then slightly decreases for the highest quantiles. The rate of change in the return to primary education is small and much lower than those to secondary and

tertiary educations (see also Table A1 in Appendix A). This result is in line with findings showing that in countries that rapidly expand access to primary education, the returns to primary education fall, while returns to higher education rise (see for instance Psacharopoulos 1994; Psacharopoulos and Patrinos 2002). In contrast, "primary education continues to be the number one investment priority in developing countries" (Psacharopoulos and Patrinos 2002).

Including age-square, the results show an overall negative effect of age on the household consumption expenditure. Its marginal effect monotonically increases across the first four quantiles and is not significant for the 90th quantile. On average, an additional year of age decreases the household consumption expenditure (in log) by approximately (0.667 0.395 0.294 0.243), respectively. For each of the quantiles (0.10, 0.25, 0.5, 0.75), these marginal effects also increase with age[14].

The marginal effects of the household's size monotonically decrease, and their rates of change across quantiles are higher for upper quantiles. Living in rural areas has a negative and significant effect on the consumption expenditures for all quantiles. Senegal's rural economy is largely agricultural, which is seasonal. The marginal effects of living in rural ares are comparatively higher than the other effects of covariates for poor households. Indeed, the urban labor force is more skilled and earns higher wages than the rural labor force.

5. Conclusions and Policy Implications

In this study, we provide a Bayesian estimation method for the unconditional quantile regression model based on the Re-centered Influence Function (RIF). The method makes use of the dichotomous structure of the RIF and estimates a non-linear probability model by a logistic regression using a Gibbs within a Metropolis-Hastings sampler. This approach performs better in the presence of heavy-tailed distributions. Applied to a nationally-representative household survey, the Senegal Poverty Monitoring Report (2005), the empirical results primarily show evidence from the heterogeneous pattern of changes in the rate of returns to education across quantiles and across the different levels of education. The marginal effects of education monotonically increase and are comparatively higher for upper quantiles (0.50, 0.75, 0.90). The return to primary education does not vary much across quantiles compared with those to secondary and tertiary education.

In most developing countries, promoting education is not only for development policy and for eradicating poverty, but it is also an argument to attract institutional financing and other forms of aid from donors. Senegal witnessed one of the largest increases in the achievement of the second priority of the MDGs. The rate of primary education in Senegal climbed from 54 percent in 1994 to over 82 percent in 2005. In Senegal, as well as in most developing countries, the quality of education in public schools has deteriorated following the increase of enrollment rates. The growing number of primary schools has partially contributed to the literacy and encouraged the education of girls. In contrast, the growing number of public primary schools disadvantages children from low-income families due to the lack of educational resources.

Author Contributions: The authors have contributed equally to this work.

Acknowledgments: We gratefully acknowledge the financial support from the MultiRisk project (ANR-16-CE26-0015) managed by the French ANR. We thank the referees for their constructive comments that have helped to improve the paper.

Conflicts of Interest: The authors declare no conflict of interest.

---

[14] Considering the three age values (30, 50, 65)/100, the following marginal effects for the four quantiles are (−0.679 −0.402 −0.300 −0.2482); (−0.667 −0.395 −0.294 −0.243) and (−0.658 −0.390 −0.290 −0.239), respectively.

Appendix A. Comparison with Conditional Quantile Regression Model

Table A1 presents the estimation results of the conditional quantile regression using Gibbs sampling[15]. The results are in line with those provided by the RIF-regression. The rate of change in the return to primary education does not vary much between the lower and the upper quantiles compared with those to secondary and tertiary education.

Table A1. Bayesian conditional quantile regression using Gibbs sampling.

	Lowest 0.10	**Lower Middle 0.25**	**Median 0.50**	**Upper Middle 0.75**	**Highest 0.90**
Intercept	12.046 (0.180)	12.447 (0.137)	12.898 (0.113)	13.368 (0.141)	13.749 (0.187)
primary	0.071 (0.049)	0.095 (0.036)	0.101 (0.029)	0.111 (0.035)	0.130 (0.047)
secondary	0.234 (0.049)	0.275 (0.036)	0.341 (0.033)	0.377 (0.035)	0.454 (0.053)
tertiary	0.648 (0.079)	0.736 (0.060)	0.749 (0.055)	0.845 (0.062)	0.970 (0.094)
age	−0.034 (0.031)	−0.046 (0.024)	−0.063 (0.020)	−0.082 (0.025)	−0.097 (0.034)
age^2	0.001 (0.001)	0.002 (0.001)	0.003 (0.001)	0.004 (0.001)	0.004 (0.001)
size	−0.029 (0.004)	−0.032 (0.002)	−0.035 (0.002)	−0.035 (0.002)	−0.031 (0.002)
female	0.097 (0.050)	0.100 (0.034)	0.122 (0.028)	0.090 (0.031)	0.093 (0.044)
rural	−0.603 (0.038)	−0.512 (0.025)	−0.473 (0.022)	−0.446 (0.025)	−0.415 (0.034)
married	0.117 (0.055)	0.102 (0.037)	0.076 (0.031)	0.030 (0.035)	0.006 (0.051)

The age variable was divided by 100. age^2 represents the square of age. Standard errors are indicated in parentheses. Bold figures correspond to posterior means for which 0 is contained in a 95% HPD interval.

References

ANSD. 2007. *ANSD, Enquête de Suivi de la Pauvreté au Sénégal, ESPS*. Technical Report. Dakar: Agence Nationale de la Statistique et de la Démographie (ANSD).

Bahadur, R. R., and Leonard J. Savage. 1956. The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *Annals of Statistics* 27: 1115–22. [CrossRef]

Boccanfuso, Dorothée, Bernard Decaluwé, and Luc Savard. 2008. Poverty, income distribution and CGE micro-simulation modeling: Does the functional form of distribution matter? *Journal of Economic Inequality* 6: 149–84. [CrossRef]

Boccanfuso, Dorothée, Antonio Estache, and Luc Savard. 2009. A macro-micro analysis of the effects of electricity reform in Senegal on poverty and distribution. *Journal of Development Studies* 45: 351–68. [CrossRef]

Buchinsky, Moshe. 1994. Changes in the U.S. Wage Structure 1963–1987: An application of Quantile Regression. *Econometrica* 62: 405–58. [CrossRef]

Chamberlain, Gary. 1994. Quantile Regression Censoring and the Structure Of Wages. In *Advances in Econometrics*. Edited by Sims Christopher. Oxford: Cambridge University Press.

Davidson, Russell. 2012. Statistical inference in the presence of heavy tails. *Econometrics Journal* 15: C31–53. [CrossRef]

Deaton, Angus. 1997. *The Analysis of Household Surveys*. Baltimore and London: The John Hopkins University Press.

Deaton, Angus, and John Muellbauer. 1980. An Almost Ideal Demand System. *The American Economic Review* 70: 312–26.

Delaunay, Karine. 2012. Education in Senegal: Inequality in Development. Discussion Paper 397. Marseille, France: Institution de Recherche Pour le Development (IRD).

[15] Details of Bayesian inference for quantile regression based on Gibbs sampling can be found in: Yu and Moyeed 2001; Kozumi and Kobayashi 2011; Yang et al. 2016.

Diawara, Barassou. 2012. Schooling and Assets Ownership. *Modern Economy* 3: 126–38. [CrossRef]

DSRP. 2005. Document de Stratégie de Réduction de la Pauvreté (2003–2005). *Ministère de l'Économie et des Finances du Sénégal, Unité de Coordination et de Suivi de la politique Economique*. Technical Report. DSRP. Available online: http://www.bameinfopol.info/IMG/pdf/DSRP.pdf (accessed on 2 May 2017).

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. Unconditional Quantile Regressions. *Econometrica* 77: 953–73.

IMF. 2007. *Republic of Senegal: Poverty Reduction Strategy Paper*. Technical Report 07/316. Washington: International Monetary Fund, IMF.

Koenker, Roger, and Gilbert Bassett. 1978. Regression Quantiles. *Econometrica* 46: 33–50. [CrossRef]

Kozumi, Hideo, and Genya Kobayashi. 2011. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation* 81: 1565–78. [CrossRef]

Lubrano, Michel, and Abdoul Aziz J. Ndoye. 2014. Bayesian Unconditional Quantile Regression: An Analysis of Recent Expansions in Wage Structure and Earnings Inequality in the US 1992–2009. *Scottish Journal of Political Economy* 61: 129–53. [CrossRef]

Lubrano, Michel, and Abdoul Aziz J. Ndoye. 2016. Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach. *Computational Statistics & Data Analysis* 100: 830–46. [CrossRef]

Machado, José A. F., and José Mata. 2001. Earning functions in Portugal 1982–1994: Evidence from quantile regressions. *Empirical Economics* 26: 115–34. [CrossRef]

Marin, Jean-Michel, and Christian P. Robert. 2001. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer-Verlag Inc.

Mincer, Jacob. 1974. *Schooling, Experience and Earnings*. New York: The Natural Bureau of Economic Research.

Newey, Whitney K. 1994. The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 62: 1349–82. [CrossRef]

Psacharopoulos, George. 1994. Returns to investment in education: A global update. *World Development* 22: 325–43. [CrossRef]

Psacharopoulos, George, and Harry A. Patrinos. 2002. Returns to Investment in Education: A Further Update. Policy Research Working paper 2881. Washington, DC, USA: Education Sector Unit, World Bank.

Ravallion, Martin. 2001. *Growth, Inequality and Poverty: Looking beyond Averages*. Washington: Development Research Group, World Bank.

Yang, Yunwen, Huixia J. Wang, and Xuming He. 2016. Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood. *International Statistical Review* 84: 327–44. [CrossRef]

Yu, Keming, and Rana A. Moyeed. 2001. Bayesian quantile regression. *Statistics & Probability Letters* 84: 437–47.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Review

Using the GB2 Income Distribution

Duangkamon Chotikapanich [1], William E. Griffiths [2,*], Gholamreza Hajargasht [3], Wasana Karunarathne [2] and D. S. Prasada Rao [4]

[1] Monash Business School, Monash University, Melbourne VIC 3145, Australia; duangkamon.chotikapanich@monash.edu
[2] Department of Economics, University of Melbourne, Melbourne VIC 3010, Australia; lakminik@unimelb.edu.au
[3] Department of Accounting, Economics and Finance, Swinburne University of Technology, Hawthorn VIC 3122, Australia; rhajargasht@swin.edu.au
[4] School of Economics, University of Queensland, St. Lucia QLD 4072, Australia; d.rao@uq.edu.au
* Correspondence: wegrif@unimelb.edu.au

Received: 9 February 2018; Accepted: 4 April 2018; Published: 18 April 2018

Abstract: To use the generalized beta distribution of the second kind (GB2) for the analysis of income and other positively skewed distributions, knowledge of estimation methods and the ability to compute quantities of interest from the estimated parameters are required. We review estimation methodology that has appeared in the literature, and summarize expressions for inequality, poverty, and pro-poor growth that can be used to compute these measures from GB2 parameter estimates. An application to data from China and Indonesia is provided.

Keywords: inequality; poverty; pro-poor growth; GMM estimation

JEL Classification: I32; O15; C13

1. Introduction

Specification and estimation of parametric income distributions has a long history in economics. Much of the literature on alternative distributions can be accessed through the book by Kleiber and Kotz (2003), and the papers in Chotikapanich (2008). A series of papers by McDonald and his coauthors (McDonald 1984; McDonald and Xu 1995; Bordley et al. 1997; McDonald and Ransom 2008; McDonald et al. 2011) carry details of many of the distributions and the relationships between them. Our focus in this paper is on the generalized beta distribution of the second kind (GB2). It is a four-parameter distribution defined over the support $(0, \infty)$, and obtained by transforming a standard beta random variable defined on $(0, 1)$. As described by McDonald and Xu (1995), it nests many popular three-parameter specifications of income distributions including the generalized gamma, beta2, Singh-Maddala and Dagum distributions. Two-parameter special cases of these distributions include the lognormal, gamma, Weibull, Lomax and Fisk distributions.[1] Parker (1999) describes a model of firm optimizing behavior that leads to a GB2 distribution for earnings. Applications have appeared in Butler and McDonald (1986), Cummins et al. (1990), Feng et al. (2006), Jenkins (2009), Graf and Nedyalkova (2014), and Jones et al. (2014). Biewen and Jenkins (2005) analyze poverty differences using Singh-Maddala and Dagum distributions, with parameters as functions of personal household characteristics, and with their choice between the Singh-Maddala and Dagum distributions based on preliminary estimates of GB2 distributions. Quintano and D'Agostino (2006)

[1] McDonald and Xu (1995) and McDonald and Ransom (2008) also consider a five-parameter generalized beta distribution which nests the GB2 and a GB1 distribution.

use the Dagum distribution and the Biewen-Jenkins methodology to examine the dependence of inequality and poverty on personal characteristics. In an extensive study examining global inequality, Chotikapanich et al. (2012) estimate special case beta2 distributions for 91 countries in 1993 and 2000. In an application involving 10 regions, Hajargasht and Griffiths (2013) find that the GB2 distribution compares favorably with the four-parameter double Pareto-lognormal distribution in terms of goodness-of-fit.

Estimation of a good-fitting parametric income distribution such as the GB2 facilitates further analysis. Once important quantities such as mean income, the Gini coefficient, the Lorenz curve, and the headcount ratio have been expressed in terms of the parameters of the distribution, they can be readily estimated from those parameters. If interest centers on a region which comprises a collection of countries or areas, a GB2 distribution can be estimated for each country/area; inequality, poverty and pro-poor growth for the region can be analyzed by computing estimates of indicators expressed in terms of the parameters of a regional distribution which will be a population-weighted mixture of the GB2 distributions. If only grouped data are available, then estimating a distribution such as the GB2 provides a means for accommodating within-group variation, an important consideration for assessing inequality and poverty.

The purpose of this paper is to collect results on measures for inequality, poverty, and pro-poor growth, expressed as functions of the parameters of the GB2 distribution and its mixtures, and to summarize various methods of estimation that have appeared in the literature for estimating GB2 parameters from single observations or from grouped data. Expressions for the inequality, poverty, and pro-poor growth measures are given in Section 2. Section 3 contains a description of the various estimation techniques. The results from an application to 4 years of data for China and Indonesia are presented in Section 4. Some concluding remarks are offered in Section 5.

2. Inequality and Poverty Measures from the GB2 Distribution

Throughout we assume that income Y for a given country or area, can be represented by a GB2 distribution whose probability density function (pdf) is given by

$$f(y|a,b,p,q) = \frac{ay^{ap-1}}{b^{ap}\mathrm{B}(p,q)\left(1+\left(\frac{y}{b}\right)^{a}\right)^{p+q}} y > 0 \tag{1}$$

where $a > 0,\ b > 0,\ p > 0$ and $q > 0$ are its parameters and $\mathrm{B}(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}dt$ is the beta function. The cumulative distribution function (cdf) corresponding to (1) is given by

$$F(y|a,b,p,q) = \frac{1}{\mathrm{B}(p,q)}\int_0^w t^{p-1}(1-t)^{q-1}dt = B(w|p,q) \tag{2}$$

where $w = (y/b)^a/\left[1+(y/b)^a\right]$. The function $B(w|p,q)$ is the cdf for the normalized beta distribution, defined on the (0, 1) interval, with parameters p and q, and evaluated at w. It is a convenient representation because both it, and its inverse, are commonly included as readily-computed functions in statistical software. Properties of the GB2 distribution and its special cases have been considered extensively by McDonald (1984) and Kleiber and Kotz (2003). Three-parameter special cases, which have been popular in the literature, are the Singh-Maddala distribution[2] where $p = 1$, the Dagum distribution where $q = 1$, and the beta2 distribution where $a = 1$. Extension to a 5-parameter GB distribution has been considered by McDonald and Xu (1995) and McDonald and Ransom (2008). Some further properties of the GB2 distribution are described by Graf and Nedyalkova (2014). In this

[2] The Singh-Maddala distribution is also commonly known as the Burr distribution, and has been described using a variety of other names. See (Kleiber and Kotz 2003, p. 198).

section, we summarize the main results from the GB2 distribution that are relevant for computing measures of inequality, poverty and pro-poor growth.

We envisage a scenario where GB2 distributions have been estimated for a number of countries, or for specific areas within a country such as urban and rural, and the objective is to evaluate inequality and poverty measures using the estimated parameters of the GB2 distributions. As well as evaluation of the measures from single GB2 distributions, we are interested in evaluating them for mixtures that arise when urban and rural GB2 distributions are combined to obtain a distribution for a country, or when country GB2 distributions are combined to obtain the distribution for a region. In most instances, we can express measures in terms of quantities such as beta and gamma functions that are readily computed by available software. Measures whose exact computation proves to be difficult can usually be written in terms of expectations which can be estimated by averaging values of the function over simulated draws from one or more of the GB2 distributions. Key quantities that are used for calculation of many measures, and for estimation of GB2 distributions, are the GB2 moments and moment distribution functions. We begin by giving expressions for them, as well as indicating how the GB2 Lorenz curve can be obtained. We then consider measures for inequality, poverty and pro-poor growth.

The k-th moment of the GB2 exists for $-ap < k < aq$ and is given by

$$\begin{aligned} \mu^{(k)} = E\left(Y^k\right) &= \frac{b^k \mathrm{B}(p+k/a, q-k/a)}{\mathrm{B}(p,q)} \\ &= \frac{b^k \Gamma(p+k/a)\Gamma(q-k/a)}{\Gamma(p)\Gamma(q)} \end{aligned} \tag{3}$$

where $\Gamma(\cdot)$ is the gamma function. The k-th moment distribution function for the GB2 is given by[3]

$$\begin{aligned} F_k(y|a,b,p,q) &= \tfrac{1}{\mu^{(k)}} \int_0^y t^k f(t)dt \\ &= F(y|a,b,p+k/a,q-k/a) \end{aligned}$$

This result—that the GB2's moment distribution functions can be written in terms of its cdf evaluated at different parameter values—is particularly useful for deriving the Lorenz curve and for setting up and computing GMM estimates from grouped data. The Lorenz curve, relating the cumulative proportion of income η to the cumulative proportion of population u is given by

$$\begin{aligned} \eta(u) &= F_1\left[F^{-1}(u|a,b,p,q)\middle|a,b,p,q\right] \\ &= F\left[F^{-1}(u|a,b,p,q)\middle|a,b,p+1/a,q-1/a\right] \\ &= B\left[B^{-1}(u|p,q)\middle|p+1/a,q-1/a\right] \qquad 0<u<1 \end{aligned}$$

where the function $B(\cdot|\cdot,\cdot)$ is defined in Equation (2).

2.1. Inequality Measures

2.1.1. Gini Coefficient

The most widely used inequality measure is the Gini coefficient. McDonald (1984) and McDonald and Ransom (2008) use hypergeometric functions to express the Gini coefficient in terms of the GB2 parameters. An algorithm for computing these functions has been proposed by Graf (2009). It has been our experience that it is easier computationally to compute the Gini coefficient via numerical integration than to numerically evaluate the hypergeometric functions. Another alternative is to

[3] See, for example, (Butler and McDonald 1989).

estimate the Gini coefficient by simulating from the GB2 distribution. Specifically, noting that the Gini coefficient is given by

$$\begin{aligned} G &= -1 + \tfrac{2}{\mu}\int_0^\infty yF(y|\boldsymbol{\Phi})f(y|\boldsymbol{\Phi})dy \\ &= -1 + \tfrac{2}{\mu}E[yF(y|\boldsymbol{\Phi})] \end{aligned}$$

where $\mu = \mu^{(1)} = E(y) = b[\Gamma(p+1/a)\Gamma(q-1/a)]/[\Gamma(p)\Gamma(q)]$ and $\boldsymbol{\Phi}' = (a, b, p, q)$, we can draw observations $(y_1, y_2, \ldots, y_M)$ from $f(y|\boldsymbol{\Phi})$ and estimate G from

$$\hat{G} = -1 + \frac{2}{\mu}\frac{1}{M}\sum_{m=1}^{M} y_m F(y_m|\boldsymbol{\Phi})$$

The number of draws M can be made as large as necessary to achieve the derived level of accuracy. To draw observations from $f(y|\boldsymbol{\Phi})$, we first draw observations $(w_1, w_2, \ldots, w_M)$ from a standard beta (p, q) distribution, defined on the $(0, 1)$ interval, and then compute $y_m = b[w_m/(1-w_m)]^{1/a}$. If interest centers on one of the special case distributions where $p = 1$, $q = 1$ or $a = 1$, then closed form expressions in terms of gamma or beta functions are available for the Gini coefficient. They are

Beta2	$a = 1$	$G = \frac{2\mathrm{B}(2p,2q-1)}{2\mathrm{B}^2(p,q)}$
Singh–Maddala	$p = 1$	$G = 1 - \frac{\Gamma(q)\Gamma(2q-1/a)}{\Gamma(q-1/a)\Gamma(2q)}$
Dagum	$q = 1$	$G = \frac{\Gamma(p)\Gamma(2p+1/a)}{\Gamma(2p)\Gamma(p+1/a)} - 1$

Suppose now we have estimated GB2 income distributions for a number of different areas, such as countries within a region or urban and rural areas within a country, and we are interested in estimating the Gini coefficient for the combined area. The combined income distribution can be written as a population-weighted mixture of the individual GB2 distributions. That is,

$$f(y|\boldsymbol{\Phi}) = \sum_{j=1}^{J} \lambda_j f\left(y\middle|\boldsymbol{\Phi}_j\right) \tag{4}$$

where $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_J)$, λ_j is the proportion of the combined population in area j, and $\boldsymbol{\Phi}'_j = (a_j, b_j, p_j, q_j)$ is the vector of parameters of the distribution for area j. As noted by Chotikapanich et al. (2007), in this case the Gini coefficient for a combination of J areas can be estimated from

$$G = -1 + \frac{2}{\mu_C}\sum_{j=1}^{J}\sum_{\ell=1}^{J}\lambda_j\lambda_\ell\tau_{j\ell}$$

where

$$\tau_{j\ell} = \frac{1}{M}\sum_{m=1}^{M} y_{j,m}F\left(y_{j,m}\middle|\boldsymbol{\Phi}_\ell\right)$$

$\mu_C = \sum_{j=1}^{J}\lambda_j\mu_j$ is the mean of the combined areas, μ_j is the mean for area j, and $y_{j,m}$ is the m-th draw from pdf $f\left(y\middle|\boldsymbol{\Phi}_j\right)$. For the empirical work in this paper we estimated separate distributions for rural and urban areas in China and Indonesia, then combined them.

2.1.2. Generalized Entropy Measures

Next we consider the generalized entropy (GE) class of inequality measures, whose expressions in terms of the parameters of the GB2 distribution were provided by Jenkins (2009). The GE index is given by

$$I(\alpha) = \frac{1}{\alpha(\alpha-1)}\left[\frac{\mu^{(\alpha)}}{\mu^{\alpha}} - 1\right] \quad \text{for } \alpha \neq 0, 1 \tag{5}$$

where, for the GB2 distribution, $\mu^{(\alpha)} = \int_0^\infty y^\alpha f(y|\boldsymbol{\phi})dy$ is given in (3), and $\mu^\alpha = \left[\mu^{(1)}\right]^\alpha$. For large positive α, the index $I(\alpha)$ is sensitive to large differences at the top of the distribution; for large negative α, it is sensitive to differences at the bottom end of the distribution. Theoretically, α can range from $-\infty$ to ∞, but values between -1 and 2 are usually considered in applications. Two popular special cases are obtained by taking limits as $\alpha \to 0$ and $\alpha \to 1$. The case where $\alpha \to 0$ is known as the mean logarithmic deviation or Theil(0) (Theil 1967, p. 127). Its general expression, and the result for the GB2 distribution, are[4]

$$\begin{aligned} I(0) &= \int_0^\infty \log\left(\frac{\mu}{y}\right) f(y|\boldsymbol{\phi})dy \\ &= \log(\mu) - E[\log(y)] \\ &= \ln(\mu/b) - \psi(p)/a + \psi(q)/a \end{aligned}$$

where $\psi(c) = d\log\Gamma(c)/dc$ is the digamma function, computable by most software. The index obtained as $\alpha \to 1$ is known as Theil(1) (Theil 1967, p. 96). Its general expression, and result for the GB2 distribution, are

$$\begin{aligned} I(1) &= \int_0^\infty \frac{y}{\mu}\log\left(\frac{y}{\mu}\right) f(y|\boldsymbol{\phi})dy \\ &= [E(y\log(y))]/\mu - \log\mu \\ &= [\psi(p+1/a) - \psi(q-1/a)]/a + \log(b/\mu) \end{aligned}$$

In the event that software is not available to compute the digamma function, draws $(y_1, y_2, \ldots, y_M)$ from $f(y|\boldsymbol{\phi})$ can be used to calculate $\sum_{m=1}^M \log(y_m)/M$ and $\sum_{m=1}^M y_m \log(y_m)/M$ as estimators for $E[\log(y)]$ and $E[y\log(y)]$, respectively.

The GE index for a mixture of income distributions and its decomposition into within and between group inequality has been considered by Sarabia et al. (2017). To obtain the GE index for a region whose income distribution is a mixture of GB2 distributions, the quantities $\mu^{(\alpha)}$ and μ^α, defined in (5) for the GB2 distribution $f(y|\boldsymbol{\phi})$, are replaced by the corresponding moments for the mixture distribution $f(y|\boldsymbol{\Phi}) = \sum_{j=1}^J \lambda_j f\left(y\middle|\boldsymbol{\phi}_j\right)$ given in (4). For $\alpha \neq 0$, 1, the resulting index is

$$\begin{aligned} I_C(\alpha) &= \frac{1}{\alpha(\alpha-1)}\left[\int_0^\infty \left(\frac{y}{\mu_C}\right)^\alpha \sum_{j=1}^J \lambda_j f\left(y\middle|\boldsymbol{\phi}_j\right)dy - 1\right] \\ &= \frac{1}{\alpha(\alpha-1)}\left[\frac{1}{\mu_C^\alpha}\sum_{j=1}^J \lambda_j E_j(y^\alpha) - 1\right] \\ &= \frac{1}{\alpha(\alpha-1)}\left[\frac{\sum_{j=1}^J \lambda_j \mu_j^{(\alpha)}}{\left(\sum_{j=1}^J \lambda_j \mu_j\right)^\alpha} - 1\right] \end{aligned} \tag{6}$$

[4] See McDonald and Ransom (2008) or Jenkins (2009) for derivations. Equation (4) in Jenkins (2009) should read $I(1)v_1/\mu - \log\mu$. Sarabia et al. (2017) give details of the Theil indices for a wide range of distributions including the GB2.

where $\mu_j^{(\alpha)} = E_j(y^\alpha)$ is the α-moment with respect to $f\left(y\middle|\boldsymbol{\Phi}_j\right)$, the distribution of the j-th component. For the case where $\alpha = 0$, we have

$$\begin{aligned} I_C(0) &= \int_0^\infty \log\left(\tfrac{\mu_C}{y}\right) \sum_{j=1}^J \lambda_j f\left(y\middle|\boldsymbol{\Phi}_j\right) dy \\ &= \log \mu_C - \sum_{j=1}^J \lambda_j E_j(\log y) \end{aligned}$$

where, for the GB2 distribution, $E_j(\log y) = \left[\psi(p_j) - \psi(q_j)\right]/a_j + \log(b_j)$. For the case where $\alpha = 1$,

$$\begin{aligned} I_C(1) &= \int_0^\infty \tfrac{y}{\mu_C} \log\left(\tfrac{y}{\mu_C}\right) \sum_{j=1}^J \lambda_j f\left(y\middle|\boldsymbol{\Phi}_j\right) dy \\ &= \tfrac{1}{\mu_C} \sum_{j=1}^J \lambda_j E_j(y \log y) - \log \mu_C \end{aligned}$$

with

$$E_j(y \log y) = \left(\mu_j/a_j\right)\left[\psi(p_j + 1/a_j) - \psi(q_j - 1/a_j)\right] + \mu_j \log b_j.$$

An attractive feature of the GE index from a mixture is that it decomposes into a GE measure of inequality within the components of the mixture and a GE measure of inequality between components. To establish this decomposition, we write the index for the j-th area as

$$I_j(\alpha) = \frac{1}{\alpha(\alpha-1)}\left[\frac{\mu_j^{(\alpha)}}{\mu_j^\alpha} - 1\right]$$

and note that

$$\frac{\mu_j^{(\alpha)}}{\mu_j^\alpha} = \alpha(\alpha-1) I_j(\alpha) + 1$$

Substituting this expression into (6) yields

$$\begin{aligned} I_C(\alpha) &= \tfrac{1}{\alpha(\alpha-1)}\left\{\sum_{j=1}^J \lambda_j \left(\tfrac{\mu_j}{\mu_C}\right)^\alpha \left[\alpha(\alpha-1) I_j(\alpha) + 1\right] - 1\right\} \\ &= \sum_{j=1}^J \lambda_j \left(\tfrac{\mu_j}{\mu_C}\right)^\alpha I_j(\alpha) + \tfrac{1}{\alpha(\alpha-1)}\left\{\sum_{j=1}^J \lambda_j \left(\tfrac{\mu_j}{\mu_C}\right)^\alpha - 1\right\} \\ &= I_C^{with}(\alpha) + I_C^{betw}(\alpha) \end{aligned}$$

where $I_C^{with}(\alpha) = \sum_{j=1}^J \lambda_j \left(\mu_j/\mu_C\right)^\alpha I_j(\alpha)$ is a weighted average of the inequalities for each area with weights given by $\lambda_j \left(\mu_j/\mu_C\right)^\alpha$, and $I_C^{betw}(\alpha) = [\alpha(\alpha-1)]^{-1}\left\{\sum_{j=1}^J \lambda_j \left(\mu_j/\mu_C\right)^\alpha - 1\right\}$ is a discrete version of the GE index for the J areas, measuring between inequality. Note that, unless $\alpha = 0$ or 1,

the weights do not sum to 1. When $\alpha = 0$, the weights are the population shares λ_j; when $\alpha = 1$, the weights are the income shares $\lambda_j \mu_j / \sum_{j=1}^{J} \lambda_j \mu_j$. The components for these two cases are

$$\begin{aligned} I_C^{with}(0) &= \sum_{j=1}^{J} \lambda_j I_j(0) \\ &= \sum_{j=1}^{J} \lambda_j \log\left(\mu_j\right) - \sum_{j=1}^{J} \lambda_j E_j(\log y) \end{aligned}$$

$$\begin{aligned} I_C^{betw}(0) &= \sum_{j=1}^{J} \lambda_j \log\left(\tfrac{\mu_C}{\mu_j}\right) \\ &= \log \mu_C - \sum_{j=1}^{J} \lambda_j \log \mu_j \end{aligned}$$

$$\begin{aligned} I_C^{with}(1) &= \sum_{j=1}^{J} \lambda_j \tfrac{\mu_j}{\mu_C} I_j(1) \\ &= \tfrac{1}{\mu_C} \sum_{j=1}^{J} \lambda_j E_j(y \log y) - \sum_{j=1}^{J} \lambda_j \tfrac{\mu_j}{\mu_C} \log\left(\mu_j\right) \end{aligned}$$

$$\begin{aligned} I_C^{betw}(1) &= \sum_{j=1}^{J} \lambda_j \tfrac{\mu_j}{\mu_C} \log\left(\tfrac{\mu_j}{\mu_C}\right) \\ &= \sum_{j=1}^{J} \lambda_j \tfrac{\mu_j}{\mu_C} \log\left(\mu_j\right) - \log \mu_C \end{aligned}$$

2.1.3. Atkinson Index

The Atkinson index is an inequality index that can be viewed as an ordinal special case of a GE index. It is given by

$$A(\varepsilon) = 1 - \tfrac{1}{\mu}\left[\mu^{(1-\varepsilon)}\right]^{1/(1-\varepsilon)} \quad \text{for } 0 < \varepsilon \neq 1$$

$$A(1) = 1 - \tfrac{\exp\{E(\log(y))\}}{\mu}$$

The parameter ε reflects the degree of aversion to inequality in a social welfare function. As $\varepsilon \to 0$, there is no aversion to inequality, and $A(\varepsilon) \to 0$. As $\varepsilon \to \infty$, social welfare is increased by redistributing income towards complete equality; $A(\varepsilon) \to 1$. To compute A from the parameters of the GB2 distribution, we note that $\mu^{(1-\varepsilon)}$ is given in Equation (3) and $E[\log(y)] = [\psi(p) - \psi(q)]/a - \log(b)$. Alternatively, and for computing $A_C(\varepsilon)$, the Atkinson index for a mixture of GB2 distributions, the relationship between $A(\varepsilon)$ and the GE index $I(\alpha)$ can be exploited. With $\alpha = 1 - \varepsilon$, and $\varepsilon > 0$, it is given by

$$A(\varepsilon) = 1 - [\alpha(\alpha - 1)I(\alpha) + 1]^{1/\alpha} \quad \text{for } 0 \neq \alpha < 1$$

$$A(0) = 1 - \exp\{-I(0)\}$$

2.1.4. Pietra Index

In contrast to the Gini coefficient, which is equal to twice the area between the Lorenz curve and the line of perfect equality, the Pietra index is equal to the maximum distance between the Lorenz curve and the perfect equality line (Kleiber and Kotz 2003), as well as twice the area of the largest triangle within the area between the Lorenz curve and line of perfect equality (Butler and McDonald 1989). Details of these results and an extensive analysis of the Pietra index, generally, and in terms of several

distributions and their mixtures, can be found in Sarabia and Jordá (2014). For a single GB2 distribution, we have

$$\begin{aligned} P &= \tfrac{1}{2\mu}\int_0^{\infty} |y-\mu| f(y|\boldsymbol{\Phi})dy \\ &= F(\mu|\boldsymbol{\Phi}) - F_1(\mu|\boldsymbol{\Phi}) \\ &= F(\mu|a,b,p,q) - F(\mu|a,b,p+1/a,q-1/a) \end{aligned}$$

For a mixture of distributions, it is given by

$$P_C = \sum_{j=1}^{J} \lambda_j F\left(\mu_C|\boldsymbol{\Phi}_j\right) - \frac{1}{\mu_C}\sum_{j=1}^{J} \lambda_j \mu_j F_1\left(\mu_C|\boldsymbol{\Phi}_j\right)$$

2.1.5. Quintile Share Ratio

Inequality is often also expressed in terms of the ratio of the income share of the richest to the income share of the poorest in the population. Graf and Nedyalkova (2014) consider the quintile share ratio (QSR), which is the ratio of the income share of the richest 20% relative to the income share of the poorest 20%. For the GB2 distribution, it is given by

$$QSR = \frac{1 - B\left[B^{-1}(0.8|p,q)|p+1/a,q-1/a\right]}{B\left[B^{-1}(0.2|p,q)|p+1/a,q-1/a\right]}$$

Noting that,

$$\begin{aligned} F_1(y|\boldsymbol{\Phi}) &= \tfrac{1}{\mu_C}\int_0^{y} t\sum_{j=1}^{J} \lambda_j f\left(t\middle|\boldsymbol{\Phi}_j\right)dt \\ &= \tfrac{1}{\mu_C}\sum_{j=1}^{J} \lambda_j \mu_j F_1\left(y\middle|\boldsymbol{\Phi}_j\right) \end{aligned}$$

the QSR for a mixture of GB2 distributions can be computed from

$$QSR_C = \frac{1 - \sum_{j=1}^{J} \lambda_j \mu_j B\left(w_{j,0.8}|p_j+1/a_j, q_j-1/a_j\right)}{\sum_{j=1}^{J} \lambda_j \mu_j B\left(w_{j,0.2}|p_j+1/a_j, q_j-1/a_j\right)}$$

where $w_{j,0.8} = \left(y_{0.8}/b_j\right)^{a_j}/\left[1+\left(y_{0.8}/b_j\right)^{a_j}\right]$ and $w_{j,0.2} = \left(y_{0.2}/b_j\right)^{a_j}/\left[1+\left(y_{0.2}/b_j\right)^{a_j}\right]$, with $y_{0.2}$ and $y_{0.8}$ being the 20th and 80th percentiles from the mixture distribution. To obtain $y_{0.2}$ and $y_{0.8}$, the mixture distribution function needs to be inverted to obtain its corresponding quantile function, something that is not possible in closed form. As alternatives, one can (1) attempt to solve the required equation numerically, or (2) generate a large number of observations from each component, combine and sort these components, choosing the 20th and 80th empirical percentiles as estimates.

2.2. Poverty Measures

Expressions for several poverty measures in terms of the parameters of the GB2 distribution have been provided by Chotikapanich et al. (2013). The first is the headcount ratio which is simply the proportion of the population with income less than or equal to a poverty line z

$$H(z) = F(z|\boldsymbol{\Phi}) = B(v|p,q) \tag{7}$$

where $v = (z/b)^a / \left[1 + (z/b)^a\right]$. Setting the poverty line at 0.6 times the median gives what Graf and Nedyalkova (2014) term the at-risk-poverty rate ($ARPR$). It can be calculated from (7) after setting the poverty line at

$$z = 0.6b\left(\frac{B^{-1}(0.5|p,q)}{1 - B^{-1}(0.5|p,q)}\right)^{1/a} \tag{8}$$

A second poverty measure used extensively in the literature is the $FGT(\alpha)$ class of measures (Foster et al. 1984) given by

$$FGT(\alpha) = \int_0^z \left(\tfrac{z-y}{z}\right)^{\alpha} f(y|\boldsymbol{\Phi})dy \quad \text{for } \alpha \geq 1$$

For integer values of α, this expression can be written in terms of incomplete moments of the GB2 distribution as well as in terms of the income gap ratio, defined as the average amount of money that must be given to each of the poor to bring them up to the poverty line, expressed relative to the poverty line. Working in this direction, we define the k-th incomplete moment for the GB2 distribution, relative to poverty line z, as

$$\begin{aligned} \mu_z^{(k)} &= E\left(y^k \middle| y < z\right) = \tfrac{1}{F(z|\boldsymbol{\Phi})}\int_0^z y^k f(y|\boldsymbol{\Phi})dy \\ &= \tfrac{\mu^{(k)} B(v|p+k/a, q-k/a)}{B(v|p,q)} \end{aligned}$$

Defining the income gap ratio as $g(z) = (z - \mu_z)/z$ where $\mu_z = \mu_z^{(1)}$ is mean income of the poor, we can write

$$\begin{aligned} FGT(1) &= B(v|p,q) - (\mu/z)B(v|p+1/a, q-1/a) \\ &= H(z)g(z) \end{aligned}$$

and

$$\begin{aligned} FGT(2) &= B(v|p,q) - (2\mu/z)B(v|p+1/a, q-1/a) \\ &\quad + \left(\mu^{(2)}/z^2\right)B(v|p+2/a, q-2/a) \\ &= H(z)\left[[g(z)]^2 + [1-g(z)]^2 \tfrac{\sigma_z^2}{\mu_z^2}\right] \end{aligned}$$

where $\sigma_z^2 = \mu_z^{(2)} - \mu_z^2$ is the variance of the income of the poor. For noninteger values of α, we can simulate values $y_1, y_2, \ldots y_M$ from the GB2 distribution and use the estimator

$$FGT(\alpha) = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{z - y_m}{z}\right)^{\alpha} I(y_m \leq z)$$

where $I(\cdot)$ is an indicator function equal to 1 if its argument is true and zero otherwise.

As an alternative to the income gap ratio $g(z) = (z - \mu_z)/z$, Graf and Nedyalkova (2014) use a concept known as the relative median poverty gap ($RMPG$). It is defined as the relative gap between a poverty line, which is 0.6 times the median income of the population, and the median income of the poor. Specifically, with z defined as in (8),

$$RMPG = \frac{z - m_{poor}}{z}$$

where the median of the poor is defined as

$$m_{poor} = b\left(\frac{B^{-1}(A/2|p,q)}{1 - B^{-1}(A/2|p,q)}\right)^{1/a}$$

with A being the at-risk-poverty rate (the headcount ratio using the poverty line in (8)).

Considering the income shortfall in log format leads to the Watts index (Watts 1968), defined as

$$\begin{aligned} W &= \int_0^{\infty} (\ln z - \ln y) f(y|\boldsymbol{\Phi}) dy \\ &= \ln\left(\tfrac{z}{b}\right) B(v|p,q) - \\ &\quad \tfrac{1}{a}\left\{ D_p B(v|p,q) - D_q B(v|p,q) + B(v|p,q)[\psi(p) - \psi(q)] \right\} \end{aligned} \tag{9}$$

where $D_p B(v|p,q)$ and $D_q B(v|p,q)$ are the derivatives of the beta cdf $B(v|p,q)$ with respect to p and q, respectively. These derivatives are available in some software (e.g., EViews), otherwise (9) can be estimated via simulation.

The last poverty measure that we describe is the Sen index (Sen 1976) where the poverty gap is weighted by a person's rank in the ordering of the poor. This index is given by

$$\begin{aligned} S &= 2\int_0^{z} \left(\tfrac{z-y}{z}\right)\left(\tfrac{H(z)-F(y|\boldsymbol{\Phi})}{H(z)}\right) f(y|\boldsymbol{\Phi}) dy \\ &= H(z)(g(z) + (1 - g(z))G(z)) \end{aligned} \tag{10}$$

where $G(z)$ is the Gini coefficient for the poor given by

$$G(z) = -1 + \frac{2}{\mu_z H^2(z)} \int_0^{z} y F(y|\boldsymbol{\Phi}) f(y|\boldsymbol{\Phi}) dy$$

The last line in (10) shows how the index can be written in terms of the headcount ratio, the aggregate income gap ratio and the inequality of the poor measured using $G(z)$. Expressing S in terms of the parameters of the GB2 distribution is more difficult than it was for the other indices. In (10) we can use $H(z) = B(v|p,q)$ and $g(z) = 1 - \mu_z/z$, but evaluation of $G(z)$ is more troublesome. If we follow the simulation approach and draw M observations y_m, $m = 1, 2, \ldots, M$ from $f(y|\boldsymbol{\Phi})$, it can be estimated using

$$G(z) = -1 + \frac{2}{\mu_z H^2(z)} \frac{1}{M} \sum_{m=1}^{M} [y_m B(w_m|p,q) I(y_m \leq z)]$$

where $w_m = (y_m/b)^a / [1 + (y_m/b)^a]$.

For aggregating poverty over a number of areas each of which has a GB2 distribution, the headcount ratio, FGT, and Watts indexes are simply population-weighted averages of the indexes for each area. That is, using obvious notation,

$$H_C(z) = \sum_{j=1}^{J} \lambda_j F\left(z \middle| \boldsymbol{\Phi}_j\right) = \sum_{j=1}^{J} \lambda_j B\left(v_j \middle| p_j, q_j\right)$$

$$FGT_C(\alpha) = \sum_{j=1}^{M} \lambda_j FGT_j(\alpha)$$

$$W_C = \sum_{j=1}^{M} \lambda_j W_j$$

This result does not hold for the at-risk-poverty rate and the relative median poverty gap where the poverty line is endogenous, nor does it hold for the Sen index, which contains the cdf. For $ARPR$ and $RMPG$, the median of the mixture is required and $RMPG$ also needs the median of the poor from the mixture distribution. These values can be estimated by simulating observations from the

component distributions and ordering them as was suggested for the QSR. For the Sen index for the mixture, we have

$$\begin{aligned} S_C &= 2\left[FGT_C(1) - \sum_{j=1}^{J}\sum_{\ell=1}^{J}\lambda_j\lambda_\ell\int_0^z\left(\tfrac{z-y}{z}\right)F(y|\boldsymbol{\Phi}_\ell)f\left(y\middle|\boldsymbol{\Phi}_j\right)dy\right] \\ &= 2\left[FGT_C(1) - \sum_{j=1}^{J}\sum_{\ell=1}^{J}\lambda_j\lambda_\ell\gamma_{j\ell}\right] \end{aligned}$$

The term $\gamma_{j\ell} = \int_0^z [(z-y)/z]F(y|\boldsymbol{\Phi}_\ell)f\left(y\middle|\boldsymbol{\Phi}_j\right)dy$ can be estimated from

$$\hat{\gamma}_{j\ell} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{z-y_m}{z}\right)F\left(y_{j,m}\middle|\boldsymbol{\Phi}_\ell\right)I\left(y_{j,m}\le z\right)$$

where the $y_{j,m}$ are draws from $f(y\big|\boldsymbol{\Phi}_j)$.

2.3. Measures of Pro-Poor Growth

In addition to examining changes in poverty incidence over time using measures such as the headcount ratio or refinements of it that take into account the severity of the poverty, it is useful to examine whether growth has favored the poor relative to others placed at more favorable points in the income distribution. Following Duclos and Verdier-Chouchane (2010), we consider three such pro-poor measures, namely, measures attributable to Ravallion and Chen (2003), Kakwani and Pernia (2000), and a "poverty equivalent growth rate" (*PEGR*) suggested by Kakwani et al. (2004).

The first step towards the Ravallion-Chen measure is the construction of a "growth incidence curve" (*GIC*), which describes the growth-rate of income at each percentile u of the distribution. Specifically, if $F_A(y)$ is the income distribution function at time A, and $F_B(y)$ is the distribution function for the new income distribution at a later point B, then

$$GIC(u) = \frac{F_B^{-1}(u) - F_A^{-1}(u)}{F_A^{-1}(u)}$$

For computing values of $GIC(u)$ from the GB2 distribution, note that

$$F^{-1}(u|\boldsymbol{\Phi}) = b\left(\frac{B^{-1}(u|p,q)}{1-B^{-1}(u|p,q)}\right)^{1/a}$$

where $B^{-1}(u|p,q)$ is the quantile function of the standardized beta distribution evaluated at u. When we have a regional distribution or a country distribution, which is a mixture of rural and urban GB2 distributions, it is no longer straightforward to compute the quantile function. In this case, we require $F^{-1}(u|\boldsymbol{\Phi})$ which is the inverse function of $F(y|\boldsymbol{\Phi}) = \sum_{j=1}^{J}\lambda_j F\left(y\middle|\boldsymbol{\Phi}_j\right)$. One needs to either solve the resulting nonlinear equation numerically or estimate $F^{-1}(u|\boldsymbol{\Phi})$ using an empirical distribution function obtained by generating observations from the relevant GB2 distributions in the mixture. We followed the latter approach in our applications.

The *GIC* can be used in a number of ways. If $GIC(u) > 0$ for all u, then the distribution at time B first-order stochastically dominates the distribution at time A. If $GIC(u) > 0$ for all u up to the initial headcount ratio H_A, then growth has been *absolutely* pro-poor. If $GIC(u) > (\mu_B - \mu_A)/\mu_A$ for all u up to the initial headcount ratio H_A, that is, the growth rate of income of the poor is greater than the growth rate of mean income (μ), then growth has been *relatively* pro-poor.

For a single measure of pro-poor growth Ravallion and Chen suggest using the average growth rate of the income of the poor. It can be expressed as

$$RC = \frac{1}{H_A}\int_0^{H_A} GIC(u)du$$

For a GB2 distribution (not a mixture), this integral can be evaluated numerically. Alternatively, we can generate observations from a GB2 distribution or a mixture and compute

$$\hat{RC} = \frac{1}{N_1}\sum_{i=1}^{N_1} GIC(i/N)$$

where N is the total number of observations generated, and $N_1 = H_A N$.

The Kakwani-Pernia measure compares the change in a poverty index such as the change in the headcount ratio, $H_A - H_B$, with the change that would have occurred with the same growth rate, but with distribution neutrality, $H_A - H_{\widetilde{B}}$. Here, $\widetilde{B}$ denotes an income distribution that would be obtained if all incomes changed in the same proportion as the change in mean income that occurred when moving from distribution A to distribution B. To obtain $\widetilde{B}$ in the context of single GB2 distributions, we can simply change the scale parameter b and leave the parameters a, p and q unchanged. The Lorenz curve and inequality measures obtained from a GB2 distribution depend on a, p and q, but do not depend on b. Thus, we have

$$a_{\widetilde{B}} = a_A \quad p_{\widetilde{B}} = p_A \quad q_{\widetilde{B}} = q_A \quad b_{\widetilde{B}} = \left(\tfrac{\mu_B}{\mu_A}\right) b_A$$

Finding $\widetilde{B}$ for a mixture of GB2 distributions—a situation that occurs when we combine rural and urban distributions to find a country distribution—is less straightforward. In this case, the scale parameters in all components of the mixture change and the other parameters are left unchanged. For example, using the superscripts r and u to denote rural and urban, respectively, and $\left(\lambda_A^r, \lambda_A^u\right)$ and $\left(\lambda_B^r, \lambda_B^u\right)$ to denote the respective population proportions at times A and B, we first compute the combined means at times A and B as

$$\mu_A = \lambda_A^r \mu_A^r + \lambda_A^u \mu_A^u \qquad \mu_B = \lambda_B^r \mu_B^r + \lambda_B^u \mu_B^u$$

Then, we obtain the distribution function for $\widetilde{B}$ as follows

$$a_{\widetilde{B}}^j = a_A^j \quad p_{\widetilde{B}}^j = p_A^j \quad q_{\widetilde{B}}^j = q_A^j \quad b_{\widetilde{B}}^j = \left(\tfrac{\mu_B}{\mu_A}\right) b_A^j \quad j = u, r$$

$$F\left(y \middle| \boldsymbol{\Phi}_{\widetilde{B}}^r, \boldsymbol{\Phi}_{\widetilde{B}}^u\right) = \lambda_A^r F\left(y \middle| \boldsymbol{\Phi}_{\widetilde{B}}^r\right) + \lambda_A^u F\left(y \middle| \boldsymbol{\Phi}_{\widetilde{B}}^u\right)$$

Thus, to obtain $\widetilde{B}$ we assume that all incomes in the rural and urban sectors increase in the same proportion as their respective mean incomes, and the distributions of income and the population proportions in each of the sectors remain the same.

The Kakwani-Pernia measure is

$$KP = \frac{H_A - H_B}{H_A - H_{\widetilde{B}}}$$

Assuming the growth in mean income has been positive, a value $KP > 0$ implies the change in the distribution has been absolutely pro-poor, and a value $KP > 1$ implies the change in distribution has been relatively pro-poor.

The third measure of pro-poor growth is the poverty-equivalent growth rate (*PEGR*) suggested by Kakwani et al. (2004). In the context of our description of the Kakwani-Pernia measure, it is the growth rate used to construct distribution $\widetilde{B}$ such that $H_B = H_{\widetilde{B}}$. In other words, it is the growth rate necessary

to achieve the observed change in the headcount ratio when distribution neutrality is maintained. In terms of the GB2 distribution, it is the value g^* that solves the following equation

$$H_B = B(u|p_B, q_B) = B(u^*|p_A, q_A)$$

where $u = (z/b_B)^{a_B}\left[1 + (z/b_B)^{a_B}\right]$ and

$$u^* = \frac{[z/(g^*+1)b_A]^{a_A}}{1 + [z/(g^*+1)b_A]^{a_A}}$$

Thus, to find g^* we have $u^* = B^{-1}(H_B|p_A, q_A)$ and

$$g^* = \frac{z}{b_A}\left(\frac{1-u^*}{u^*}\right)^{1/a_A} - 1$$

As was the case with previous calculations, for a mixture of GB2 distributions, this procedure is less straightforward. As an alternative, to find an approximate g^* for a combined rural–urban distribution, we computed separate growth rates and g_u^* for the two sectors and found a weighted average of them using weights from period B.

$$g^* = \lambda_B^r g_r^* + \lambda_B^u g_u^*$$

If $g^* < g = (\mu_B/\mu_A - 1)$, then, under distribution neutrality, the growth rate required to achieve the same outcome for the headcount ratio is less than realized growth rate, implying that the change in the distribution has not favored the poor. Conversely, when $g^* > g$, a higher growth rate is required under distributional neutrality to equate the two headcount ratios. In this case, the distributional effect must have favored the poor.

3. Estimation

All the required quantities—the means of the distributions, the density and distribution functions, the Gini coefficients, the poverty measures, and the pro-poor growth measures—depend on the unknown parameters $\boldsymbol{\phi}_j$ of the GB2 distributions. Potential methods of estimation of these parameters depend on whether the available data are in the form of single observations or are grouped, and, if they are grouped, whether information on group means, as well as the number of observations in each group, is available.

3.1. Estimation with Single Observations

For single observations, say a sample of observations $(y_1, y_2, \ldots, y_T)$, maximum likelihood estimation can be used with the log-likelihood given by

$$L(\boldsymbol{\phi}) = \sum_{t=1}^{T} \log f(y_t|\boldsymbol{\phi})$$

For samples where sampling weights are available, a pseudo log-likelihood can be maximized to provide consistent parameter estimates, and their precision can be assessed with a sandwich covariance matrix estimator. Details of this estimation procedure are described by Graf and Nedyalkova (2014). With income equivalized over all household members, and sampling weights w_i attached to each household, their pseudo log-likelihood is given by

$$L(\boldsymbol{\phi}) = \sum_{i=1}^{h} w_i n_i \log f(y_i|\boldsymbol{\phi})$$

where h is the number of households and n_i is the number of persons in household i.

A further estimation method has been suggested by Graf and Nedyalkova (2014). This method minimizes a weighted sum of squared distance between sample quantities for (*ARPR*, *RMPG*, *QSR*, Gini), and these quantities are expressed in terms of GB2 parameters. This method has some similarities to the grouped data methods of estimation we describe in the next subsection, where a weighted squared distance between empirical and theoretical quantiles and group means is minimized. One difference is that, for using quantiles and group means, an optimal weight matrix can be derived. Deriving an optimal weight matrix for the Graf-Nedyalkova proposal would appear to be a more difficult problem.

3.2. Estimation with Grouped Data

Suppose now that the observations $(y_1, y_2, \ldots, y_T,)$ have been grouped into N income classes $(x_0, x_1), (x_1, x_2), \cdots, (x_{N-1}, x_N)$ with $x_0 = 0$ and $x_N = \infty$. Let c_i be the proportion of observations in the i-th group, let $\overline{y}_i$ be mean income for the i-th group, and let $\overline{y}$ be overall mean income. In some instances, where income share data for each group $(s_1, s_2, \ldots s_N)$ are available, the group means may need to be calculated from $\overline{y}_i = s_i\overline{y}/c_i$. Choice of an estimation method depends on how much of the information just described is available. If the c_i and x_i are available, but the $\overline{y}_i$ are not, then the multinomial likelihood is a natural choice. In this case the log-likelihood is given by

$$L(\boldsymbol{\phi}) \propto \sum_{i=1}^{N} c_i \log[F(x_i|\boldsymbol{\phi}) - F(x_{i-1}|\boldsymbol{\phi})]$$

Another possibility is the minimum chi-squared estimator described in McDonald and Ransom (2008).

For the scenario where one also has data for the group means $\overline{y}_i$, and when the group bounds x_i may or may not be available, estimators based on moment conditions have been suggested by Chotikapanich et al. (2007), Hajargasht et al. (2012) and Griffiths and Hajargasht (2015). To describe the objective functions that are minimized to obtain these estimators, we need the moments of each group up to order 2, expressed in terms of $\boldsymbol{\phi}$ and $\mathbf{x}' = (x_1, x_2, \ldots, x_{N-1})$. Working in this direction, we define

$$\begin{aligned} k_i &= F(x_i|\boldsymbol{\phi}) - F(x_{i-1}|\boldsymbol{\phi}) \\ \mu_i &= \mu[F_1(x_i|\boldsymbol{\phi}) - F_1(x_{i-1}|\boldsymbol{\phi})] \\ \mu_i^{(2)} &= \mu^{(2)}[F_2(x_i|\boldsymbol{\phi}) - F_2(x_{i-1}|\boldsymbol{\phi})] \end{aligned}$$

where $F_1(x_i|\boldsymbol{\phi})$ and $F_2(x_i|\boldsymbol{\phi})$ are the moment distribution functions defined in Section 2. Further, we define $v_i = k_i\mu_i^{(2)} - \mu_i^2$. Then, Hajargasht et al. (2012) show that the GMM estimator that uses moments for c_i and $\widetilde{y}_i = c_i\overline{y}$, and the optimal weight matrix, can be written as

$$GMM_1(x, \boldsymbol{\phi}) = \sum_{i=1}^{N} w_{1i}(c_i - k_i)^2 + \sum_{i=1}^{N} w_{2i}(\widetilde{y}_i - \mu_i)^2 - 2\sum_{i=1}^{N} w_{3i}(c_i - k_i)(\widetilde{y}_i - \mu_i) \tag{11}$$

where $w_{1i} = \mu_i^{(2)}/v_i$, $w_{2i} = k_i/v_i$ and $w_{3i} = \mu_i/v_i$. $GMM_1(\mathbf{x}, \boldsymbol{\phi})$ can be minimized with respect to both $\mathbf{x}$ and $\boldsymbol{\phi}$, or, if observations on $\mathbf{x}$ are available, with respect to $\boldsymbol{\phi}$ only. Because the weights depend on $(\mathbf{x}, \boldsymbol{\phi})$, a variety of estimators can be used, depending on whether $GMM_1(\mathbf{x}, \boldsymbol{\phi})$ is minimized directly or a two-step or iterative procedure is employed. In a two-step procedure, initial estimates with weights that are not dependent on the parameters are obtained, and then estimates that minimize $GMM_1(\mathbf{x}, \boldsymbol{\phi})$, with weights computed from the initial estimates, are computed. Iterating this process leads to an iterative estimator.

An estimator that uses weights that do not depend on $(\mathbf{x}, \boldsymbol{\phi})$, and which is useful for obtaining starting values for a two-step or iterative estimator from (11), is that proposed by Chotikapanich et al. (2007). In contrast to (11), they considered moment conditions for c_i and $\overline{y}_i$

instead of c_i and $\widetilde{y}_i = c_i\overline{y}_i$. Although they focused on the special case beta 2 distribution, their results also hold for the more general GB2 distribution. The function that they minimized is

$$GMM_2(\mathbf{x}, \boldsymbol{\phi}) = \sum_{i=1}^{N}\left(\frac{c_i - k_i}{c_i}\right)^2 + \sum_{i=1}^{N}\left(\frac{\overline{y}_i - \mu_i/k_i}{\overline{y}_i}\right)^2 \tag{12}$$

The weights used for this estimator $\left(c_i^{-2}$ and $\overline{y}_i^{-2}\right)$ are not optimal, but they have the intuitive appeal of minimizing the sum of squares of percentage errors. Also, computation of the second moment $\mu_i^{(2)}$ is not required.

A third GMM estimator is that described by Griffiths and Hajargasht (2015). Like (12), this estimator considers the moment conditions for c_i and $\overline{y}_i$, but uses the optimal weight matrix.[5] It is given by

$$GMM_3(\mathbf{x}, \boldsymbol{\phi}) = k_i^{-1}\sum_{i=1}^{N}(c_i - k_i)^2 + k_i^3 v_i^{-1}\sum_{i=1}^{N}(\overline{y}_i - \mu_i/k_i)^2 \tag{13}$$

Relative to the other optimal weight formulation in (11), this objective function avoids the term with the cross product of the moment conditions.

4. Applications

A major source of data for the cross-country study of income distributions, inequality and poverty is from the World Bank PovcalNet website. We used data from China and Indonesia, two Asian countries with relatively large populations. The years considered were 1999, 2005, 2010 and 2013 for China and 1999, 2005, 2010 and 2016 for Indonesia[6]. The data available are in grouped form comprising population shares and corresponding expenditure shares for a number of classes, together with mean monthly expenditure that has been reported from surveys, and then converted to purchasing power parity (PPP) using the World Bank's 2011 PPP exchange rates for the consumption aggregate for national accounts. Also available are the data on population size. Throughout the paper we use the generic term *income* distributions, although our example distributions are for expenditure. For both countries, separate data were available for rural and urban populations and so distributions were estimated for each of these components. Data for China were in the form of 20 groups, with the exception of China-rural 1999 (19 groups) and 2005 (17 groups), while those for Indonesia were available in 100 groups. To make the data for both countries relatively consistent for estimation, we aggregated the Indonesian data into 20 groups. The distributions were estimated by minimizing the objective function $GMM_3(\mathbf{x}, \boldsymbol{\phi})$ given in (13). Initial estimates were obtained by minimizing $GMM_2(\mathbf{x}, \boldsymbol{\phi})$, those initial estimates were used to compute the weights for $GMM_3(\mathbf{x}, \boldsymbol{\phi})$, the estimates from were then used to compute a new set of weights, and the process was continued for 10 iterations. Parameterizing the objective function in terms of (a, μ, p, q) instead of (a, b, p, q) facilitated convergence.

Parameter estimates for each of the distributions are presented in Table 1, along with corresponding estimates for mean income and the populations for each region. The density functions for China and Indonesia, obtained as mixtures of the urban and rural densities, are plotted in Figures 1 and 2, respectively. A striking feature of the parameter estimates is the very large estimates for p (and correspondingly small estimates for b) for Indonesia-urban in 2010 and 2016. As $p \to \infty$, the GB2 distribution approaches the 3-parameter inverse generalized gamma distribution,[7] and so the results

[5] It may be better to describe the estimators that minimize $GMM_2(\mathbf{x}, \boldsymbol{\phi})$ and $GMM_3(\mathbf{x}, \boldsymbol{\phi})$ as minimum distance estimators rather than GMM estimators because the "moment condition" for $\overline{y}_i$ is plim $\overline{y}_i = \mu_i/k_i$ not $E(\overline{y}_i) = \mu_i/k_i$. The asymptotic distribution is the same, however. See, for example, (Greene 2012, chp. 13).

[6] The version of the data that was used was downloaded on 9 March 2018 at http://iresearch.worldbank.org/PovcalNet/povOnDemand.aspx.

[7] See (McDonald and Xu 1995, p. 139).

suggest this special-case distribution would be adequate for these two cases. Its density function is given by

$$f(y|a,q,\beta) = \frac{a\beta^{aq}}{\Gamma(q)} y^{-aq-1} \exp\left(-\frac{\beta}{y}\right)^{a}$$

The figures show that, for both countries, there is an improvement over time in the sense that the distribution shifts to the right, and mean income increases, with the most dramatic improvements being from 1999 to 2005, and after 2010.

Table 1. Parameter estimates, mean income and population.

Country/Year	*a*	*b*	*p*	*q*	μ	Population (Millions)
China rural						
2013	1.5806	101.3579	3.8613	2.1609	190.23	635.69
2010	1.2063	21.4069	11.6780	2.2025	131.52	697.21
2005	1.3443	32.0352	7.0416	2.3558	100.07	749.35
1999	2.0243	30.1693	3.3733	1.3113	67.78	815.97
China urban						
2013	1.6455	261.4467	2.3392	1.9792	373.92	721.69
2010	1.8842	187.8696	2.3745	1.5884	306.81	658.50
2005	1.8294	144.7708	2.4059	1.7919	217.11	554.37
1999	1.6302	95.0994	3.2433	2.5261	134.70	436.77
Indonesia rural						
2016	2.0275	55.8739	3.8536	1.3660	129.11	118.90
2010	2.1389	36.6977	4.4602	1.2132	96.63	121.45
2005	2.7720	52.1883	2.5501	1.1926	85.84	122.57
1999	3.0994	49.6466	2.0371	1.2727	67.62	123.52
Indonesia urban						
2016	0.7417	0.0010	25,914.0	4.0699	208.00	142.22
2010	0.9107	0.0094	15,488.0	3.2802	156.25	121.08
2005	2.0275	55.8746	3.8535	1.3660	129.11	104.15
1999	2.0737	35.6598	4.7873	1.2719	96.37	85.10

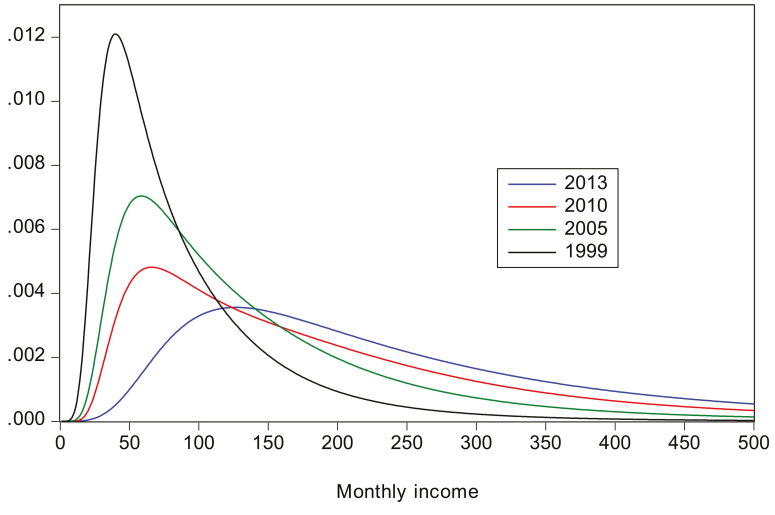


Figure 1. Income distributions for China.

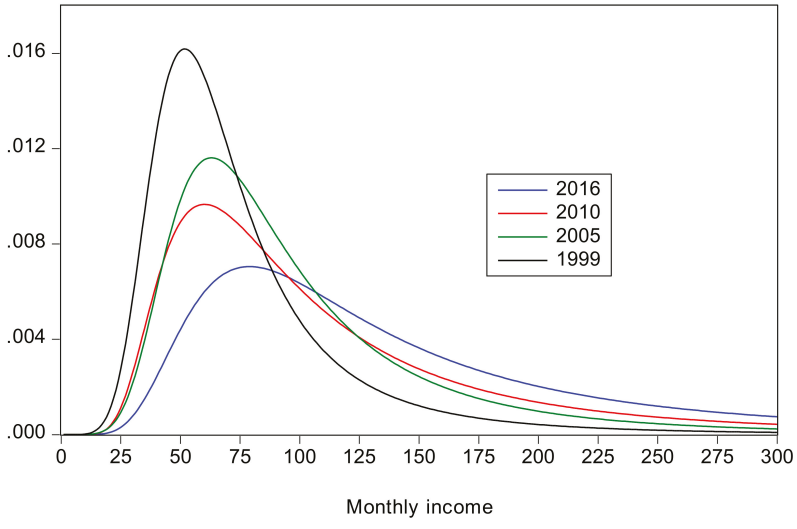


Figure 2. Income distributions for Indonesia.

Inequality measures for the rural and urban areas and their combined distributions are presented in Table 2. We computed the Gini coefficient, the Pietra index, QSR, $I(0)$ and $I(1)$. The within and between urban and rural components for $I_C(0)$ and $I_C(1)$ are reported in Table 3. Tables 4 and 5 contain poverty measures and pro-poor growth measures, respectively. For poverty measures, the headcount, $FGT(1)$, $FGT(2)$ and Sen indices were computed using a poverty line of \$57.8 per month, equivalent to \$1.9 per day. Pro-poor growth measures, RC, KP and $PEGR$ were computed for the combined distributions; the GIC's for each time interval are depictured in Figures 3–8. From the tables and figures, we can make the following observations about China.

1. All inequality measures indicate that inequality increased from 1999 to 2010, and then declined from 2010 to 2013. The recent decline is attributable to a decline in rural inequality; there was an increase in urban inequality in the same period. Also, there is no clear conclusion about how rural inequality changed from 1999 to 2005; the Gini and $I(1)$ suggest a slight decrease, whereas QSR, $I(0)$ and Pietra suggest a slight increase.
2. Inequality is much greater in the combined distribution than in its components, reflecting the large discrepancy in mean incomes between the rural and urban areas. Within inequality remains greater than between inequality, however.
3. The changes in inequality have been accompanied by large increases in mean income and large decreases in poverty. The decline in poverty was particularly dramatic for rural China where the headcount ratio declined from 57% in 1999 to 3.7% in 2013. Poverty in rural China is uniformly greater than that in urban China.
4. The GIC curves show that, from 1999 to 2010, growth has favored the rich more than the poor, but from 2010 to 2013, growth has strongly favored the poor relative to the rich, a result consistent with the decline in inequality over this period. The scalar measures of pro-poor growth are also consistent with this observation. Growth has favured the poor in an absolute sense from 1999 to 2010 $(0 < RC < g, 0 < KP < 1, PEGR < g)$, and in a relative sense after 2010 $(RC > g, KP > 1, PEGR > g)$.

Examining the results for Indonesia, we find:

1. Urban inequality changed very little from 1999 to 2005, increased dramatically from 2005 to 2010, and then increased more moderately from 2010 to 2016. Rural inequality increased from 1999 to 2010, but declined thereafter. The combined results reflect these changes, with increasing inequality overall, but with Gini coefficients approximately the same in 2010 and 2016.
2. Poverty declined from 1999 to 2005, remained roughly constant from 2005 to 2010, when there were large increases in inequality, and then declined again from 2010 to 2016. From 2005 to 2010 a decline in urban poverty was offset by an increase in rural poverty.
3. The *GIC* curves show that growth has favored the rich relative to the poor in all time intervals. From 2005 to 2010 the poor faired very badly; the growth rate for the bottom 15% of the population was negative. This period was also one where the growth in mean incomes was low relative to that in the other two periods. The scalar pro-poor growth measures are in line with the conclusions from the *GIC* curves. Growth was absolutely but not relatively pro-poor in the first and third time intervals; in the second interval it was not absolutely pro-poor according to the *RC* measure, and only slightly absolutely pro-poor using the *KP* measure.

Table 2. Inequality measures.

Country/Year	Gini	*QSR*	*I*(0)	*I*(1)	Pietra
China rural					
2013	0.3349	5.4526	0.1903	0.2086	0.2424
2010	0.3959	7.1456	0.2664	0.3189	0.2901
2005	0.3519	5.8464	0.2097	0.2375	0.2563
1999	0.3638	5.6579	0.2083	0.2495	0.2545
China urban					
2013	0.3735	6.5286	0.2291	0.2454	0.2628
2010	0.3540	5.9757	0.2126	0.2370	0.2545
2005	0.3436	5.7017	0.1992	0.2163	0.2460
1999	0.3185	4.9247	0.1649	0.1731	0.2246
China combined					
2013	0.4010	8.1998	0.2659	0.2864	0.2874
2010	0.4323	9.5593	0.3274	0.3451	0.3155
2005	0.4052	6.4547	0.2796	0.2979	0.2959
1999	0.3941	4.5101	0.2495	0.2683	0.2825
Indonesia rural					
2016	0.3343	5.2640	0.1912	0.2270	0.2442
2010	0.3502	5.2808	0.1962	0.2412	0.2480
2005	0.2756	3.9165	0.1275	0.1448	0.1980
1999	0.2352	3.3989	0.1002	0.1087	0.1746
Indonesia urban					
2016	0.4154	7.9453	0.2920	0.3409	0.3044
2010	0.4070	6.7226	0.2493	0.2930	0.2818
2005	0.3444	5.2640	0.1912	0.2270	0.2442
1999	0.3368	5.2471	0.1939	0.2370	0.2467
Indonesia combined					
2016	0.4027	7.6873	0.2737	0.3286	0.2963
2010	0.4042	6.5792	0.2513	0.3013	0.2842
2005	0.3297	4.6841	0.1776	0.2117	0.2357
1999	0.2959	4.0104	0.1539	0.1879	0.2169

Table 3. Between and within inequality.

Country/Year	$I_C(\mathbf{0})$	$I_C^{with}(\mathbf{0})$	$I_C^{betw}(\mathbf{0})$	$I_C(\mathbf{1})$	$I_C^{with}(\mathbf{1})$	$I_C^{betw}(\mathbf{1})$
China combined						
2013	0.2659	0.2109	0.0550	0.2864	0.2340	0.0523
2010	0.3274	0.2399	0.0875	0.3451	0.2622	0.0829
2005	0.2796	0.2053	0.0743	0.2979	0.2244	0.0735
1999	0.2495	0.1932	0.0563	0.2683	0.2101	0.0582
Indonesia combined						
2016	0.2737	0.2461	0.0276	0.3286	0.3019	0.0267
2010	0.2513	0.2227	0.0286	0.3013	0.2732	0.0281
2005	0.1776	0.1568	0.0208	0.2117	0.1910	0.0207
1999	0.1539	0.1385	0.0154	0.1879	0.1723	0.0156

Table 4. Poverty measures.

Country/Year	*HC*	*FGT*(1)	*FGT*(2)	SEN
China rural				
2013	0.0374	0.0070	0.0021	0.0099
2010	0.2042	0.0489	0.0171	0.0713
2005	0.2998	0.0786	0.0296	0.1057
1999	0.5702	0.1907	0.0844	0.2568
China urban				
2013	0.0077	0.0017	0.0006	0.0020
2010	0.0085	0.0017	0.0005	0.0023
2005	0.0294	0.0062	0.0021	0.0088
1999	0.1064	0.0233	0.0080	0.0324
China combined				
2013	0.0216	0.0042	0.0013	0.0083
2010	0.1079	0.0256	0.0089	0.0496
2005	0.1848	0.0478	0.0179	0.0901
1999	0.4084	0.1324	0.0577	0.2289
Indonesia rural				
2016	0.1267	0.0243	0.0073	0.0348
2010	0.3033	0.0700	0.0234	0.0995
2005	0.2917	0.0613	0.0193	0.0883
1999	0.4647	0.1117	0.0385	0.1526
Indonesia urban				
2016	0.0649	0.0122	0.0035	0.0174
2010	0.1142	0.0221	0.0065	0.0313
2005	0.1267	0.0243	0.0073	0.0353
1999	0.3031	0.0700	0.0234	0.0941
Indonesia combined				
2016	0.0931	0.0177	0.0052	0.0345
2010	0.2089	0.0461	0.0150	0.0863
2005	0.2159	0.0443	0.0138	0.0828
1999	0.3988	0.0947	0.0324	0.1659

Table 5. Pro-poor growth measures.

Country/Year	Growth Rate	Growth Rate for the Poor (*RC*)	*KP*	*PEGR*
China				
2010–2013	0.3218	0.6245	1.4251	0.3245
2005–2010	0.4536	0.2331	0.6503	0.2839
1999–2005	0.6446	0.5281	0.8702	0.4504
Indonesia				
2010–2016	0.3614	0.2836	0.8622	0.2414
2005–2010	0.1956	–0.0107	0.0709	0.0079
1999–2005	0.3323	0.2449	0.8049	0.2575

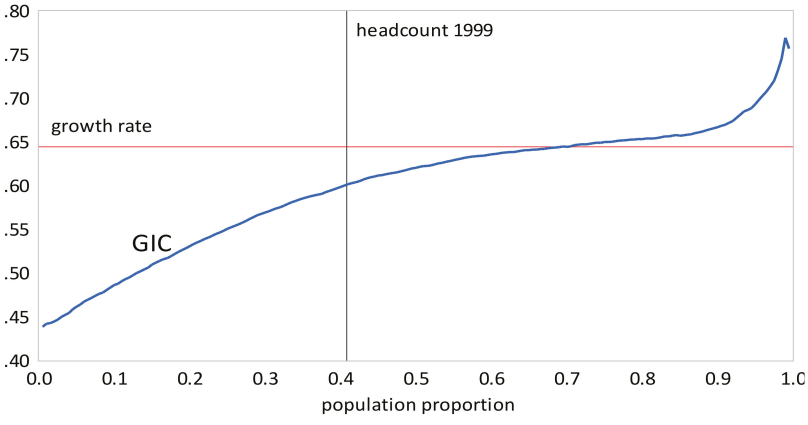


Figure 3. Growth incidence curve, China 1999–2005.

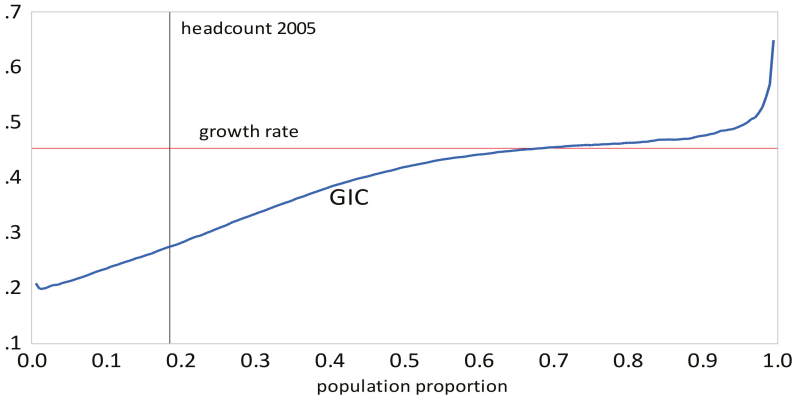


Figure 4. Growth Incidence Curve, China 2005–2010.

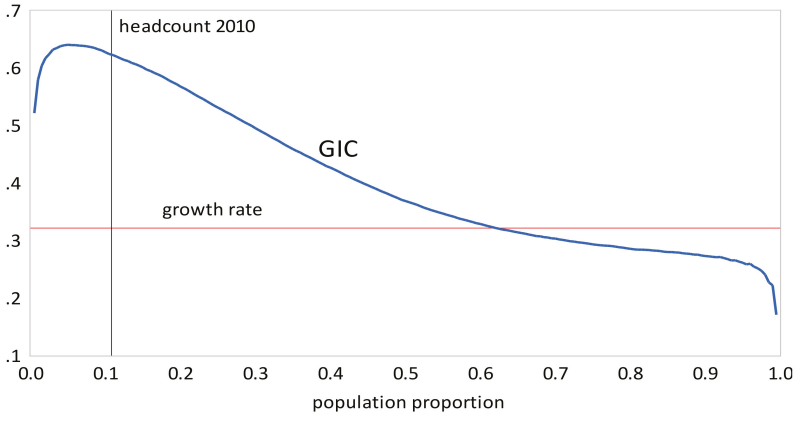


Figure 5. Growth Incidence Curve, China 2010–2013.

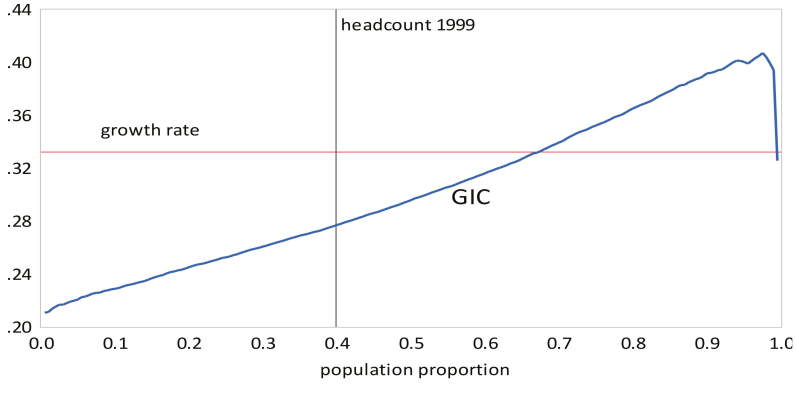


Figure 6. Growth Incidence Curve, Indonesia 1999–2005.

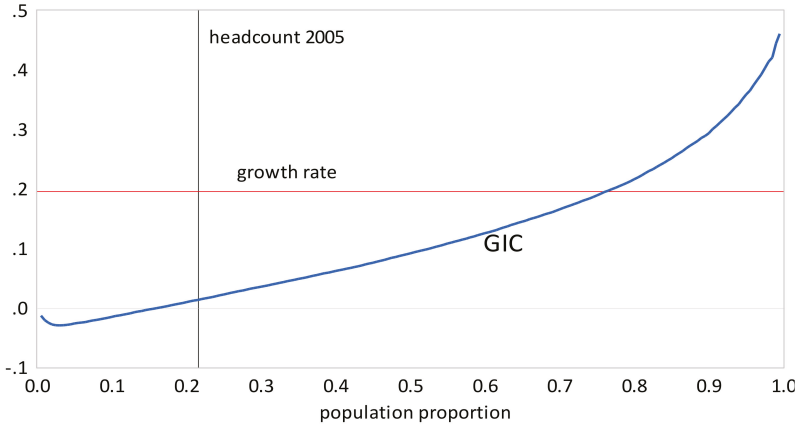


Figure 7. Growth Incidence Curve, Indonesia 2005–2010.

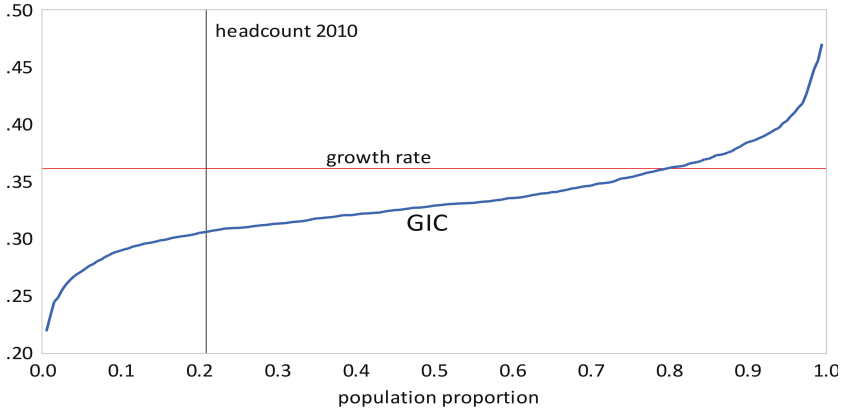


Figure 8. Growth Incidence Curve, Indonesia 2010–2016.

5. Concluding Remarks

Studying income distributions can provide valuable information about important aspects of a society's welfare such as the degree of inequality, the incidence of poverty, and whether there have been improvements in welfare over time. The GB2 is a popular and versatile distribution well suited to this purpose. We have reviewed some of the common indexes for measuring inequality, poverty and pro-poor growth, and described how values for these indexes can be computed from estimates of the parameters of the GB2 distribution. Optimal techniques for estimating the parameters using either single observations or grouped data are also reviewed. It is our hope that the bringing together of all these results into a single source will facilitate and promote use of the GB2 distribution.

Acknowledgments: The authors acknowledge support from ARC Grant DP140100673. Comments from two referees and the editor have led to substantial improvements in the paper.

Author Contributions: William Griffiths conceived and wrote the paper. Duangkamon Chotikapanich supplied the data, converted it into a form suitable for estimation, and computed inequality, poverty and pro-poor growth measures. Duangkamon Chotikapanich and Wasana Karunarathne developed the material on poverty and pro-poor growth measures. Gholamreza Hajargasht developed the software for GMM estimation and estimated the income distributions. Prasada Rao provided the necessary background information.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Biewen, Martin, and Stephen P. Jenkins. 2005. A Framework for the Decomposition of Poverty Differences with an Application to Poverty Differences between Countries. *Empirical Economics* 30: 331–58. [CrossRef]

Bordley, Robert F., James B. McDonald, and Anand Mantrala. 1997. Something New, Something Old: Parametric Models for the Size of Distribution of Income. *Journal of Income Distribution* 6: 91–103.

Butler, Richard J., and James B. McDonald. 1986. Income Inequality in the U.S.: 1948–80. *Research in Labor Economics* 8: 85–140.

Butler, Richard J., and James B. McDonald. 1989. Using Incomplete Moments to Measure Inequality. *Journal of Econometrics* 42: 109–20. [CrossRef]

Chotikapanich, Duangkamon, ed. 2008. *Modeling Income Distributions and Lorenz Curves*. New York: Springer.

Chotikapanich, Duangkamon, William Griffiths, Wasana Karunarathne, and D. S. Prasada Rao. 2013. Calculating Poverty Measures from the Generalized Beta Income Distribution. *Economic Record* 89: 48–66. [CrossRef]

Chotikapanich, Duangkamon, William E. Griffiths, and D. S. Prasada Rao. 2007. Estimating and Combining National Income Distributions Using Limited Data. *Journal of Business and Economic Statistics* 25: 97–109. [CrossRef]

Chotikapanich, Duangkamon, William E. Griffiths, D. S. Prasada Rao, and Vicar Valencia. 2012. Global Income Distributions and Inequality, 1993 and 2000: Incorporating Country-level Inequality Modeled with Beta Distributions. *The Review of Economics and Statistics* 94: 52–73. [CrossRef]

Cummins, John, Georges Dionne, James McDonald, and B. Michael Pritchett. 1990. Applications of the GB2 Family of Distributions in Modeling Insurance Loss Processes. *Insurance: Mathematics and Economics* 9: 257–72. [CrossRef]

Duclos, Jean-Yves, and Audrey Verdier-Chouchane. 2010. *Analyzing Pro-Poor Growth in Southern Africa: Lessons from Mauritius and South Africa*. Working Papers Series No. 115. Tunis: African Development Bank.

Feng, Shuaizhang, Richard Burkhauser, and J. S. Butler. 2006. Levels and Long-Term Trends in Earnings Inequality: Overcoming Current Population Survey Censoring Problems Using the GB2 Distribution. *Journal of Business and Economic Statistics* 24: 57–62. [CrossRef]

Foster, James, Joel Greer, and Erik Thorbecke. 1984. A Class of Decomposable Poverty Measures. *Econometrica* 52: 761–66. [CrossRef]

Graf, Monique. 2009. An Efficient Algorithm for the Computation of the Gini Coefficient of the Generalised Beta Distribution of the Second Kind. In *JSM Proceedings, Business and Economic Statistics Section*. Alexandria: American Statistical Association, pp. 4835–43.

Graf, Monique, and Desislava Nedyalkova. 2014. Modeling of Income and Indicators of Poverty and Social Exclusion Using the Generalized Beta Distribution of the Second Kind. *Review of Income and Wealth* 60: 821–42. [CrossRef]

Greene, William H. 2012. *Econometric Analysis*, 7th ed. New York: Prentice Hall.

Griffiths, William, and Gholamreza Hajargasht. 2015. On GMM Estimation of Distributions from Grouped Data. *Economics Letters* 126: 122–26. [CrossRef]

Hajargasht, Gholamreza, and William E. Griffiths. 2013. Pareto-Lognormal Distributions: Inequality, Poverty, and Estimation from Grouped Income Data. *Economic Modelling* 33: 593–604. [CrossRef]

Hajargasht, Gholamreza, William E. Griffiths, Joseph Brice, D. S. Prasada Rao, and Duangkamon Chotikapanich. 2012. Inference for Income Distributions Using Grouped Data. *Journal of Business of Economic Statistics* 30: 563–76. [CrossRef]

Jenkins, Stephen P. 2009. Distributionally-Sensitive Inequality Indices and the GB2 Income Distribution. *Review of Income and Wealth* 55: 392–98. [CrossRef]

Jones, Andrew M., James Lomas, and Nigel Rice. 2014. Applying Beta-Type Size Distributions to Healthcare Cost Regressions. *Journal of Applied Econometrics* 29: 649–70. [CrossRef]

Kakwani, Nanak, and Ernesto M. Pernia. 2000. What is Pro-Poor Growth. *Asian Development Review* 18: 1–16.

Kakwani, Nanak, Shahidur R. Khandker, and Hyun Son. 2004. *Pro-Poor Growth: Concepts and Measurement with Country Case Studies*. Working Paper No. 1. Brasilia: International Poverty Centre, United Nations Development Programme.

Kleiber, Christian, and Samuel Kotz. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. New York: John Wiley and Sons.

McDonald, James B. 1984. Some Generalized Functions for the Size Distribution of Income. *Econometrica* 52: 647–63. [CrossRef]

McDonald, James B., and Michael Ransom. 2008. The Generalized Beta Distribution as a Model for the Distribution of Income: Estimation of Related Measures of Inequality. In *Modeling Income Distributions and Lorenz Curves*. Edited by Duangkamon Chotikapanich. New York: Springer, pp. 147–66.

McDonald, James B., Jeff Sorensen, and Patrick A. Turley. 2011. Skewness and Kurtosis Properties of Income Distribution Models. *Review of Income and Wealth* 59: 360–74. [CrossRef]

McDonald, James B., and Yexiao J. Xu. 1995. A generalization of the beta distribution with applications. *Journal of Econometrics* 66: 133–52. Erratum in *Journal of Econometrics* 69: 427–28. [CrossRef]

Parker, Simon C. 1999. The Generalized Beta as a Model for the Distribution of Earnings. *Economics Letters* 62: 197–200. [CrossRef]

Quintano, Claudio, and Antonella D'Agostino. 2006. Studying Inequality in Income Distribution of Single-Person Households in Four Developed Countries. *Review of Income and Wealth* 52: 525–46. [CrossRef]

Ravallion, Martin, and Shaohua Chen. 2003. Measuring Pro-Poor Growth. *Economics Letters* 78: 93–99. [CrossRef]

Sarabia, José María, and Vanesa Jordá. 2014. Explicit Expressions of the Pietra Index for the Generalized Function for the Size Distribution of Income. *Physica A* 416: 582–89. [CrossRef]

Sarabia, José María, Vanesa Jordá, and Lorena Remuzgo. 2017. The Theil Indices in Parametric Families of Income Distributions—A Short Review. *Review of Income and Wealth* 63: 867–80. [CrossRef]
Sen, Amartya K. 1976. Poverty: An Ordinal Approach to Measurement. *Econometrica* 44: 219–31. [CrossRef]
Theil, Henri. 1967. *Economics and Information Theory*. Amsterdam: North Holland.
Watts, Harold W. 1968. An Economic Definition of Poverty. In *On Understanding Poverty*. Edited by Daniel P. Moyniham. New York: Basic Books, pp. 316–29.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Top Incomes, Heavy Tails, and Rank-Size Regressions

Christian Schluter [1,2]

[1] Aix-Marseille School of Economics, 5 Boulevard Maurice Bourdet CS 50498, 13205 Marseille CEDEX 01, France; christian.schluter@univ-amu.fr; Tel.: +33-413-55-25-72

[2] Department of Economics, University of Southampton, Highfield, Southampton SO17 1BJ, UK

Received: 19 November 2017; Accepted: 20 February 2018; Published: 2 March 2018

Abstract: In economics, rank-size regressions provide popular estimators of tail exponents of heavy-tailed distributions. We discuss the properties of this approach when the tail of the distribution is regularly varying rather than strictly Pareto. The estimator then over-estimates the true value in the leading parametric income models (so the upper income tail is less heavy than estimated), which leads to test size distortions and undermines inference. For practical work, we propose a sensitivity analysis based on regression diagnostics in order to assess the likely impact of the distortion. The methods are illustrated using data on top incomes in the UK.

Keywords: top incomes; heavy tails; rank size regression; extreme value index; regular variation

JEL Classification: D31; C13; C14

1. Introduction

Income distributions exhibit, like many other size distributions in economics and the natural science, upper tails that decay like power functions (see e.g., Schluter and Trede 2017). The recent and rapidly growing literature on top incomes focuses on this upper tail, and its presence has important consequences for the measurement of inequality.[1] However, estimating the heaviness of the upper tail is challenging, since real world size distributions usually are Pareto-like (i.e., tails are regularly varying) rather than strictly Pareto.

To be precise, let $X_1, \ldots, X_n$ be a sequence of positive independent and identically distributed random variables (e.g., incomes) with distribution function F that is regularly varying, so for large x

$$1 - F(x) = x^{-\frac{1}{\gamma}} l(x), \qquad \gamma \in (0, \infty), \tag{1}$$

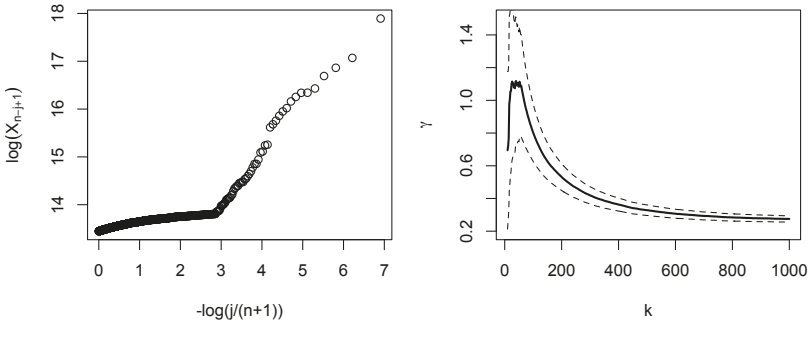
where l is slowly varying at infinity, i.e., $l(tx)/l(x) = 1$ as $x \to \infty$. The parameter γ, usually referred to as extreme value index (and $1/\gamma$ as the tail exponent), is unknown and needs to be estimated. Many estimators have been proposed in the statistical literature (see e.g., the textbook treatments in Embrechts et al. 1997 or Beirlant et al. 2004).

An estimator popular among economists is based on a simple ordinary least squares (OLS) regression of log sizes on log ranks (e.g., Jenkins 2017 and Atkinson 2017, and references therein, in the income distribution and top incomes literature, this regression is ubiquitous in the city size literature). The enduring popularity of the OLS estimator is partly due to its simplicity, and partly due

[1] See e.g., Schluter and Trede (2002) in the contexts of Lorenz curves, Davidson and Flachaire (2007) who propose a semi-parametric bootstrap, Cowell and Flachaire (2007) who advocate semi-parametric methods, or Burkhauser et al. (2012) who seek to reconcile survey and tax return data. Also observe that the p^{th} moment of the income distribution is finite only if $p < 1/\gamma$, so very heavy tails can directly affect the validity of some standard inequality measurement tools. For instance, statistical inference for the Generalised Entropy index with parameter 2 requires the existence of the fourth moment (Cowell 1989).

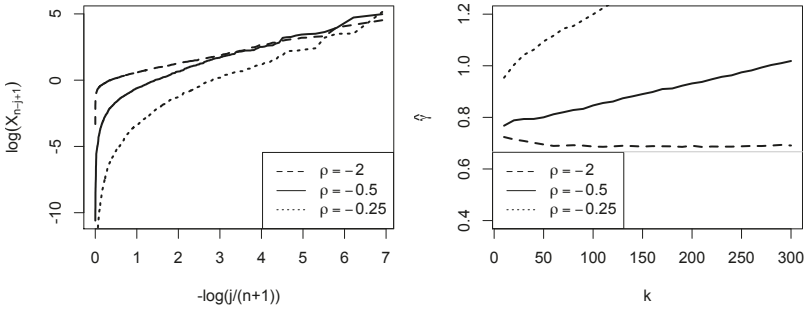
to a powerful intuition based on a Pareto quantile-quantile (QQ)-plot, the regression estimating its slope coefficient. However, if the tail of the distribution varies regularly, the Pareto QQ-plot will become linear only *eventually*. In particular, (1) can be expressed equivalently, using the tail quantile function $U(x) = \inf\{t : \Pr(X > t) = 1/x\}$ where $x > 1$, as $U(x) = x^{\gamma}\tilde{l}(x)$ where $\tilde{l}(x)$ is a slowly varying function. Hence, as $x \to \infty$, $\log U(x) \sim \gamma \log(x)$ since then $\log \tilde{l}(x) \to 0$. Replacing these population quantities with their empirical counterparts gives the Pareto QQ-plot, and γ is its ultimate slope. This qualification (usually ignored by practitioners in economics) has important consequences for the behaviour of the estimator: Since the OLS estimator estimates the slope parameter of this QQ-plot, deviations from the strict Pareto model -captured by the nuisance function l- will induce distortions.

The empirical importance of this is illustrated in Figure 1, which depicts the Pareto QQ-plot for our administrative income data for the UK (the subject of our empirical application developed in Section 4 below), using the 1000 largest incomes. The plot exhibits a pronounced kink, and approximate linearity of the QQ-plot only holds for the very highest upper order statistics. Panel (b) shows the consequences for the OLS estimates: As we move in the QQ-plot from the right to the left, the departures from linearity become progressively more severe, and the OLS estimates progressively fall. Based on this first diagnostic QQ-plot, once the lower upper order statistics have been discarded as a source of downward bias, the subsequent analysis can then more clearly focus on the approximate linear part, the remaining distortions, and the choice of the number of order statistics. Figure 2 provides a further illustration for three Burr (Singh-Maddala) distributions (examined in detail in Section 3 below, being the leading parametric income distribution model) possessing the same γ. Here, the speed of decay of the nuisance function l is parametrised by the absolute value of the parameter ρ. The smaller the magnitude of ρ, the greater the initial curvature and steepness of the Pareto QQ-plot, and the larger the induced positive distortions of the OLS estimator of the slope coefficient.



(**a**) (**b**)

Figure 1. Pareto quantile-quantile (QQ)-Plot: Top incomes in the UK. Based on administrative income tax return for the UK in 2009/10. The Survey of Personal Incomes (SPI) is described in Section 4. Panel (**a**): The Pareto QQ-plot (see Section 1.1) is based on the largest 1000 incomes. Panel (**b**): Estimates of γ for the k upper order statistics using the OLS regression (solid lines), and pointwise 95% symmetric confidence intervals (dashed lines). The distributional theory is stated in Equation (8).



(**a**) (**b**)

Figure 2. Pareto QQ-Plots: The Burr distribution. Based on the Burr distribution given by $F_{(\gamma,\rho)}(x) = 1 - (1 + x^{-\rho/\gamma})^{1/\rho}$ with $\gamma = 2/3$, and $\rho \in \{-2, -0.5, -0.25\}$. Panel (**a**): Pareto QQ-plots for 3 random samples drawn from the Burr distribution. Sample size is 1000. To aid comparison across cases, the points of each QQ-plot have been connected and rendered as lines. Panel (**b**): Mean of estimates $\hat{\gamma}$ across 1000 Monte Carlo simulations for given ρ, drawing samples of size 1000 in each iteration. The faint horizontal line is the population value $\gamma = 2/3$.

In this paper, we examine the asymptotic distortions of the OLS estimator that arise in these circumstances, caused by the slow decay of the nuisance function l and modeled here as higher order regular variation. The theory is presented in Section 2 (proofs are collected in Appendix A), and numerical illustrations and quantifications of the distortions are provided in Section 3, as well as of the stark consequence for inference. More specifically, we show formally that the OLS estimator over-estimates the true value in the leading heavy-tailed model (i.e., the Hall class, which includes the Burr (Singh-Maddala) distribution, as well as the student, Fréchet, and Cauchy distributions). An empirical illustration in the context of top incomes in the UK using data on tax returns is the subject of Section 4.

1.1. The Log-Log Rank-Size Regression

We briefly review the rank size regression. Let $X_{1,n} \leq \cdots \leq X_{n,n}$ denote the order statistics of $X_1, \cdots, X_n$, and consider the k upper order statistics. Let ranks be shifted by a constant $\eta < 1$. The regression of sizes on ranks leads to the minimisation of the least squares criterion

$$\sum_{j=1}^{k} \left(\log \frac{X_{n-j+1,n}}{X_{n-k,n}} - g \log \frac{k+1}{j-\eta} \right)^2 \tag{2}$$

with respect to g, where $\eta < 1$ and $1 \leq j \leq k < n$. The classic case is $\eta = 0$. However, since the OLS estimator of the slope coefficient is not invariant to shifts in the data, it is conceivable that a purposefully chosen shift could yield an asymptotic refinement (Gabaix and Ibragimov 2011 consider this in the strict Pareto model $1 - F(x) = cx^{-1/\gamma}$). The analysis below allows for this possibility.

The justification of considering regression (2) is based on a Pareto QQ-plot (Beirlant et al. 1996): For a sufficiently high threshold $X_{n-k,n}$ where $k < n$, the Pareto quantile plot in model (1) with coordinates $(-\log(j/(n+1)), \log X_{n-j+1,n})_{j=1,\cdots,k}$ becomes ultimately linear. The line through point $-\log((k+1)/(n+1)), \log X_{n-k,n})$ with slope g is thus given by $y = \log X_{n-k,n} + g[x + \log((k+$

$1)/(n+1))]$ and the data points are $(x,y) = (-\log(j/(n+1)), \log X_{n-j+1,n})_{j=1,\cdots,k}$. The regression estimator estimates this slope parameter. In particular, the OLS estimator of the slope coefficient g is

$$\hat{\gamma} \;=\; \frac{\frac{1}{k}\sum_{j=1}^{k}\log\left(\frac{k+1}{j-\eta}\right)\left[\log X_{n-j+1,n} - \log X_{n-k,n}\right]}{\frac{1}{k}\sum_{j=1}^{k}\left[\log\frac{k+1}{j-\eta}\right]^2} \equiv \frac{N_{n,k}}{D_k}, \qquad \eta < 1. \tag{3}$$

Note that the denominator D_k is a Riemann approximation to $\int_0^1 \log^2 x \mathrm{d}x = 2$. An asymptotic expansion of the denominator reveals that

$$D_k = 2 + O\left(\frac{\log^2 k}{k}\right) \tag{4}$$

From Kratz and Resnick (1996, proof of their Equation 2.4, p. 704) we know that the numerator $N_{n,k}$ converges in probability to 2γ, hence the estimator is weakly consistent: $\hat{\gamma} \to^P \gamma$ as $k \to \infty$ and $k/n \to 0$. We proceed in the next Section to refine this result by obtaining higher order expansions of the estimator in (3).

The literature contains several variants of regression (2). Rather regressing log sizes on log ranks, one could regress log ranks on log sizes, thus obtaining the 'dual' regression. In view of (3), our asymptotic analysis of the numerator carries immediately over to this dual regression. Another variant of (2) includes the additional estimation of a regression constant: $\log X_{n-j+1,n}$ is regressed on a constant and $\log j$. Kratz and Resnick (1996) obtain the distributional theory for this alternative estimator and show that its asymptotic variance is $2\gamma^2/k$, which exceeds, as will be shown below, the asymptotic variance of $\hat{\gamma}$ given by (3). Hence this regression variant is less efficient. Schultze and Steinebach (1996) also prove weak consistency of the estimator in this setting.

2. Asymptotic Expansions and Distributional Theory

2.1. Preliminaries: Higher Order Regular Variation

In order to obtain our asymptotic expansions, we use an equivalent representation of model (1) based on regular variation and extreme value theory. First we recall the definition of first-order regular variation, and then proceed to model the slowly varying nuisance function l in (1) by a refinement to second-order regular variation. We then show that most heavy-tailed distributions of interest (in the income, finance and urban literature) satisfy this condition.

It is well known that model (1) has the equivalent (first-order regular variation) representation (e.g., Dekkers et al. 1989)

$$\lim_{t\to\infty}\frac{\log U(tx) - \log U(t)}{a(t)/U(t)} = \log x, \tag{5}$$

for all $x > 0$ where a is a positive norming function with the property $a(t)/U(t) \to \gamma$. The problem for estimating the extreme value index γ is the behaviour of the slowly varying function l in (1). It is, therefore, common practice in the extreme value literature to model such second-order behaviour, thus strengthening model (1), by strengthening the first-order regular representation (5) to second-order regular variation. Following De Haan and Stadtmüller (1996), we assume that the following refinement of (5) holds

$$\lim_{t\to\infty}\frac{\frac{\log U(tx)-\log U(t)}{a(t)/U(t)} - \log x}{A(t)} = H_{\gamma,\rho}(x) \tag{6}$$

for all $x > 0$, where $H_{\gamma>0,\rho<0}(x) = \frac{1}{\rho}(\frac{x^\rho-1}{\rho} - \log x)$ with $\rho < 0$. This parameter ρ is the so-called second-order parameter of regular variation, and $A(t)$ is a rate function that is regularly varying with index ρ, with $A(t) \to 0$ as $t \to \infty$. As ρ falls in magnitude, the nuisance part of l in (1) decays more slowly. Our numerical illustrations will thus consider small magnitudes for ρ.

Examples. Most heavy-tailed distributions of interest satisfy representation (6). Consider the Hall class of distributions (Hall 1982), given by, for large x,

$$F(x) = 1 - ax^{-1/\gamma}[1 + bx^{\beta} + o(x^{\beta})]$$

with $\gamma, a > 0$, $b \in R$, $\beta < 0$. In this class, the nuisance function l in model (1) converges to a constant at a polynomial rate. The Hall class nests, for instance, the Burr (Singh-Maddala), Student, Fréchet, and Cauchy distributions.[2] The tail quantile function is $U(x) = cx^{\gamma}[1 + dx^{\rho} + o(x^{\rho})]$ where $c = a^{\gamma}$, $d = b\gamma a^{\gamma\beta}$. This Hall class satisfies the second order representation (6) with $\rho = \gamma\beta < 0$, and rate function

$$A(t) = \frac{\rho^2}{\gamma} dt^{\rho}.$$

Figure 2 illustrates the role of ρ for the Burr distribution (examined in greater detail in Section 3) in terms of the Pareto QQ-plot, and the implications for the estimator $\hat{\gamma}$ of its slope parameter. For $\rho = -2$ the plot is close to linear, and the estimates close to the population value. However, as ρ falls in magnitude, the initial curvature increases, and the slope estimates consequently becomes more positively distorted as the number of upper order statistics k entering the estimator increases.

2.2. The Main Results

We first state the higher order asymptotic expansion of the numerator $N_{n,k}$. We then obtain the distributional theory for our estimator $\hat{\gamma}$, before returning to the distortions induced by deviations from the strict Pareto model (captured by second order regular variation).

Asymptotic expansion. In theAppendix A we prove the following higher order expansion of the numerator $N_{n,k}$ under the assumption of second-order regular variation (6). Throughout, we will consider an intermediate sequence $k = k_n$ of positive integers such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. It is then true that, for $\gamma > 0$ and $\rho < 0$,

$$\begin{aligned} N_{n,k}/\gamma &= 2 - \left(\frac{1}{2} - \eta\right)\frac{\log(k-\eta)}{k} - \left(\frac{1}{2} - \eta\right)\frac{\log^2 k}{2k} \\ &+ O_p\left(\frac{1}{k^{1/2}}\right) + O\left(\frac{1}{k}\right) + O_p\left(\frac{\log k}{k^{1/2}}\right) \\ &+ A\left(\frac{n}{k}\right)\frac{1}{\rho}\left[\frac{2-\rho}{(1-\rho)^2}\right] + O_p\left(\frac{\log k}{k}\right) + o_p(A(n/k)) \end{aligned} \tag{7}$$

A few comments are in order. The first two lines of this expression characterise the first-order behaviour of the numerator. It can be seen that setting the regression shift factor η to $1/2$ eliminates the second and third term. However, the term $O_p\left(\log k/k^{1/2}\right)$ is still present. The asymptotic refinement due to second-order regular variation is given by the terms of line 3. Although $A(t) \to 0$ as $t \to \infty$, this decay might be slow: $A(t)$ is regularly varying with index ρ, and as ρ falls in magnitude the nuisance part of l in (1) decays more slowly. A slow decay then introduces a noticeable distortion in finite samples. We examine these distortions after stating the distributional theory for the estimator.

---

[2] The Burr distribution $F_{(\gamma,\rho)}(x) = 1 - (1 + x^{-\rho/\gamma})^{1/\rho}$ is a member of the Hall class with parameters γ and $\rho < 0$, $c = 1$ and $d = \gamma/\rho$, as is the Student t_δ distribution with δ degrees of freedom where $\gamma = 1/\delta$, $\rho = -2/\delta$, $d = \gamma BC^{-2\gamma}$, $B = -0.5\delta^2(\delta+1)/(\delta+2)$, and $C = \Gamma((\delta+1)/2)\delta^{(\delta-1)/2}/(\delta\pi)^{1/2}\Gamma(\delta/2)$ (valid for $\delta > 2$); so is the Fréchet distribution $F_\gamma(x) = \exp(-x^{-1/\gamma})$ with $\rho = -1$, $c = 1$, and $d = -.5\gamma$, and the Cauchy distribution with $\gamma = 1$, $\rho = -2$, $c = 1/\pi$, and $d = -0.5\pi^2$.

Distributional theory. Beirlant et al. (1996) observe that our slope estimator $\hat{\gamma}$, given by (3), is (to first order) a member of the class of kernel estimators discussed in Csorgo et al. (1985) with kernel $K(t) = 1 - \log t$. Since $\int_0^1 K(t)\mathrm{d}t = 2$ and not unity, a scale correction is required. Since $\int_0^1 K^2(t)\mathrm{d}t = 5$, the following result obtains as $k \to \infty$ and $k/n \to 0$, and if $\sqrt{k}A(n/k) \to 0$

$$\sqrt{k}(\hat{\gamma} - \gamma) \to^d N\left(0, \frac{5}{4}\gamma^2\right) \tag{8}$$

Higher order distortions. Asymptotically, the estimator is thus unbiased if $\sqrt{k}A(n/k) \to 0$. If this decay is slow, however, the estimator will suffer from a higher order distortion in finite samples. By (7), this distortion equals, for $\gamma > 0$ and $\rho < 0$,

$$b_{k,n} \equiv \frac{1}{2}\frac{\gamma}{\rho}\frac{2-\rho}{(1-\rho)^2}A(n/k) \tag{9}$$

In particular, in the Hall model, $A(t) = (\rho^2/\gamma)dt^\rho$. The sign of the higher order distortion of $N_{n,k}$ and hence $\hat{\gamma}$ is, since $\rho < 0$, then given by -sgn(d). For the Burr (Singh-Maddala), student, Fréchet, and Cauchy distributions it can be shown that $d < 0$, leading to a positive higher order distortion. We conclude that the higher order distortion induced by higher order regular variation is positive for many popular distribution -i.e., for which the nuisance function l in model (1) converges to a constant at a polynomial rate- leading to an overestimation of γ.[3]

Simulation evidence for these theoretical results is presented next. We also quantify the higher order distortions and the consequences for statistical inference about γ.

3. Numerical Illustrations

We illustrate numerically several of our results in a Monte Carlo study. First, we verify the distributional theory, then show that most of the empirical distortion is captured by the bias function $b_{k,n}$. At the same time, we show that the distortions can be sizeable, leading to substantial test size distortions, while a bias correction using $b_{k,n}$ would reconcile nominal and actual test sizes.

Our Monte Carlo study is based on the Burr distribution, a member of the Hall class, parametrised here as $F_{(\gamma,\rho)}(x) = 1 - (1 + x^{-\rho/\gamma})^{1/\rho}$ with parameters γ and $\rho < 0$. In the income distribution and inequality literature, this distribution is also know as the Singh-Maddala distribution, and used frequently in parametric income models. Specifically, we set $\gamma = 2/3$, and $\rho = -1/2$ to begin with. Qualitatively similar results are obtained for the student, Fréchet, and Cauchy distributions, all of which are members of the Hall class, and therefore not reported here. Since $1 < 1/\gamma < 2$ we consider a situation of fairly heavy tails (as second moments of the distribution do not exist). However, the qualitative insights depend little on the actual choice of γ. We have chosen $\rho = -1/2$ as our leading example since we are interested in the consequences of deviating from a strict Pareto model. As ρ falls in magnitude the nuisance part of l in (1) decays more slowly. This is illustrated in Figure 2, where we depict three Pareto QQ-plots for different ρ. For $\rho = -2$, the plot is almost linear throughout. The deviations from the strict Pareto model become increasingly more pronounced in the left part of the plot as ρ falls in magnitude.

For the simulation study, we draw $R = 1000$ samples of size $n = 10,000$ at first (then $n = 1000$), and consider the upper k order statistics. In order to choose a particular k, we follow standard practice and

[3] De Haan and Ferreira (2006) consider the merit of shifting the tail for tail quantile functions $U(t) = c_0 + c_1 t^\gamma + c_2 t^{\gamma+\tau} + o(t^{\gamma+\tau})$ where c_0 and c_2 are not zero, $c_1 > 0$, and $\gamma > 0$ and $\tau < 0$. It can then be shown that if $\tau < -\gamma$, the second order parameter satisfies $\rho = -\gamma$. A data shift that eliminates c_0 then results in $\rho = \tau$, so the post shift second order parameter has increased in magnitude, leading to a decrease in the induced distortion. However, the reverse reasoning also applies. In particular, the Hall model is $U(x) = cx^\gamma[1 + dx^\rho + o(x^\rho)]$. A data shift by c_0 yields $U(x) + c_0 = cx^\gamma[1 + (c_0/c)x^{-\gamma} + dx^\rho + o(x^\rho)]$, and increases the distortion if $\gamma \leq |\rho|$.

minimise the theoretical asymptotic Mean Squared Error (AMSE) (e.g., Hall 1982, or Beirlant et al. 1996), given by $b_{k,n}^2 + (1/k)(5/4)\gamma^2$, trading off distortion and dispersion. The theoretical higher order bias in $\hat{\gamma}$ induced by higher order regular variation in this Burr case is

$$b_{k,n} = \frac{1}{2}\gamma \frac{2-\rho}{(1-\rho)^2}\left(\frac{n}{k}\right)^{\rho}$$

which is, of course, increasing in k. The theoretical AMSE is minimised around $k^* = 200$, which also corresponds to the minimiser of the empirical AMSE based on the R samples. The mean of $\hat{\gamma}$ at this k^* is 0.739, and exceeds, as predicted by the theory, the population value $\gamma = 2/3$.

Figure 3 depicts the results. In panel (a) we illustrate the distributional theory, given by (8), for k^*, by plotting a kernel density estimate of $\sqrt{k^*}\hat{\gamma}$ (solid line), as well as a normal density with variance $(5/4)\gamma^2$, centered on the empirical mean of the simulated data. The two are in close agreement. The figure also implies that any inferential problems are due to location shifts. In panel (b) we contrast the empirical distortions (solid line) with $b_{k,n}$ (dashed line). $\hat{\gamma}$ overestimates γ, and the distortion increases in k. It is evident that most of the distortion is captured by $b_{k,n}$. In panel (c) we illustrate the consequences of the distortions for statistical inference, by plotting the empirical coverage error rates of the usual 95% symmetric confidence intervals. The higher order distortions lead to undermining inference because of the considerable size distortions. For instance, at k^*, the empirical coverage error rate is 30% for a nominal 5% rate. Shifting the estimate by $b_{k^*,n}$ reduces the coverage error rate to 7%.

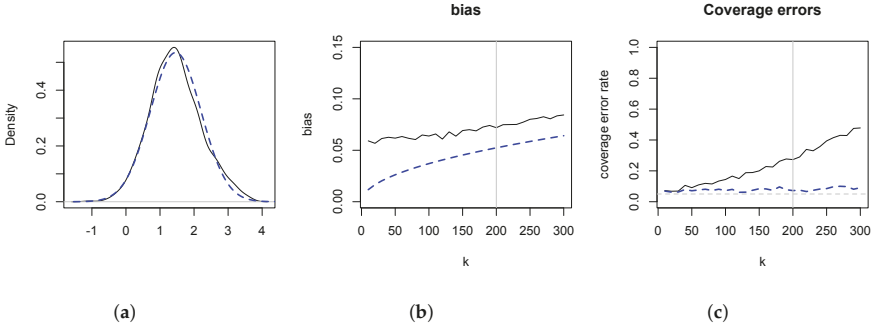


Figure 3. Bias and Inference: Burr. Monte Carlo study for the Burr distribution with parameters $\gamma = 2/3$ and $\rho = -0.5$. Based on samples of size n=10,000 and R=1000 repetitions. $k^* = 200$ minimises the asymptotic Mean Squared Error (AMSE), and is depicted by the vertical lines in panels b and c. Panel (**a**): Density plot of $\sqrt{k^*}\hat{\gamma}$ (solid line) and shifted normal density with variance $(5/4)\gamma^2$ (dashed line). Panel (**b**): empirical bias (solid line) and higher order bias function $b_{k,n}$ (dashed line). Panel (**c**): Coverage error rate of the usual 95% symmetric confidence intervals for nominal rate of 5%, with no bias correction (solid line) and correction by the theoretical $b_{k,n}$ (dashed line).

Next, we consider the role of the sample size n. Reducing the sample sizes in the Monte Carlo to $n = 1000$ yields results that are in line with the above theory, and therefore not depicted. The bias of $\hat{\gamma}$ increases by a factor predicted by the theory, namely $b_{k,1000}/b_{k,10,000} = 10^{1/2} = 3.16$. The optimal k^* shrinks by a factor of 4, as now $k^* = 50$. The density of $\sqrt{k^*}\hat{\gamma}$ is in good agreement with the theory, and empirical coverage error rates at this k^* are 32% for the uncorrected and 11% for the corrected estimator. The empirical coverage error rate for the uncorrected estimator rises steeply after k^*, reaching 64% at $k = 100$. Reducing the sample sizes further to 100 results in $k^* = 20$, and an empirical coverage error rate for the uncorrected estimator of 46% at this k^*. Biases are increased by a factor $b_{k,100}/b_{k,10,000} = 10$.

Finally, we illustrate the importance of the speed of decay in the nuisance function l of model (1). As ρ falls in magnitude, the nuisance function l decays more slowly. For the Burr case with $\gamma = 2/3$,

we depict in Figure 4 $b_{k,n}$ as ρ falls in magnitude for $n = 1,000$ and selected k. While for $\rho = -2$ the distortions are negligible (in line with Figure 2, it is evident that for small magnitudes of ρ the higher order distortions cannot be ignored).

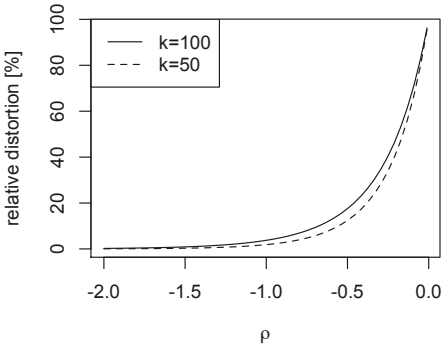


Figure 4. Relative distortions in the Burr model. Burr model with $\gamma = 2/3$ and $n = 1000$. Depicted is $100 * b_{k,n}/\gamma$ as ρ varies.

As the purpose of our simulation study is the provision of numerical evidence for our theory, we have used the theoretical bias function $b_{k,n}$ in the Burr case. When no such external knowledge is available, estimating the bias function requires non-parametric estimates of the second order parameter ρ and the function $A(\cdot)$. However, existing methods perform poorly, yielding excessively volatile estimates. The theory then informs a sensitivity analysis which is described in Section 4.1 in the context of our empirical application.

4. Empirical Illustration: Top incomes in the UK

Our empirical application uses administrative income tax return data are from the public-release files of the Survey of Personal Incomes (SPI) for the year 2009/10 (see e.g., Jenkins 2017 for a detailed description, and an analysis that includes rank size regressions). The SPI data underlie the UK top income share estimates in the World Top Incomes Database (WTID), and is a stratified sample of the universe of tax returns. The unit of taxation is the individual, and we use total taxable income as the income variable. The file contains 674,715 individuals, and we consider the n largest incomes.

In Figure 1 panel (a), we have depicted the Pareto QQ-plot for the 1000 largest incomes. It is evident that the data clearly reject a strict Pareto model: The plot exhibits a pronounced kink, and approximate linearity of the QQ plot only holds for the very highest upper order statistics. The function l in (1) captures this significant departure from the strict Pareto model. The Pareto QQ-plot thus conveys crucial information that is usually ignored by practitioners in economics, making it a key diagnostic device. For instance, a common mechanical approach is to set k by choosing 'blindly' (i.e., without reference to the Pareto QQ-plot) e.g., the top 1% or the top 1000 observations. Since the approximate linearity only obtains for about the 70 largest observations, the estimate of the slope parameter of the Pareto QQ-plot, i.e., the OLS estimator (3), will be severely biased if k is set to 1000 or higher. This is illustrated in panel (b) of the figure: The estimates fall for higher values of k, since the estimation procedure then attributes increasing weights to the left of the kink in the Pareto QQ-plot.

In the light of these observations, we restrict our subsequent analysis to the range of k in which the Pareto QQ-plot is approximately linear. We confirm this in Figure 5 panel (a), having restricted the plot to the $n = 70$ highest incomes. The plot now appears fairly linear. In panel (b), we depict the regression estimates $\hat{\gamma}$ and the 95% symmetric pointwise confidence intervals. One first visual way of

choosing an estimate is to consider an area of the plot where the estimate is fairly stable (as is done by inspecting Hill or so-called alternative Hill plots) and picking the largest such k since the variance of the estimate falls in k. Such subjective choice would be around $k = 60$ with an estimate of $\hat{\gamma} = 1.070$ (indicated by the horizontal faint line in the figure).[4] Overall, the visual method would suggest an estimate of γ between 0.9 and 1, implying very heavy tails. Taking into consideration the variability of the estimate, one cannot reject the hypothesis that the tail index be unity, i.e., Zipf's law. Returning to panel (a) we have also plotted the line with slope 1. This line does well in describing the data. We turn to a method that permits an objective choice of a particular k, and examine the remaining distortions in the estimate of γ.

4.1. Sensitivity Analysis, and the Choice of k

The preceding analysis has shown that $\hat{\gamma}$ is likely to suffer from positive higher order distortions, captured by $b_{k,n}$. Estimating this bias function requires non-parametric estimates of the second order parameter ρ and the function $A(\cdot)$, but existing methods perform poorly, yielding excessively volatile estimates. Hence we limit ourselves to a sensitivity analysis, taking ρ as a sensitivity parameter, whose objective is to gauge plausible values of the potential distortions based on diagnostics of the rank size regression. This approach is sketched next.

Following Beirlant et al. (1996), we observe that the mean weighted theoretical squared deviation

$$\frac{1}{k}\sum_{j=1}^{k} w_{j,k} E\left(\log\left(\frac{X_{n-j+1,n}}{X_{n-k,n}}\right) - \gamma \log\left(\frac{k+1}{j}\right)\right)^2$$

equals, to first order,

$$c_k Var(\hat{\gamma}) + d_k(\rho) b_{k,n}^2 \tag{10}$$

for some coefficients c_k depending only on k, and $d_k(\rho)$ depending on k and ρ (these are stated explicitly in the Appendix A). Set $w_{j,k} \equiv 1$. An estimate of the mean theoretical deviation is the mean of the squared residuals $k^{-1}SSR_k$ of the rank size regression. In view of the usual bias-variance trade-off for our estimator $\hat{\gamma}$ for fixed n, we ascribe all the measured deviation $k^{-1}SSR_k$ to the bias, thereby defining a very conservative bound, and let

$$\tilde{b}_{k,n}(\rho) = [k^{-1}SSR_k / d_k(\rho)]^{1/2}$$

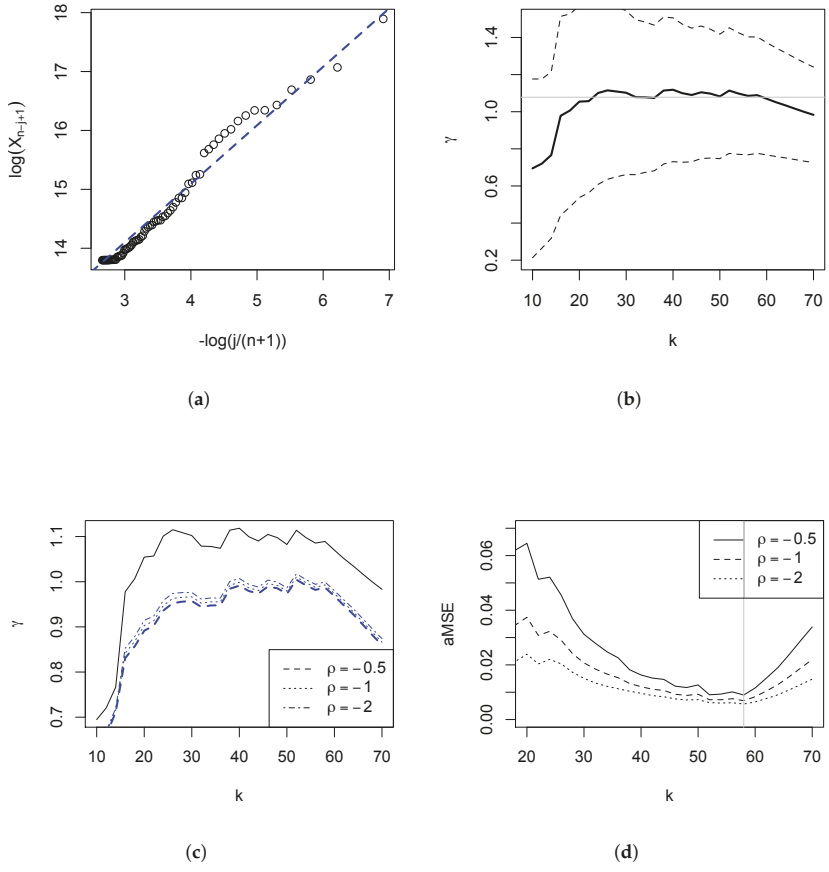
This conservative sensitivity analysis then consists of examining $\hat{\gamma} - \tilde{b}_{k,n}(\rho)$ for a range of values of ρ.

Figure 5 panel (c) reports the results of such a sensitivity analysis for k being restricted to the $n = 70$ highest incomes. Since under this restriction the Pareto QQ-plot is approximately linear, we expect that the remaining distortions are fairly modest. This is borne out in the sensitivity plot, as the precise value of ρ now plays only a minor role.

Should a researcher wish to choose a particular k by minimising an approximation to the AMSE, Equation (10) is the basis of the procedure proposed in Beirlant et al. (1996): Apply two weighting schemes $w_{j,k}^{(i)}$ $(i = 1, 2)$, estimate the corresponding two mean weighted theoretical deviations using the residuals, and compute a linear combination thereof such that $Var(\hat{\gamma}) + b_{k,n}^2$ obtains. We have carried out this programme (see Appendix A for further details) for weights $w_{j,k}^{(1)} \equiv 1$ and $w_{j,k}^{(2)} = j/(k+1)$ for given ρ, and Figure 5 panel (d) depicts the results. Minimising this approximation to the AMSE yields $k^*(\rho)$, which, for $\rho \in \{-2, -1, -0.5\}$, resulted in $k^* = 58$ across the selected ρ, for which $\hat{\gamma}_{k^*} = 1.089$ obtains. In view of the results depicted in panel (c) it is not surprising that changing ρ has only a small

[4] Alternative estimators lead to similar conclusions. For instance, using the classic Hill estimator, at $k = 60$ an estimate of γ of 1.017 is obtained. The plot (not displayed here) is fairly stable around this value for $k \in [20, 60]$.

effect. This estimate of γ is very close to the subjective visual choice of $\hat{\gamma}$ of 1.075, reported above, based on Figure 5b.



(**a**) (**b**)

(**c**) (**d**)

Figure 5. Estimates of γ. Based on Survey of Personal IncomeS (SPI) data for 2009/10. Panel (**a**) Pareto QQ plot for the largest 70 incomes. The dashed line has slope 1. Panel (**b**): Estimates of $\hat{\gamma}$ (solid line) and the 95% symmetric pointwise confidence interval (dashed line). The faint horizontal line at 1.075 is subjectively chosen. Panel (**c**): Sensitivity analysis. Plot of $\hat{\gamma}$ (solid line) and $\hat{\gamma} - \tilde{b}_{k,n}(\rho)$ for a different values of ρ. Panel (**d**): Approximation to the AMSE for different values of ρ. Minimising AMSE yields $k^* = 58$ (vertical line) across the selected ρ, for which $\hat{\gamma}_{k^*} = 1.089$ obtains.

5. Conclusions

The OLS estimator of the slope coefficient in the rank size regression (shifted or unshifted) can suffer significant higher order distortions that arise from the slow decay of the nuisance function l in the model $1 - F(x) = x^{-\frac{1}{\gamma}} l(x)$ for $\gamma > 0$. Modeling the tail as second order regular variation, we have shown that the estimator over-estimates the true value in models in which l converges to a constant at a polynomial rate (i.e., in the leading heavy-tailed distributions). Our numerical illustrations have shown that these distortions can be dramatic, leading to test size distortions in which actual error rates are multiples of nominal error rates. The empirical illustration based on the Pareto QQ-plot has revealed a further distortion, namely the presence of a pronounced kink. Figure 1 has revealed that

using the common rule to choose 1% of the observation for tail estimation would lead to a severe under-estimation of how heavy the tail is.

The higher order distortions are functions of $A(\cdot)$ and the second order regular variation parameter ρ. Since existing methods usually result in poor estimates of these, reliable bias corrections are not feasible. In view of this we have proposed a sensitivity analysis based on diagnostics from the rank size regression. When applied to our data on top incomes, we still cannot reject the hypothesis γ be unity, a situation often described in several fields as Zipf's law (e.g., Schluter and Trede 2017).

The simplicity of the regression estimator is undoubtedly the principal reason for its popularity among practitioners in economics. This paper has shown that in many situations the naive (i.e., 'blind') use of this estimator should be considered with care: Pareto QQ-plot, the sensitivity plot and the AMSE plot convey jointly important information about the behaviour of the estimator.

Acknowledgments: I thank the referees for their constructive comments that have helped to improve the paper.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Proofs

Before proving the main result given by (7), we consider first the behaviour of the numerator $N_{n,k}$ under first-order regular variation (5). We then refine the asymptotic expansion by assuming that the second-order regular variation (6) holds.

First-order asymptotic expansion of the numerator $N_{n,k}$. Assume that (5) holds, and consider an intermediate sequence $k = k_n$ of positive integers such that $k_n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$. It will be shown that

$$\begin{aligned} N_{n,k}/\gamma &= 2 - \left(\frac{1}{2} - \eta\right)\frac{\log(k-\eta)}{k} - \left(\frac{1}{2} - \eta\right)\frac{\log^2 k}{2k} \\ &+ O_p\left(\frac{1}{k^{1/2}}\right) + O\left(\frac{1}{k}\right) + O_p\left(\frac{\log k}{k^{1/2}}\right) \end{aligned} \tag{A1}$$

Remark: The term $O_p\left(\frac{\log k}{k^{1/2}}\right)$ dominates $(\log k)/k$, and is not eliminated by setting the shift factor η to $1/2$.

In the proof of (A1) we will make use of the following Euler Maclaurin formulae (e.g., Gabaix and Ibragimov 2011, Equations A.4 and A.5)

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k}\log^2(i-\eta) &= 2 + \frac{k-\eta}{k}\log^2(k-\eta) - 2\frac{k-\eta}{k}\log(k-\eta) + \frac{\log^2(k-\eta)}{2k} + O\left(\frac{1}{k}\right) \\ &= 2 + \log(k-\eta)(\log(k-\eta)-2) + O\left(\frac{\log^2 k}{k}\right) \end{aligned} \tag{A2}$$

and

$$\frac{1}{k}\sum_{i=1}^{k}\log(i-\eta) = -1 + \log(k-\eta) + \left(\frac{1}{2} - \eta\right)\frac{\log(k-\eta)}{k} + O\left(\frac{1}{k}\right) \tag{A3}$$

Proof of (A1). We adapt the proofs of Kratz and Resnick (1996) (KR henceforth) of their Equations 2.4 and 2.8. The key is the use of Renyi's representation of exponential order statistics, which implies (e.g., KR, p. 705)

$$E_{n-k+i,n} - E_{n-k,n} =^d \sum_{j=k-i+1}^{k}\frac{E_j}{j}$$

where E_j ($j = 1, \cdots, n$) denote iid unit exponential random variables, and $E_{n-k,n}$ denotes the $(n-k)$-th order statistic. We obtain an asymptotic refinement by using, instead of KR's Lemmas 2.2 and 2.3, the above Euler Maclaurin formulae, and Lyapunov's central limit theorem (CLT). Our numerator is denoted there by A_n, and the indices are mapped by setting $i = k+1-j$. From KR (pp. 704-707), we have

$$\begin{aligned} N_{n,k}/\gamma &=^d \frac{1}{k}\sum_{i=1}^{k} - \log(\frac{i-\eta}{k+1})\sum_{j=i}^{k}\frac{E_j}{j} + o_p(1/k) \\ &= \log(k+1)\bar{E}_k - \frac{1}{k}\sum_{j=1}^{k} E_j\left[\frac{1}{j}\sum_{i=1}^{j}\log(i-\eta)\right] + o_p(1/k) \end{aligned} \tag{A4}$$

where $\bar{E}_k = (1/k)\sum_j^k E_j$.

We first show that KR's result (A4) can also be derived from the first order regular variation condition under the stated assumptions. Let Y denote a standard Pareto random variable, and denote the $(n-k)$-th order statistic by $Y_{n-k,n}$. Consider the scaled log excesses

$$\frac{\log X_{n-i+1,n} - \log X_{n-k,n}}{a(Y_{n-k,n})/U(Y_{n-k,n})}$$

where $a(.)$ and $U(.)$ are defined in representation (5). Then, noting that $X_{i,n} =^d U(Y_{i,n})$, and using (5) with $t = Y_{n-k,n}$ and $x = Y_{n-i+1,n}/Y_{n-k,n}$, the scaled log excesses satisfy as $n \to \infty$ and $n/k \to \infty$

$$\begin{aligned} \frac{\log X_{n-i+1,n} - \log X_{n-k,n}}{a(Y_{n-k,n})/U(Y_{n-k,n})} &=^d \frac{\log U(Y_{n-i+1,n}) - \log U(Y_{n-k,n})}{a(Y_{n-k,n})/U(Y_{n-k,n})} \\ &= \log\left(\frac{Y_{n-i+1,n}}{Y_{n-k,n}}\right) + o_p(1) \end{aligned}$$

By Renyi's representation of exponential order statistics, we have $Y_{n-i+1,n}/Y_{n-k,n} =^d Y_{k-i+1,k}$, so

$$\log\left(\frac{Y_{n-i+1,n}}{Y_{n-k,n}}\right) =^d \log\left(Y_{k-i+1,k}\right) =^d E_{k-i+1,k} =^d \sum_{i=j}^{k}\frac{E_i}{i}$$

since, using Renyi's representation again, $E_{k-j+1,k} =^d E_{1,k} + \sum_{i=j}^{k-1}\frac{E_i}{i} = \sum_{i=j}^{k}\frac{E_i}{i}$. From Wellner (1978), we know that $\frac{k}{n}Y_{n-k,n} \to^p 1$, so $a(Y_{n-k,n})/U(Y_{n-k,n}) \to \gamma$. Using the definition of $N_{n,k}$, on combining the results we thus obtain

$$N_{n,k}/\gamma =^d \left[\frac{1}{k}\sum_{j=1}^{k}(-1)\log\left(\frac{j-\eta}{k+1}\right)\sum_{i=j}^{k}\frac{E_i}{i}\right] + o_p(1/k).$$

as claimed.

We proceed to examine (A4). Using (A3) yields

$$\begin{aligned} N_{n,k}/\gamma &= \log(k+1)\bar{E}_k + \bar{E}_k - \frac{1}{k}\sum_{j=1}^{k} E_j\log(j-\eta) \\ &- \left(\frac{1}{2}-\eta\right)\frac{1}{k}\sum_{j=1}^{k}E_j\frac{\log(j-\eta)}{j} - \frac{1}{k}\sum_{j=1}^{k}E_j O\left(\frac{1}{j}\right) \end{aligned}$$

By Lyapunov's CLT,

$$\frac{k^{1/2}}{\log k}\left[\frac{1}{k}\sum_{j=1}^{k}(E_i-1)\log(j-\eta)\right]\to^d N(0,1)$$

so $(1/k)\sum_{j=1}^{k}E_i\log(j-\eta)=O_p\left(\frac{\log k}{k^{1/2}}\right)+(1/k)\sum_{j=1}^{k}\log(j-\eta)$. Using again (A3) and substituting the result, we obtain

$$\begin{aligned}N_{n,k}/\gamma &= \bar{E}_k+1+\log(k+1)\bar{E}_k-\log(k-\eta)\\ &- \left(\frac{1}{2}-\eta\right)\frac{\log(k-\eta)}{k}+O\left(\frac{1}{k}\right)+O_p\left(\frac{\log k}{k^{1/2}}\right)\\ &- \left(\frac{1}{2}-\eta\right)\frac{1}{k}\sum_{j=1}^{k}\frac{\log(j-\eta)}{j}-\frac{1}{k}\sum_{j=1}^{k}O\left(\frac{1}{j}\right)\\ &- \left(\frac{1}{2}-\eta\right)\frac{1}{k}\sum_{j=1}^{k}(E_j-1)\frac{\log(j-\eta)}{j}-\frac{1}{k}\sum_{j=1}^{k}(E_j-1)O\left(\frac{1}{j}\right)\end{aligned}$$

Note that $\bar{E}_k+1=2+O_p(k^{-1/2})$ and $\log(k+1)(\bar{E}_k-1)=O_p\left(\frac{\log k}{k^{1/2}}\right)$. $(1/k)\sum_{j=1}^{k}\frac{\log(j-\eta)}{j}$ is a Riemann approximation to the integral $k^{-1}\int_1^k(\log x-\eta)/x\mathrm{d}x=(\log^2 k)/2k$, and $(1/k)\sum_{j=1}^{k}\frac{1}{j}$ is a Riemann approximation to the integral $k^{-1}\int_1^k(1/x)\mathrm{d}x=(\log k)/k$. By Lyapunov's CLT, $(k/\sqrt{2})[(1/k)\sum_{j=1}^{k}(E_j-1)\frac{\log(j-\eta)}{j}]\to^d N(0,1)$, so $(1/k)(E_j-1)\frac{\log(j-\eta)}{j}=O_p(1/k)$. Similarly, the last term is $O_p(1/k^2)$. Hence

$$\begin{aligned}N_{n,k}/\gamma &= 2-\left(\frac{1}{2}-\eta\right)\frac{\log(k-\eta)}{k}-\left(\frac{1}{2}-\eta\right)\frac{\log^2 k}{2k}\\ &+ O_p\left(\frac{1}{k^{1/2}}\right)+O\left(\frac{1}{k}\right)+O_p\left(\frac{\log k}{k^{1/2}}\right),\end{aligned}$$

which is Equation (A1), as claimed.

Before refining the asymptotic expansion, we briefly consider:

Proof of (4). $D_k=2+O\left(\frac{\log^2 k}{k}\right)$. Expanding the quadratic in the definition of D_k

$$D_k=\log^2(k+1)-2(\log k+1)\frac{1}{k}\sum_{i=1}^{k}\log(i-\eta)+\frac{1}{k}\sum_{i=1}^{k}\log^2(i-\eta)$$

and using the Euler Maclaurin formulae yields the stated result.

We are now in a position to examine the behaviour of the numerator $N_{n,k}$ under second order regular variation.

Proof of the higher order expansion (7). Consider the scaled log excesses again, this time using representation (6) instead of (5). Set again $t=Y_{n-k,n}$ and $x=Y_{n-i+1,n}/Y_{n-k,n}$, and recall $Y_{n-i+1,n}/Y_{n-k,n}=^d Y_{k-i+1,k}$. Hence we obtain the higher order expression

$$\begin{aligned}\frac{\log X_{n-i+1,n}-\log X_{n-k,n}}{a(Y_{n-k,n})/U(Y_{n-k,n})} &=^d \sum_{i=j}^{k}\frac{E_i}{i}\\ &+ A(Y_{n-k,n})H_{\gamma,\rho}(Y_{k-i+1,k})+o_p(1).\end{aligned}\tag{A5}$$

The role of the first term on the right for $N_{n,k}$ has already been described above. In what follows, we consider the higher order term. Since $H_{\gamma>0,\rho<0}(x) = \frac{1}{\rho}(\frac{x^\rho-1}{\rho} - \log x)$, the higher order expansion of $N_{n,k}$ requires the analysis of

$$\frac{1}{k}\sum_{i=1}^{k}(-1)\log\left(\frac{i-\eta}{k+1}\right)\left[\frac{Y_{k-i+1,k}^{\rho}-1}{\rho}\right] = \log(k+1)\bar{Y}_\rho - \frac{1}{k}\sum_{i=1}^{k}(\log i-\eta)\left[\frac{Y_{k-i+1,k}^{\rho}-1}{\rho}\right]$$

where $\bar{Y}_\rho = (1/k)\sum_{i=1}^{k}\frac{Y_i^\rho-1}{\rho}$. $\bar{Y}_\rho$ has expectation $(1-\rho)^{-1}$, so by the CLT $\bar{Y}_\rho = (1-\rho)^{-1} + O_p(k^{-1/2})$. To handle the last sum, note that $Y_{k-j+1,k} =^d \exp(E_{k-j+1,k}) =^d (V_{j,k})^{-1}$ where V denotes a standard uniform random variable, and we replace the order statistic $V_{j,k}$ by its expectation, $V_{j,k} = j/(k+1) + O_p(k^{-1/2})$. A Taylor series expansion then gives $(V_{j,k}^{-1})^\rho = (k+1/j)^\rho + O_p(k^{-1/2})$. Then

$$\begin{aligned}\frac{1}{k}\sum_{i=1}^{k}(-1)\log\left(\frac{i-\eta}{k+1}\right)\left[\frac{Y_{k-i+1,k}^{\rho}-1}{\rho}\right] &= \frac{\log(k+1)}{1-\rho} - \frac{1}{\rho}\frac{(k+1)^\rho}{k}\sum_{i=1}^{k}(\log i-\eta)j^{-\rho} \\ &+ \frac{1}{\rho}\frac{1}{k}\sum_{i=1}^{k}\log(i-\eta) + O_p\left(\frac{\log k}{k^{1/2}}\right)\end{aligned}$$

For the third term on the rhs, we use the Euler Maclaurin (A3), for the second term on the rhs we have the following Euler Maclaurin

$$\frac{1}{k}\sum_{j=1}^{k}j^{-\rho}\log(j-\eta) = \frac{1}{1-\rho}k^{-\rho}\log(k-\eta) - \left(\frac{1}{1-\rho}\right)^2 k^{-\rho} + o(k^{-\rho})$$

Combing these two Euler Maclaurin formulae, we can simplify to get

$$\begin{aligned}\frac{1}{\rho}\frac{(k+1)^\rho}{k}\sum_{i=1}^{k}(\log i-\eta)j^{-\rho} - \frac{1}{\rho}\frac{1}{k}\sum_{i=1}^{k}\log(i-\eta) &= \frac{\log k-\eta}{1-\rho} - \frac{1}{\rho}\frac{1}{(1-\rho)^2} + \frac{1}{\rho} + O\left(\frac{\log k}{k}\right) \\ &= \frac{\log k-\eta}{1-\rho} - \frac{2-\rho}{(1-\rho)^2} + O\left(\frac{\log k}{k}\right)\end{aligned}$$

Therefore[5]

$$\frac{1}{k}\sum_{i=1}^{k}(-1)\log\left(\frac{i-\eta}{k+1}\right)\left[\frac{Y_{k-i+1,k}^{\rho}-1}{\rho}\right] = \frac{2-\rho}{(1-\rho)^2} + O_p\left(\frac{\log k}{k}\right) \tag{A6}$$

We are now in a position to combine the results. In order to simplify notation, denote the first order expansion of the numerator $N_{n,k}/\gamma$ by $N_{1,n,k}/\gamma$, given by the rhs of (A1). Then substituting the higher order expression for the scaled excesses (A5) into the formula for $N_{n,k}$, recalling that $\frac{k}{n}Y_{n-k,n} \to^p 1$ (Wellner 1978), and using (A6) yields

$$N_{n,k}/\gamma = N_{1,n,k}/\gamma + A\left(\frac{n}{k}\right)\frac{1}{\rho}\left[\frac{2-\rho}{(1-\rho)^2}\right] + O_p\left(\frac{\log k}{k}\right) + o_p(A(n/k)).$$

Proof of (8). The class of kernel estimator considered in Csorgo et al. (1985) is of the form

$$\hat{\gamma}_{kernel} = \frac{\sum_{j=1}^{k}(j/k)K(j/k)[\log X_{n-j+1} - \log X_{n-j}]}{\int_0^1 K(t)\mathrm{d}t}$$

[5] To support this key expression, numerical evidence from a Monte Carlo with $k = 1000$, 1000 samples, and $\eta = 0$ yielded for the lhs of (15) v. $(2-\rho)/(1-\rho)^2$ the following: $\rho = -1/2$: 1.105 v. 1.111, $\rho = -1$: .746 v. 0.75, $\rho = -2$: 0.443 v. 0.444.

Their Theorem 2 (or Theorem 1.1 in Beirlant et al. 1996) states the following. Under general conditions on kernel K and distribution function F so that there exists a nonrandom sequence C_n such that $C_n(\hat{a}-\gamma)$ converges weakly to a limiting $N(0,1)$ distribution for some sequence $k=k_n\to\infty$ with $k_n/n\to 0$, it is necessary and sufficient that

$$\lim_{n\to\infty}\sqrt{k}\int_0^1 b(kw/n)K(w)\mathrm{d}w=0$$

where b is a function such that $b(1/x)\to 0$ as $x\to\infty$, and that for the tail quantile function $U(x)=x^{\gamma}\tilde{l}(x)$ with

$$\tilde{l}(x)=c(x)\exp\left(\int_{1/x}^1\frac{b(u)}{u}\mathrm{d}u\right)$$

where $c(x)\to c$ as $x\to\infty$. If this condition is satisfied, then as $k=k_n\to\infty$ and $k_n/n\to 0$

$$\sqrt{k}\left(\int_0^1 K^2(v)\mathrm{d}v\right)^{-1/2}(\hat{\gamma}_{kernel}-\gamma)\to^d N(0,\gamma^2).$$

Beirlant et al. (1996) observe that our slope estimator $\hat{\gamma}$, given by (3), is (to first order) a member of the class of kernel estimators $\hat{\gamma}_{kernel}$ with kernel $K(t)=1-\log t$, and that the above condition holds under the regular variation hypothesis. Turning to the specific kernel $K(t)=1-\log t$, since $\int_0^1 K(t)\mathrm{d}t=2$ and not unity, a scale correction is required. As $\int_0^1 K^2(t)\mathrm{d}t=5$, the stated result (8) follows.

Proof of (10). Consider the mean weighted theoretical squared deviation

$$\frac{1}{k}\sum_{j=1}^k w_{j,k}E\left(\log\left(\frac{X_{n-j+1,n}}{X_{n-k,n}}\right)-\gamma\log\left(\frac{k+1}{j}\right)\right)^2$$

for some weights $w_{j,n}$. Using (A5) this equals, to first order,

$$\frac{\gamma^2}{k}\sum_{j=1}^k E\left(\left(\sum_{i=j}^k\frac{E_i}{i}-\log\left(\frac{k+1}{j}\right)\right)+A(Y_{n-k,n})H_{\gamma,\rho}(Y_{k-i+1,k})\right)^2$$

Then, recalling that $Y_{k-j+1,k}=^d (V_{j,k})^{-1}$ and proceeding as in Beirlant et al. (1996, Section 4), which involves approximating expectations $E(f(V_{j,k}))$ by the leading term $f(j/(k+1))$ when applying the delta method yields, to first order,

$$\frac{\gamma^2}{k}\tilde{c}_k+d_k(\rho)b_{k,n}^2$$

with

$$\tilde{c}_k=\sum_{j=1}^k w_{j,k}\left(\sum_{l=1}^{k-j+1}\left(\frac{1}{k-l+1}\right)^2+\left(\sum_{l=1}^{k-j+1}\frac{1}{k-l+1}-\log\left(\frac{k+1}{j}\right)\right)^2\right)$$

and $d_k(\rho)=(\frac{1}{2}\frac{2-\rho}{(1-\rho)^2})^{-2}\tilde{d}_k(\rho)$ with

$$\tilde{d}_k(\rho)=\frac{1}{k}\sum_{j=1}^k w_{j,k}\left(\frac{(j/(k+1))^{-\rho}-1}{\rho}\right)^2$$

Finally, we set $c_k=(4/5)\tilde{c}_k$ and $w_{j,k}\equiv 1$ to arrive at (10).

Remark: In order to obtain an estimate of the AMSE, Beirlant et al. (1996) use two weighting schemes,

namely $w_{j,k} \equiv 1$ leading to coefficients, say, c_k^1 and d_k^1 and mean weighted squared residuals $k^{-1}SSR_k^1$, and $w_{j,k} = j/(k+1)$ leading to c_k^2, d_k^2, and $k^{-1}SSR_k^2$. Then a linear combination of two approximate MSE expressions (with coefficients, say, x and y) is sought that yields $Var(\hat{\gamma}) + b_{k,n}^2$, which is achieved by solving simultaneously the equations

$$\begin{aligned} xc_k^1 + yc_k^2 &= 1 \\ xd_k^1 + yd_k^2 &= 1. \end{aligned}$$

References

Atkinson, Anthony Barnes. 2017. Pareto and the upper tail of the income distribution in the UK: 1799 to the present. *Economica* 84: 129–56.

Beirlant, Jan, Petra Vynckier, and Jozef L. Teugels. 1996. Tail index estimation, Pareto quantile plots, and regression diagnostics. *Journal of the American Statistical Association* 9: 1659–67.

Beirlant, Jan, Yuri Goegebeur, Johan Segers, and Jozef L. Teugels. 2004. *Statistics of Extremes*. Wiley Series in Probability and Statistics. Chichester: Wiley.

Burkhauser, Richard V., Shuaiz hang Feng, Stephen Jenkins, and Jeff Larrimore. 2012. Recent trends in top income shares in the USA: Reconciling estimates from March CPS and IRS tax return data. *Review of Economics and Statistics* 94: 371–88.

Cowell, Frank A. 1989. Sampling variances and decomposable inequality measures. *Journal of Econometrics* 42: 27–41.

Cowell, Frank A., and Emmanuel Flachaire. 2007. Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics* 141: 1044–72.

Csorgo, Sandor, Paul Deheuvels, and David Mason. 1985. Kernel Estimates of the Tail Index of a Distribution. *The Annals of Statistics* 13: 1050–77.

Davidson, Russell, and Emmanuel Flachaire. 2007. Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics* 141: 141–66.

De Haan, Laurens, and Ana Ferreira. 2006. *Extreme Value Theory*. New York: Springer.

De Haan, Laurens, and Ulrich Stadtmüller. 1996. Generalized regular variation of second order. *Journal of the Australian Mathematical Society (Series A)* 61: 381–95.

Dekkers, Arnold L. M., John H. J. Einmahl, and Laurens de Haan. 1989. A moment estimator for the index of an extreme-value distribution. *Annals of Statistics* 17: 1833–55.

Embrechts, Paul, Claudia Kluppelberg, and Thomas Mikosch. 1997. *Modelling Extremal Events*. Berlin: Springer.

Gabaix, Xavier, and Rustam Ibragimov. 2011. Rank - 1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business and Economic Statistics* 29: 24–39.

Hall, Peter. 1982. On some simple estimate of an exponent of regular variation. *Journal of the Royal Statistical Society Ser. B* 44: 37–42.

Kratz, Marie, and Sidney I. Resnick. 1996. The QQ-estimator and heavy tails. *Communications in Statistics. Stochastic Models* 12: 699–724.

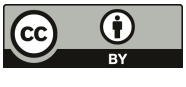
Jenkins, Stephen P. 2017. Pareto models, top incomes and recent trends in UK income inequality. *Economica* 84: 261–89.

Schluter, Christian, and Mark Trede. 2002. Tails of Lorenz curves. *Journal of Econometrics* 109: 151–66.

Schluter, Christian, and Mark Trede. 2017. Size distributions reconsidered. *Econometric Reviews*, forthcoming.

Schultze, J., and J. Steinebach. 1996. On least squares estimates of an exponential tail coefficient. *Statistics and Decisions* 14: 353–72.

Wellner, Jon A. 1978. Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Zeitschrift fuer Wahrscheinlichkeitstheorie und verwandte Gebiete* 45: 73–88.



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC Data

Vladimir Hlasny [1,*] and Paolo Verme [2]

[1] Department of Economics, Ewha Womans University, Seoul 03760, Korea
[2] The World Bank, Washington, DC 20433, USA; pverme@worldbank.org
* Correspondence: vhlasny@ewha.ac.kr

Received: 1 January 2018; Accepted: 23 May 2018; Published: 4 June 2018

Abstract: It is sometimes observed and frequently assumed that top incomes in household surveys worldwide are poorly measured and that this problem biases the measurement of income inequality. This paper tests this assumption and compares the performance of *reweighting* and *replacing* methods designed to correct inequality measures for top-income biases generated by data issues such as unit or item non-response. Results for the European Union's Statistics on Income and Living Conditions survey indicate that survey response probabilities are negatively associated with income and bias the measurement of inequality downward. Correcting for this bias with *reweighting*, the Gini coefficient for Europe is revised upwards by 3.7 percentage points. Similar results are reached with *replacing* of top incomes using values from the Pareto distribution when the cut point for the analysis is below the 95th percentile. For higher cut points, results with *replacing* are inconsistent suggesting that popular parametric distributions do not mimic real data well at the very top of the income distribution.

Keywords: top incomes; inequality measures; survey non-response; Pareto distribution; parametric estimation; EU SILC

JEL Classification: D31; D63; N35

1. Introduction

Thanks to the wide public attention that top incomes have received in the aftermath of the global financial crisis, it is now acknowledged that top incomes have grown disproportionally faster than other incomes in industrialized countries over the past several decades. The fact that these top incomes are difficult to capture in household surveys potentially leads to biases in the estimation of income inequality related to the representation and precision of reported top incomes, even though the direction of the bias is not a priori clear (Deaton 2005, p. 11). These range from issues related to sampling, to issues related to data collection, data preparation or data analysis. The European Union Survey of Income and Living Conditions, for example, suffers from data issues such as under-representation of the highest incomes (Bartels and Metzing 2017; Törmälehto 2017). Most countries in Europe suffer from very high non-response rates reaching up to 50 percent of the sample. Income measurement issues including surveying, interview methods and post-survey treatment also explain differences in inequality measurements across data sources (Frick and Krell 2010).

Two types of in-survey methods have been proposed to address the question of correcting inequality in the presence of top-income biases while relying on survey microdata only. The first method, which we call *reweighting*, attempts to correct the sampling weights of existing observations using information on unit or item non-response rates across demographic cells such as geographical areas (Mistiaen and Ravallion 2003; Korinek et al. 2006, 2007). The approach exploits the relationship

between response rates and shapes of income distributions across national regions to estimate the gradient of households' response probability by income level. It then uses the estimated response probabilities to reweight the observed incomes by the mass of nonresponding households in order to correct the measure of inequality. The second method, which we call *replacing* attempts to replace top-income observations with observations generated from known theoretical distributions. This method can be used to correct for issues such as top coding, trimming or censoring but can also mitigate the problem of unit or item non-responses if these non-responses are concentrated among top incomes (Cowell and Victoria-Feser 2007; Jenkins et al. 2011). Several distributions have been suggested as candidates, including Pareto type I or type II, or generalized beta.[1] Hlasny and Verme (2018) have combined the reweighting and replacing methods, and studied the contribution of each method to the composite correction of an inequality index.

It is evident that both the reweighting and replacing methods have their advantages and disadvantages, as the information available within surveys has its limits even if used creatively to correct for top-income problems. Proper reweighting and replacing depend on the appropriateness of parametric assumptions imposed on a particular national distribution of incomes at hand. Using alternative methods based on out of survey information such as tax records or national accounts data to inform the measurement of top incomes has its own measurement problems. Good tax or macro data are only available in a few countries and data may not be comparable across countries, whereas household survey data of reasonable quality are now available in most countries worldwide.

This paper compares the reweighting and replacing methods using the European Union's Statistics on Income and Living Conditions (SILC) survey data, taking into account heterogeneity of income distributions, differences in sampling designs and definitions of non-response rates across EU member states. We find survey non-response probabilities to be negatively and significantly associated with income indicating that measures of inequality are downward biased. Correcting for this bias with reweighting, the Gini coefficient for Europe is revised upwards by 3.7 percentage points. Similar results are reached with replacing of top incomes using values from the Pareto distribution when the cut point for replacing is set below the 95th percentile. For higher cut points, results with replacing are inconsistent suggesting that popular parametric distributions do not mimic well real data at the very top of the income distribution.

The paper is organized as follows. The next section discusses measurement issues related to top incomes. The following section outlines the main methods used to correct for top-income biases related to unit non-response. Section 3 describes the data. Section 4 presents main results and Section 5 concludes.

2. Materials and Methods

Problems related to top-income data may be due to sample design, data collection, data preparation or data analysis. We introduce these four typologies of errors in turn clarifying the type of error we address in this paper.

Sample design issues emerge when the sampling is designed in such a way that top incomes cannot be captured by design. This can occur, for example, when the sampling is done poorly or when the population census is old or the master sample has not been updated to capture newly constructed wealthy areas. If detected, some of these issues can be corrected post-survey by reweighting the sample, but either detecting or correcting these problems post-survey is not simple. It is important to note here that we should not expect exceptionally high incomes to be captured in household sample

[1] Similar methods include Lakner and Milanovic (2013) who combined corrections for unit non-response with corrections for measurement errors among top incomes, and calibrated the estimated Pareto distribution among top incomes using aggregate income information from national accounts data. Bartels and Metzing (2017) replaced the top one percent of incomes in the EU Statistics on Income and Living Conditions (SILC) surveys with Pareto estimates obtained using World Wealth and Income Database information.

surveys. Billionaires are a very rare characteristic in any population. There are less than 3000 people worldwide with this characteristic and most countries have only one or two billionaires at the most. If one wishes to study billionaires, sample surveys are not the right instrument. It would also be unwise to add billionaires in survey income statistics partly because they are billionaires in wealth, not income, and partly because most of their wealth is generated globally rather than in a particular country. Including billionaires in income statistics would simply bias survey population statistics. Therefore, when we consider the very top income earners in this paper we are considering millionaires in wealth whose income is counted in the hundreds of thousands of euros annually. This is the class of people we want properly represented in household sample surveys at the top of the distribution.

Data collection issues mostly arise from respondents' or interviewers' non-compliance to survey instructions and may result in unit non-response, item non-response, item underreporting or generic measurement errors:

Unit non-response. Unit non-response refers to households that were selected into the sample but did not participate in the survey. The reasons for non-participation can be many such as a change of address or non-interest on the part of the household. Interviewers generally have lists of addresses that can be used to replace the missing household but this practice is not always sufficient to complete the survey with the full expected sample. Most of the available household survey data suffer from unit non-response.[2] In some surveys, the reason for non-response is recorded but in others it is not. Unit non-response bias results if non-response is not random but systematically driven by specific factors. This paper will address unit non-response issues using reweighting.

Item non-response. Item non-response occurs when households participating in the survey do not reply to an item of interest (income or expenditure in our case). Item non-response biases results if it is non-random and related to specific factors. Non-response may be related to households' characteristics such as wealth or education, and this may bias statistics constructed with income or expenditure variables. As compared to unit non-response, it is possible to correct for item non-response using information on the reasons for non-response (when available) or by means of imputation using household and individual socio-economic characteristics to predict income. The reweighting method proposed in this paper also corrects for item non-response.

Item underreporting. Consistent underreporting of variables on the part of respondents can lead to poor estimates of inequality. For example, if the degree of underreporting rises with income, the measurement of inequality could be affected. Even if underreporting applies equally across respondents, the measurement of inequality may change if the income inequality measure used is not scale invariant. Over-reporting is also possible although extremely rare with income and expenditure data, particularly at the top end of the distribution. The replacing method used in this paper helps to correct for item underreporting.

Generic measurement errors. Any variable including income or expenditure can be subject to measurement error. This error is typically expected to be random, distributed normally and with zero mean. For example, extreme observations in an income distribution can result from data input errors, but if they are very large they bias sample statistics significantly. Statistical agencies are usually quite thorough on this issue and clear data of errors before providing the data to researchers. This issue will not be treated in this paper explicitly but these errors are implicitly treated when replacing observations.

Data preparation issues are mostly a consequence of statistical agencies' compliance with rules and regulations governing data confidentiality and data use, and may result in top coding, sample trimming, or the provision of limited subsamples to researchers.

Topcoding. Top coding is the practice adopted by some statistical agencies such as the US Census Bureau to modify intentionally the values of some variables to prevent identification of households or

[2] Notable exception is that of income surveys based on national tax registers (Burricand 2013; Jäntti et al. 2013).

individuals. It can take various forms, from replacing values above a certain threshold with means or medians of top cells to swapping incomes across top observations. In some cases and for research purposes, statistical agencies provide restricted access to the original values. However, in most cases researchers are left with the problem of having to correct sample statistics for top coding. In this paper, we use EU-SILC data which are not subject to top coding on the part of Eurostat, although it is possible that some countries apply some form of topcoding to their data before transmitting these data to Eurostat. Replacing corrects for topcoding but only for the segment of data replaced whereas reweighting is unlikely to correct for topcoding.

Trimming. Trimming is the practice of cutting off some observations from the sample. This may be done for confidentiality reasons or for observations that appear unreliable. Researchers may not be informed whether statistical agencies have trimmed data, why trimming was performed, or both. A related issue is that of trimming through sampling weights. Statistical agencies sometimes trim sampling weights to bring them within a narrow range of values or to limit their influence if their variable values may have been mismeasured. The overarching objective is to control the influence of units that are rare in the sampling frame. Trimming observations or weights biases statistical measurement and should be corrected for. Trimming is similar to unit or item non-response in that we are missing income observations. Reweighting can help to address this issue if trimmed income observations come from within the support in the observed sample.

Provision of subsamples. Some statistical agencies cannot provide the entire data sets to researchers for confidentiality or national-security reasons or simply to prevent others from replicating official statistics. In many countries, statistical agencies provide 20% to 50% of their samples to researchers. These subsamples are usually extracted randomly so that statistics produced from these subsamples may be reasonably accurate. As we know from sampling theory, random extraction is the best option for extracting a subsample in the absence of any information on the underlying population. However, only one subsample is typically extracted from the full sample and given to researchers and this implies that a particularly "unlucky" random extraction can potentially provide skewed estimates of the statistics of interest. Hlasny and Verme (2018) have tested the margins of error in inequality measurement that can arise from the provision of subsamples instead of full samples and found significant margins of error. This issue is not treated in this paper because EU-SILC data are provided in full.

Data analysis issues may arise from an inadvertently wrong choice of statistical estimators on the part of researchers. Some estimators are more sensitive than others to the issues listed above so that one choice of estimator may lead to greater errors than others. For example, Cowell and Victoria-Feser (1996) have found that the Gini index is more robust to contamination of extreme values than two members of the generalized entropy family, a finding later confirmed by Cowell and Flachaire (2007). Based on these findings, we will focus on the Gini index and leave the discussion of alternative inequality estimators aside. Also important to note is that many researchers routinely trim outliers or problematic observations or apply top coding with little consideration of the implications for the measurement of inequality.

2.1. Reweighting

Unlike the case of item non-response, unit non-response cannot be dealt with by inferring households' unreported income from their other reported characteristics, because we do not observe any information for the non-responding households. In an effort to address this problem, Atkinson and Micklewright (1983) used information on non-response rates across regions to uniformly 'gross up' the mass of respondents in a region by the regional non-response rate. This is the approach taken by several national statistical agencies in adjusting sampling weights for regional unit non-response. This approach is inadequate, as it accounts only for inter-regional differences in non-response rates, and not for systematic differences in response probability across units within individual regions.

Mistiaen and Ravallion (2003), and Korinek et al. (2006, 2007) proposed a probabilistic model that uses information on non-response rates across geographic regions as well as information about the

distribution within regions. They estimated the response probability of each household, and used the inverse of this estimate to adjust each household's weight. Each household's weight is thus 'grossed up' non-uniformly to match the mass of all respondents to the size of the underlying population.

The central tenet of the method is that the probability of a household i in a region j to respond to the survey, P_{ij}, is a deterministic function of its arguments. Logistic functional form is used for its simplicity and its robustness properties:

$$P_{ij}\left(x_{ij},\theta\right) = \frac{e^{g(x_{ij},\theta)}}{1+e^{g(x_{ij},\theta)}}, \tag{1}$$

Here $g(x_{ij},\theta)$ is a stable function of x_{ij}, the observable demographic characteristics of responding households that are used in estimations, and of θ, the corresponding vector of parameters. Variable-specific subscripts are omitted for conciseness. $g(x_{ij},\theta)$ is assumed to be twice continuously differentiable. Equation (1) thus imposes several restrictions on the modeled behavioral relationship between households' characteristics x_{ij} and their response probability: the relationship is deterministic and dictated by the logistic functional form and the functional form of $g(x_{ij},\theta)$, differentiable at all levels of x_{ij}, and identical across all households and regions. These restrictions are strong, but several facts help to justify them. One, the logistic function is well-accepted as a robust form to model probabilistic relations. Two, Korinek et al. (2006, 2007), and Hlasny and Verme (2018) have evaluated alternative forms of $g(x_{ij},\theta)$ including non-monotonic functions on US and Egyptian data, and have concluded that some of the most parsimonious functions provide very good fit, compared to both uncorrected income distributions and compared to external information on the true degree of inequality in those countries. Three, nonlinear forms of $P(x_{ij},\theta)$ and $g(x_{ij},\theta)$ allow for response differences between poorer and richer households in a realistic way. Four, a comparative study of US, EU and Egyptian data led to similar estimation results across countries, suggesting that the behavioral tendencies exhibit a high degree of consistency across regions (Hlasny and Verme 2015). Five, supplementing $g(x_{ij},\theta)$ with indicators for subsets of regions helps to attenuate any systematic behavioral differences across parts of the country.

The number of households in each region ($\hat{m}_j$) is imputed as the sum of inverted estimated response probabilities of responding households in the region ($\hat{P}_{ij}$) where the summation is over all N_j responding households.

$$\hat{m}_j = \sum_{i=1}^{N_j} \hat{P}_{ij}^{-1}\left(x_{ij},\hat{\theta}\right). \tag{2}$$

The parameters θ can be estimated by fitting the estimated and actual number of households in each region using the generalized method of moments estimator:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_j \left[\left(\hat{m}_j - m_j\right) w_j^{-1} \left(\hat{m}_j - m_j\right)\right] \tag{3}$$

where m_j is the number of households in region j according to sample design, and w_j is a region-specific analytical weight proportional to m_j.[3] The asymptotic variance of $\hat{\theta}$ can be estimated as the ratio of the model objective value (the weighted sum of squared region-level residuals), and the squared partial derivative of this objective value with respect to $\hat{\theta}$ (equal to $-\sum_j w_j^{-1} \sum_i \left(x_{ij}/e^{x_{ij}\hat{\theta}}\right)$ under the assumed

[3] An illustration is in order. Suppose there are two income groups residing in two national regions. Region 1 has a higher share of the richer income group, and correspondingly a higher unit non-response rate, as the richer households are less likely to participate. As a result, mean income and income inequality index may or may not differ across the two regions. To correct the mean incomes and inequality indexes in each region as well as nationally, we wish to give more weight to each richer household until the sum of weights equals the underlying regional population, because behind each responding rich household there are more non-responding rich households. Equation (2) 'blows up' the weight of each responding household systematically, under the household-level behavioral rules specified in Equation (1), to fit the joint weighted mass of the responding households to the underlying regional population (Equation (3)). In one region the weighted mass of the responding households may exceed the underlying population, while in the other region it may fall short (because of the restrictions imposed in Equation (1)), but the nationwide sum of the weighted masses equals the underlying national population.

logistic functional form), both weighted by region-specific analytical weights w_j (Equations (11)–(14) in Korinek et al. 2007).

Under the assumptions of random sampling within and across regions, representativeness of the sample for the underlying population in each region, and stable functional form of $g(x_{ij},\theta)$ for all households and all regions, the estimator $\hat{\theta}$ is consistent for the true θ. Estimated values of $\hat{\theta}$ that are significantly different from zero would serve as an indication of a systematic relationship between household demographics and household response probability, and of a non-response bias in the observed distribution of the demographic variable. In that case, we could reweight observations using the inverted estimated household response probabilities to correct for the bias.

Applying the model in Equations (1)–(3) involves making several decisions regarding the delineation of regions, and choosing parametric forms for the functions $P(x_{ij},\theta)$ and $g(x_{ij},\theta)$. The choice of regional delineation involves a trade-off between the number of j data points for the model loss function (Equation (3)), and the number and distribution of within-j observations vis-à-vis the underlying population to achieve consistency for the underlying distribution of incomes. The sample in each region should encompass the entire range of values of relevant characteristics of the underlying population, calling for a higher geographic level at which sample stratification was performed.

Properties of the data at hand thus call for different degrees of data aggregation, but there is presently little guidance for arbitrary national surveys. For the United States CPS, Korinek et al. (2006, 2007) used state-level aggregation, because geographic identifiers are consistently reported only at that level whereas county or metropolitan statistical area identifiers are missing for some responding as well as non-responding households. Hlasny and Verme (2017) considered various degrees of geographic aggregation, from the level of 185 metropolitan statistical areas (MSAs) to that of 7 census divisions. They concluded that an intermediate level of aggregation, at the level of states or groups of 1–2 MSAs, performed more consistently than extreme aggregation or disaggregation. Using the Egyptian Household Income, Expenditure and Consumption Survey (HIECS), Hlasny and Verme (2018) assessed the degrees of regional aggregation from a high administrative level (governorate by urban–rural areas, 50 areas with 939.7 observations on average) down to the level of primary sampling units (PSUs, 2526 areas with 18.6 observations on average). These alternative approaches yielded different corrections for unit non-response, but the more detailed level of disaggregation was deemed conceptually more appropriate. It gave rise to a higher number of data points used in optimization (Equation (3)). Moreover, the observed range of household characteristics in each Egyptian PSU likely comprised the values of non-responding households, while higher levels of geographic aggregation would make behavioral responses less stable across households within areas j.

For the set of national surveys in the SILC, this paper uses regional aggregation to the highest level of nomenclature of territorial units for statistics (NUTS-1) level. With the exception of a handful of countries, non-response rates are not available at more detailed levels of disaggregation. At the same time, heterogeneity of non-response rates reported by national statistical agencies puts aggregation to the level of EU member states into question. In a similar vein, to satisfy the assumption of stability of $g(x_{ij},\theta)$ across all regions, functional form and covariates x_{ij} are selected to make households across all regions behaviorally similar, in the sense that households with similar values of demographic variables should have similar response probabilities across all regions. To effectively neutralize the cross-country heterogeneity in households' response probabilities, logarithmic specification of $g(x_{ij},\theta)$, and country indicators are used in $g(x_{ij},\theta)$. On the margins, we will report how the addition of regional indicators affects the correction for the unit non-response bias.[4]

For the covariates in x_{ij}, Korinek et al. (2006, 2007) evaluated several variables affecting households' response probability, including income, gender, race, age, education, employment status, household

[4] Exclusion of influential regions and EU member states was also tried, but is not discussed here, as it prevents the estimation of inequality for EU member states and EU at large. (These results are available on request.)

size and an urban–rural indicator. Hlasny and Verme (2018) compared income and expenditures, and indicators for survey rounds. These studies concluded that univariate models controlling for expenditures or income are the most efficient. Because this paper focuses on equivalized disposable income as the welfare aggregate, and because arbitrary household surveys worldwide may not consistently report any additional household characteristics, equivalized disposable income is used as the only explanatory variable.[5]

Finally worth noting, SILC surveys already provide a limited correction for unit non-response through sampling weights. This method accounts for differences in response rates across regions but not for systematic differences across demographic groups within regions. Unfortunately, these sampling weights cannot be decomposed into weights for unit non-response and weights for other issues with unit representativeness. We could either double-correct for unit non-response by using the available sampling weights, or ignore other sample representativeness issues by not using the weights. In the United States CPS (Korinek et al. 2006, 2007; Hlasny and Verme 2017) and the Egyptian HIECS (Hlasny and Verme 2018), the correction for non-response (through $\hat{P}_i^{-1}$) affected inequality estimates substantially more than the corrections for other sample representativeness issues (through sampling weights), and so the non-response correction weights should be used with or without the survey sampling weights. These findings may not apply to surveys with less prevalent or less systematic non-responses, and with graver sampling design issues. In the case of the SILC, the great heterogeneity in sample representativeness across EU member states, and the modest role of non-response correction in the available sampling weights are thought to favor the usage of the non-response correction weights ($\hat{P}_i^{-1}$) in tandem with the sampling weights. To accommodate all these options, alternative estimates of inequality are produced: on uncorrected data, data corrected with non-response-bias weights, data corrected with statistical agency weights, and data corrected with both sets of weights simultaneously. Estimates obtained without sampling weights are reported on the margins.

2.2. Replacing

An alternative approach to correct for poorly reported top incomes is to remove the top end of the distribution and replace it with synthetic values under some parametric assumptions. Cowell and Victoria-Feser (1996), Cowell and Flachaire (2007) and a large body of following studies combined estimates from a Pareto distribution (Pareto 1896) for the top of the income distribution with non-parametric statistics for the rest of the distribution. Atkinson et al. (2011) summarize this literature, and model the historic distribution of top incomes in several countries. Testing this method on US CPS data, Hlasny and Verme (2017) find that replacing actual top incomes with Pareto parametric estimates has a small positive effect on the computed Gini, implying that the reported top incomes are distributed more narrowly than the predicted values. However, the effect is smaller than a correction for unit non-response alone using the reweighting method, suggesting that top-income biases operating in opposite directions may be at play. Burkhauser et al. (2010) compared four alternative parametric estimators for replacing of topcoded incomes and combined the estimates with those from non-topcoded incomes. Alvaredo and Piketty (2014) have recently proposed to use synthetic data for the entire income distribution, and estimate inequality using a mix of Pareto distributions for top incomes and log-normal distributions for the rest of incomes. Alvaredo et al. (2017) improve on this methodology by collecting survey micro-data from several countries, and replacing top incomes with values from the Pareto distribution benchmarked using administrative income tax data from a highly unequal paragon country, Lebanon. Using uncharacteristically high parametric values for the distributions in the Middle East countries, these approaches yielded higher inequality measures

[5] This decision also can be viewed as upholding the anonymity axiom that inequality measures be based only on the welfare aggregate itself, and independent of other household characteristics (Litchfield 1999).

than those using raw survey data or using the Pareto replacement of top incomes alone (estimated by Hlasny and Intini 2015; Hlasny and Verme 2018).

Beside the Pareto distribution, other parametric forms have been suggested in recent literature as providing superior fit to income distributions in particular countries. A generalized beta distribution of the second kind (GB2), also known as the Feller-Pareto distribution, is a suitable functional form representing well a large extent of the income distribution (McDonald 1984). The upper tail of the distribution can be modeled as heavy and decaying similar to a power function, while the lower end of the distribution can be short-tailed. The lognormal, Fisk, Singh-Maddala (Singh and Maddala 1976) and Dagum (1980) distributions have also been suggested as candidates for modeling income distributions, being themselves limiting cases of the GB2 distribution with some parameters held fixed (McDonald 1984). However, their fit was not consistent or universally good across various waves of European and US income surveys (Butler and McDonald 1989; Brachmann et al. 1996; Jenkins 2007; Jenkins 2009a; Brzezinski 2013), and so the more flexible GB2 distribution may be preferred.

This study uses the parametric properties of the Pareto and GB2 distributions to evaluate how representative are the top-income observations in our sample to the corresponding expected income distribution, and which parametric form provides the best fit for SILC data. Following Cowell and Flachaire (2007) and Davidson and Flachaire (2007) we correct the Gini coefficient by replacing highest-income observations with values drawn from a parametric distribution and combining the corresponding parametric inequality measure for these incomes with a non-parametric measure for lower incomes. The following sections discuss the mechanics of fitting the alternative parametric forms to the data at hand.

2.2.1. Pareto Distribution

For the past century, the Pareto distribution has been applied to various socio-economic phenomena and is thought to be suitable to model the distribution of upper incomes. The Pareto distribution can be described by the following cumulative density function:

$$F(x) = 1 - \left(\frac{L}{x}\right)^{\alpha},\ L \le x \le \infty, \tag{4}$$

where α is a fixed parameter called the Pareto coefficient and x is the variable of interest (income in our case) and L is the lowest value allowed for x in the case of left censoring. The corresponding probability density function, allowing for right-censoring at H (separating potentially contaminated top-income observations, $H \le x \le \infty$, from reliable bottom observations, $0 \le x \le H$), is

$$f(x) = \frac{\alpha L^{\alpha}}{x^{\alpha+1}} \Big/ \left(1 - \left(\frac{L}{H}\right)^{\alpha}\right),\ L \le x < H \tag{5}$$

This density function is decreasing, tending to zero as x tends to infinity and has a mode equal to the minimum value, L. As income becomes larger, the number of observations declines following a law dictated by the constant parameter α. Clearly, this distribution function does not suit perfectly all incomes under all income distributions, but it should be thought of as one alternative in modeling the right-hand tail of a general income distribution.

Parameter α in Equations (4) and (5) can be estimated using maximum likelihood from a right-truncated Pareto distribution, which also provides robust standard errors (Jenkins and Van Kerm 2015).

The Gini among the top k households can be derived from the expression of the corresponding Lorenz curve as follows

$$Gini = 1 - 2\int_0^1 1 - [1 - F(x)]^{1-\frac{1}{\alpha}}\, dF(x) = \frac{1}{2\alpha - 1} \tag{6}$$

with a standard error composed of a sampling error in the estimation of the Pareto distribution, and an error in the estimation of the Gini coefficient. The sampling standard error under the Pareto distribution is equal to $4\alpha(\alpha-1)/\left[\eta(\alpha-2)(2\alpha-1)^2(3\alpha-2)\right]$ (Modarres and Gastwirth 2006), where η is the estimation sample size ($L \le x < H$). The estimation error due to the potentially imprecise estimates of α is equal to $\epsilon/\left(2\alpha^2 - 2\alpha - 2\alpha\epsilon + \epsilon + 0.5\right)$, where ϵ is the standard error of $\hat{\alpha}$.

2.2.2. Generalized Beta Type 2 Distribution

Because the Pareto distribution is not representative of incomes in the middle or bottom of the income distribution, and because even among top incomes in some countries it may not follow the dispersion of incomes accurately, more flexible parametric distributions have been considered in recent literature. The four-parameter Generalized Beta distribution type 2 (GB2) has been suggested as providing better and more consistent fit for the distribution in various EU and US income surveys (Jenkins et al. 2011). It has the cumulative distribution function

$$F(x) = I\left(p, q, \frac{(x/b)^a}{1+(x/b)^a}\right) \tag{7}$$

In this equation, $I(p,q,y)$ is the regularized incomplete beta function, in which the last argument, y, is income normalized to be in the unit interval. Parameters a, b, p, and q are parameters estimable with their standard errors by maximum likelihood. Because the right tail may be contaminated by top-income issues, right-truncation may be applied in the calculation of the GB2 density and model likelihood functions.

Moreover, as with the Pareto distribution, the GB2 distribution itself may not approximate well the bottom-most incomes, even though it tends to perform well in the middle and the top of the distribution. Jenkins et al. (2011, p. 69) propose left-truncating the distribution at the 30th percentile, a suggestion that this paper follows.[6] Finally worth noting, the Gini under the estimated left- and right-truncated GB2 distribution could be computed by evaluating the corresponding generalized hypergeometric function ${}_3F_2\left(\hat{a}, \hat{b}, \hat{p}, \hat{q}\right)$ (McDonald 1984; Jenkins 2009b).

2.3. Corrected Gini for EU States and EU-Wide

Replacing of observed top incomes with fixed Pareto or GB2 fitted values has the problem that it does not account for parameter-estimation error and sampling error in the available sample. The resulting Gini carries an artificially low standard error. An and Little (2007), and Jenkins et al. (2011) account for sampling error by drawing random values from the estimated distribution for all top incomes.

In the case of the EU SILC, we derive a corrected Gini coefficient across all EU member states as follows. The cumulative parametric distributions in Equations (4) and (7) are estimated at the level of each member state, and top incomes observed in each member state are replaced with random draws from the corresponding state-specific parametric distribution, as proposed by An and Little (2007), and Jenkins et al. (2011). Combining the observed lower-income values and the imputed top incomes across all EU member states allows us to derive a non-parametric estimate of the aggregate EU-wide Gini. Finally, repeating the exercise (bootstrapping) we obtain a *quasi-nonparametric* EU-wide Gini with its standard error (Reiter 2003).

As compared to the semi-parametric approach conventionally used in countries with homogeneous populations, this procedure allows the EU-wide distribution to include observations from both tails of state-level distributions, and preserve the original number of observations for each country. It also allows modalities such as custom truncation of state samples used for parametric estimation and for inequality measurement. Estimating the parametric distributions at the level of EU member states and replacing top incomes according to the estimated country-specific distributions ensures that each state will have true lower incomes as well as replacement top incomes in the EU-wide data.[7] The random draws of incomes ($x > H$) from the parametric distributions (estimated on incomes between L and H) can be combined with true lower incomes (up to H) as well as with incomes across

---

[6] Indeed, during GB2 estimation on the SILC with Eurostat sampling weights, the algorithm could not converge due to the bottommost income observations (2.50 Euro/year or less). This indicates atypical distribution of the bottommost incomes. Indeed, there are over 100 observations in the SILC with annual income less than 100 Euro, suggesting measurement errors.

[7] Conversely, if all EU-wide incomes were used for estimation and replacement, this estimation and replacement would be largely done on the richest member states. Poorer states would then be represented with largely true incomes, while the

EU states. Such flexible estimation of the EU-wide Gini and its standard error would generally not be possible with parametric estimates of the top-income Ginis.

Comparing the corrected state Ginis from the replacing analysis with the observed non-parametric Ginis would indicate whether the observed high incomes have been generated by Pareto- or GB2-like statistical processes, or whether the observed Gini is affected by top-income issues such as missing or non-representative values. A quasi-nonparametric Gini that is lower than the nonparametric Gini can be interpreted as evidence that some top incomes are extreme compared to those predicted under the parametric distribution. A higher quasi-nonparametric Gini would indicate that the observed top incomes are distributed more narrowly than would be predicted parametrically, potentially implying under-representation, censoring, or measurement errors in relation to high-income units in the sample.

An important decision in applying the replacing method relates to the range of incomes that should be replaced as potentially nonrepresentative or contaminated. Cowell and Flachaire (2007) choose a threshold at the 90th percentile of incomes. On the basis of the quality of fit in the United Kingdom income surveys, Jenkins (2017) advocates setting the threshold at top 1% or 5% incomes. We consider replacing between the top 1% and the top 10% of incomes with synthetic values contaminated only by randomness of the draw from the parametric distributions.

In conclusion, the reweighting and replacing methods differ in several respects and address different types of problems related to top incomes. Reweighting considers the entire income support and reweights all observations throughout the support according to the probability of non-response estimated with real data. Replacing keeps all observations up to the cut point unaltered while replacing all observations above the cut point with observations drawn from a theoretical distribution. Reweighting uses a probabilistic model drawing information from within and between regions' non-response rates to estimate the probability of non-response. Replacing does not make use of non-response rates or probabilistic models and uses instead estimated parameters from theoretical distributions to replace observations at the top. Reweighting is suited to address issues related to unit and item non-response and trimming whereas replacing is suited to address issues related to item underreporting, generic measurement errors, topcoding, and undue sensitivity of inequality measurement to the inclusion of rare extreme income observations.

3. Data

The methodologies outlined in the above section are evaluated using the set of national household surveys included in the 2011 round of the EU Statistics on Income and Living Conditions (SILC). This is a challenging set of surveys with different types of problems related to measurement issues that affect top incomes and inequality estimates.[8]

The SILC surveys, coordinated by a Directorate-General of the European Commission, Eurostat, cover one of the most heterogeneous and largest common markets, including some of the world's most affluent nations as well as former socialist economies. All European Union member states as well as Iceland, Norway and Switzerland are included. The data include relatively large sample sizes for each state but suffer from very different non-response rates across member states, and from limited potential for regional disaggregation. Average national non-response rates range from 3.3 to 50.7 percent across member states in the 2011 wave, and from 3.5 to 48.1 percent in 2009 (Tables S1 and S2 in the online Supplementary Material). These features allow for a limited number of model specifications to be used to reevaluate inequality under various measurement issues.[9]

---

richest states would be largely replaced, a dubious exercise. Moreover, while the Pareto law may hold for each EU member state, there is no guarantee that it would hold on incomes EU wide.

[8] This analysis cannot be performed across multiple waves of SILC for several reasons: SILC was first collected only in 2004; Availability of countries has varied by wave; member states are not required to collect or publish sub-national non-response rates, and some statistical agencies have declined to compute them for the authors of this study citing lack of resources.

[9] For more information on the SILC see: http://ec.europa.eu/eurostat/web/income-and-living-conditions.

SILC data are rarely used as one dataset for cross-country analysis in the same fashion as one would do cross-region analysis in a specific country. That is because SILC data are derived from country specific surveys which take different forms in different countries. However, in our case, they are an interesting set of data in that they are characterized by substantial diversity compared to other national surveys (Hlasny and Verme 2015). They are therefore a good benchmark to test how different top incomes correction methodologies perform under such diversity, provided that systematic cross-country differences are controlled for.[10] One challenge is that incomes exhibit substantial cross-nation inequality, but relatively less inequality within nations, as evidenced by the difference between state-specific and EU-wide Gini indexes (refer to Tables S1 and S2). In fact, decomposition of the EU-wide Gini reveals that 67 percent of inequality arises solely from income differences between EU member states, and only 4 percent arises solely from within-state inequality, while 29 percent is due to an overlap of the between and within state inequality (2009 SILC shows analogous results).

With little overlap between income distributions in the richest and the poorest member states, when the reweighting correction method is run at the level of states (rather than within-state regions), it would effectively adjust the mass of entire member states in the calculation of the Gini. The vast majority of households in rich states would be assigned higher weights, and the majority of households in poor states would be assigned lower weights. This suggests that the analysis performed at a more geographically disaggregated level is warranted. To that end, we have collected unit non-response rates for NUTS-1 regions, that is geographic divisions, provinces or states of EU member countries.[11] Refer to Tables S2 and S3 in the online Supplementary Material. In what follows, we will primarily make use of the 2011 round of the SILC, and we will report on the 2009 round only on the margins. When not noted explicitly, the discussion refers to the 2011 round.

Household non-response rates (*NRh*) in SILC surveys are computed using Eurostat notation as:

$$NRh = 1 - \underbrace{\frac{\sum 1(db120 = 11)}{\sum 1(db120 \neq \varnothing) - \sum 1(db120 = 23)}}_{\text{Address contact rate}} \underbrace{\frac{\sum 1(db135 = 1)}{\sum 1(db130 \neq \varnothing)}}_{\substack{\text{Rate of complete} \\ \text{interviews accepted}}}, \tag{8}$$

where $1(\cdot)$ is a binary indicator function, db120 is the record of contact at the address, db130 is the household questionnaire result and db135 is the household interview acceptance result. Addresses that could not be located or accessed (db120 $\leq$ 22) are accounted for in the address contact rate, while non-existing, non-residential, non-occupied and non-principal residence addresses (db120 = 23) are omitted. Rate of complete interviews accepted is the accepted interviews (i.e., at least one personal interview in household accepted) among all households completing, refusing to cooperate, temporarily absent, or unable to respond due to illness, incapacity, language or other problems.

Sampling weights available in SILC (db090) account for units' probability of selection, limited correction for the probability of non-response by different population subgroups, and calibration of sample representativeness vis-à-vis the distribution of households and persons in the target population, including by sex, age, household size and composition and NUTS-2 region (European Commission 2006).

The income variable that is best comparable across SILC national surveys is the equivalized disposable income, hx090. The equivalized household size is computed as hx050 $= 1 + 0.5 \times$ (adults $- 1$) $+ 0.3 \times$ children, where adults are those aged 14 or over at the end of the income

[10] Sampling weights in the SILC are distributed very widely, from essentially zero to 38,357.27 (mean 901.89, standard deviation 1050.31) in the 2011 round. This also suggests that comparing unweighted, SILC weighted, and our non-response probability weighted statistics may yield very different estimates. Moreover, sampling weights in the SILC are trimmed from below and from above to limit the extent to which individual observations can influence sample-wide statistics. To evaluate how much this trimming affects survey-wide results, we could compare results across alternative weighting schemes, or replace the trimmed weights with imputed values.

[11] For Cyprus, Estonia, Germany, Iceland, Latvia, Lithuania, Luxembourg, Malta, Portugal, Slovenia and Switzerland, non-response rates are available by the degree of urbanization (db100 variable): dense, intermediate or thin level of population density. In 2009 for Slovakia and the UK, only nationwide non-response rates are available.

reference period and children are those aged 13 or less.[12] Income is not adjusted for cost-of-living differences across EU member states for conceptual and empirical reasons. First, workers in the European Single Market can spend their income in any jurisdiction as well as on Internet purchases, circumventing local price differentials. Second, it is unclear which single cross-country price index should be applied to workers' earnings, consumption and savings, and the SILC database does not provide such a price index. The income aggregate across countries may also have a different capacity to capture capital income either by design or by practice.

Finally, many of the EU statistical agencies combine survey and administrative information such as tax and social security records to estimate income (refer to individual chapters of Jäntti et al. (2013)). This may result in a more accurate estimation of incomes as compared to countries that do not adopt this strategy. If this is the case, both the reweighting and replacing methods should show (correctly) a lower bias as for any survey with better quality data. However, these techniques vary across countries and can play a role when comparing estimated biases across countries. Considering the fact that the original survey instruments differ and that the income aggregates are not identical in their composition, estimations presented in this paper are not strictly comparable across countries. Moreover, the influence of each country in the overall estimation for the EU Gini is also affected by these factors.

4. Results

4.1. Reweighting

Table 1 presents the benchmark results for the *reweighting* correction method described in Equations (1)–(3). Equivalized disposable income is used as the outcome variable whose inequality is being measured, as well as the main element of x_{ij} (in logarithmic form). Binary indicators for European countries are also included as element of x_{ij} in light of the high heterogeneity in incomes, inequalities and non-response rates across Europe.[13]

The main finding is that households' survey response probability is related negatively to disposable income. The estimated coefficient on log income ($\hat{\theta}_2$) is negative and significantly different from zero, an indication that unit non-response is related to incomes and is therefore expected to bias our measurement of inequality. As a consequence, the corrected Ginis are consistently higher than the non-corrected Ginis. The unweighted corrected Gini coefficient is 48.34. This is higher than the uncorrected and unweighted Gini by 3.25 percentage points, statistically highly significant. Making use of the sampling weights provided by national statistical agencies does not affect these findings. The correction for unit non-response in this case amounts to 3.70 percentage points of the Gini.[14]

---

[12] There are two editions of the EU-SILC survey produced by Eurostat. The Production Data Base (PDB) includes all available variables for responding and nonresponding households, while a Users Data Base (UDB) excludes nonresponding units and variables that could potentially allow identification of households. Related to our analysis, the PDB includes variables *DB120*, *DB130* and *DB135*, defining responding and non-responding households, *DB060*-*DB062*, identifying primary sampling units, and *DB075*, separating the traditional non-response rate (households interviewed for the first time) from the attrition rate (households from the 2nd to the 4th interview). Unfortunately, the PDB is not shared with users for confidentiality reasons, so in this study we rely on the UDB datasets.

[13] This includes 27 country indicators, with Hungary and Slovakia; Denmark and Norway; and Ireland and Island respectively sharing single indicators due to their empirical similarities, and The Netherlands serving as a baseline country. Alternatively, 12 regional indicators plus a baseline were considered, in agreement with geopolitical division of Europe and with empirical distribution of incomes, inequalities and non-response rates across countries. Refer to Table S16 in the online Supplementary Material. However, this less parameterized specification still produced inconsistent results due to the remaining systematic heterogeneity within the 13 European regions.

[14] Note that applying the sampling weights to the distribution of incomes uncorrected for unit non-response reduces the Gini in the SILC by 5.7 percentage points. This happens because sampling weights in the SILC (correcting for various sampling issues including region-level non-response) and the estimated non-response correction weights are related negatively for most households. SILC sampling weights are higher among households with atypical incomes, and lower among households in the middle of the national income distributions. Hence, combination of the two sets of weights serves to dampen the effect of inflating the representation of atypical units with very low incomes. This dampening—which lowers the estimate of inequality—overshadows the double-correction for unit non-response among top-income households.

To the extent that applying the statistical agency weights amounts to some double-correcting for non-response and these corrections interact with each other arbitrarily, we can estimate a quasi-difference-in-difference type of effect of weighting. The stand-alone correction for non-response is estimated at 3.60 percentage points of the Gini (48.34–45.10). The stand-alone correction for non-representative sampling is estimated at −6.19 percentage points of the Gini (38.91–45.10). Adding these effects to the uncorrected Gini, we conclude that the robust Gini is 42.15. This figure is slightly lower than the original estimate of 42.61, suggesting that the double-correction of non-response is responsible for a 0.46 percentage-point inflation of the Gini. In conclusion, *reweighting* is consistent in finding an upward correction of the Gini of between 3.25 and 3.70 percentage points.

Table 1. Benchmark results of Gini correction for unit non-response bias.

Variable	Coefficient Estimate
Intercept	12.377 (1.306)
Log(income)	−1.047 (0.127)
AT	−0.571 (0.156)
BE	−1.386 (0.134)
BG	−1.360 (0.414)
CH	−0.112 (0.164)
CY	0.146 (0.311)
CZ	−1.212 (0.227)
DE	0.042 (0.175)
EE	−2.221 (0.232)
EL	−1.611 (0.169)
ES	−0.381 (0.187)
FI	−0.248 (0.158)
FR	−0.452 (0.145)
HR	−3.035 (0.219)
IE, IS	−0.794 (0.155)
IT	−0.866 (0.133)
LT	−1.790 (0.289)
LU	−0.982 (0.144)
LV	−2.249 (0.251)
MT	−0.533 (0.294)
DK, NO	−1.289 (0.135)
PL	−1.583 (0.241)
PT	−0.259 (0.348)
RO	−0.869 (0.719)
SE	−1.229 (0.133)
SI	−1.284 (0.165)
HU, SK	−1.330 (0.265)
UK	−0.972 (0.141)
Regions *j*	31 member states
Households *i*	238,383
Uncorrected Gini	45.10 (0.08)
Gini using stat. agency weights	38.91 (0.13)
Gini corrected for unit non-response bias	48.34 (0.84)
Gini corrected for unit non-resp. bias, with sampling wts.	42.61 (0.83)
Unit non-response bias	3.25
Bias (using sampling wts.)	3.70

The model is estimated on an unweighted sample, and the uncorrected or corrected weights are only applied in the calculation of the Ginis. Only incomes $\geq$1 are retained. Benchmark region is The Netherlands. Standard errors are in parentheses. Ginis and their bootstrap standard errors are multiplied by 100.

Using the results in Table 1 and the estimated non-response correction weights, we can re-estimate the Ginis for each EU member state (Table 2, last column). The corrected Gini increases by 0.2–6.5 percentage points, with the exception of Belgium and Slovakia (20.0 and 9.3 pc.pt. correction, respectively). The corrected Ginis for Belgium and Slovakia carry high standard errors and should be viewed

with great caution.[15] Across the 29 EU member states (excluding the two outliers, and without accounting for states' population or sample sizes), the estimated Gini correction is strongly positively associated with states' mean income (correl. +0.541), mean non-response rate (correl. +0.219) and the count of regions used for sub-national disaggregation (correl. +0.488).[16] Finally, refer to the discussion on the survey instruments, income aggregates and combination with administrative data to understand other potential sources of cross-country differences in estimated biases.

Table 2. Non-response rate and income distribution by member state, 2011 SILC.

Member State	Sub-National Regions	Househds.	National Non-Response Rate (%)	Mean Equivalized Disposable Income (Euro)	State Gini, SILC Weighted Households	Pure within-Region Contrib. (%)	Pure between-Region Contrib. (%)	State Gini, SILC Weighted & Non-Response Corrected
Austria	3	6183	22.6	23,713.37	27.59 (0.40)	34.2	10.4	29.54 (0.75)
Belgium	3	5897	36.7	21,622.14	27.63 (0.91)	39.5	10.7	47.61 (18.45)
Bulgaria	2	6548	7.5	3415.42	35.99 (0.58)	49.3	14.5	37.88 (1.03)
Croatia	1	6403	43.3	5981.46	32.07 (0.36)	–	–	32.81 (0.57)
Cyprus	3	3916	10.2	20,084.84	31.65 (1.02)	44.3	16.8	36.41 (4.35)
Czech Rep.	8	8865	17.1	8402.77	25.91 (0.37)	12.4	20.7	27.53 (0.57)
Denmark	1	5306	44.4	28,441.21	27.45 (0.55)	–	–	31.00 (1.30)
Estonia	2	4980	26.0	6475.47	32.62 (0.55)	54.4	12.3	34.15 (0.82)
Finland	4	9342	18.1	23,870.09	26.83 (0.37)	24.7	20.3	29.71 (1.92)
France	21	11,348	18.0	24,027.78	30.84 (0.45)	7.2	20.3	36.99 (1.72)
Germany	3	13,473	12.6	21,496.55	30.21 (0.33)	41.0	7.1	32.41 (0.77)
Greece	4	5969	26.5	12,704.72	32.92 (0.57)	27.6	17.0	35.67 (1.10)
Hungary	3	11,680	11.2	5146.29	26.86 (0.26)	34.1	22.0	27.58 (0.31)
Iceland	2	3008	24.8	20,668.26	24.99 (0.64)	53.8	6.0	28.00 (1.68)
Ireland	8	4333	19.6	39,831.65	32.92 (0.56)	14.9	23.8	34.82 (1.10)
Italy	5	19,234	25.0	18,353.37	31.72 (0.29)	21.6	23.7	35.56 (1.00)
Latvia	2	6549	18.9	5048.72	34.98 (0.39)	49.0	17.0	36.46 (0.48)
Lithuania	2	5157	18.6	4588.81	33.02 (0.57)	50.0	16.6	33.95 (0.65)
Luxembourg	3	5442	43.3	37,232.63	27.32 (0.47)	35.5	12.7	29.42 (0.86)
Malta	2	4070	11.8	12,167.55	28.29 (0.44)	81.5	1.9	28.95 (0.52)
The Netherlands	1	10,469	14.5	22,726.06	25.66 (0.34)	–	–	27.01 (0.56)
Norway	1	4621	50.7	38,616.14	24.98 (0.59)	–	–	29.39 (3.05)
Poland	6	12,861	14.9	5849.61	32.10 (0.39)	17.5	10.1	34.32 (0.73)
Portugal	3	5740	7.9	10,462.34	35.07 (0.57)	32.8	19.2	36.35 (0.72)
Romania	4	7614	3.3	2447.42	32.37 (0.39)	25.0	13.3	32.58 (0.41)
Slovakia	4	5200	14.5	6983.48	27.30 (1.26)	28.5	15.0	36.58 (9.42)
Slovenia	1	9246	23.8	12,714.07	25.84 (0.29)	–	–	26.54 (0.38)
Spain	19	12,900	37.2	14,584.40	32.67 (0.26)	6.7	23.6	33.03 (0.29)
Sweden	1	6694	36.5	23,727.45	25.76 (0.36)	36.8	9.0	28.65 (2.52)
Switzerland	3	7502	24.0	39,327.92	30.28 (0.49)	42.6	12.0	34.82 (1.60)
UK	37	8009	27.3	20,843.59	32.85 (0.57)	3.1	24.5	39.32 (2.88)
Wtd. Mean [EU wide]	5.23 [162]	7695 [238],[559]	23.9	17,929.58	29.61 [38.91]	–	–	32.99 [42.61]

Note: Non-response rate is reported in the member-states' Intermediate/Final Quality Reports at the state level as *NRh* for total sample. Incomes less than 1 are omitted. Mean incomes may not be representative of those for the entire states, as they omit non-responding households. For clarity of presentation, Ginis are multiplied by 100. Source: EU-SILC data in World Bank database; Ireland data from Luxembourg Income Study database.

4.2. Replacing

Next, we use a methodology first proposed by Cowell and Victoria-Feser (2007) to test the sensitivity of the Gini coefficients to extreme or non-representative observations on the right-hand side of the distribution. We correct for the influence of potentially contaminated top incomes using

[15] The high corrections of the Ginis in Belgium and Slovakia are not due to atypical distributions of incomes across national regions—Gini decomposition shows similar within- and between-region components (Table 2, two columns before the last column). Instead, it is due to exceptionally thin top-income distributions with rare extreme incomes. Tables S4–S9 show that the Pareto coefficients estimated among the highest quartile of incomes in Belgium (particularly from the 75–80th percentile to the 92–94th percentile) are the highest or among the highest of all EU member states. Pareto coefficients estimated for Slovakia are also above average, but not exceptionally high. These thin top ends of the income distribution suggest that the few observed extreme incomes, when reweighted, can have great influence on the measurement of inequality. This also explains the high standard errors on the Ginis.

[16] The number of regions j selected for the estimation of Equation (3) determines the weight that the model attributes to *within*-region as opposed to *between*-regions information and this choice leads to significantly different estimations of the correction bias. Analyses using finer degrees of disaggregation have been found to typically yield lower corrections for unit non-response (Hlasny 2016; Hlasny and Verme 2017, 2018). In Tables 1 and 2, however, the estimates come from a model on the entire set of 31 member states, using a fixed degree of disaggregation into 162 regions.

an estimated Pareto or generalized beta distribution as discussed in the methodological section. The analysis is performed at the level of individual EU member states, so that the replaced income values would come from all states rather than just from a handful of the richest states. Table 3 presents quasi-nonparametric estimates of the Gini coefficients obtained by replacing the highest top 1–10 percent of income observations in each state with values imputed from the estimated Pareto distribution left-truncated at the 75–85th percentile of incomes and right-truncated at the 92–99th percentile of incomes. (Tables S4–S9 in the online Supplementary Material show the results for each member state.) Lower right-truncation, such as at the 90th percentile, could not be performed because it would leave small national sample sizes for estimation (say, 85–90th percentile incomes), particularly compared to the range of incomes for replacing (say, 91st percentile incomes and above), and would yield volatile or excessively high Ginis. Recall that the estimation is performed at the national level, and national samples are not large (Table 2). By the same token, lower left-truncation would compromise the quality of fit of the Pareto distribution.

The choice of right truncation is a critical parameter because it affects which observations will be classified as uncontaminated and will be used to estimate the parametric distribution, and which observations labeled as suspect will be replaced with values drawn from the distribution. The corrected inequality index will be based on the actual income observations to the left of the right-truncation point, and only on synthetic values to the right of that point. Since there is no theoretically favored point for left- or right-truncation, and there is limited empirical guidance on how to set them particularly in a new dataset for a group of countries such as the EU-SILC, we consider a range of cutoff points. Values of the estimated parameters, measures of model fit, and the estimated corrections for the Gini can be used to determine which ranges of incomes are best suited for estimation and for out-of-sample prediction.

Results are shown in Table 3. The table has three sets of rows, for left-truncation set at the 85th, 80th and 75th percentile of national incomes. We find that the choice over left truncation in estimation does not affect the measurement of inequality significantly. The Ginis are corrected by -0.2 to +4.4 percentage points regardless of the left-truncation point. On the other hand, right truncation affects the measurement systematically. This should not be surprising, because right-truncation in this exercise affects not only the estimation of the Pareto distribution, but also the extent of replacing observed top incomes with values drawn from the national parametric distributions. When only 1% of top incomes are replaced, the Gini typically falls by 0.02 to 0.20 points, suggesting that the observed topmost incomes are extreme and over-represent the incomes of the richest 1 percent in the population as predicted by the estimated national Pareto distributions.[17] However, when 5–8 percent of observed top incomes are replaced, the Gini rises by 0.39 to 4.36 percentage points, suggesting that in this group (and particularly in the second ventile of the national distributions) the observed incomes typically underrepresent the incomes in the population due to unit non-response and other biases. These latter results are consistent with the results provided by reweighting potentially suggesting that the Pareto distribution mimics rather well the top decile of the real income distribution but not the very top of the distribution (top 1 or up to top 5 percent).

[17] Analogous replacement was also done for the top 0.2, 0.5 and 0.7 percent of incomes. The effects of these replacements are smaller than those in Table 3, as they reflect the replacement of individual outlying observations.

Table 3. Correction by replacing incomes with random draws from national Pareto distributions.

Correction of Extreme Observations	Sampling Correction	Sample Size η k obs. Replaced	Gini	Bias in Original Gini (pc.pt.)
Estimation on top 15–hth percentile of incomes				
Semi-param. estimation, h = 1%	Unweighted	η = 33,380 k = 2400	44.92 (0.07)	−0.18
	Eurostat weights	η = 34,475 k = 2587	38.71 (0.13)	−0.20
Semi-param. estimation, h = 5% [i]	Unweighted	η = 23,841 k = 11,939	45.49 (0.11)	+0.39
	Eurostat weights	η = 24,517 k = 12,545	38.85 (0.14)	−0.06
Semi-param. estimation, h = 6% [i]	Unweighted	η = 21,463 k = 14,317	45.87 (0.16)	+0.77
	Eurostat weights	η = 21,994 k = 15,068	43.27 (11.01)	+4.36
Estimation on top 20–hth percentile of incomes				
Semi-param. estimation, h = 1%	Unweighted	η = 45,295 k = 2400	44.98 (0.07)	−0.12
	Eurostat weights	η = 46,702 k = 2587	38.81 (0.13)	−0.10
Semi-param. estimation, h = 5%	Unweighted	η = 35,756 k = 11,939	45.72 (0.10)	+0.62
	Eurostat weights	η = 36,744 k = 12,545	39.66 (0.16)	+0.75
Semi-param. estimation, h = 8%	Unweighted	η = 23,860 k = 19,086	47.26 (0.18)	+2.16
	Eurostat weights	η = 29,302 k = 19,987	42.15 (0.47)	+3.24
Estimation on top 25–hth percentile of incomes				
Semi-param. estimation, h = 1%	Unweighted	η = 57,218 k = 2400	45.04 (0.08)	−0.06
	Eurostat weights	η = 58,841 k = 2587	38.89 (0.14)	−0.02
Semi-param. estimation, h = 5%	Unweighted	η = 47,679 k = 11,939	46.09 (0.17)	+0.99
	Eurostat weights	η = 48,883 k = 12,545	40.12 (0.19)	+1.21
Semi-param. estimation, h = 8%	Unweighted	η = 40,532 k = 19,086	47.89 (0.20)	+2.79
	Eurostat weights	η = 41,441 k = 19,987	42.41 (0.46)	+3.50

Notes: Pareto coefficients are estimated on non-contaminated income observations (sample size η; $L \leq x < H$; H is income corresponding to the 100-hth percentile) using maximum likelihood, and are then used to impute values for the k top-income observations. Parametric replacement is done at the national level. Europe-wide Ginis and their standard errors are computed across all national quasi-nonparametric income distributions, and are bootstrapped. For clarity, Ginis and their standard errors are multiplied by 100. Sampling weights are adopted from Eurostat. [i] Right-truncation here is higher than in the models below. Any lower right-truncation point than this leads to overly large and erratic Gini estimates due to small national estimation samples (i.e., range of income quantiles on which Pareto distribution is fit) and comparatively large national prediction samples (i.e., quantiles for which Pareto estimates are drawn). Refer to Table S5.

Tables S4–S9 in the online Supplementary Material present the Pareto coefficients α and semiparametric Ginis estimated for each EU member state.[18] Like under the reweighting approach,

[18] The parametric Gini estimates among top incomes in Tables S4–S9 were calculated under smooth fitted Pareto curves rather than from any observations or fitted values per se. As a robustness check, we have re-estimated these Ginis by replacing top incomes with numbers drawn randomly from the corresponding Pareto distributions, and bootstrapping the exercise.

the corrections of the Ginis across individual states are in line with the EU-wide corrections with some notable exceptions. Estimated Pareto coefficients are low in several states—notably Cyprus, Estonia, Ireland and Latvia—on account of a narrow dispersion of top incomes and rare extreme incomes, leading to high corrected Ginis in those states. On the other end of the spectrum, Belgium, Iceland, Norway, Slovenia and Sweden have high estimated Pareto coefficients leading to lower corrected Ginis. The effects of top-income replacement on the Ginis are dampened by the fact that replacement is applied only to top incomes, while original values are used for the rest of incomes. In comparison, the reweighting method affected the contribution of all income observations, leading to even larger corrections of the Gini.

The variation in the Pareto coefficients across model specifications indicates that the estimated α depends systematically on the way income observations are weighted, and on the range of top incomes under analysis. Pareto coefficients are estimated somewhat higher in the income distribution weighted by Eurostat sampling weights than in the unweighted distribution. Moreover, the higher the values of incomes evaluated in terms of the left and right truncation points, the higher the Pareto coefficient, and thus the lower the corresponding inverted Pareto coefficient β, the estimated top income share and the Gini. The highest Pareto coefficients are obtained when the national distributions are left-truncated at the 85th percentile and right-truncated at the 99th percentile. That suggests that extreme income dispersion may be a problem among the topmost 1% of incomes and between the 75th and 85th percentile, but not as much between the 85th and 99th percentiles.

One potential criticism of the Pareto distribution is that it relies on only one parameter to fit true top incomes. The fit of the one-parameter Pareto distribution to European and other income distributions has been questioned (Jagielski and Kutner 2013; Jenkins 2017). In the following paragraphs we re-estimate the semi-parametric Gini coefficients assuming top incomes to be distributed as under the generalized beta distribution. To do this, we estimate the generalized beta distribution that provides the best fit for the distribution of top 70 percent of incomes in each state, and then use predicted values to compute a parametric Gini coefficient for the state. To derive an EU-wide Gini, we use values drawn randomly from the parametric distributions to replace topmost incomes in each state, and combine these replacement top-income values with actual lower incomes to derive the Gini quasi-nonparametrically.[19]

Table 4 reports the main results for the EU at large, and Tables S10–S15 in the online Supplementary Material report model coefficients and parametric Ginis for individual EU member states. Comparing the Ginis in Table 4 to the nonparametric estimates in Table 1, we find that the quasi-nonparametric Ginis under the assumed generalized beta distribution are systematically lower, implying that actual incomes may be distributed more unequally than incomes predicted under that distribution. The downward correction of the Gini is up to 3.3 percentage points and 1.4 percentage points on average across the 6 model specifications reported.

Compared to the Pareto distribution, the corrections to the Gini coefficients under the generalized beta distribution are consistently negative, but of a similar magnitude in absolute value. This indicates that the estimated generalized beta distributions predict a narrower dispersion of top incomes than the estimated Pareto distributions, but both estimations give rise to concerns about top-income biases of a similar magnitude, 0–4 percentage points of the Gini.

These Ginis from random draws are very similar to the smooth-distribution Ginis in Tables S4–S9, but have slightly higher standard errors due to sampling errors.

[19] To validate the procedure, we again compare the parametric and quasi-nonparametric Ginis in each state (refer to the previous footnote). Indeed, using random income draws from a generalized beta distribution produces a similar correction of the Gini as numerical inference of the Gini under a smooth distribution.

Table 4. Correction by replacing incomes with random draws from national GB2 distributions.

Correction of Extreme Observations	Sampling Correction	Sample Size η k obs. Replaced	Gini	Bias in Original Gini (pc.pt.)
Estimation on top 70–hth percentile of incomes				
Semi-param. estimation, h = 1%	Unweighted	η = 164,423 k = 2400	43.89 (0.05)	−1.21
	Eurostat weights	η = 167,932 k = 2587	37.64 (0.08)	−1.27
Semi-param. estimation, h = 5%	Unweighted	η = 154,944 k = 11,939	44.86 (0.08)	−0.24
	Eurostat weights	η = 158,093 k = 12,545	36.80 (0.09)	−2.11
Semi-param. estimation, h = 10%	Unweighted	η = 143,233 k = 14,317	44.73 (0.14)	−0.37
	Eurostat weights	η = 145,699 k = 15,068	35.57 (0.07)	−3.34

Notes: GB2 coefficients are estimated on non-contaminated income observations (sample size η; $L \leq x < H$; L is income corresponding to the 30th percentile; H is income corresponding to the 100-hth percentile) using maximum likelihood. Quasi-nonparametric Ginis and their standard errors are bootstrap estimates, and are multiplied by 100.

Coefficient estimates presented in Tables S10–S15 carry for the most part acceptable standard errors and are rather consistent across model specifications with different sampling weights and right-truncation points. There are unclear patterns in the estimated coefficients between the analyses performed under alternative weighting schemes (unweighted versus Eurostat weighted) and alternative sample cutoff points (90th, 95th, 99th percentile). The higher the range of incomes included in estimation (up to the 95th or the 99th percentile), the systematically lower the distributional shape parameter a, but the other shape parameters (p, q) and the scale parameter (b) vary non-systematically. As a byproduct of our analysis, we can confirm that the generalized beta distribution cannot be easily approximated by Singh-Maddala or Dagum distributions as $\hat{p}$ and $\hat{q}$, respectively, are significantly different from unity across most EU member states, under all weighting schemes and sample-truncation points in the analysis.

The estimated parametric Ginis vary greatly across EU member states, due to heterogeneous distributions of incomes and sampling weights across states, different sample sizes, and different quality of fit of the parametric GB2 distributions. Like in the case of reweighting and Pareto-replacing estimation, several states end up with outlying parametric estimates of their Ginis subject to high standard errors. Across multiple runs of the analysis (in Tables S10–S15), Belgium, Bulgaria, Finland, Greece, Ireland, Latvia, Norway and Slovenia end up with unreasonably high parametric estimates of their Ginis, while Denmark, Germany, Iceland, Slovakia and Sweden end up with unreasonably low Ginis.

5. Discussion

This study has evaluated two methods—*reweighting* and *replacing*—for correcting top-income biases generated by known data issues including unit and item non-response and more generally representativeness issues of top-income observations. The joint use of two distinct statistical methods for correcting top-income biases, sensitivity analysis of their technical specifications, and analysis of their performance on a challenging heterogeneous household survey were methodological contributions of this study.

Using the reweighting approach and the 2011 wave of the SILC, the paper finds a significant 3.3–3.7 percentage point downward bias in the Gini index.[20] The weighted Europe-wide Gini index is estimated at 42.61 percent as compared to a non-corrected Gini of 38.91 percent. The average Gini for the 31 European countries considered is estimated at 32.99 percent as compared to an uncorrected Gini of 29.61 percent.

Similar results are found using the replacing method with the Pareto distribution but only when the cutoff point for replacing is below 95 percent. The use of higher cutoff points yields very low biases, below 1 percentage point of the Gini. Given that top-income biases are expected to be higher at the very top, it is possible that the Pareto distribution does not mimic well the European income distribution at the very top. This may be due to the limited flexibility offered by the one-parameter Pareto distribution.

Repeating the replacing exercise with the four-parameter GB2 distribution does not improve our findings. Our estimates of inequality fall by 0.2–3.3 percentage points of the Europe-wide Gini, while the Ginis for individual member states are estimated very widely and often unreasonably low or high. We conclude that the popular 1–4 parameter distributions such as the Pareto and the GB2 distributions are not well suited to model the topmost incomes across a heterogeneous sample of distributions, and that alternative distributions should be sought to model the very top ends. The fact that these distributions were proposed and initially tested in the 20th century combined with the sharp growth of incomes at the very top of the distribution in the 21st century in Europe and elsewhere may contribute to explain this shortcoming.

Another problem with the replacing methods, similarly to the traditional treatments for item nonresponse, is that they rely on an assumption that other income observations are valid and accurate. Replacing methods assume away measurement issues below the cutoff point. At the same time, the parametric distributions proposed yield a wide range of empirical results (in Tables 3 and 4), indicating that parameters calibrated with the lower parts of the income distributions do not offer insights of any accuracy about the very top.

In perspective of the findings from the reweighting and parametric replacing exercises, we also conclude that the systematic under-representation of top-income households due to unit nonresponse is a more worrying problem than other potential contaminations of the top-income distribution for inequality measurement. Unit non-response leads to a systematic downward bias in the measurement of the Gini coefficient by 3–4 percentage points, while the balance of other top-income biases remains unclear, and has been estimated in this study widely at between a -3 and a $+4$ percentage point adjustment to the Gini.

Supplementary Materials: The following are available online at http://www.mdpi.com/2225-1146/6/2/30/s1. Tables S1–S3 show summary statistics and unit nonresponse rates in national surveys. Tables S4–S9 show additional results of Pareto replacement for individual EU member states. Tables S10–S15 show additional results of replacement using Generalized Beta II distribution also for individual EU member states. Table S16 shows the considered delineation of 13 European regions.

Author Contributions: Both authors contributed equally to the paper.

Funding: This research received no external funding.

Acknowledgments: The authors are grateful to Ragui Assaad and Roy Van der Weide for peer reviewing an earlier draft of the paper, and to participants at the ECINEQ conference (Luxembourg 2015) for helpful discussion. All remaining errors are ours.

Conflicts of Interest: The authors declare no conflict of interest.

---

[20] This analysis cannot be performed across multiple waves of SILC for several reasons: SILC was first collected only in 2004; Availability of countries has varied by wave; member states are not required to collect or publish sub-national non-response rates, and some statistical agencies have declined to compute them for the authors of this study citing lack of resources.

References

Alvaredo, Facundo, and Thomas Piketty. 2014. Measuring top incomes and inequality in the Middle East: Data limitations and illustration with the case of Egypt. Working Paper 832, Economic Research Forum, Giza, Egypt.

Alvaredo, Facundo, Lydia Assouad, and Thomas Piketty. 2017. Measuring Inequality in the Middle East 1990–2016: The World's Most Unequal Region? WID. Available online: http://wid.world/document/alvaredoassouadpiketty-middleeast-widworldwp201715/ (accessed on 29 May 2018).

An, Di, and Roderick J. A. Little. 2007. Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society A* 170: 923–40. [CrossRef]

Atkinson, Anthony Barnes, and John Micklewright. 1983. On the reliability of income data in the family expenditure survey 1970–1977. *Journal of the Royal Statistical Society Series A* 146: 33–61. [CrossRef]

Atkinson, Anthony B., Thomas Piketty, and Emmanuel Saez. 2011. Top incomes in the long run of history. *Journal of Economic Literature* 49: 3–71. [CrossRef]

Bartels, Charlotte, and Maria Metzing. 2017. An Integrated Approach for Top-Corrected Ginis. IZA DP #10573. Available online: https://www.iza.org/publications/dp/10573/an-integrated-approach-for-top-corrected-ginis (accessed on 29 May 2018).

Brachmann, Klaus, Andreas Stich, and Mark Trede. 1996. Evaluating parametric income distribution models. *Allgemeines Statistisches Archiv* 80: 285–98.

Brzezinski, Michał. 2013. Parametric modelling of income distribution in Central and Eastern Europe. *Central European Journal of Economic Modelling and Econometrics* 3: 207–30.

Burkhauser, Richard V., Shuaizhang Feng, and Jeff Larrimore. 2010. Improving imputations of top incomes in the public-use current population survey by using both cell-means and variances. *Economic Letters* 108: 69–72. [CrossRef]

Burricand, Carine. 2013. Transition from survey data to registers in the French SILC survey. In *The Use of Registers in the Context of EU-SILC: Challenges and Opportunities*. Edited by Markus Jäntti, Veli-Matti Törmälehto and Eric Marlier. Luxembourg: European Union.

Butler, Richard J., and James B. McDonald. 1989. Using incomplete moments to measure inequality. *Journal of Econometrics* 42: 109–19. [CrossRef]

Cowell, Frank A., and Emmanuel Flachaire. 2007. Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics* 141: 1044–72. [CrossRef]

Cowell, Frank A., and Maria-Pia Victoria-Feser. 1996. Poverty measurement with contaminated data: A robust approach. *European Economic Review* 40: 1761–71. [CrossRef]

Cowell, Frank, and Maria-Pia Victoria-Feser. 2007. Robust Lorenz curves: A semiparametric approach. *Journal of Economic Inequality* 5: 21–35. [CrossRef]

Dagum, Camilo. 1980. The generation and distribution of income, the Lorenz curve and the Gini ratio. *Economie Appliquee* 33: 327–67.

Davidson, Russell, and Emmanuel Flachaire. 2007. Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics* 141: 141–66. [CrossRef]

Deaton, Angus. 2005. Measuring poverty in a growing world (or measuring growth in a poor world). *Review of Economics and Statistics* 87: 20–22. [CrossRef]

European Commission. 2006. *EU-SILC User Database Description. EU-SILC/BB D(2005).* Luxembourg: European Commission.

Frick, Joachim R., and Kristina Krell. 2010. Measuring Income in Household Panel Surveys for Germany: A Comparison of EU-SILC and SOEP. SOEP paper 265, German Institute for Economic Research (DIW), Berlin, Germany.

Hlasny, Vladimir. 2016. Unit Nonresponse Bias to Inequality Measurement: Worldwide Analysis Using Luxembourg Income Study Database. LIS Technical Paper, Luxembourg Income Study, Luxembourg City, Luxembourg.

Hlasny, Vladimir, and Vito Intini. 2015. Representativeness of Top Expenditures in Arab Region Household Surveys. UN-ESCWA/EDID Working Paper #11, United Nations Beirut Economic and Social Commission for Western Asia, Beirut, Lebanon.

Hlasny, Vladimir, and Paolo Verme. 2015. Top Incomes and the Measurement of Inequality: A Comparative Analysis of Correction Methods Using Egyptian, EU and US Survey Data. ECINEQ Conference Paper 2015-145, Society for the Study of Economic Inequality, Verona, Italy.

Hlasny, Vladimir, and Paolo Verme. 2017. The Impact of Top Incomes Biases on the Measurement of Inequality in the United States. ECINEQ Working Paper #2017-452, Society for the Study of Economic Inequality, Verona, Italy.

Hlasny, Vladimir, and Paolo Verme. 2018. Top incomes and the measurement of inequality in Egypt. *The World Bank Economic Review* 32: 428–55.

Jagielski, Maciej, and Ryszard Kutner. 2013. Modelling of income distribution in the European Union with the Fokker–Planck equation. *Physica A: Statistical Mechanics and Its Applications* 392: 2130–38. [CrossRef]

Jäntti, Markus, Veli-Matti Törmälehto, and Eric Marlier, eds. 2013. *The Use of Registers in the Context of EU–SILC: Challenges and Opportunities, Eurostat, European Commission*. Luxembourg: Publications Office of the European Union. Available online: http://ec.europa.eu/eurostat/documents/3888793/5856365/KS-TC-13-004-EN.PDF (accessed on 29 May 2018).

Jenkins, Stephen P. 2007. Inequality and the GB2 Income Distribution. IZA Discussion Papers 2831, Institute for the Study of Labor (IZA), Bonn, Germany.

Jenkins, Stephen P. 2009a. Distributionally-sensitive inequality indices and the GB2 income distribution. *Review of Income and Wealth* 55: 392–98. [CrossRef]

Jenkins, Stephen P. 2009b. *GB2LFIT: Stata Module to Fit a GB2 Distribution to Unit Record Data*. Colchester: Institute for Social and Economic Research, University of Essex.

Jenkins, Stephen P. 2017. Pareto models, top incomes and recent trends in UK income inequality. *Economica* 84: 261–89. [CrossRef]

Jenkins, Stephen P., and Philippe Van Kerm. 2015. Paretofit: Stata Module to Fit a Type 1 Pareto Distribution. April 2007. Available online: https://ideas.repec.org/c/boc/bocode/s456832.html (accessed on 29 May 2018).

Jenkins, Stephen P., Richard V. Burkhauser, Shuaizhang Feng, and Jeff Larrimore. 2011. Measuring inequality using censored data: A multiple-imputation approach to estimation and inference. *Journal of the Royal Statistical Society* 174: 63–81. [CrossRef]

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion. 2006. Survey nonresponse and the distribution of income. *Journal of Economic Inequality* 4: 33–55. [CrossRef]

Korinek, Anton, Johan A. Mistiaen, and Martin Ravallion. 2007. An econometric method of correcting for unit nonresponse bias in surveys. *Journal of Econometrics* 136: 213–35. [CrossRef]

Lakner, Christoph, and Branko Milanovic. 2013. Global Income Distribution from the Fall of the Berlin Wall to the Great Recession. World Bank Policy Research Working Paper Series #6719, World Bank, Washington, DC, USA.

Litchfield, Julie A. 1999. Inequality: Methods and Tools. Article for World Bank's Web Site on Inequality, Poverty, and Socio-Economic Performance. Available online: http://siteresources.worldbank.org/INTPGI/Resources/Inequality/litchfie.pdf (accessed on 29 May 2018).

McDonald, James B. 1984. Some generalized functions for the size distribution of income. *Econometrica* 52: 647–63. [CrossRef]

Mistiaen, Johan A., and Martin Ravallion. 2003. Survey Compliance and the Distribution of Income. Policy Research Working Paper #2956, The World Bank, Washington, DC, USA.

Modarres, Reza, and Joseph L. Gastwirth. 2006. A cautionary note on estimating the standard error of the Gini index of inequality. *Oxford Bulletin of Economics and Statistics* 68: 385–90. [CrossRef]

Pareto, Vifredo. 1896. La courbe de la repartition de la richesse, Ecrits sur la courbe de la repartition. de la richesse, (writings by Pareto collected by G. Busino, Librairie Droz, 1965), 1–15. Available online: https://www.cairn.info/ecrits-sur-la-courbe-de-la-repartition-de-la-riche--9782600040211.htm (accessed on 29 May 2018).

Reiter, Jerome P. 2003. Inference for partially synthetic, public use microdata sets. *Survey Methodology* 29: 181–88.

Singh, S. K., and Gary S. Maddala. 1976. A function for the size distribution of income. *Econometrica* 44: 963–70. [CrossRef]

Törmälehto, V.-M. 2017. High Income and Affluence: Evidence from the European Union Statistics on Income and Living Conditions (EU-SILC). Statistical Working Papers, Eurostat, Publications Office of the European Union, Luxembourg.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Polarization and Rising Wage Inequality: Comparing the U.S. and Germany

Dirk Antonczyk [1], Thomas DeLeire [1,2,3] and Bernd Fitzenberger [1,4,5,6,7,8,*]

1 Research Fellow, IZA, 53113 Bonn, Germany; dirk.antonczyk@gmail.com
2 Georgetown University, Washington DC, 20057, USA; Thomas.DeLeire@georgetown.edu
3 National Bureau of Economic Research, 1050 Massachusetts Ave., Cambridge, MA 02138, USA
4 School of Business and Economics, Humboldt University Berlin, Spandauer Strasse 1, 10099 Berlin, Germany
5 Institute For Fiscal Studies, London WC1E 7AE, UK
6 CESifo, 81679 München, Germany
7 Research Centre for Education and the Labour Market (ROA), 6211 LM Maastricht, The Netherlands
8 Centre for European Economic Research (ZEW), 68161 Mannheim, Germany
* Correspondence: fitzenbb@hu-berlin.de

Received: 31 January 2018; Accepted: 22 March 2018; Published: 11 April 2018

Abstract: Since the late 1970s, wage inequality has increased strongly both in the U.S. and Germany but the trends have been different. Wage inequality increased along the entire wage distribution during the 1980s in the U.S. and since the mid 1990s in Germany. There is evidence for wage polarization in the U.S. in the 1990s, and the increase in wage inequality in Germany was restricted to the top of the distribution before the 1990s. Using an approach developed by MaCurdy and Mroz (1995) to separate age, time, and cohort effects, we find a large role played by cohort effects in Germany, while we find only small cohort effects in the U.S. Employment trends in both countries are consistent with polarization since the 1990s. The evidence is consistent with a technology-driven polarization of the labor market, but this cannot explain the country specific differences.

Keywords: wage inequality, polarization, international comparison, cohort study, quantile regression

JEL Classification: J30; J31

1. Introduction

A substantial body of research has documented increasing wage inequality in industrialized countries (see the surveys by Acemoglu and Autor 2011; Katz and Autor 1999). Since the late 1970s, and continuing through the mid-2000s, wage inequality has been increasing in the U.S. (e.g., Autor 2014; Autor et al. 2008; Lemieux 2006a) and Germany (e.g., Biewen et al. 2017; Card et al. 2013; Dustmann et al. 2009). Moreover, rising wage inequality has been identified as a key driver of the rise in income inequality (OECD 2016; Piketty and Saez 2014). This paper provides a comparative analysis of the rise in wage inequality in the U.S. and Germany[1] focusing on the role of cohort effects. Most existing literature on wage inequality has ignored the role of cohort effects.[2]

Skill-biased technological change (SBTC) is the leading hypothesis in the literature to explain the rise in wage inequality.[3] It is often argued that for SBTC to be a compelling explanation of labor

1 In line with the literature on longer-term changes in wage inequality, this paper focuses on West Germany. As a notable exception, Biewen and Seckler (2017) analyzed changes in wage inequality in Germany as a whole since the mid 1990s.
2 Notable exceptions are—among others—Beaudry and Green (2000); Card and Lemieux (2001); Fitzenberger and Wunderlich (2002); Glitz and Wissmann (2017); Gosling et al. (2000); MaCurdy and Mroz (1995).
3 See, e.g., Acemoglu and Autor (2011); Autor (2014); Autor et al. (2008); Biewen and Seckler (2017); Katz and Autor (1999).

market trends, the trends have to be similar across different countries having access to the same technology (Card and Lemieux 2001), provided institutions—or other developments—do not cause different trends. SBTC explains a rise in wage dispersion both between education groups (rising education premia) and within education groups (growing returns to experience and unobserved skills). In addition to SBTC, the literature on the rise in wage inequality discusses the role of the supply of skilled workers,[4] changes in institutions such as the decline in unionization and changes in the minimum wage,[5] the rise of international trade[6] or the increase in workplace heterogeneity (Card et al. 2013).

In contrast to technology based explanations for the U.S., DiNardo et al. (1996) and Lemieux (2006a) argued that increasing wage inequality in the 1980s and the early 1990s can be explained to an important part by changing labor market institutions, i.e., falling real minimum wages and deunionization, and a changing composition of the workforce. If institutional differences matter, one would not necessarily expect to see similar patterns in wage growth and polarization for different countries. The evidence for Germany regarding the role of the decline in unionization in explaining the increase in wage inequality is mixed. Dustmann et al. (2009) and Biewen and Seckler (2017) found a strong role of deunionization on explaining the rise in wage inequality, while Card et al. (2013) and Antonczyk et al. (2010) pointed to the growing heterogeneity in wage setting at the firm level as key drivers. Dustmann et al. (2014) pointed out that wage inequality rose strongest after 1995 among workers covered by collective bargaining. The study attributes the rise in wage inequality to a decentralization of wage setting to the firm level among firms covered by collective bargaining. Autor et al. (2008) argued that changing minimum wages and institutions in the U.S. are unlikely to explain the continuing trend of increasing wage inequality in the upper part of the wage distribution.

SBTC has been refined by the task approach introduced by Autor et al. (2003) which implies polarization of employment and which may also be consistent with polarization of wages (Autor 2013; Autor and Dorn 2013; Autor and Handel 2013; Autor et al. 2008). Autor et al. (2003) proposed as a nuanced version of SBTC that technological change can have a "polarizing" effect on the labor market rather than uniformly favoring skilled workers. That is, technological change—for example, computerization—favors more highly skilled workers relative to less skilled routine-manual and routine-cognitive workers. At the same time, various studies find a disproportionate growth of employment for low-wage jobs (often involving non-routine-manual work) relative to medium-skilled jobs. Altogether, starting in the 1990s, the distribution of jobs has been "polarizing" with faster employment growth in the highest and lowest-paying jobs and slower growth in the middling jobs.[7] However, the relationship between changes in employment and wages is less clear. Autor and Dorn (2013) developed a theoretical model where the wage effects at the bottom of the wage distribution are ambiguous, because they depend upon whether low-skilled jobs are complements or substitutes of high-skilled jobs. Thus, technology driven polarization in employment may also be consistent with rising wage inequality at the bottom of the wage distribution.

From the late 1970s to the mid-1990s, trends in wage inequality differed strongly between the U.S. and Germany. The U.S. experienced in the 1980s a uniform increase of wage inequality along the entire wage distribution (Katz and Autor 1999), while the increase in wage inequality in Germany was restricted to the upper part of the wage distribution (Dustmann et al. 2009; Fitzenberger and Wunderlich 2002). In Germany, the increase in wage inequality in the lower half of the wage distribution began in the mid-1990s (Dustmann et al. 2009;

---

4 See, e.g., Biewen et al. (2017); Lemieux (2006a).
5 See, e.g., Antonczyk et al. (2010); DiNardo et al. (1996); Dustmann et al. (2009, 2014).
6 See, e.g., Autor et al. (2013); Biewen and Seckler (2017); Felbermayr et al. (2012).
7 Recent empirical work has provided evidence for polarization in employment in the U.S. (Acemoglu and Autor 2011; Autor and Dorn 2013; Autor et al. 2008; Lemieux 2008), Germany (Dustmann et al. 2009; Spitz-Oener 2006), Nordic countries (Asplund et al. 2011), various other European countries (Goos et al. 2014), and to a degree in Canada (Green and Sand 2015).

Dustmann et al. 2014; Biewen et al. 2017) and there was a uniform increase of wage inequality along the entire wage distribution until 2010 (Biewen et al. 2017; Dustmann et al. 2014). Autor et al. (2008) provided evidence for a polarization of wages in the U.S. during the 1990s such that wage inequality only continued to rise in the upper part of the wage distribution. Furthermore, Autor and Dorn (2013) found that employment and wages in low-skill service jobs, which involve non-routine manual tasks and which pay low wages, have grown considerably since the early 1990s. In contrast, Dustmann et al. (2009) for Germany and Goos et al. (2014) for 16 EU countries found no evidence for this. Hence, despite similar employment changes, wage inequality has been changing differently in the U.S. compared to European countries. These differences motivate our paper which takes a fresh look at the comparison of trends in wage inequality in the U.S. and in Germany using a unified framework of analysis.

As its key contribution, our study accounts for cohort effects. We define these as effects which are associated with the time a specific cohort was born and which have a permanent effect on this specific cohort. Card and Lemieux (2001) allowed for imperfect substitutability between younger and older workers to explain the fact that the large increase of the wage gap between young college- and high-school graduates is mainly driven by a slowdown in the growth of college graduates in the U.S. during the 1980s. This resulted in a stronger rise of the college–high-school wage gap for younger workers compared to older workers. In a similar vein, Glitz and Wissmann (2017) argued for Germany that the slowdown in the decline of the share of low-education relative to medium-education employment can explain the the rise in the wage differential between these two education groups. Carneiro and Lee (2011) reanalyzed the rising college–high-school premium and provide evidence that about half of the increase reported may be explained by an increased quality of college graduates during this period, which again reflects a cohort effect. Even though SBTC may have a bias in the age/cohort dimension, most of the recent literature on trends in wage inequality (see, e.g., Autor et al. 2008; Dustmann et al. 2009) restricts itself to a comparison of cross-sectional age or experience profiles in different years.

This paper builds on the empirical framework developed by MaCurdy and Mroz (1995) to investigate the importance of age, time, and cohort effects on wages in light of the linear relation between the three variables. We implement a test of the separability of the three effects using standard errors robust against correlation across time and cohort, and we discuss identification of the linear effects. Based on this approach, we examine trends in wage inequality within and across cohorts of full-time working men in the U.S. and Germany by describing a set of quantiles. Wage dispersion in both countries has been rising since the end of the 1970s. While there is strong evidence of rising wage inequality in both economies, we confirm wage polarization only for the U.S. after 1985 and for Germany prior to 1985.

Our main findings are as follows: Based on the estimated conditional time trends, we confirm widening wage dispersion in both the U.S. and Germany between 1979 and 2004. This is the case if we consider trends for wages at the median between education groups as well as quantile specific time trends within education groups. However, there are various distinct patterns. For the U.S., we find that time-trends at the median are more positive for high-education workers than for less educated workers throughout the entire period—the medium-low-education gap ceases to increase during the 1990s. Moreover, time-trends within both the group of low- and medium-education workers start polarizing at the end of the 1980s, while within wage dispersion for high-education workers steadily increases. Trends in Germany are more difficult to interpret. We find little evidence for wage polarization in Germany and growing inequality among low- and medium-education workers after 1985. Moreover, we see a large role played by cohort effects in Germany—suggesting a role for supply-side effects or an interaction with institutions in Germany—while we find smaller cohort effects of opposite sign in the U.S. In addition to wage trends, we analyze the changes in the skill composition of the workforce and find strong parallel movements between the U.S. and Germany.

The remainder of the paper proceeds as follows: Section 2 describes the two data-sets. The third section presents the basic facts of wage growth and wage dispersion for the U.S. and Germany. Section 4 introduces our version of the MaCurdy and Mroz (1995) approach.

The corresponding empirical results are presented in Section 5. Finally, Section 6 provides our conclusions. The Supplementary Material contains graphical illustrations of our estimation results. Detailed estimation results are available upon request.

2. Data

The data we use for our analysis are the U.S. Current Population Survey (CPS) and the German IAB employment subsample (IABS) [while our paper refers to Germany, recall that our analysis is restricted to West Germany]. We focus on male workers who are between 25 and 55 years old. This avoids interference with ongoing education and early retirement.

2.1. CPS for U.S.

The U.S. data used for this analysis are from the Current Population Survey, Outgoing Rotation Groups (CPS-ORG) from 1979–2004. The CPS-ORG data contain wage and salary information for respondents during the month they leave the basic (monthly) survey. Wages are inflated to 2004 dollars using the CPI-U-RS. Workers' calculated hourly wage rates are either the reported hourly wage (for the 60 percent of workers paid on that basis) or weekly earnings divided by weekly hours (for the other 40 percent of workers). For the latter group, earnings per week divided by the usual hours per week was used, unless information on usual hours per week was missing (in 2004, for example, the figures were missing for 5 percent of workers not paid on an hourly basis). In that case, the analysis used the number of actual hours worked in the previous week to construct hourly wages. While that procedure minimizes the number of workers excluded from the analysis, it introduces some noise into the calculated hourly rate of pay because the actual hours worked last week may differ from usual hours worked per week. For roughly 15 percent of workers not paid on an hourly basis, the number of actual hours worked the previous week was different from the usual hours per week. Most often, those workers indicated that they worked part time in the previous week for various reasons, but usually worked full time. The U.S. Census Bureau imputed data on hourly wage rates, usual weekly earnings, and usual hours worked per week were used in the analysis. Over the sample period, the percentage of workers with imputed wage data has increased and was 31 percent in 2004.

We consider male workers from the sample who (normally) work full time. The education level between 1979 and 1989 is measured as a categorical variable with three values regarding the years of schooling completed:

(U) 12 years or less of schooling	(low-education)
(M) 13 to 15 years of schooling	(medium-education)
(H) 16 years or more of schooling	(high-education).

These categories are defined in a slightly different way after 1990 due to changes in the CPS: (U) having a high school diploma or less and not having attended college; (M) having attended college but not having received a degree; and (H) having at least a college degree. Age is measured continuously (in years). Observations are weighted by a person-weight variable and by the hours worked in the preceding week. There is topcoding in labor earnings in the CPS but the share of topcoded observations is very small compared to the data used for Germany (Burkhauser and Larrimore 2009) so this is unlikely to affect the 80%-quantile regressions which we undertake in our subsequent analysis.

2.2. IABS for Germany

The German data used in the empirical analysis are the version of the IABS (IAB employment subsample) ending in 2004.[8] Even though the IABS starts in 1975, we only use data starting from 1979, consistent with the time period available in the CPS[9], and we also inflate wages to 2004 euros using the German CPI. The IABS involves a randomly drawn 2% sample of employees subject to social security taxation. The IABS is provided by the *Institute for Employment Research*. It contains about 400,000 individuals in each annual cross-section and it covers about 80% of the German employees. Different versions of this data set have been used in the literature (see, e.g., Fitzenberger and Wunderlich 2002; Dustmann et al. 2009; Card et al. 2013; Dustmann et al. 2014). The IABS is an earlier version of the SIAB data used, e.g., in Biewen et al. (2017).

There are two important advantages of using data from the IABS. First, the IABS is a very large sample compared to survey data such as the German Socioeconomic Panel, which is also often used in the analysis of wage trends. Second, the IABS remains representative for the workers contributing to the social security system. There are three important disadvantages of the IABS. First, there exists censoring of wages from above. When the daily gross wage exceeds the upper social security threshold ("Beitragsbemessungsgrenze"), the daily social security threshold is reported instead. This censoring affects roughly the top 10%–14% of the workers in the wage distribution.[10] Among university graduates, censoring from above can affect about half of the population. This is one of the reasons why we estimate quantile regressions of wages, which are robust against right censoring. Second, there exists a structural break in 1984. Since that year, one-time payments and other bonuses have been included in the reported earnings leading to an increase in the observed inequality of wages at that time. The correction suggested by Fitzenberger (1999) is used as a conservative correction (see also, among others, Dustmann et al. (2009); Fitzenberger and Wunderlich (2002); Glitz and Wissmann (2017), who use such a correction).[11] Third, the IABS does not provide detailed information on hours worked, but it provides an indicator for full-time work. As we restrict the analysis to full-time working males, our results are likely to be robust and comparable to the U.S.-data. Recall that the studies mentioned at the end of the previous paragraph are based on data as reported in the IABS.

Workers are grouped by their skills according to the following formal education levels given in the IABS:

(U) without a vocational training degree	(low-education)
(M) with a vocational training degree	(medium-education)
(H) with a technical college ("Fachhochschule") or a university degree	(high-education)

The education groups for the U.S. are defined by years of schooling while the grouping for Germany is based on educational degrees because of the importance of the vocational training system in Germany. A number of medium-education degrees in Germany would rather correspond to tertiary degrees in the U.S. e.g., technical degrees. We follow the common definitions taken in the literature on wage inequality for the two countries to make our results comparable to the literature.

In light of the polarization hypothesis, our choice of education groups was driven by the desire to be able to analysis non-monotonic wage trends/profiles with regard to higher education, i.e., whether the medium-education group is losing ground relative to the high- and low-education

---

[8] This study uses the factually anonymous IAB Employment Sample (IABS) (Years 1975–2004), see Drews (2008). Data access was provided via a Scientific Use File supplied by the Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB).

[9] Between 1975 and 1979, a slight increase of wage dispersion in the upper part of the distribution takes place and virtually no change in wage-dispersion in the lower part, as measured by the 80%–50% and 50%–20% differences in log-wages.

[10] The value of this threshold changes annually.

[11] This correction amounts to correcting wages before 1984, which are above the median, by the estimated disproportionate wage growth between 1983 and 1984. This disproportionate wage growth is estimated as a linear function of the rank difference from the median upwards. No correction is implemented below the median.

groups. If polarization in wages or employment is relevant, the precise definition of the education groups would not matter as long as the medium group covers the middle of the education distribution, which clearly is the case for both countries (see evidence on employment shares in Figure 1 as discussed in Section 3.2), even though the size of different education groups differ by country.

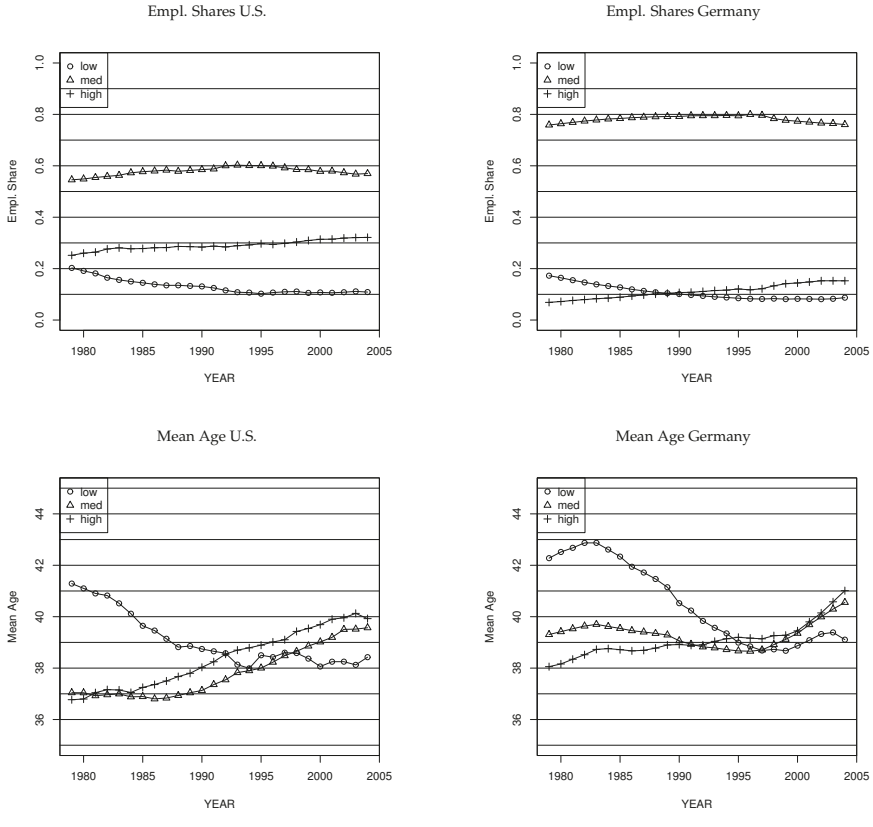


Figure 1. Employment Shares and Mean Age 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

2.3. Construction of Cohort-Year-Education Cells

Our level of analysis are wage quantiles by year, cohort/age, and education level, where cohort is defined by year of birth. For each cell, we calculate different quantiles for the real wage. Applying the approach proposed by Fitzenberger (1999), this is done for the German data in the following way. The IABS contains information on the social security insurance spells comprising the starting point and the end point as well as the average daily gross wage[12] (excluding employer's distribution) for this spell.

An annual wage observation for one individual is calculated as the weighted average of the wages he earned during his different spells within one year, where the spell lengths are used as the weights. The sum of the spell lengths for all individuals in one cell is used to calculate the number of employed workers within this cell. This variable is used as a weight in the regressions.

[12] The daily social security threshold is reported instead if the daily gross wage exceeds the upper social security threshold, see above.

The next step consists of calculating the 20%, 50%, and 80% quantile for the cells, where again the spell lengths are used as weights. We also record the sum of spell lengths as cell weights. In the case of Germany, when the quantile coincides with the threshold, it is recorded as being censored. These information are sufficient for our empirical analysis to estimate quantile regressions based on cell data. The cohort year-skill cell data for the CPS are constructed in an analogous way as for the German data, using the weights described above.

3. Basic Empirical Facts

3.1. Unconditional Wage Growth

Figure 2 depicts the wage growth jointly for all education groups between 1979 and 2004. For the U.S., wages at the three quantiles fall until 1996, with the largest decline at the 20% quantile being −13 log points. Wages at the median decline 10 log points and those at the 80% quantile decline 4 log points. This implies rising wage dispersion both in the upper and the lower part of the U.S. wage distribution. Between 1996 and 2004, wages grow at all quantiles, whereby wages at the 20% quantile and at the 80% quantile rise about 9 log points, which is 1–2 log points more than the rise of the wages at the median. This is evidence for a polarization of wages between 1996 and 2004. Overall, however, between 1979 and 2004 the wage dispersion increased both in the upper half and the lower half of the distribution—as measured by the 80–50 and the 50–20 difference of log-wages, respectively.

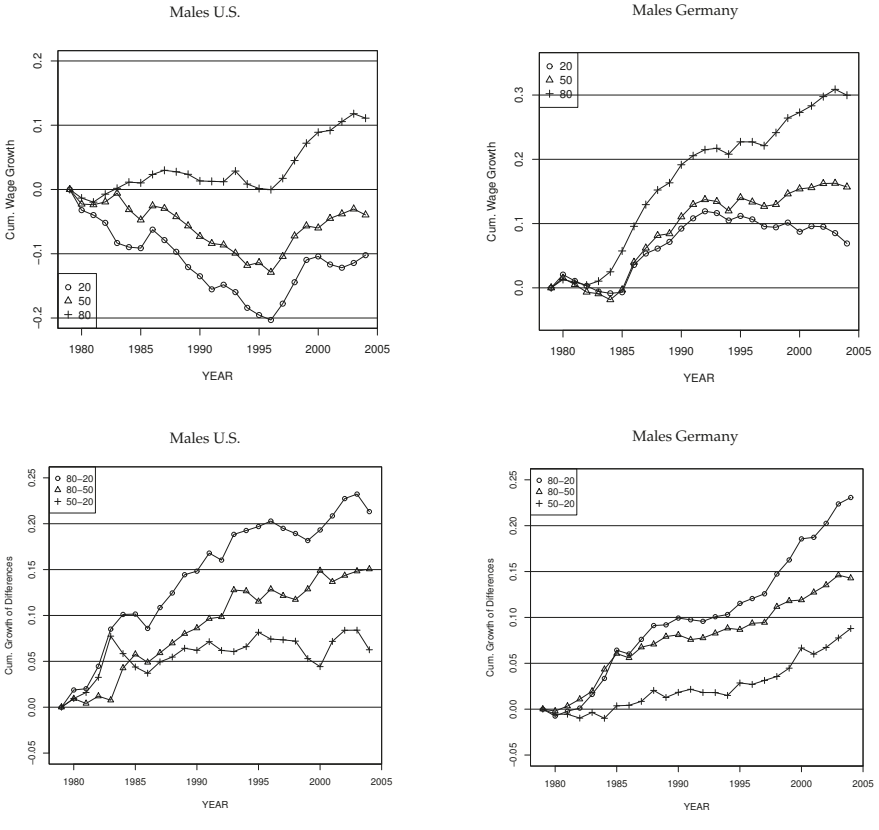


Figure 2. Total Unconditional Cumulated Wage growth at 20%, 50%, 80% quantiles and quantile differences, 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

In Germany, wages throughout the distribution start to grow in the mid-1980s, and wages at the 80% quantile exhibit larger growth rates than those at the median and the 20% quantile. Wage inequality in the upper part of the wage distribution keeps rising steadily since the beginning of the 1980s, while wage dispersion in the lower part of the wage distribution only starts to increase in the mid-1990s. These results are in line with Dustmann et al. (2009) and Biewen et al. (2017). Between 1979 and 2004, the 20% quantile, the median, and the 80% quantile increase by 9, 15, and 20 log points, respectively, i.e., real wage growth is considerably higher in Germany compared to the U.S. Finally, in Germany, the 20% quantile and the 80% quantile only grow both faster than the median during the early 1980s, which is evidence for polarization of wages.

Turning to education group specific trends, Figure 3 shows the unconditional wage growth at different quantiles conditional on education and Figure 4 summarizes overall wage dispersion. Between 1979 and 1996, real wages of low-education workers in the U.S. fell by about 32–34 log points. After 1996, real wages recovered for this group and there was a sharp decline in wage inequality below the median. Wages of medium-education workers also increased after a low in 1996 and a clear pattern of polarization is observable since the early 1990s, as the 80–50 difference keeps increasing and the 50–20 difference starts to decrease. In the U.S., only the group of high-education workers experienced real wage gains between 1979 and 2004. Wage inequality steadily increased for this group since the late 1980s. Our findings are similar to, e.g., Autor et al. (2008).

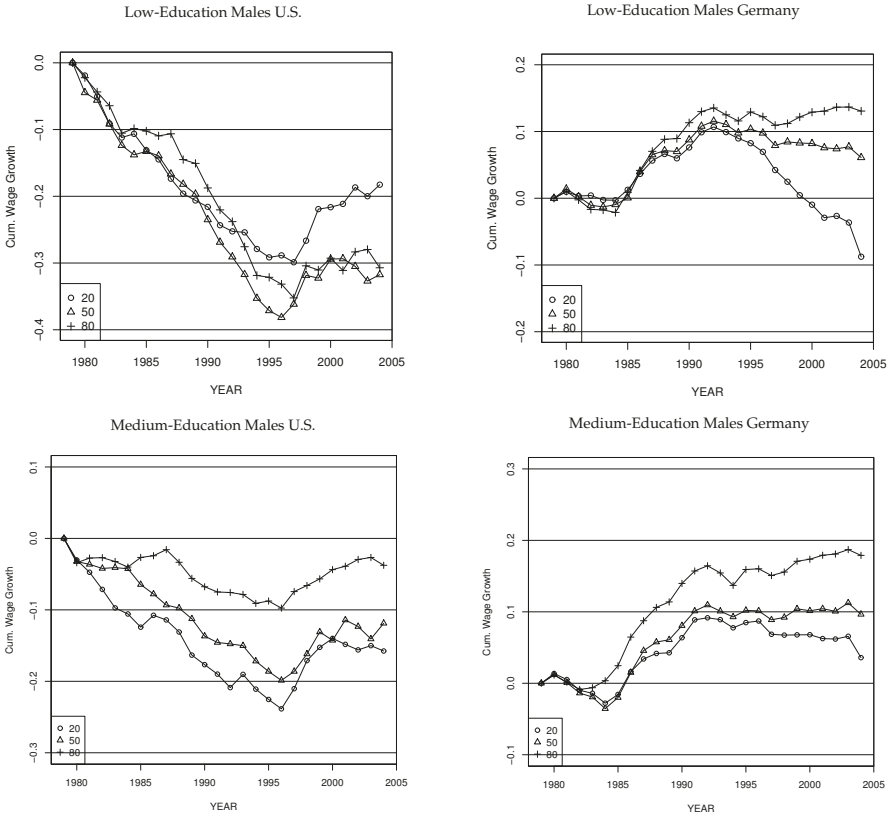


Figure 3. *Cont.*

Figure 3. Unconditional Cumulated Wage growth 19799–2004 for males: (**left**) U.S.; and (**right**) Germany.

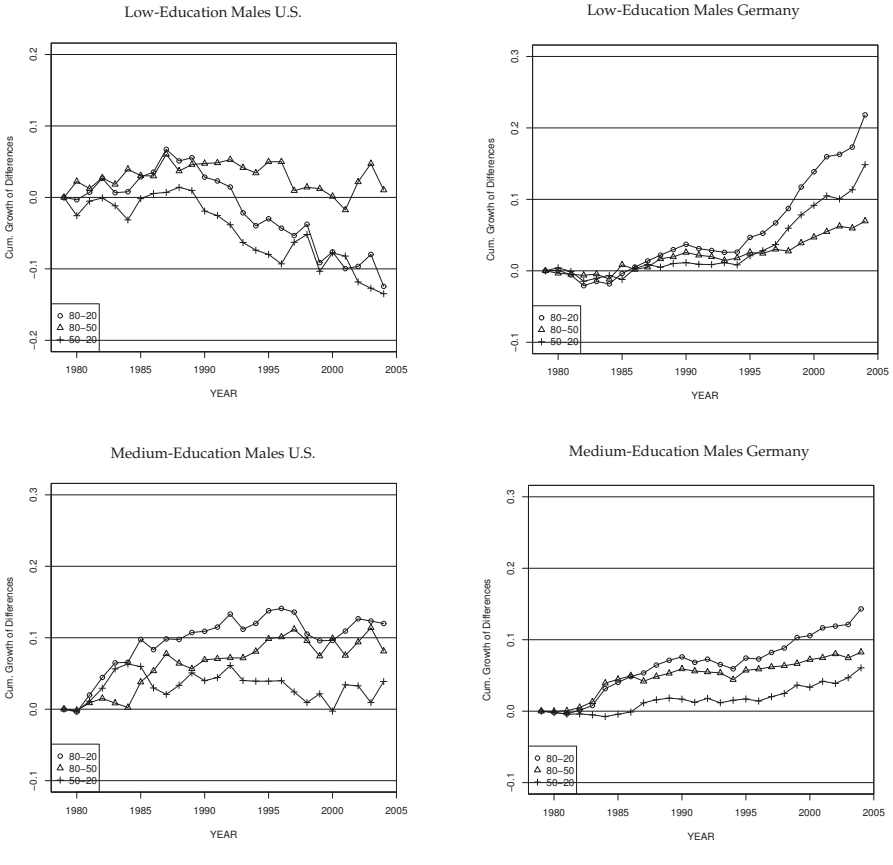


Figure 4. *Cont.*

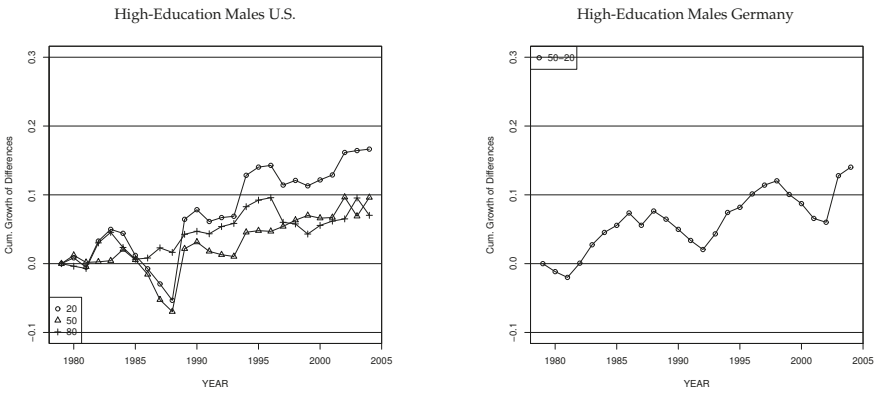


Figure 4. Unconditional Cumulated wage dispersion 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

In Germany, only low-education workers at the 20% quantile had lower real wages in 2004 than in 1979 (a 10 log points cumulative decline) and the decline began in the early 1990s. During the last twelve years of observation, the 20% quantile of wages of the low-education fell by 20 log points. Wages at the median fell to a lesser extent, while wages at the 80% quantile have been flat since the early 1990s. Up until 1991/1992, wage trends were quite uniform but after the severe recession in 1992/1993, wage dispersion has been increasing along the entire distribution. Medium-education workers in Germany, making up the major part of the entire German workforce, experience quite similar movements as described above for the overall wage distribution not conditioning on educational-level—rising wage dispersion in the upper part beginning in the 1980s and increasing wage inequality in the lower part of the distribution since the mid-1990s. Furthermore, similar to the development of the entire wage-distribution, we observe a polarizing pattern of wages until 1984. German high-education workers experience considerable gains since the early 1980s: wages rose by 17 log points and 30 log points for workers at the 20% quantile and the median respectively.

Figure 5 displays the skill premia (measured at the median). In the U.S., the premia for high-education workers relative to medium-education workers and for medium-education workers relative to low-education workers increased steadily from 1979 to 2004. By contrast, the medium-to-low premium in Germany fell during the early 1980s and grew slowly between the mid-1980s and 2004 (See Glitz and Wissmann 2017) for related evidence. The high-to-medium premium grew considerably during the late-1980s and again during the late-1990s and early 2000s.

Figure 5. Unconditional Wage Premia: (**left**) U.S.; and (**right**) Germany.

3.2. Employment Changes

Figure 1 plots the employment shares of the different education groups. Incidentally, in both the U.S. and Germany, the share of low-education workers ceased to decline in the mid-1990s, i.e., skill upgrading from low-education workers stopped at that time. For both countries, increased immigration might help to explain these trends.[13] Medium-education workers in both countries are the largest education group. Their employment shares grew slightly until the mid-1990s and fell slightly afterwards in both the U.S. and Germany. The share of high-education workers rises monotonically in both countries, while the relative rise is more pronounced in Germany, doubling from 8% in 1979 to 16% in 2004, whereas over the same period the share in the U.S. rises from 16% to 22%.

To investigate changes in the age structure of employment, Figure 1 plots the mean age of the workers in the different education groups. The average age of U.S. medium-education and high-education workers has been increasing since the mid-1980s. The mean age of low-education workers in the U.S. decreased strongly until the mid-1990s and remained constant afterwards. For Germany, the mean age of medium-education and high-education workers has been rising continuously since the mid-1990s. Similar to the U.S., the average age of low-education workers fell strongly until the middle of the 1990s and grew slightly afterwards. This finding may be explained by immigration and by the fact that older low-education workers tend to leave the workforce disproportionately.

4. Empirical Approach

This section presents the empirical framework developed by MaCurdy and Mroz (1995) to investigate the movement of the entire wage distribution for synthetic cohorts over time. A cohort is defined by the year of birth. We estimate various quantile regressions to decompose between- and within-group shifts in the wage distribution. We allow that wage trends differ across cohorts indicating the presence of *cohort effects* and across quantiles indicating a trend towards changing within group wage dispersion.

A cohort effect designates a movement of the entire life-cycle wage profile for a given cohort relative to other cohorts. In providing a parsimonious representation of trends in the entire wage distribution, we are able to pin down precisely the differences in wage trends across groups of workers defined by education level.

Due to the inherent identification problem between age, cohort, and time effects, wage profiles based on cross-section relationships between age and wages over a sequence of years and movements of life-cycle wage profiles faced by successive cohorts are mathematically the same mapping. However, considering the wage growth experienced by a particular cohort over time or over age, it can be tested whether apart from the differential age effect, different cohorts exhibit the same time trend.

4.1. Characterization of Wage Profiles

We denote the age of an employee by α and calendar time by t. A cohort c can be defined by the year of birth. The variables age, cohort and calendar year are linked by the relation $t = c + \alpha$, which causes the age-period-cohort identification problem. Studies of wage trends often investigate movements of "age-earnings profiles"

$$ln[w(t, \alpha)] = f(t, \alpha) + u. \quad (1)$$

[13] For Germany, following the reunification in 1990, a large inflow of ethnic Germans as well as a wave of immigration of workers from East Germany (the former German Democratic Republic, GDR) is well documented in the literature (see, e.g., Bundesamt fur Migration und Fluchtlinge 2005; Fuchs-Schundeln and Schundeln 2009). D'Amuri et al. (2010) and Glitz and Wissmann (2017) concluded that this immigration waves had only a small impact on wages among the natives.

The deterministic function $f(t,\alpha)$ measures the systematic variation in wages (modeling a certain location measure—in our case, conditional quantiles) and u reflects cyclical or transitory variations. For a fixed year t, the function $f(t,\alpha)$ yields the conventional cross-sectional wage profiles. Movements of f as a function of t describe how cross-sectional wage profiles shift over time. The cross-sectional relation f as a function of age does not describe "life-cycle" wage growth for any cohort or, put differently, the cross-sectional relation may very well be the result of "cohort effects". In fact, "cohort earnings profiles" $g(c,\alpha)$ given by

$$g(c,\alpha) \equiv g(t-\alpha,\alpha) \equiv f(t,\alpha) \tag{2}$$

represent mathematically the same function as "age-earnings profiles" $f(t,\alpha)$, i.e., $g(c,\alpha)$ and $f(t,\alpha)$ represent the same mapping of (c,α,t) to log wages.

The function g describes how age-earnings wage profiles differ across cohorts. Holding age constant, $g(c,\alpha)$ describes the profiles of wages earned by different cohorts over time. Holding the cohort constant yields the profile experienced by a specific cohort over time and age. The latter is referred to as the "life-cycle profile", because it reflects the wage movements over the life-cycle of a given cohort. Without further assumptions, "pure life-cycle effects" due to aging or "pure cohort effects" cannot be identified, because, e.g., variations in wages due to aging may be associated with changing time based on $g(c,\alpha)$ or changing cohort based on $f(t,\alpha)$.

4.2. Testing for Uniform Insider Wage Growth

Our analysis investigates whether wage trends are uniform across cohorts in the sense that every cohort experiences the same time trend in wages and the same age-specific wage growth (life-cycle effect). Despite the age-period-cohort identification problem, the existence of a uniform (separable) time trend across cohorts is a testable implication.

The following notion of wage growth proves useful: Wage growth for a given cohort in the labor market over time ("Insider Wage Growth"), given by

$$\frac{\partial g}{\partial t}|_c = \frac{\partial g}{\partial \alpha}|_c \equiv g_\alpha(c,\alpha) \equiv g_\alpha, \tag{3}$$

comprising the simultaneous change of time and age. Alternatively, holding age constant yields the change of wages earned by different cohorts at specific ages. For the age at labor market entry, α_e, entry wage growth is given by

$$\frac{\partial g}{\partial t}|_{\alpha=\alpha_e} = \frac{\partial g}{\partial c}|_{\alpha=\alpha_e} \equiv g_c(c,\alpha_e) = g_c(t-\alpha_e,\alpha_e) \equiv e(t), \tag{4}$$

again comprising two effects, namely a change of cohort and time. Equation (4) describes entry wage growth.

If wage growth is separable in age and time, i.e., is the sum of a pure aging effect and a pure time effect as follows

$$g_\alpha = a(\alpha) + b(t) = a(\alpha) + b(c+\alpha), \tag{5}$$

then life-cycle wage growth $a(\alpha)$ is the same for each year t. Condition (5) is designated as the "uniform insider wage growth hypothesis" (H_{UI}). If H_{UI} holds, we can construct a "life-cycle wage profile" independent of the calendar year and a macroeconomic time trend independent of age. We will investigate H_{UI} by testing for the significance of interaction terms of α and t in the specification of g_α.

Integrating back condition (5) on the derivative g_α with respect to α yields an additive form for the systematic component of the wage function $g(c,\alpha)$:

$$g(c,\alpha) = G + K(c) + A(\alpha) + B(c+\alpha), \tag{6}$$

where $G + K(c)$ is the cohort specific constant of integration. At a given point in time, the wages of cohorts differ only by the age-effect, given by $A(\alpha)$, and by a cohort specific level, given by $K(c)$. The "uniform insider wage growth hypothesis" H_{UI} can be tested by investigating whether "interaction terms" $R(\alpha, t)$ enter specification (6) which are constructed as integrals of interaction terms of α and t in g_α.

4.3. Empirical Implementation

We specify the wage function $g(c, \alpha)$ for individual i in the sample year t using a fairly flexible functional form:

$$ln[w_{i,t}] = g(c_i, \alpha_{i,t}) + \overline{u}_t + u_{i,t} \tag{7}$$

where $\alpha_{i,t}$ and c_i denote the age of individual i at time t and the cohort of individual i, respectively. $g(c, \alpha)$ is specified as a smooth function of c and α. We further decompose the error term into a period specific fixed effect $\overline{u}_t$ and a stochastic error term u_{it}. In the empirical analysis, we take 25 years to be the age of entry into the labor market and we define $\alpha = (\text{age} - 25)/10$ and therefore $\alpha_e = 0$. Analogously, since the observation period starts in 1979, we define time $t = (\text{calendar year} - 1979)/10$. For each cohort, c corresponds to the time t at which α equals zero. For the cohort of age 25 in the year 1979, c equals zero and older cohorts have negative values for c.

As a flexible empirical approximation of the wage profile imposing the hypothesis of uniform insider wage growth, we use polynomials in age, cohort, and time:

$$\begin{aligned} A(\alpha) &= A_1\alpha + A_{(2)}(\alpha) = A_1\alpha + A_2\alpha^2 + A_3\alpha^3 \\ B(t) &= B_1 t + B_{(2)}(t) = B_1 t + B_2 t^2 + B_3 t^3 + B_4 t^4 + B_5 t^5 \\ K(c) &= K_1 c + (1-\delta)K_b(c) + \delta K_a(c) \\ \text{with } \delta &= 1 \quad \text{for} \quad c \geq 0 \quad \text{and } \delta = 0. \end{aligned} \tag{8}$$

We include year dummies that are orthogonalized with respect to $B(t)$ to estimate period specific fixed effects $\bar{u}_t$, i.e., the estimated year effects are uncorrelated with the estimated smooth time trend $B(t)$, see Fitzenberger and Wunderlich (2002) for details. Altogether, we fully saturate the time dimension, i.e., the model is estimated as if a complete set of year dummies is used. However, $B(t)$ is estimated just as if no further year effects $\bar{u}_t$ where included. Thus, $\bar{u}_t$ represents the year specific deviation from the smooth trend $B(t)$, which we interpret as cyclical year effect. We estimate a fifth order polynomial in time for $B(t)$, yielding a satisfactory decomposition of trend and cycle.[14] Equation (8) allows for a third order polynomial in age, a third order polynomial in $K_b(c)$, and a second order polynomial in $K_a(c)$.[15]

To test H_{UI}, we consider in the derivative g_α the following four interaction terms of age and time αt, αt^2, $\alpha^2 t$, and $\alpha^2 t^2$. The implied non-separable variant of $g(c, \alpha)$ expands (6) by incorporating the integrals of these interaction terms, denoted by R_1-R_4, see MaCurdy and Mroz (1995) and Fitzenberger and Wunderlich (2002) for details, and we test for significance of R_1-R_4.[16]

---

[14] Antonczyk et al. (2017) suggest a nonparametric estimator for a model with $g(c, \alpha) = A(\alpha) + B(t) + K(c)$ where $t = c + \alpha$ and apply the model to the sample of the low-education and the medium-education in Germany. The findings for this restricted model are very similar to the findings based on the flexible parametric model specification used here.

[15] We did investigate the robustness with regard to different specifications with regard to higher order terms in age or cohort (results are available upon request). Higher order terms did not affect the main results but they could add wiggly behavior in the ends of the estimated age/cohort profiles.

[16] For instance, $R_1 = \int \alpha(c + \alpha)d\alpha = (c\alpha^2/2) + (\alpha^3/3)$.

Only if H_{UI} holds, it is meaningful to construct an index of a life-cycle wage profile as a function of pure aging *and* a macroeconomic trend index. Otherwise, a different wage profile would apply for each cohort. Under H_{UI}, the life-cycle (L) is given by

$$ln[w_L(\alpha)] = (A_1 - K_1)\alpha + A_{(2)}(\alpha) \tag{9}$$

and the macroeconomic (m) wage trend index is given by

$$ln[w_m(t)] = (B_1 + K_1)t + B_{(2)}(t). \tag{10}$$

When interpreting these indices, it is important to recognize that neither the level nor the coefficient on the linear term are identified in an econometric sense. In fact, identification relies on the assumption that the coefficient on the linear cohort term is equal to zero. To motivate this, we argue that setting the linear cohort term to zero is quite natural. If, for instance, also entry wages grow at the same rate as the time effect $b(t)$ before and during the sample period, the entire cross-section profile $f(\alpha, t)$ exhibits purely parallel shifts over time, a situation, one would not naturally characterize by the existence of "cohort effects". Our notion of a cohort effect requires a situation where the differences in starting points of the common life-cycle profile differ from the macroeconomic wage growth experienced by the cohorts in the labor market. For this reason, we also orthogonalize our polynomial specifications for $K_a(c)$ and $K_b(c)$ with respect to the linear cohort effect.

Quantile regressions provide a useful tool to study wage differences across and within groups of workers with different socio-economic characteristics. We estimate conditional quantiles of wages

$$q_\theta(ln[w_{i,t}]|c, \alpha, \beta^\theta) = g^\theta(c, \alpha, \beta^\theta), \tag{11}$$

where $q_{\theta,t}(ln[w_{i,t}]|c, \alpha, \beta^\theta)$ denotes the θ-quantile of the wage in cohort age-cell (c, α) ($\equiv$ cohort year-cell (c, t) where $t = c + \alpha$). The vector β^θ comprises the coefficients relating to the set of regressors ($\equiv$ powers of c, α and t; year dummies). In the empirical analysis, we model the following quantiles: $\theta = 0.2, 0.5, 0.8$ (20%, 50%, and 80% quantile).

We use the minimum-distance approach proposed by Chamberlain (1994) or MaCurdy and Mroz (1995) for the estimation of quantile regressions when the data on the regressors can be grouped into cells and censoring is not too severe. The approach consists of calculating the respective cell quantiles in a first stage and regressing (by weighted least squares) those empirical quantiles, which are not censored, on the set of regressors in the second stage. For the dataset used in this study, the cell sizes are large enough for making this a fruitful approach. However, for Germany, we do not estimate the 80% quantile for males in education group (H) since censoring is too severe in this case. When applying the minimum-distance approach, we use the cell sizes as weights.

The error terms are allowed to be dependent across individuals within cohort year-cells and across adjacent cohort year-cells. We use a flexible moving block bootstrap approach allowing for standard error estimates which are robust against fairly arbitrary heteroscedasticity and autocorrelation of the error term. The block bootstrap approach employed here extends the standard bootstrap procedure in that it draws blocks of cell observations, including the cell weights, to form the resamples. While the goal to capture dependence across observations is similar to a cluster bootstrap, we do not rely on fixed cluster but rather moving clusters which overlap. Specifically, we draw a two-dimensional block of observations with block length eight in the cohort and block length six in the time dimension with replacement until the resample has become at least as large as the resample size, see Fitzenberger and Wunderlich (2002) for details. Contrasting the results using the moving-blocks-bootstrap approach with conventional standard error estimates indicates that allowing for correlation between the error terms within and across cohort year-cells (when forming the blocks) changes the estimated standard errors considerably (detailed results are available upon request).

5. Results

Based on the empirical framework introduced above, this section discusses the estimated specifications and then presents the empirical results.

5.1. Estimated Specifications for Wage Equations

We estimate two specifications for the 20%, 50%, and 80% quantile for males by education groups (U), (M), and (H). The high degree of censoring allows only a meaningful estimation of the more restrictive specification for the 20% and the 50% quantile in the case of high-education (H) males in Germany.

The more general specification (Model 1), which does not impose the uniform insider wage growth hypothesis (H_{UI}) introduced in Section 4.2, is given by

$$\begin{aligned} g(c,\alpha)+\bar{u}_t &= G + a_1\alpha + a_2\alpha^2 + a_3\alpha^3 + b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5 \\ &\quad + \gamma_{b2} c_b^2 + \gamma_{b3} c_b^3 + \gamma_{a2} c_a^2 + \gamma_{a2} c_a^3 + \sum_{j=1}^{4} \rho_j R_j + \sum_{i=1979}^{2004-N_b-1} \kappa_i YD_i, \end{aligned} \tag{12}$$

where the age polynomial is of order 3, the time polynomial of order 5, and $c_b = (1-\delta)c$ and $c_a = \delta c$ are the cohort terms before and after 1979, orthogonalized with respect to the linear cohort term. All specifications include the cyclical year dummies YD_i which are orthogonalized with respect to the time trend, thus $N_b = 5$ (we lose six degrees of freedom $[2004 - N_b - 1]$ because an intercept is included).

Model 2 is the restricted version of Model 1 in Equation (12):

$$\text{Model 2: Specification (12) with } \rho_j = 0, \text{ for } j = 1, ..., 4\,. \tag{13}$$

Model 2 imposes H_{UI}, i.e., separability of wage growth into age and time effects. Statistical tests imply that for all education groups at the three quantiles considered both life-cycle profiles [Equation (9)] and macro-trends [Equation (10)] are the same across cohorts, because H_{UI} cannot be rejected at a 1% significance levels (detailed results are available in the Supplementary Material, Tables S1 and S2). In all cases except two, we also do not find significance at the 5% significance level. Note that we do not perform the test for high-education workers in Germany because of the high degree of censoring in this group (see Section 2). Because of the evidence in favor of H_{UI}, we only report in the following estimation results for Model 2. For this model, the estimation of time trends and life-cycle profiles is thus meaningful. Even though, further hypothesis tests would suggest to use a more parsimonious specification of model (see again Tables S1 and S2 in the Supplementary Material), we do not further restrict the model specifications because we do not want to base the comparison across worker groups on differences in specification choices. We also include estimates of Model 2 for high-education workers in Germany for the 20%- and the 50%-quantile.

5.2. Life-Cycle Profiles

Figure 6 shows the estimated life-cycle profiles. Note that wage growth over the life-cycle at the median wage, which closely relates to a standard human capital wage equation (Gosling et al. 2000), is positively correlated with educational level—i.e., the descriptive returns to experience are increasing with education. For most cases, life-cycle wage growth is higher at higher quantiles (and mostly significantly so, detailed results are available upon request), i.e., inequality within education group typically increases with age. There are three exceptions: At the 50% and 80% quantile for medium-education in the U.S. and for low-education in Germany, life-cycle profiles basically coincide which implies that upper-tail wage dispersion does not increase with age. For low-education in Germany, life-cycle wage growth is even higher at the 20% quantile than at the median, i.e., for this education group lower-tail wage dispersion falls with age.

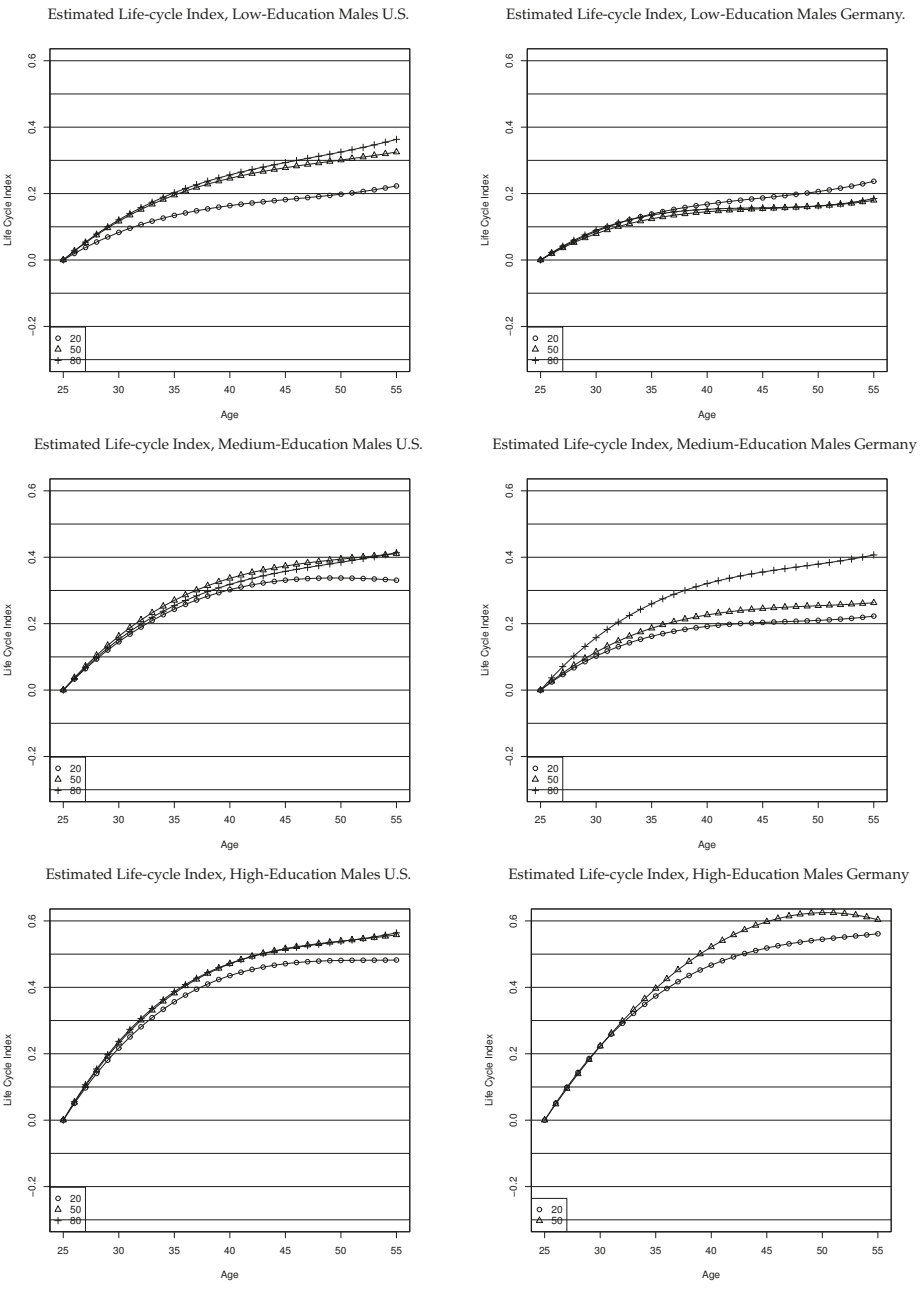


Figure 6. Life-cycle indices 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

Despite the similar concave profiles, the amount of within-cohort life cycle wage growth differs both across education groups and across countries. Life-cycle wage growth increases with education in both countries and is higher in the U.S. for low-education at the two higher quantiles and for medium-education at the two lower quantiles. Life-cycle wage growth is very similar in both countries at the upper quantile for medium education and at the two lower quantiles for high education.

The decreasing within-cohort wage dispersion when workers age for low-education in Germany may be due to a selection process. Older German low-education workers at the bottom of the skill-specific wage distribution might drop out of the labor-market as they get older, e.g., due to layoffs, if their productivity lies below the wages set by union wage agreements. Contrary to low education, the increase of within-cohort wage dispersion associated with aging is twice as strong for German medium-education workers compared to U.S. medium-education workers. This may reflect the larger heterogeneity of the medium education group in Germany, which comprises a higher share of workers compared to the medium education group in the U.S.

The development of wage dispersion over the life-cycle for the U.S. is in line with findings for the UK (Gosling et al. 2000). Wage dispersion over the life-cycle grows less for higher education levels, i.e., in the U.S. low-education workers experience the highest increase in wage dispersion over the life-cycle, while for Germany dispersion increases most strongly for medium-education in the upper part of the wage distribution.

5.3. Time-Trends

Figure 7 depicts trends in real wages due to macroeconomic shifts in the U.S. and Germany. Time-trends in the U.S. were more positive for workers with higher educational attainment than for low- and medium-education workers. Comparing low- and medium-education workers in Germany at the different quantiles, we see that time-trends in wages were roughly the same across education groups. Time-trends for German high-education workers were similar to those of less skilled workers until the early 1990s, but wage growth was stronger thereafter. Finally, our estimates suggest that time-trends in wages developed more positively for German workers than for U.S. workers.

The mid-1990s mark a turning point in the development of the macro wage indices of both low-education and medium-education worker in the U.S. Until that point in time, workers in both subgroups experienced real wage losses throughout the entire wage distribution, being stronger for the low-education (-30 log points at the 80% and 20% quantile and -32 log points at the median). Medium-education workers incurred losses of -11, -20, and -22 log points at the 80%, 50%, and 20% quantile, respectively. Between 1996 and 2004, however, wages grew considerably at all considered quantiles of both low- and medium-education workers. Wages for low-education at the 20% quantile grew by 10 log points, wages at the median and at the 80% quantile by 5 log points. For medium-education, the wage growth starting in the mid-1990s was less pronounced. Wage growth was about 4 log points at both the 20% and the 80% quantile and about 3 log points at the median. Time trends are most positive for high-education workers in the U.S., with a cumulated wage growth of -1, 8, and 17 log points at the 20%, 50%, and 80% quantile, respectively, between 1979 and 2004.

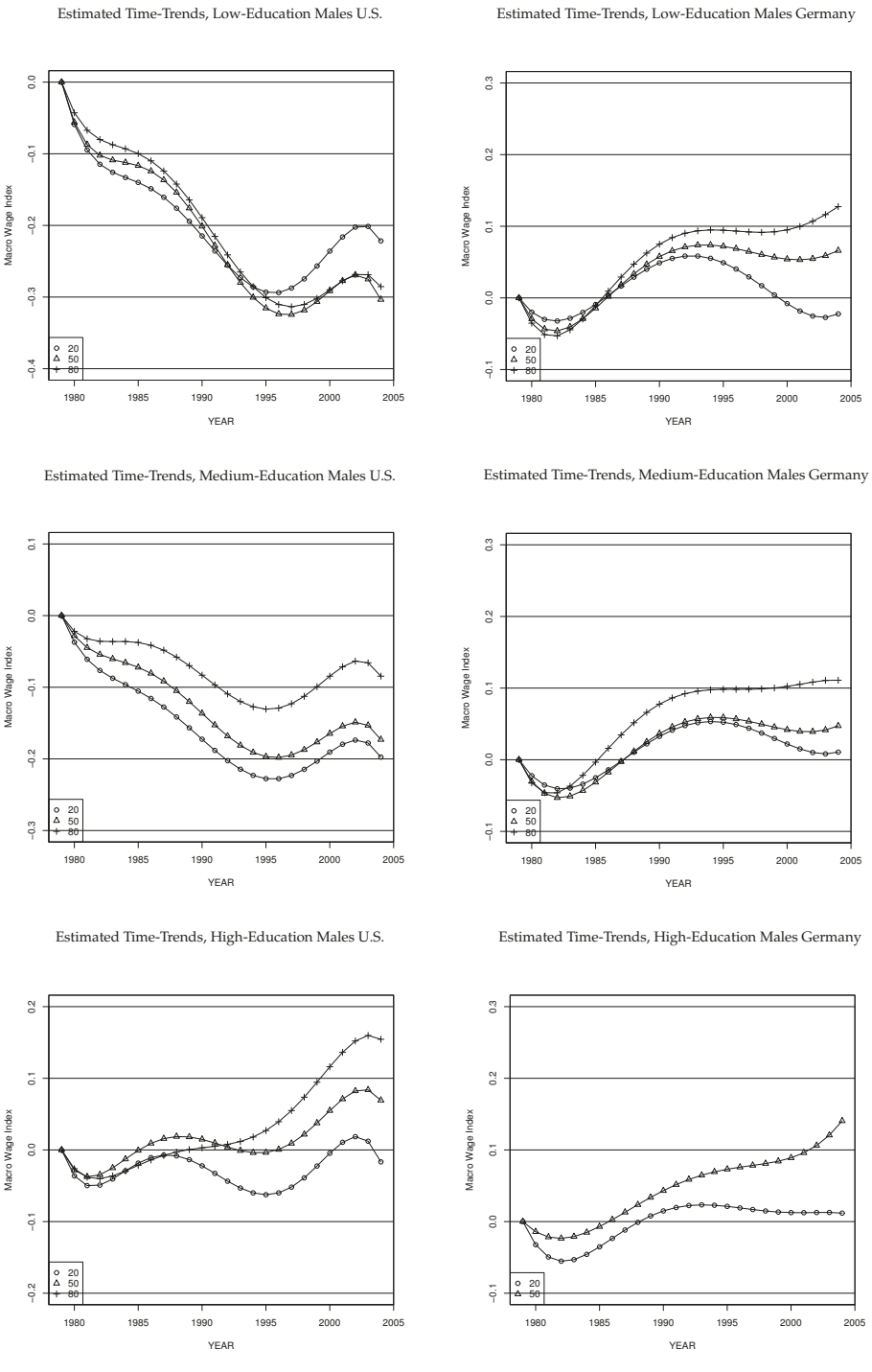


Figure 7. Time-Trends 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

For low-education workers in Germany, the 20%, 50%, and 80% quantiles of the wage distribution move in a parallel manner between 1979 and 1992, resulting in an uniform gain of about 8 log points along the entire distribution. Thereafter, wages at the 80% quantile exhibit small gains, while the wages at the 20% quantile decrease, resulting in real wage losses of 5 log points between 1992 and 2004. Wages at the median remain flat during this period.[17] Medium-education workers in Germany do slightly better than low-education workers, in terms of time-trends at the lower end of the skill-specific wage-distribution. Time-trends for wages at and above the median are fairly similar. Cumulated wage growth at the 20% quantile for German medium-education workers is slightly above zero, compared to real wage losses of about 2 log points in the group of the low-education. However, this masks the fact that since the beginning of the 1990s, real wage losses are more pronounced among low-education workers in the lower part of the distribution. Wages at the 20% quantile of German high-education workers were staying flat since the beginning of the 1990s. Over the entire period, cumulated wage growth is about 1 log points for this group at the 20% quantile. The time-trend for German high-education workers at the median starts to increase monotonically in the early 1980s, at an annual rate of about 0.5 log points. Wages at the 20% quantile were rising between the early 1980s and the early 1990s, but then started to flatten out.

5.4. Cohort Effects

There are a number of reasons for the existence of cohort effects. Not being exhaustive, we discuss three. First, Card and Lemieux (2001) argued that the increasing wage premium between college graduates and high-school graduates is due to a slowdown in the growth of supply of higher-skilled workers. Second, cohort effects may reflect changes in educational policy, or more generally, any pre-labor market conditions (Carneiro and Lee 2011). Third, cohort effects may reflect labor market conditions at labor market entry which may have lasting effects over the life course (see, e.g., Berger (1985) for the effect of cohort size). In a labor market with frictions, cohort effects may be implied by wage adjustments which are strongest among younger workers at labor market entry and which may persist over the life cycle. These mechanisms may operate at the same time and our analysis will not be able to distinguish between them.

Figure 8 plots the estimated cohort effects for the different groups in both economies. These are quadratic and cubic terms for cohorts that enter the labor market before and after 1979, orthogonalized to the linear cohort term. For both medium- and high-education workers in the U.S., negative cohort effects are estimated for the oldest cohorts and positive effects for the youngest cohorts. For low-education workers, we find positive cohort effects for the youngest cohorts and negative ones for the oldest cohorts at the 80% quantile. Interestingly, we find that during the 1980s cohort effects had a positive effect on medium-education and high-education workers—this is the period for which Card and Lemieux (2001) observed increasing skill premia among younger workers for the U.S. [18] For Germany, for all education groups, both the youngest and the oldest cohorts exhibit negative cohort effects, relative to the cohorts entering the labor market between the mid-1960s and mid-1980.[19] Furthermore, the youngest cohorts experience higher within-cohort wage dispersion due to these effects.

---

[17] One possible cause for the declines in wages among low-education workers at the lower end of this wage distribution (and therefore at the lower end in the overall wage distribution) may be the large inflow (immigration) of low-education workers into West-Germany after the reunification, resulting in an higher supply of low-education workers, in combination with the recession that took place in Germany in 1992/1993, see Section 3. However, recall that D'Amuri et al. (2010) and Glitz and Wissmann (2017) concluded that immigration waves in Germany had only a small impact on wages among the natives.

[18] Increasing wage dispersion due to cohort effects across education groups may also indicate selection effects, i.e., the "ability" of workers within education groups can change over time (Carneiro and Lee 2011).

[19] Due to the severe censoring, we find only cohort effects for the younger German high-education workers. The youngest high-education workers are also negatively affected by cohort effects.

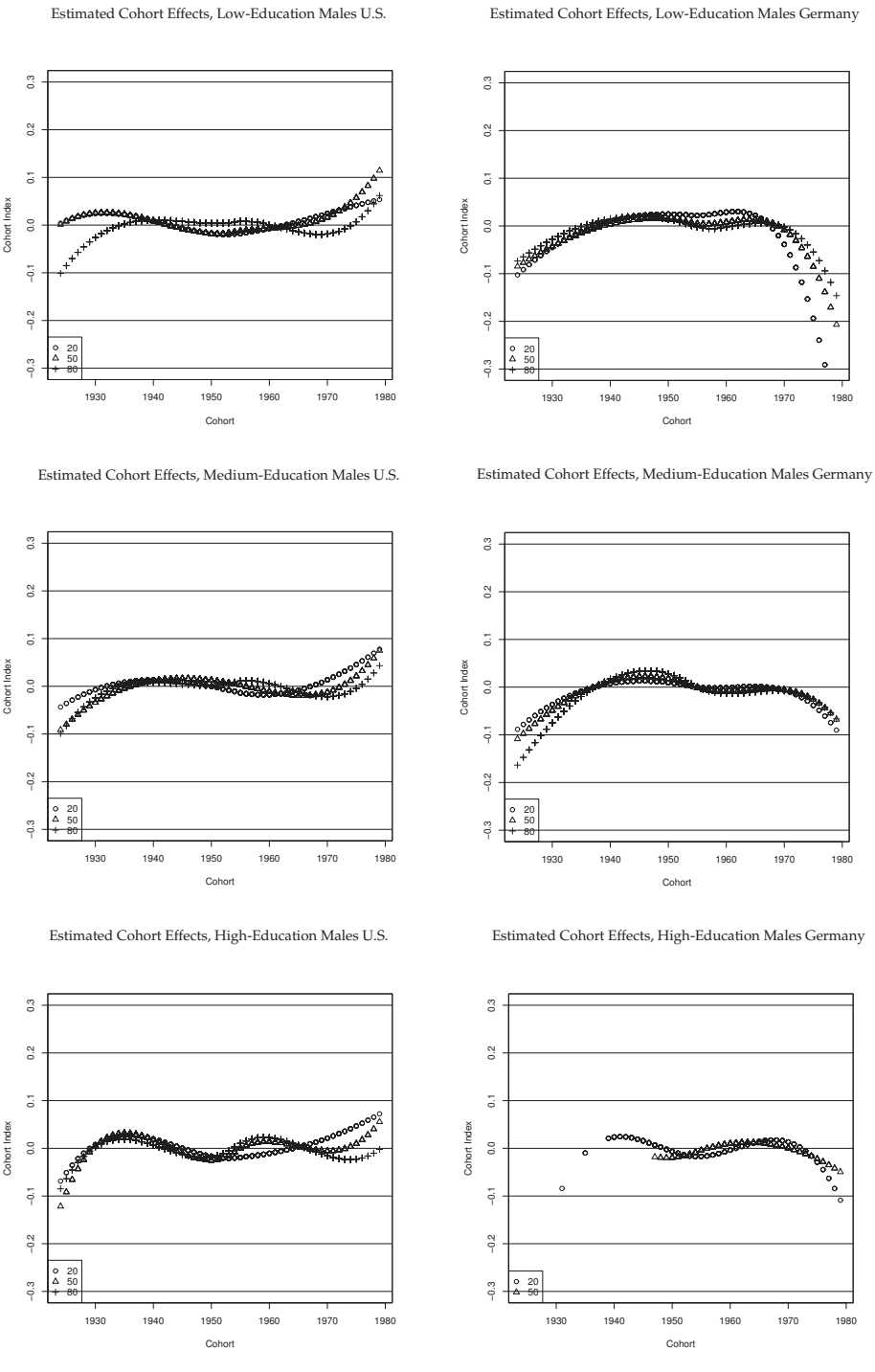


Figure 8. Cohort Effects 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

The trend in entry wages is the sum of cohort effects and the macroeconomic time-trend (see Section 4; Figure S1 in the Supplementary Material shows the estimated entry wages over time). In the U.S., positive cohort effects for the youngest cohorts partly reverse the decline in real wages for low- and medium education. Furthermore, entry-wages become more dispersed among medium- and high-education workers, and less dispersed for low-education workers. In Germany, the negative cohort effects and the increase in within-group wage dispersion for younger cohorts since the 1990s add to the rising wage dispersion, especially within education groups. The least-paid low-educated workers show the largest real wage loss at labor market entry. While cohort effects mitigate within-group inequality of entry wages in the U.S. cohort effects are more sizeable the low-educated in Germany and strongly increase inequality of entry wages, especially in the lower tail of the wage distribution.

5.5. Uniformly Rising Wage Dispersion or Wage Polarization?

5.5.1. Development of Skill Premia due to Macroeconomic Shifts

How much of the increase in wage dispersion in the U.S. and Germany is due to rising skill premia across educational groups? Some studies have suggested that this part is substantial. For example, Lemieux (2006b) found that almost half of the increase in wage inequality in the U.S. can be explained by changes in skill premia. For Germany, our descriptive results in Section 3 show that the rise of the low-to-medium education premium and the increase in dispersion in the lower part of the German wage distribution in the 1990s take place during the same period. Most of the rise of the education premium between high- and medium-education workers occurs also after 1990.

Figure 9 depicts the estimated time-trends of median wages across education groups. Cumulated wage growth over time in the U.S. at the median is much better the higher the education level, i.e., the education premium is increasing strongly which is consistent with the SBTC. Note, however, that the medium-to-low wage premium started to decrease slightly since the mid-1990s, which reflects a weak (albeit not significant) tendency towards wage polarization (see Autor and Dorn 2013; Autor et al. 2008) for further evidence of wage polarization during the 1990s in the U.S.). For Germany, until the mid-1990s, median-wages across education groups move in a parallel fashion. Since then, wages of the high-education exhibit higher growth rates than those of low- and medium-education workers, while the education premium across medium- and low-education German workers does not change over time. The latter observation is somewhat surprising, as the unconditional dispersion between those two groups at the median is clearly increasing since the end of the 1980s (see Figure 5). What can explain these differences between the unconditional development of the education premium and the time-trends? Below, we provide evidence that negative cohort effects for young low-education workers have contributed to the increasing education premium observed unconditionally[20]— which could have been caused by the inflow of young low-education workers into West Germany after the fall of the iron curtain (note that D'Amuri et al. (2010) and Glitz and Wissmann (2017) provided evidence against this). Moreover, and at least as important, we find that the decline in average age of low-education workers and changes in the age-structure of the group of the medium-education (Figure 1) contributed to the rising education premium in Germany, as Figure 10 reveals. Mechanically, this happens because the median wage of the medium-education (low-education) workers increases (decreases) as medium-education (low-education) workers become older (younger). Finally, unions may have successfully counteracted an increasing education premium between medium- and low-education workers, which otherwise would have prevailed due to technological change. The same mechanical compositional effects account for roughly 40% of the sharp increase of 17 log points in the education premium between medium- and high-education workers

[20] Section 5.5.3 summarizes compositional effects on wage growth and wage dispersion both across and within education groups.

in Germany during the early 1990s and 2004, which is observed unconditionally. During the early 1980s, time-trends seem to play no substantial role in explaining the somewhat increasing education premium between medium- and high-education German workers observed unconditionally.

Figure 9. Medians of Educational Groups 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

For the U.S., the time-trends describe the same qualitative but attenuated patterns for the skill premia as the ones observed unconditionally. During the 1980s, when the education premium between medium- and low-education U.S. workers increased, negative cohort effects for the low-education were at work. The declining age of low-education workers also contributed to the rising wage premium, while the age-structure of medium-education workers was quite stable during the 1980s. Regarding the wage premium between high-education and medium-education in the U.S., we see that the aging of the high-education contributed to an increasing premium during the 1980s. Altogether, we find somewhat similar patterns regarding the compositional effects on the wage premia for the U.S. and Germany.

Macroeconomic shifts are likely to be smooth functions of SBTC, institutional factors,[21] and supply-side factors. Given that we observe two industrialized countries that arguably have access to the same technologies, our evidence regarding the different developments of education premia is unlikely to be explained by technological change alone. In fact, supply-side and institutional factors seem to play a key role in explaining the rise of unconditional wage differences between education groups in Germany. This suggests to consider the interaction between labor market institutions, supply-side effects, and SBTC.[22] Note that trends in relative labor-supply across education groups as well as the age-pattern within skill groups show very similar trends in both countries. This indicates that institutional factors—and their interaction with SBTC—may be more important than supply-side factors in explaining the differences across countries.

[21] Besides deunionization and the minimum wage, institutional factors can reflect social norms and incentives set by tax-systems.

[22] This point has also been made by Lemieux (2008).

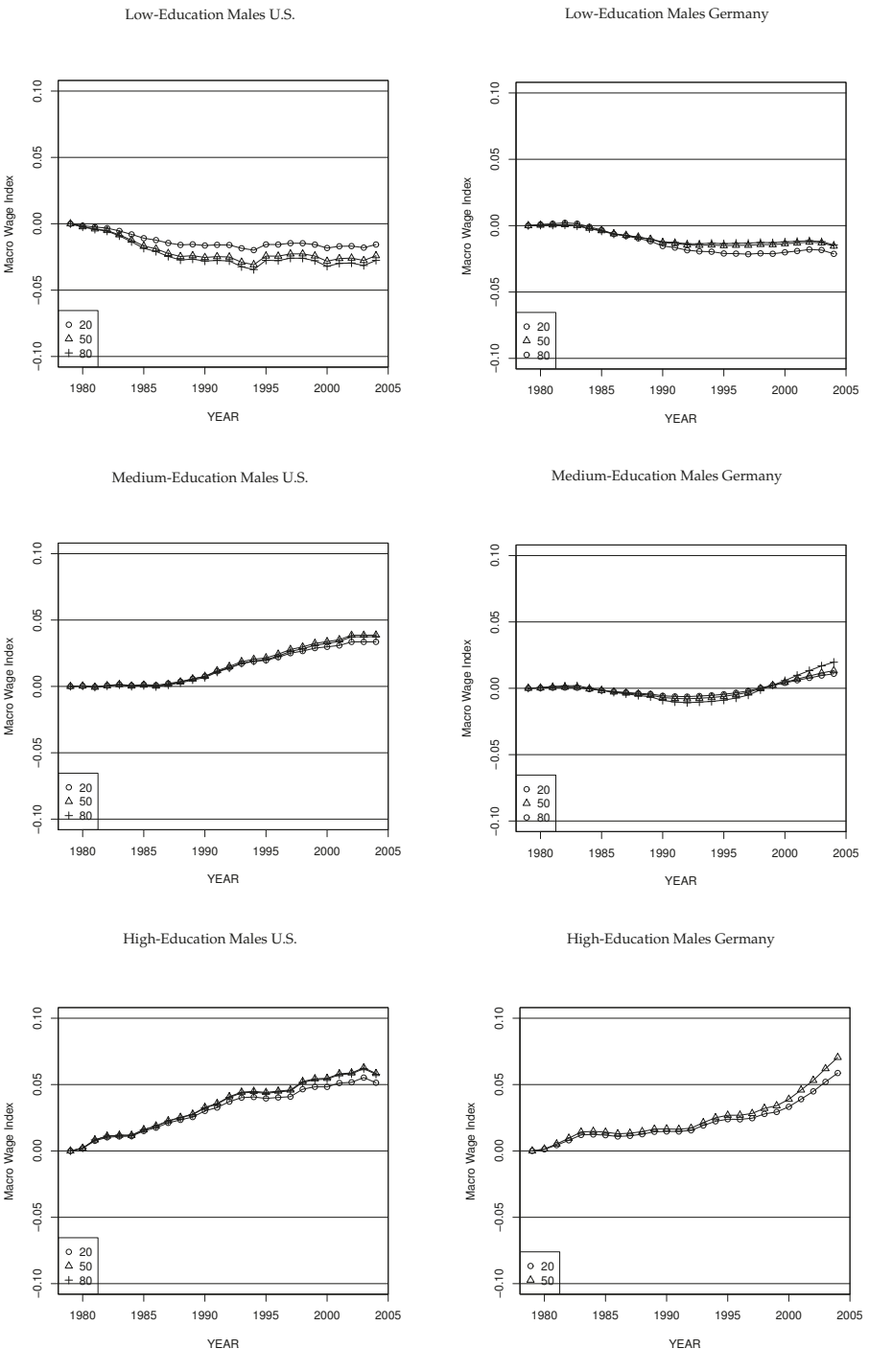


Figure 10. Effect of change in the age structure on wage growth: 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

5.5.2. Wage Dispersion within Skill-Groups

Since the macroeconomic changes in education premia are very small in Germany, most of the increase of wage dispersion in Germany is therefore likely to be due to diverging time-trends within education groups. For low-education workers, Figure 11 depicts the estimated macroeconomic changes in within-group inequality due to macroeconomic shifts within education groups, as measured by the difference of the time trends at the three quantiles. We focus on the strong differences across countries found for low-education workers. The trends for medium- and high-education are fairly similar across countries generally showing a similar increase in within-group wage dispersion over time (see Figure 7 in the main text as well as Figure S2 in the Supplementary Material).

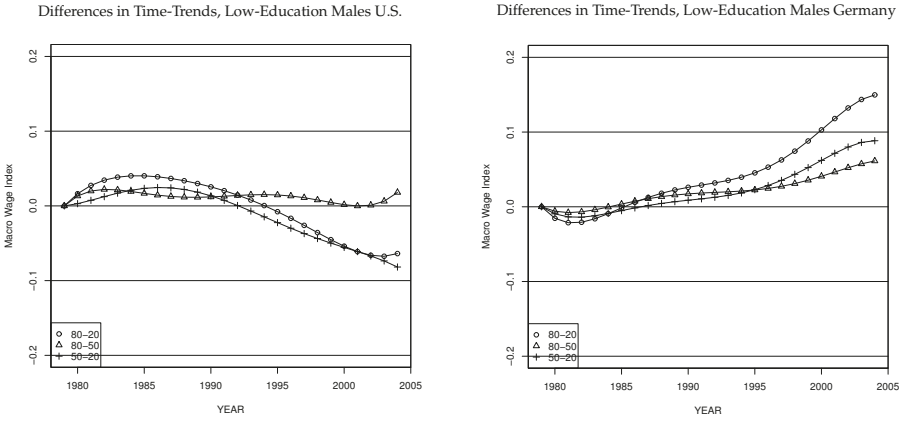


Figure 11. Differences in Time-Trends 1979–2004 for low-education males: **(left)** U.S.; and **(right)** Germany.

In the U.S., low-education workers experienced an astonishing decline in wage dispersion in the lower part of the wage distribution starting in the mid-1980s. After a short period with a rise of the 50%–20% difference by 2 log points, wages at the median dropped more sharply then wages at the 20% quantile until 1996 (and thereafter increased more slowly), resulting in a decreasing dispersion of the lower part of the wage-distribution. Moreover, this decrease is the driving force behind the decline of overall decreasing wage inequality, as measured by the 80%–20% difference, as the inequality in the upper part was quite stable between 1980 and the end of the 1990s (thereafter wage inequality in the upper part decreased by about 2 log points).[23] Increasing wage inequality among U.S. medium-education workers since the early 1990s masks a weak polarization pattern which starts as early as the end of the 1980s, because from then onwards wage growth is slightly higher at the 20% quantile compared to the median. Inequality increases above the median during the 1980s and afterwards basically stays constant.

Our results regarding wage inequality of U.S. low-education workers and the lower part of U.S. medium-education workers for the 1980s may reflect "episodic events", such as the declining real minimum wage and deunionization, and are thus in line with Card and DiNardo (2002).[24] The polarization of wages, beginning at the end of the 1980s, has also been documented by Autor and Dorn (2013), who argued that the low-skill service sector is the driving force.

---

[23] During the first half of the 1990s we find some support for within-education group polarization of wages, as the 80%–50% difference slightly increases while the 50%–20% difference sharply drops during that period.

[24] Chernozhukov et al. (2013), building upon DiNardo et al. (1996), showed that minimum wage seems to play a larger role for the increase of the 50%–10% difference than deunionization. Autor et al. (2008), in the same line, concur that the decline of the minimum wage contributed to the rising lower tail wage-inequality.

The highest increase of overall wage dispersion, as well as dispersion in the lower part of the distribution, is observed for the group of high-education workers in the U.S. for whom neither unions nor minimum wages are likely to play an important role. It is rather likely that technological change had heterogeneous effects among the group of college-graduates (see Lemieux 2006b). Moreover, changes in social norms might have played a certain role especially for this group (see Piketty and Saez 2003).

After a short period of decreasing overall wage inequality in Germany between 1979 and 1982, low-education workers experience a large increase in wage dispersion, where the rise in the 50%–20% difference dominates after the mid-1990s. Unemployment rates in Germany are high among those workers, hence there might also be selection processes driving these developments.

Until the mid-1990s, the 50%–20% difference of medium-education workers in Germany remained almost unchanged, compared to 1979. The rise in overall wage inequality until then was purely driven by an increasing dispersion in the upper part of the wage distribution. Since the mid-1990s wage dispersion is increasing monotonically both in the lower- and upper part of the distribution (this is similar to findings in Dustmann et al. (2009)).

The 50%–20% difference of high-education workers is quite flat until the early 1990s, when it starts to increase monotonically until the end of our observed period. The late increase in wage dispersion among German high-education workers is interesting considering the fact that unconditional wage dispersion in Germany at the top already started to increase during the 1980s. Apparently this was not caused by an increasing within-wage dispersion among high-education workers below the median.

What explains these differences in the development of polarization between the U.S. and Germany? For the U.S., we see patterns of polarization due to macroeconomic shifts both within and across education groups.[25] For Germany, we find little evidence after the early 1980s for polarization of unconditional wages and of wage inequality within education groups.

Similar to Fitzenberger and Wunderlich (2002) and Dustmann et al. (2009), the development of the German wage structure is consistent with the SBTC story, if one allows for institutional factors (including the effects of social norms Piketty and Saez (2003)), such as unions and implicit minimum wages implied by the welfare state, which, in comparison to the U.S., delayed the widening of the German wage dispersion in the lower part for about ten years. A further explanation might be that social norms in Germany have been different, an explanation which is put forward by Piketty and Saez (2003) for other continental European countries as well. Similar to the argument made by Chernozhukov et al. (2013) that the decline of unions and of the minimum wage in the 1980s in the U.S. counteracted the polarization of wages during that period, the increasing flexibility in Germany due to a higher decentralization of wage setting during the 1990s and the early 2000s (see Dustmann et al. 2014) may have counteracted a polarization of wages.

5.5.3. Compositional Effects on Wage growth and Inequality

Figure 6 depicts the life-cycle profiles of wage growth conditional on education, showing that inequality varies by age. To illustrate this, Figure 10 plots the effect of the changing age structure on wage growth and (implicitly) on wage dispersion. This is done by using the estimates of the life-cycle profile of wages and the changing distribution of ages to calculate the implied change in wages. The increase of the mean ages both of medium- and high-education workers in the U.S. reflect the changes of the age structure which result in increasing wages in these two subgroups. However, wage inequality within education groups only slightly increases due to the changing age structure. The trend for low-education workers in the U.S. is reversed: The mean age decreases between 1979 and 2004, and changes in the age-structure lead to decreasing wages as well as less wage-inequality over time, being mainly driven by declining wage dispersion in the lower part of the wage distribution. Comparing the development of the wages at the medians across education groups, it is clear that,

[25] Autor et al. (2008) also documented this pattern of polarization both within and between education groups.

first, throughout the entire period the changing age structure among low- and medium-education U.S. workers led to an increasing education premium between medium and low, second, that during the 1980s, the aging of high-education workers led to an increasing education premium between high and medium.

For Germany, the results differ for the low-education. Although the age-pattern is qualitatively the same between 1979 and 2004 compared to the U.S., the rejuvenation of this education group, indicated by a decrease of the mean age, leads to an increasing within wage dispersion over time, as the 20% quantile in this group experiences the largest life-cycle wage growth. The changing age-structure of medium- and high-education workers in Germany, indicated by the rise of the mean age starting in the late 1990s, mechanically leads to increasing wages for both groups. The age-decomposition effect only plays a minor role in explaining changes of wage dispersion conditional on education though. The aging of German medium-education workers since the early 1990s led to an increasing education premium between low- and medium-education workers, which, as we have shown above, is not due to macro-economic shifts. Similarly, differences in the pattern of aging between medium- and high-education workers led to an increasing education premium between those two groups.

Figures 12 and 13 depict the impact of the inflow and outflow of the cohorts on skill-specific wage growth and dispersion, respectively. The latter graphs show that starting in the early 1990s, the change in the cohort structure supports the catching-up process of both wages at the median and the 20% quantile to wages at the 80% quantile in the group of low-education workers in the U.S. The 80–50 and 80–20 difference of wages had increased before, though, due to cohort effects. Contrary to that, cohort effects in Germany for the group of low-education led to an increasing wage dispersion of about 5 log points throughout the entire wage-distribution between 1992 and 2004, while before the early 1990s, cohort effects led to a decreasing wage dispersion, with the movements of the 80-20 difference mainly being driven by changes of the wage dispersion in the lower part. Cohort effects for medium- and high-education workers affect wage dispersion somewhat less in both countries. Relatively to the oldest and the youngest cohorts, those in the middle seem to exhibit higher cohort specific wage dispersion, driven mostly by positive cohort effects at the median and the 80% quantile. In the middle of the observation period, the presence of these cohorts in the middle is strongest, resulting in the strongest increase in wage dispersion within skill groups. Based on Figure 12, the sharp drop of cohort effects among low-education German workers mechanically increases the wage premium between low- and medium-education workers in Germany. Compositional effects regarding the cohort structure also seem to increase the education premium between high- and medium-education workers in Germany since the early 1990s. For the U.S., such compositional effects play only a minor role.

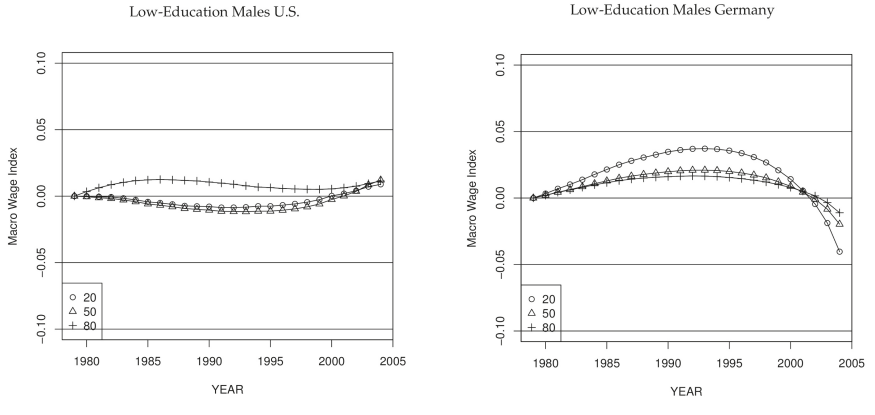


Figure 12. *Cont.*

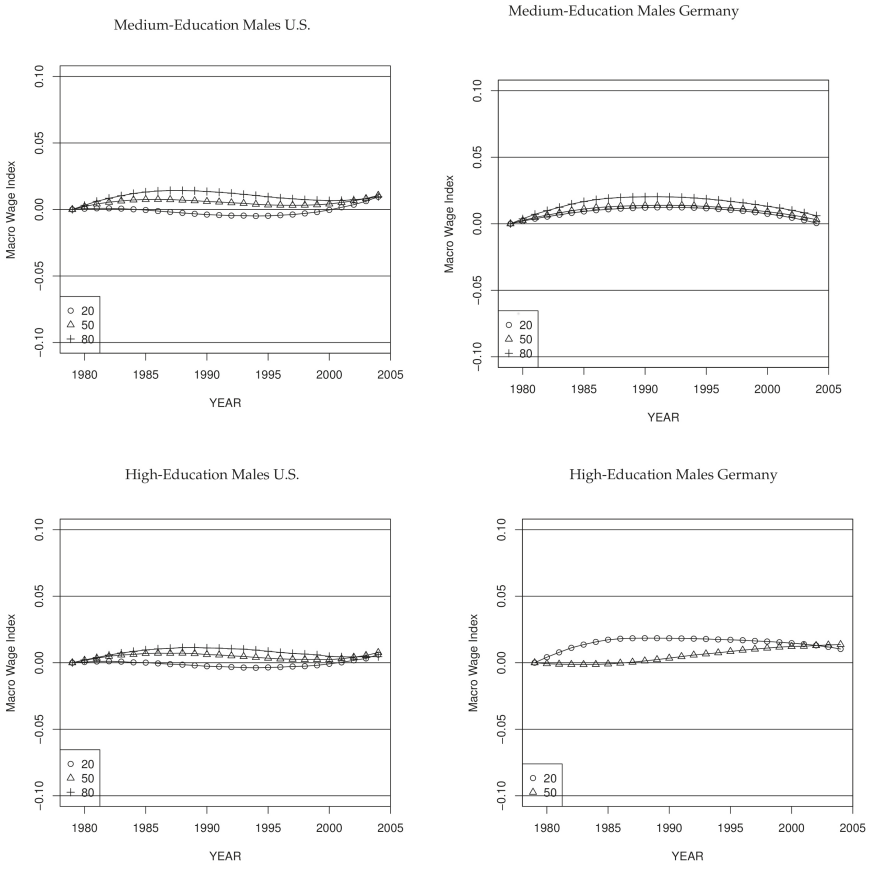


Figure 12. Effect of change in the cohort structure on wage growth: 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

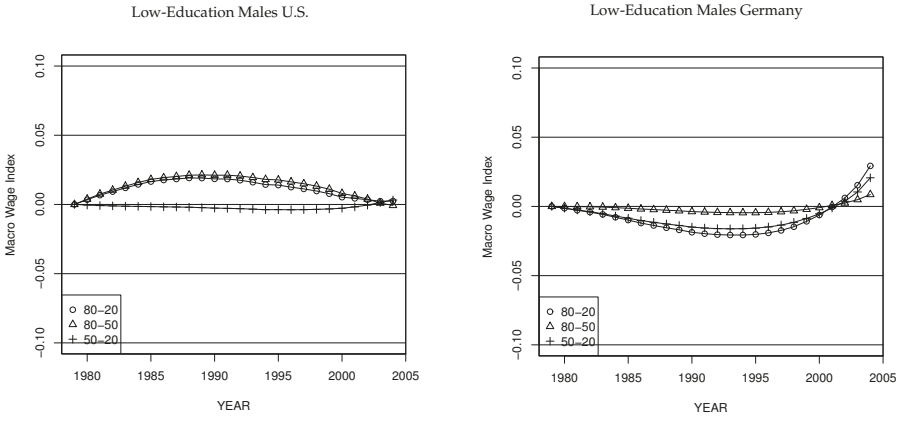


Figure 13. *Cont.*

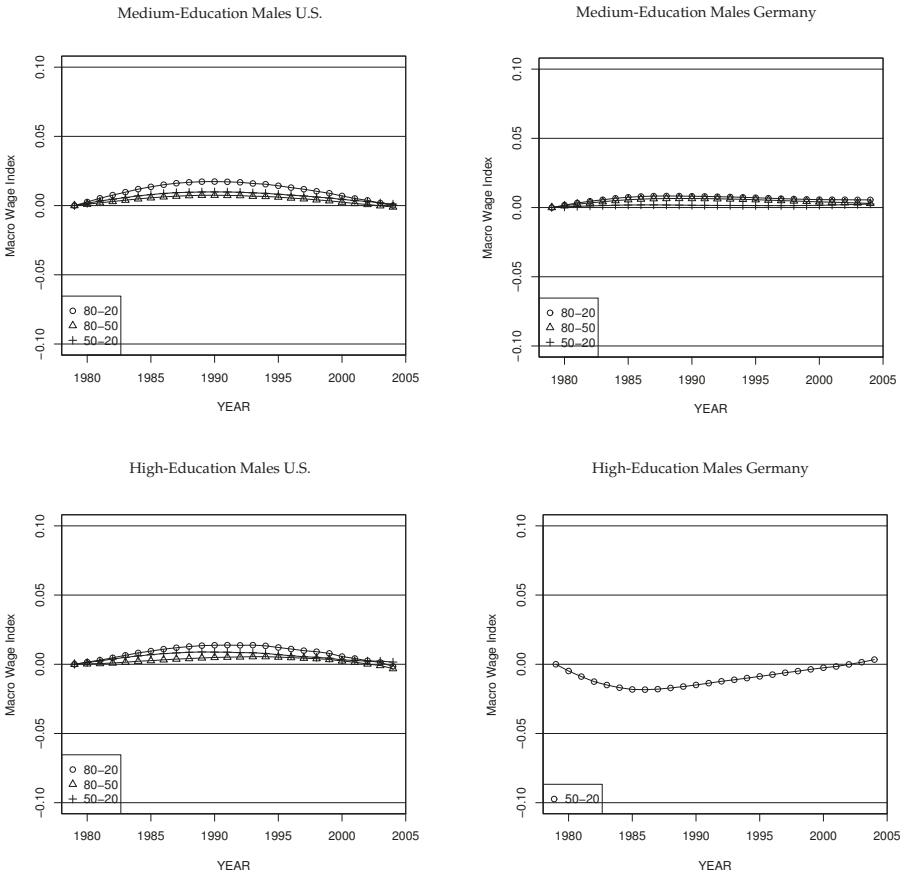


Figure 13. Effect of change in the cohort structure on wage dispersion: 1979–2004 for males: (**left**) U.S.; and (**right**) Germany.

5.6. Employment Growth

A large literature finds polarization of employment in both Germany and the U.S. since the 1980s based on employment trends for occupations (see, e.g., Autor et al. 2003; Spitz-Oener 2006). We complement these findings based on data for age-education cells. Using a method similar to Card et al. (1999), we rank the age-education cells across education groups for a base year according to the cells unconditional median wages, which we normalized by the estimated age-specific life-cycle wage growth of the specific cells, i.e., we do not use the unconditional wage level as in the polarization literature (e.g., Autor and Dorn 2013).[26] Then, we calculate the cumulated relative employment growth of each cell over the next ten years.[27] Our age variable is discrete, ranging between 25 and 55, and we distinguish between three educational levels, which yields 93 cells for this analysis,

[26] Cells whose median wages are top-coded, which happens frequently for the group of high-education German workers, are given the highest ranks, whereby the general pattern of the graphs is not affected by the chosen order. We thus draw random numbers to determine the order of the ranks at the top-end.

[27] Note that in the latter period different worker cohorts are in these cells.

which Card et al. (1999) interpreted as "education groups". The base years we choose are 1979, 1984, 1989, and 1999. The results are depicted in Figures 14 and 15.

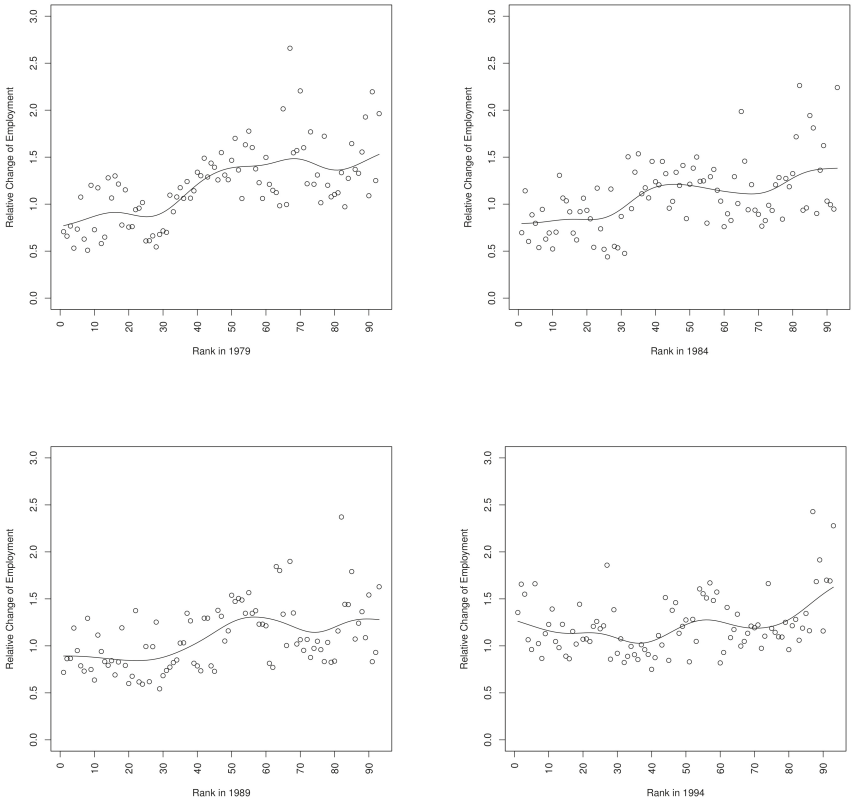


Figure 14. Employment Changes by age-education group, U.S.

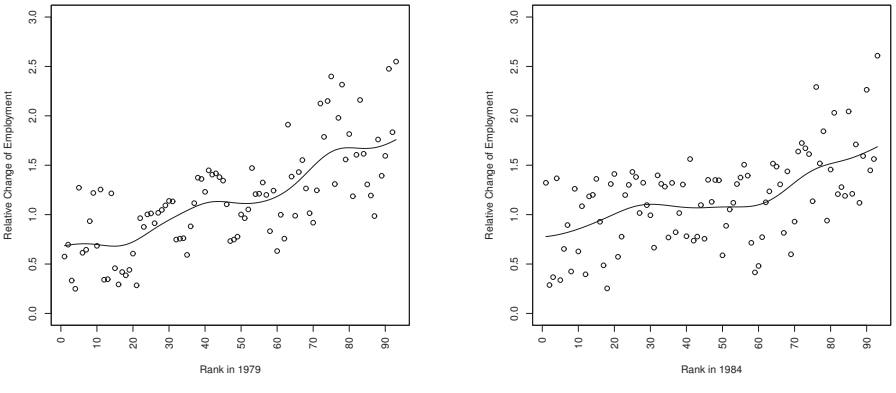


Figure 15. *Cont.*

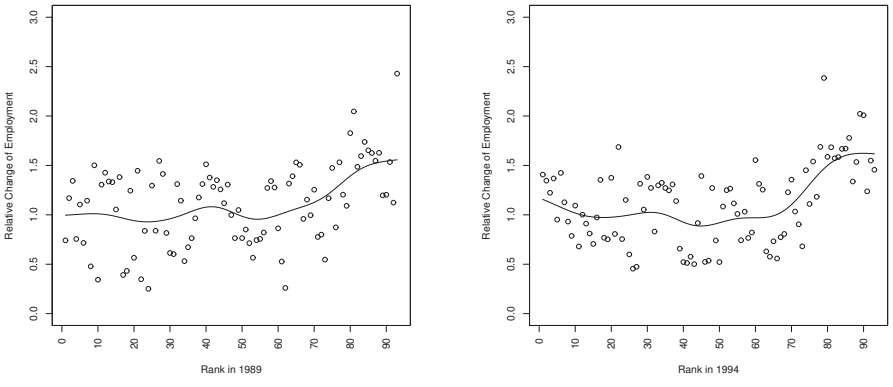


Figure 15. Employment Changes by age-education group, Germany.

For the base years 1979 and 1984, relative changes in employment in both economies is a monotonically increasing function of the rank of wages in the base year. We find evidence of employment polarization in Germany starting with the base year 1989, which becomes more pronounced for the base year 1994. This means that for the latter two base years, age-education cells which are ranked at the bottom exhibit higher growth rates than those at the middle, while the highest ranked age-education exhibit the largest growth rates. For the U.S., we observe a similar pattern of polarization starting in the second half of the 1990s. There are striking similarities in the four graphs between the U.S. and Germany.

This simple analysis helps us to separate demand-side vs. supply-side stories. In the U.S., we observe polarization in wages and employment, as a nuanced version of the SBTC-story would suggest, while in Germany we observe polarization in employment but little evidence for polarization in wages.

6. Conclusions

This paper revisits the rise in wage inequality in both the U.S. and Germany. A technology-based explanation for a widening wage gap between high-education workers and low-education workers should apply to both economies, while episodic changes and institutional differences may imply cross-country differences. The methods we employ enable us to separately identify life-cycle wage profiles, time-trends in wages (due to macroeconomic shifts), and cohort wage effects. Our analysis applies quantile regressions of wage focussing on three representative quantiles (20%, median, and 80%).

We find that there is increasing wage inequality over the life cycle in both countries and for all education groups, with one exception. For low-education workers in Germany, there is decreasing wage inequality over the life cycle. The changing age structure of the workforce has important implications for trends in wage inequality in both the U.S. and Germany. There exist important cohort effects for Germany. Both the old and the young cohorts of workers have sizeable negative cohort effects. These effects could be the result of supply-side factors such as immigration, cohort size, or selection into education group. However, D'Amuri et al. (2010) and Glitz and Wissmann (2017) argued that immigration waves in Germany had only a small impact on wages among the natives. In the U.S., by contrast, the size of the cohort effects is considerably smaller.

The time trends in wages favor high-education workers in both the U.S. and in Germany, but rising skill premia are much more important in the U.S. In the U.S., there were secular declines in wages until

the mid-1990s for low- and medium-education workers when these trends reversed. In Germany, we see the opposite pattern—rising secular trends in wages until the mid-1990s and a flattening (at the median) or a decline (at the 20% quantile) in wages afterwards. After the mid-1990s, wage inequality increases among the low-education workers in Germany and declines among the low-education workers in the U.S. In Germany, the rising premium between medium- and low-education workers is entirely due to cohort and aging effects. In the U.S., there is faster wage growth both at the top and the bottom of the distribution. We see basically no evidence of wage polarization in Germany after the early 1980s.

Summing up, on the one hand, there are some similarities in trends in wage inequality—and in particular in employment—between the U.S. and Germany which is consistent with a technology-based explanation of labor market trends since the late 1970s. On the other hand, various patterns in wage inequality differ strongly between the two countries, which makes it unlikely that technology effects alone can explain the empirical findings. Episodic changes resulting from changes in institutional factors such as deunionization, decentralization of wage setting to the firm level, or the minimum wage may explain the differences, which are partly reflected in the cross-country differences in cohort effects. SBTC may interact in important ways with institutional differences between the U.S. and Germany. The decentralization of wage setting in Germany may have lowered in particular wages of less skilled workers in the youngest cohorts, whose entry wages are less protected by the institutions in Germany.

Supplementary Materials: The following are available online at http://www.mdpi.com/2225-1146/6/2/20/s1, Figure S1: Cumulated growth of Entry Wages, Figure S2: Differences in Time-Trends 1979–2004 for males, Table S1: Model specification and results for full-time working men, U.S., Table S2: Model specification and results for full-time working men, Germany.

Acknowledgments: We thank David Autor, Thomas E. MaCurdy, Salvador Navarro, Timothy Smeeding, and Christopher Taber for useful discussions. We benefitted from valuable comments received at workshops in Berlin and Freiburg. Parts of this paper were written while Dirk Antonczyk was visiting the Institute for Research on Poverty at the University of Wisconsin—Madison. He would like to thank the center for its hospitality. Financial support by the German Research Foundation (DFG) (project "Collective Bargaining and the Distribution of Wages: Theory and Empirical Evidence" in SPP 1169 and project "Accounting for Selection Effects in the Analysis of Wage Inequality in Germany" in SPP 1764), the German Academic Exchange Service (DAAD), and the "Wissenschaftliche Gesellschaft Freiburg" is gratefully acknowledged. The responsibility for all errors is, of course, ours.

Author Contributions: All authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

Acemoglu, Daron, and David H. Autor. 2011. Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics* 4: 1043–171.

Antonczyk, Dirk, Bernd Fitzenberger, Enno Mammen, and Kyusang Yu. 2017. A Nonparametric Approach to Identify Age, Time, and Cohort Effects. Unpublished Manuscript, University of Heidelberg, Heidelberg.

Antonczyk, Dirk, Bernd Fitzenberger, and Katrin Sommerfeld. 2010. Rising wage inequality, the decline of collective bargaining, and the gender wage gap. *Labour Economics* 17: 835–47.

Asplund, Rita, Erling Barth, Per Lundborg, and Kjersti Misje Nilsen. 2011. Polarization of the nordic labour markets. *Finnish Economic Papers* 24: 87–110

Autor, David H. 2013. The "task approach" to labor markets: An overview. *Journal for Labour Market Research* 46: 185–99.

Autor, David H. 2014. Skills, education, and the rise of earnings inequality among the "other 99 percent". *Science* 344: 843–51.

Autor, David H., and David Dorn. 2013. The growth of low-skill service jobs and the polarization of the US labor market. *The American Economic Review* 103: 1553–97.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. The China syndrome: Local labor market effects of import competition in the United States. *The American Economic Review* 103: 2121–68.

Autor, David H., and Michael J. Handel. 2013. Putting tasks to the test: Human capital, job tasks, and wages. *Journal of Labor Economics* 31, S59–S96.

Autor, David H., Larry F. Katz, and Melissa S. Kearney. 2008. Trends in U.S. Wage Inequality: Re-Assessing the Revisionists. *Review of Economics and Statistics* 90: 300–23.

Autor, David H., Frank Levy, and Richard J. Murnane. 2003. The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics* 118: 1279–333.

Beaudry, Paul, and David A. Green. 2000. Cohort patterns in canadian earnings: Assessing the role of skill premia in inequality trends. *Canadian Journal of Economics/Revue Canadienne D'économique* 33: 907–36.

Berger, Mark C. 1985. The effect of cohort size on earnings growth: A reexamination of the evidence. *Journal of Political Economy* 93: 561–73.

Biewen, Martin, Bernd Fitzenberger, and Jakob de Lazzer. 2017. Rising Wage Inequality in Germany: Increasing Heterogeneity and Changing Selection into Full-Time Work. IZA Discussion Paper, No. 11072, IZA, Bonn, Germany.

Biewen, Martin, and Matthias Seckler. 2017. Changes in the German Wage Structure: Unions, Internationalization, Tasks, Firms, and Worker Characteristics. IZA Discussion Paper, No. 10763, IZA, Bonn, Germany.

Bundesamt für Migration und Flüchtlinge. 2005. *Der Einfluss von Zuwanderung auf die deutsche Gesellschaft*. Nürnberg: Bundesamt für Migration und Flüchtlinge.

Burkhauser, Richard V., and Jeff Larrimore. 2009. Using internal cps data to reevaluate trends in labor-earning gaps. *Monthly Labor Review* 132: 3.

Card, David, and John E. DiNardo. 2002. Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles. *Journal of Labor Economics* 20: 733–82.

Card, David, Jörg Heining, and Patrick Kline. 2013. Workplace heterogeneity and the rise of west german wage inequality. *The Quarterly Journal of Economics* 128: 967–1015.

Card, David, Francis Kramarz, and Thomas Lemieux. 1999. Changes in the Relative Structure of Employement: A Comparison of the United States, Canada and France. *The Canadian Journal of Economics* 32: 843–77.

Card, David, and Thomas Lemieux. 2001. Can Falling Supply Explain the Rising Return to College Education for Younger Men? A Cohort Based Analysis. *The Quarterly Journal of Economics* 116: 705–46.

Carneiro, Pedro, and Sokbae Lee. 2011. Trends in quality-adjusted skill premia in the united states, 1960–2000. *The American Economic Review* 101: 2309–49.

Chamberlain, Gary. 1994. Quantile Regression, Censoring and the Structure of Wages. *Advances in Econometrics* 1: 171–209.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. Inference on counterfactual distributions. *Econometrica* 81: 2205–68.

D'Amuri, Francesco, Gianmarco I. P. Ottaviano, and Giovanni Peri. 2010. The labor market impact of immigration in western germany in the 1990s. *European Economic Review* 54: 550–70.

DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. Labor Markets Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach. *Econometrica* 64: 1001–44.

Drews, Nils. 2008. Das Regionalfile der IAB-Beschäftigtenstichprobe 1975–2004—Handbuch-Version 1.0.2. FDZ-Datenreport, 02/2008 (de), FDZ, Nuremberg, Germany.

Dustmann, Christian, Bernd Fitzenberger, Uta Schönberg, and Alexandra Spitz-Oener. 2014. From sick man of europe to economic superstar: Germany's resurgent economy. *The Journal of Economic Perspectives* 28: 167–88.

Dustmann, Christian, Johannes Ludsteck, and Uta Schönberg. 2009. Revisiting the German Wage Structure. *The Quarterly Journal of Economics* 124: 843–81.

Felbermayr, Gabriel, Daniel Baumgarten, and Sybille Lehwald. 2012. Increasing Wage Inequality in Germany: What Role Does Global Trade Play? Technical report, Global Economic Dynamics, Bertelsmann Stiftung, Gütersloh, Germany.

Fitzenberger, B. 1999. *Wages and Employment Across Skill Groups: An Analysis for West Germany*. Heidelberg: Springer/Physica.

Fitzenberger, Bernd, and Gaby Wunderlich. 2002. Gender Wage Differences in West Germany: A Cohort Analysis. *German Economic Review* 3: 379–414.

Fuchs-Schündeln, Nicola, and Matthias Schündeln. 2009. Who stays, who goes, who returns? East-West migration within Germany since reunification. *Economics of Transition* 17: 703–38.

Glitz, Albrecht, and Daniel Wissmann. 2017. Skill Premiums and the Supply of Young Workers in Germany. IZA Discussion Paper, No. 10901, IZA, Bonn, Germany.

Goos, Maarten, Alan Manning, and Anna Salomons. 2014. Explaining job polarization: Routine-biased technological change and offshoring. *American Economic Review* 104: 2509–26.

Gosling, Amanda, Stephen Machin, and Costas Meghir. 2000. The Changing Distribution of Male Wages in the U.K. *Review of Economic Studies* 67: 635–66.

Green, David A., and Benjamin M. Sand. 2015. Has the canadian labour market polarized? *Canadian Journal of Economics/Revue Canadienne D'économique* 48: 612–46.

Katz, Larry F., and David H. Autor. 1999. Changes in the Wage Structure and Earnings Inequality. *Handbook of Labor Economics* 3A: 1463–55.

Lemieux, Thomas. 2006a. Increased Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill. *American Economic Review* 96: 461–98.

Lemieux, Thomas. 2006b. Postsecondary Education and Increasing Wage Inequality. *American Economic Review Papers and Proceedings* 96: 195–99.

Lemieux, Thomas. 2008. The Changing Nature of Wage Inequality. *Journal of Population Economics* 21: 21–48.

MaCurdy, Thomas, and Thomas Mroz. 1995. Measuring Macroeconomic Shifts in Wages from Cohort Specifications. Unpublished Manuscript, Stanford University and University of North Carolina.

OECD. 2016. *Income Inequality Update: Income Inequality Remains High in the Face of Weak Recovery*. Paris: OECD.

Piketty, Thomas, and Emmanuel Saez. 2003. Income Inequality in the United States, 1913–1998. *The Quarterly Journal of Economics* 118: 1–39.

Piketty, Thomas, and Emmanuel Saez. 2014. Inequality in the long run. *Science* 344: 838–43.

Spitz-Oener, Alexandra. 2006. Technical Change, Job Tasks, and Rising Educational Demands: Looking Outside the Wage Structure. *Journal of Labor Economics* 24: 235–70.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

The Wall's Impact in the Occupied West Bank: A Bayesian Approach to Poverty Dynamics Using Repeated Cross-Sections

Tareq Sadeq [1,2] and Michel Lubrano [2]*

[1] Department of Economics, Birzeit University, Birzeit, West Bank 627, Palestine; tsadeq@birzeit.edu
[2] Aix-Marseille Univ., CNRS, EHESS, Centrale Marseille, AMSE, Ilot Bernard Dubois, 5 Bd Maurice Bourdet, 13001 Marseille, France
* Correspondence: michel.lubrano@univ-amu.fr; Tel.: +33-413-55-25-57

Received: 3 December 2017; Accepted: 22 May 2018; Published: 30 May 2018

Abstract: In 2002, the Israeli government decided to build a wall inside the occupied West Bank. The wall had a marked effect on the access to land and water resources as well as to the Israeli labour market. It is difficult to include the effect of the wall in an econometric model explaining poverty dynamics as the wall was built in the richer region of the West Bank. So a diff-in-diff strategy is needed. Using a Bayesian approach, we treat our two-period repeated cross-section data set as an incomplete data problem, explaining the income-to-needs ratio as a function of time invariant exogenous variables. This allows us to provide inference results on poverty dynamics. We then build a conditional regression model including a wall variable and state dependence to see how the wall modified the initial results on poverty dynamics. We find that the wall has increased the probability of poverty persistence by 58 percentage points and the probability of poverty entry by 18 percentage points.

Keywords: Bayesian inference; pseudo panels; data augmentation; walls; poverty dynamics

JEL Classification: C11; C33; I32; O15

1. Introduction

Speaking about the Israeli-Palestinian conflict, Atran et al. (2007) concluded their article devoted to exploring sacred values and conflict resolution by the following words: "*We urgently need more scientific research to inform better policy choices*". The aim of the present paper is to shed some light on the economic consequences in terms of poverty dynamics on the Palestinian population of what is called by the Israeli government *a separation wall*, a *security fence*. The building of this wall was decided in 2002, after many political discussions, in order to prevent terrorist attacks starting from the West Bank. It was presented by the Israeli Government as being temporary, meant to be destroyed after peace negotiations. However, as underlined in Leuenberger (2016), "*other factors were equally, if not more, influential, such as: demography, location of water aquifers, as well as the inclusion of (what under international law are considered illegal) Jewish settlements within the Occupied Palestinian Territories, inside Israeli-controlled territory. Assessments of the Barrier's function can thus quickly become mired in controversy.*" As a matter of fact, this *security fence* is called *a wall of apartheid* by the Palestinians while the simple term *wall* is used by the International Court of Justice.[1] Where does the controversy comes from?

[1] International Court of Justice (2004) Legal Consequences of the Construction of a Wall in the Occupied Palestinian Territory: Advisory Opinion (9 July), quoted by Leuenberger (2016).

In fact, the wall does not follow the *Green Line*, the international separation between the State of Israel and the occupied West Bank. With a total length of 708 kilometers, the Wall is more than double the length of the Green Line and at times runs 18 kilometers deep inside the West Bank. As described in Cohen (2006), "*many of the portions of the wall that comprise imposing concrete slabs are located in the heart of Palestinian communities, splitting towns, villages, streets, and even extended families (Usher 2006). In other places the fence portions separate Palestinian farmers from their fields, jobs, or schools, cresting visible and acute disruption of normal life.*" A detailed map produced by the *Applied Research Institute Jerusalem* is reproduced in Leuenberger (2016, Figure 3), showing that about 85% of the route followed by the wall runs inside the West Bank, well away from the internationally recognised Green Line.

This wall was built not only to separate Palestinians from Israelis, but also to separate them from Israeli settlements inside the West Bank. As a result, around 943 square kilometres of land located between the wall and the Green Line 1967 border (16.8% of the total West Bank area) has been declared by the Israeli army as a military zone known as the *Seam Zone*. This area has become inaccessible to Palestinians living on the eastern side of the wall and having no special or seasonal permits. The permits allow accessing the *Seam Zone* only through a few gates in the wall and during short time periods, usually 15 min in the morning and 15 min in the evening (Hareuveni 2012).

In the post-Oslo peace agreements period (1993–1995) and prior to the wall construction (2002), the occupied West Bank and the Gaza Strip have been subject to a closure policy imposed by the Israeli army. Total closure made movements between the West Bank and the Gaza Strip almost impossible; then internal closure in the West Bank prevented Palestinians from accessing Jerusalem. The checkpoints system installed between West Bank localities made movements and access to land more and more difficult.[2]

Many academic studies were led to measure the consequences of the wall. However, those were mainly concerned about law and politics. Hassan (2005) discusses the implications of the advice of the International Court of Justice on the relationships between Israel and the US Government. Kattan (2007) discusses the examination by the Supreme Court of Israel of the Advisory Opinion rendered by the International Court of Justice concerning the legality of the Wall. Malone (2004) examines how the route followed by the Wall affects water access for Palestinian villages of the West Bank when Reynolds (2015) details environmental damages and reduction of bio-diversity.

We focus here on the economic consequences of the Wall on the Palestinian society. As underlined in Roy (2000), "*Given the extreme dependence of the Palestinian economy on Israel, the impact of closure-restricting the jobs and income of Palestinians working in Israel, reducing Palestinian trade levels, lowering production levels, and so on, has been to heighten poverty*". Palestinians in West Bank found themselves separated from lands behind the wall and from the economic resources they represent. Land behind the wall is known to be more fertile for agriculture and to contain rich natural resources. This zone used to offer scope for future economic development of the occupied Palestinian territory, as well as urban expansion (World Bank 2008).

In addition to the segregation of Palestinians from agricultural land, the wall deprives Palestinians from employment opportunities in Israel. Employment in Israel, that concerns mostly unskilled workers and which is paid higher than in the local Palestinian market, represented an unstable but important income source for low-income and low-asset households. According to the Palestinian Central Bureau of Statistics labour force surveys (PCBS 2000, 2005), around 26% of total Palestinian employment in the West Bank was located in Israel in 1999, but this share declined to around 12% in 2003 and 2004 due to the decline in the number of permits issued by Israel and also due to the wall and checkpoints in the West Bank. Adnan (2015) finds, using the Palestinian Labour Force Survey, that closures and living on the West Bank side of the wall deter out-migration to Israel and increase the

[2] A record of the consequences of the wall is made by many institutions and in particular by the UN Office for the Coordination of Humanitarian Affairs (UN OCHA) which also produces neutral maps as underlined in Leuenberger (2016). https://www.ochaopt.org/theme/west-bank-barrier.

probability of being unemployed. In an anthropological paper, Bornstein (2001) argues that the wall does not make it impossible for Palestinian workers to illegally enter Israel, but it makes it impractical on a daily basis. The segregation policy forces Palestinian workers without permits to stay hidden for weeks on construction sites and in factories, due to the risk of being arrested. This increased risk either discourages Palestinians from working in Israel or may imply long term life deterioration, including reduction in income and consumption, for those subject to this risk. In both cases, households are pushed into chronic poverty traps.

Evidence for eastern and southern African countries (Jayne et al. 2003) shows that land distribution among smallholders is related to income poverty. In the occupied Palestinian territory, land access restrictions and land confiscations render land prices excessively steep (World Bank 2008). This resulted in higher asset-value inequality and income inequality between Palestinians who own or work in lands behind the wall and Palestinians who own highly demanded lands on the eastern side of the wall.

Negative consumption and income shocks in conflict areas may have long-term effects on school drop-out, displacement, nutrition and health status deterioration, which may imply a chronic poverty status (Carter and Barrett 2006; Ibañez and Moya 2010). Moreover, if shocks persistently result in asset losses or in inaccessibility to their location, income can fall below the critical threshold for several periods and households will be more likely to fall into chronic poverty (Dercon 1998). Households subject to shocks usually refer to credit markets or sell part of their non-productive assets as a strategy of adjustment to shocks, but credit markets exclude low-income and low-asset households. Thus, the initial condition of a low consumption level with insufficient asset-base pushes households into poverty traps (Zimmermann and Carter 2003; Carter and Barrett 2006; Reynolds 2015).

Evidence is found in the literature for a downward spiral of poverty and resource degradation. Poor people over-use existing accessible resources due to high population growth, limited access to resources and inequality in resource allocation. Overuse leads again to resource degradation and increasing poverty (Cleaver and Schreiber 1994; Forsyth and Scoones 1998; Scherr 2000).

A similar case under colonisation and resource access restrictions in recent history is South Africa, where black South Africans had no access to certain resources including land and water. Five years after the fall of the apartheid regime, poverty prevalence was still increasing among black South Africans with a high probability of chronic poverty (Carter and May 2001). Candidates for chronic poverty in South Africa are mostly black, female, rural, people with health problems, elderly and farm workers (Aliber 2003). Moreover, education is found to be an important factor in poverty dynamics determination (Jalan and Ravallion 2000; Fuwa 2007).

The objective of this paper is to measure poverty dynamics and income mobility in the occupied Palestinian territory and its determinants. We focus on the West Bank region, excluding the Gaza Strip, but including the Jerusalem area behind the wall. The focus on Palestinians in the West Bank only is for two reasons. First, the wall exists neither in the Gaza Strip nor in Jerusalem. The wall does not prevent Palestinians living in Jerusalem from working in Israel. Second, poverty dynamics and patterns are different in the Gaza Strip and in Jerusalem from what they are in the West Bank. In the Gaza Strip, poverty is due to closure and wars. In Jerusalem, Palestinians are more likely to work in Israel, but they are constrained by Israeli fiscal policies and they are consuming at Israeli prices.

To quantify the impact of the wall on poverty dynamics, we use the model of Cappellari and Jenkins (2004) which provides a convincing approach to modelling poverty entry and poverty persistence. However, the only data we have are provided by the *Palestinian Expenditure and Consumption Survey* (PECS) collected for the years 1998, 2004 and 2011, which are repeated cross-sections. For the years 2004 and 2011, the PECS contains a geographical variable indicating if a household is located or not in a zone impacted by the wall. We introduce a new Bayesian method of generating pseudo panels, treating the question as an incomplete data problem. Inside the loop of a Gibbs sampler, we explain the income-to-needs ratio using time invariant data for 2004 and 2011 to generate the missing values. Then we use both observed and latent variables to explain the income-to-needs ratio for 2011, this time conditionally on being poor in 2004 and being affected or not

by the wall. We have thus two ways of measuring poverty dynamics and the final effect of the wall on poverty dynamics is determined by a difference between a marginal probability and a conditional probability taking into account the effect of the wall.

The paper is organised as follows. After this introduction, Section 2 describes the *Palestinian Expenditure and Consumption Survey*, and discusses the definition of the poverty line. Section 3 proposes a first measure of the impact of the wall on poverty and shows how a naive strategy would provide wrong results. Section 4 presents a new model for poverty dynamics and shows how this model can be adapted for a repeated cross-section in a Bayesian framework and how the impact of the wall can be measured. Then, Section 5 presents our empirical results. Conclusions and recommendations are presented in the last section.

2. The Palestinian Expenditure and Consumption Survey

We use three waves (1998, 2004 and 2011) of the *Palestinian Expenditure and Consumption Survey* as provided by the Palestinian Central Bureau of Statistics (PCBS). Details of the main variables are given in Appendix A. We focus in this section on income, consumption and the definition of poverty.

2.1. Income, Consumption and Family Composition

In low-income countries, a definition of poverty is usually based on household's consumption level instead of income. The adoption of this definition is mainly due to the consumption of self-produced goods. Poverty status is defined as an indicator variable if the household's consumption is below the poverty line. The two variables, household consumption and poverty line have to be made compatible in terms of household composition. In 1998 and 2004, the average household composition according to 1996 census was 2 adults and 4 children. In 2011, the average household composition according to 2007 census was 2 adults and 3 children. The poverty line is defined by the PCBS for a representative household and thus has not exactly the same content in 1998 and in 2004 on the one hand and 2011 on the other. We report the values of the poverty line in Table 1 for an average household composition.

Table 1. Comparing poverty lines for a representative household.

Year	Official Palestine	50% Mean	60% Median	Official Israel
1998	1,460	1,627	1,610	-
2004	1,934	1,954	1,899	-
2011	2,293	2,509	2,377	5,301

Household consumption adjusted for family composition. Figures correspond to monthly consumption in NIS (New Israeli Shekel). The poverty line in Israel is defined as half the median disposable income, weighted by household size. Source: National Insurance Institute, November 2012.

2.2. Poverty Lines

We have to make the household consumption level compatible with that used in official poverty lines. For doing so, we adopted the Oxford (old OECD) equivalence scale. The elasticity of consumption is usually high using the Oxford equivalence scale, and well adapted to the economic situation of the West Bank where, in 2004, 20% of our sample concerns households living in camps.[3] The Oxford equivalence scale is $N_i = 1 + 0.7N_a + 0.5N_C$ where N_a is the number of adults other than the household head and N_c the number of children under 15 years in household i. To obtain a figure compatible with official poverty line, household consumption is divided by N_i (its equivalised size) and then

[3] For instance, Lanjouw and Ravallion (1995) estimate the elasticity of consumption in Pakistan to be around 0.6 and Dreze and Srinivasan (1997) assume it to be 0.85 in rural India.

multiplied by $(1.7 + 0.5 \times 4)$ for 1998 and 2004 and by $(1.7 + 0.5 \times 3)$ for 2011. It is interesting to compare, in Table 1, these official poverty lines to other definitions of relative poverty lines. The official poverty line is close to relative poverty lines, representing roughly 55–60% of the median consumption. It is much lower than the corresponding Israeli poverty line of 5,301 NIS in 2011. The relative poverty line of 50% of the mean is well below the official line in 2011 due to an increase in inequality. The Gini index is 0.328 for 1998 and 2004 and is 0.339 in 2011.

3. Stylised Facts and Empirical Strategy

The wall's effect is represented by a variable provided by the PCBS. This is a geographical variable documenting the location of a household, using three criteria. The first criterion is Jerusalem and those households are un-impacted. The two other criteria are if a household located in the West Bank is situated or not in a geographical zone that has lost lands because of the building of the wall. This variable is available only for 2004 and 2011. We shall see that the measured effect of this variable can be misleading if not treated properly. An appropriate empirical strategy is required.

3.1. Stylised Facts around the Wall

The wall variable is a limited definition of the effect of the wall as it does not take into account the limitations concerning access to the labour market. Households which are located in Jerusalem have different consumption patterns and the wall does not prohibit them from working in Israel. An indicator variable for Palestinians living in Jerusalem has to be used as a control variable.

3.1.1. Poverty Rates

In Table 2, we report overall poverty rates using the official poverty line in the first column, and in the next columns we decompose the same head count indicator for three sub-populations. The overall poverty rate has increased after the Israeli closure policy in 2000 and during the construction of the wall (between 2002 and 2004). Poverty then decreased in 2011, without returning to its level of 1998. For 2004 and 2011, the indicator variable *wall* allows us to assess the difference between Jerusalem and the rest of the West Bank. Poverty is very low in Jerusalem, much higher in the West Bank, but significantly lower in the region where the wall was built and said to have impacted the population. This simply means that the wall was built in the richer part of the West Bank, confirming the analysis reported in the introduction. If we use a simple differences-in-differences approach, poverty has diminished by $22.9 - 18.0 = 4.9$ percentage points in the un-impacted region while it diminished by only $15.6 - 12.2 = 3.4$ percentage points in the impacted region, which makes an excess of $4.9 - 3.4 = 1.5$ percentage points. The wall had clearly an impact on the dynamics of poverty, a fact that requires further investigation.

Table 2. Poverty rates decomposition.

		Poverty Rates		
Date	**Total**	**Jerusalem**	**Impacted**	**Un-Impacted**
1998	13.2	NA	NA	NA
2004	18.0	1.1	15.6	22.9
2011	14.7	1.1	12.2	18.0
		Sample Sizes		
1998	1965	NA	NA	NA
2004	1934	272	486	1176
2011	2909	271	847	1791

The official poverty line was used for computing the poverty rates. The variable *wall* can take three values: 0 Jerusalem, 2 impacted, 3 not impacted. This variable is available only for 2004 and 2011. Total population can be exactly decomposed according to these three characteristics.

3.1.2. Access to Land and to Public Water

Table 3 explores the effects of the wall on households in terms of access to land and water resources. Obviously, all households have reduced agricultural land cultivation, but households impacted by the wall have reduced land use in a larger proportion with diff-in-diff equal to $21.50 - 8.80 = 12.70$. This is due to difficulties in access to land behind the wall.

Access to a public water network has increased for all categories, and because the wall was built in the richer region, average access to water is greater in the impacted region than in the un-impacted region. However, the increase for un-impacted households is greater than for impacted households, leading to a diff-in-diff of $6.7 - 0.3 = 6.4$.

Table 3. Access to resources (percentage of households by wall's effect).

Population Group	1998	2004	2011	Diff
Land ownership and cultivation for un-impacted households	NA	36.4	27.6	−8.8
Land ownership and cultivation for impacted households	NA	43.8	22.3	−21.5
Land ownership and cultivation for Jerusalem	NA	7.4	0.4	−7.1
Total land ownership and cultivation rate	31.8	34.2	23.5	−10.7
Connection to water for un-impacted households	NA	79.0	85.7	6.7
Connection to water for impacted households	NA	94.0	94.3	0.3
Connection to water in Jerusalem	NA	99.6	100.0	0.4
Total connection rate to a public water network	78.9	85.7	89.5	3.8

The table indicates the percentage of households that own a land for cultivation and the percentage of households that have access to public water network. These figures are ventilated according to the value of the variable wall, that can take three values: 0 Jerusalem, 2 impacted, 3 not impacted. This variable is available only for 2004 and 2011.

3.1.3. Poverty and Access to Natural Resources

As indicated in the introduction, access to land and water resources may have an important impact on poverty and notably on chronic poverty status. Table 4 illustrates the poverty rates in 1998, before the wall's construction, in 2004 just after the wall construction, and seven years later in 2011, as a function of access to natural resources. These rates follow the same general pattern as those depicted in Table 2 (rise between 1998 and 2004 and a slight decrease in 2011, but without going back to the level of 1998 most of the time). However, we can note two major facts. Poverty is higher when there is no access to natural resources (land and public water), but the fluctuations are greater for those who had a profitable economic activity related to land ownership and access to public water network, which could mean that the effect of the wall was greater for those households.

Table 4. Poverty rates by access to resources.

Population Group	1998	2004	2011
Non-connected to the public water network	22.4	23.1	22.0
Connected to the public water network	10.7	17.1	13.9
Do not own an agricultural land	13.7	17.2	15.2
Own and cultivate an agricultural land	12.0	19.7	13.2

Land ownership is a simple dichotomous variable with 1 for TRUE. Access to water has varying items. For 1998, the variable *WAT* meant 1 = piped supply, 2 = public tap, 3 = well, 4 = tanker, 5 = other. For 2004, the variable *h*12*a* meant 1 = public network 2 = private system 3 = no piped water. Finally, for 2011, the variable *h*9*a* indicated water connection as to 1 = local public network, 2 = Israeli network, 3 = rain water, 4 = bridges, 5 = tank, 6 = others.

3.2. Empirical Strategy

Using the example of India, Bertrand et al. (2004) have shown that it is not at all an easy task to measure the impact of a dam on poverty rates, simply because dams are not built at random.

There are configurations where it is easier to build a dam, thus the decision to build a dam can be made endogenous because richer regions have more funds. In addition, the effect of a dam is positive downstream and negative upstream.

We have a somewhat similar case here. The wall was not built at random, but mainly in rich areas. So, even after the wall construction, poverty is still less significant in the impacted districts than in the un-impacted districts. Stylised facts have shown that the population was affected by the wall in very diverse ways. So simply introducing a dummy variable in a regression explaining the log of the income-to-needs ratio is not the correct way to proceed.[4] If we introduce that variable in such a regression as displayed in Table 5, we get an estimated positive coefficient both for 2004 and for 2011. The wall variable simply confirms that the wall was built in richer regions. We had a first very simple evaluation, using the evolution of poverty as given in Table 2. What needs to be shown is the effect that the wall had on poverty dynamics, which means poverty persistence and poverty exit. We have to develop a specific model for this purpose.

Table 5. Misleading Wall effect on income-to-needs ratio.

	2004			2011		
Regressors	**Estimate**	**Sd. Er.**	***t*-Value**	**Estimate**	**Sd. Er.**	***t*-Value**
Intercept	0.482	0.120	4.03	0.509	0.122	4.19
Wall	0.082	0.027	3.05	0.079	0.023	3.47
sex	0.249	0.041	6.15	0.162	0.038	4.27
age	−0.022	0.005	−4.39	−0.024	0.005	−4.81
age^2	0.027	0.005	5.46	0.031	0.005	6.29
Jerusalem	0.698	0.034	20.74	0.714	0.036	20.13
educ	0.038	0.003	13.88	0.039	0.003	14.53

Regression 2004: $\hat{\sigma} = 0.496$ on 1927 DF, $R^2 = 0.28$; Regression 2011: $\hat{\sigma} = 0.524$ on 2679 DF, $R^2 = 0.21$.

4. A Model of Poverty Dynamics in Repeated Cross-Sections

We propose a model for inference on poverty dynamics when the data are only in the form of repeated cross-sections, using a Bayesian approach. The Bayesian approach is particularly suited here because repeated cross-sections can be seen as an incomplete data problem. We first review the classical literature on poverty dynamics and on repeated cross-sections in order to formulate a model in terms of latent variables. We then present inference procedures to measure the impact of the wall on poverty dynamics.

4.1. Literature

The literature on poverty dynamics has much more recent roots than that of income dynamics. The founding papers were Lillard and Willis (1978) and Bane and Ellwood (1986). The former proposes a Markov chain model of income transitions. The latter is attached to the modelling of the length of poverty spells and the probability of exiting poverty. However, it is concerned only with head-count poverty and requires rather long panels. Rodgers and Rodgers (1993) still rely on long panels, but can distinguish between the three aspects of poverty (incidence, intensity and inequality) and propose a decomposition between chronic and transitory poverty. Cappellari and Jenkins (2004) tackle the question of attrition when measuring poverty dynamics, building a three equation model. The first equation explains the probability that an individual observed at time $t-1$ can still be observed at time t. The second equation explains the marginal probability of being in poverty at time $t-1$. The last equation explains the conditional probability of being in poverty at time t when in poverty at time $t-1$.

[4] The income-to-needs ratio is defined as the ratio between equivalised household total consumption and the official poverty line.

A key parameter is the correlation ρ of the error terms between t and $t-1$. This model is essentially a dynamic probit model with selection bias. It serves to explain poverty persistence and exit from poverty. Dang et al. (2014) have the same concern, but using a rather different strategy based on linear regressions (and not on probit models) explaining consumption or income. Their main originality is that they deal with repeated cross-sections instead of true panels, using time invariant explanatory variables to link two periods. However, Dang et al. (2014) could only provide bounds for poverty persistence and poverty exit probabilities as ρ is not identified in their model.

4.2. Modelling Poverty Dynamics Using Panel Data

We first develop a model inspired by Cappellari and Jenkins (2004), estimating poverty persistence and exit from poverty, assuming provisionally that we have a balanced panel over two periods, 1 and 2 (2004 and 2011 in our case). In a next sub-section, we shall adapt this model to repeated cross-sections. The main variable that we have to explain is the log of the income-to-needs ratio $\log(y_{i1}/z_1)$ where y_{i1} is the income of individual i in period 1 and z_1 is the poverty line applicable in time period 1. The following regression provides information on the initial state:

$$\log(y_{i,1}/z_1) = x'_{i,1}\beta + u_{i,1}. \tag{1}$$

Individual i is in a state of poverty if $\log(y_{i1}/z_1) < 0$. If the error term is Gaussian with zero mean and variance σ_1^2, the marginal probability of being poor in the initial period for individual i is equal to $\Phi(-x'_{i,1}\beta/\sigma_1) = 1 - \Phi(x'_{i,1}\beta/\sigma_1)$. As we are interested in a transition probability between periods 1 and 2, given the observed state in period 1, we define the dummy variable $d_{i,1}$:

$$d_{i,1} = \mathbb{1}(\log(y_{i,1}/z_1) < 0), \tag{2}$$

where $\mathbb{1}(a)$ is the indicator function equal to 1 if a is true and 0 otherwise. The income-to-needs ratio for the second period can now be explained by the observed past state of poverty $d_{i,1}$ and by some other exogenous variables $w_{i,1}$ related to the initial state and influencing the next state. The effect of $w_{i,1}$ should be allowed to be different, depending on the nature of the previous state. So the equation explaining the income-to-needs ratio in period 2 is:

$$\log(y_{i2}/z_2) = d_{i,1}w'_{i,1}\gamma_1 + (1-d_{i,1})w'_{i,1}\gamma_2 + u_{i,2}, \tag{3}$$

where $w_{i,1}$ is a set of exogenous variables observed in period 1, containing $x_{i,1}$ and at least another variable. The error term $u_{i,2}$ is assumed to be Gaussian with zero mean and variance σ_2^2. The two error terms are correlated over time with $\mathrm{Corr}(u_{i,2}, u_{i,1}) = \rho$ for the same individual i and independent between two different individuals. In the model of Cappellari and Jenkins (2004), ρ is identified only if $w_{i,1}$ has an element which is not in $x_{i,1}$.

Let us define a second dummy variable for period 2:

$$d_{i,2} = \mathbb{1}(\log(y_{i,2}/z_2) < 0). \tag{4}$$

Following Cappellari and Jenkins (2004), poverty persistence is defined as the state of being poor in period 2 while having being poor in period 1. Its probability is given by:

$$s_{i2} = \Pr(d_{i,2} = 1 | d_{i,1} = 1) = \Phi_2\left(-\frac{w'_{i,1}\gamma_1}{\sigma_2}, -\frac{x_{i,1}\beta}{\sigma_1}; \rho\right) \Big/ \Phi\left(-\frac{x_{i,1}\beta}{\sigma_1}\right), \tag{5}$$

which corresponds to the ratio between a joint probability and a marginal probability, Φ_2 being the bivariate Gaussian cumulative distribution. Poverty entry is defined in a similar way as:

$$e_{i2} = \Pr(d_{i,2} = 1 | d_{i,1} = 0) = \Phi_2\left(-\frac{w'_{i,1}\gamma_2}{\sigma_2}, \frac{x_{i,1}\beta}{\sigma_1}; -\rho\right) \Big/ \Phi\left(\frac{x_{i,1}\beta}{\sigma_1}\right). \quad (6)$$

Estimating model (1)–(3) is quite simple if we observe the same individual over the two periods. Poverty persistence and poverty exit are just simple transformations of the estimated parameters. However, of course we need a true panel, which of course is not the case here. We have only repeated cross-sections, which makes the problem more complex.

4.3. Poverty Dynamics and Repeated Cross-Sections

Feasible panel data sets are not so common, especially in developing countries and so techniques have been found in order to retrieve information from repeated cross-sections. Deaton (1985) first proposed to take means inside cohort clusters defined by time invariant instrumental variables (following the terminology used in Verbeek 2008). However, this approach can lead to a substantial loss of information. An alternative approach was taken in a series of papers, mainly Dang et al. (2011); Dang and Lanjouw (2013) and Dang et al. (2014), following an initial idea of out-of-sample imputation of Elbers et al. (2003). Let us consider again two periods, and two samples coming from the same population. The starting point is a two-equation model explaining the log of the income-to-needs ratio as before, denoted y here for simplicity of notation, for period 1 and period 2:

$$y_{i,1} = x_{i,1}\beta_1 + u_{i,1} \quad (7)$$
$$y_{j,2} = x_{j,2}\beta_2 + u_{j,2}. \quad (8)$$

There are $n = n_1 + n_2$ individuals with $i = 1, \ldots, n_1$ and $j = n_1 + 1, \ldots, n$, so that there is no overlapping between the two periods or if there is, we do not know which individuals are present in the two periods. So initially it would seem there is no clear link between these two equations, apart from the fact that the two samples are drawn from the same population, while i and j do not concern the same observed individuals or households. The first link that is introduced between the two samples is that $x_{i,1}$ and $x_{j,2}$ are time invariant exogenous variables and so it is clear that $x_{i,1} = x_{i,2}$ and $x_{j,1} = x_{j,2}$. The idea used in Dang et al. (2014) is to simulate values for the missing individuals in one of the two periods. Because both y_i and y_j are drawn from the same population and are functions of the same time invariant exogenous variables, we can simulate for instance the unobserved $\tilde{y}_{i,2}$, using $x_{i,1}\beta_2$. So having estimated separately the two equations in (7) and (8), Dang et al. (2014) define the event of entering a state of poverty as the joint event of not being poor at time 1 and being poor at time 2. This joint probability is given by:

$$\Pr(y_{j,2} < 0 \text{ and } y_{i,1} > 0) = \Pr(u_{j,2} < -x_{j,2}\beta_2 \text{ and } u_{i,1} > -x_{i,1}\beta_1). \quad (9)$$

This probability is a function of the joint distribution of $(u_{i,1}, u_{j,2})$ with a coefficient of association $\rho \geq 0$. If $\rho = 0$, then mobility attains its upper bound. If $\rho = 1$, then mobility reaches its lower bound. A positivity assumption for ρ is justified on the basis of household fixed effect and the persistence of shocks. However, apart from this prior restriction, ρ is not identified because the two equations in (8) are totally symmetric. Without further assumptions, Dang et al. (2014) can propose only bounds for poverty transition probabilities. Assuming a Gaussian distribution for the error terms, the probability of entering poverty becomes:

$$\Pr(y_{j,2} < 0 \text{ and } \tilde{y}_{j,1} > 0) = \Phi_2\left(\frac{-x'_{j,2}\beta_2}{\sigma_2}, \frac{x'_{j,2}\beta_1}{\sigma_1}, -\rho\right), \quad (10)$$

where the four quantities β_1, σ_1, β_2 and σ_2 have been estimated directly from the two equations of the initial model. A set of values for ρ has to be picked in the interval $[0, 1]$ in order to compute (10).

We should point out the main differences between the model of Dang et al. (2014) and our first model (1)–(3) based on Cappellari and Jenkins (2004). Our first model is fundamentally asymmetric whereas the model of Dang et al. (2014) assumes a perfect symmetry. Entering poverty in (6) is normalised by the probability of the initial state. The probability of entering poverty in (10) depends essentially on the differences between β_1 and β_2 and the differences between σ_1 and σ_2. There is no way of including the effect of the wall, except in a symmetric way in $x_{j,2}$. In addition, in Section 3.2, we have seen that to do so was not a reasonable solution.

4.4. Repeated Cross-Section as an Incomplete Data Problem

We want of course to treat the problem in a different way and show that a repeated cross-section is fundamentally an incomplete data problem. In a full data model, unobserved individuals in period 1, $\tilde{y}_{j,1}$ and the unobserved individuals in period 2, $\tilde{y}_{i,2}$ are treated as latent variables so that we have $n = n_1 + n_2$ individuals (observed and unobserved) for each period. Formally, this means:

$$\textit{Period 1} \begin{cases} \tilde{y}_{j,1} &= x_{j,2}\beta_1 + \tilde{u}_{j,1}, & \tilde{u}_{j,1} \sim \mathrm{N}(0, \sigma_1^2), \\ y_{i,1} &= x_{i,1}\beta_1 + u_{i,1}, & u_{i,1} \sim \mathrm{N}(0, \sigma_1^2), \end{cases} \tag{11}$$

$$\textit{Period 2} \begin{cases} y_{j,2} &= x_{j,2}\beta_2 + u_{j,2}, & u_{j,2} \sim \mathrm{N}(0, \sigma_2^2), \\ \tilde{y}_{i,2} &= x_{i,1}\beta_2 + \tilde{u}_{i,2}, & \tilde{u}_{i,2} \sim \mathrm{N}(0, \sigma_2^2). \end{cases} \tag{12}$$

For instance $y_{j,2}$ represents the observed group in period 2 while $\tilde{y}_{j,1}$ represents the same group being unobserved in period 1. The assumption that x does not vary over time (variables such as age, sex, religion, localisation,...) implies that $x_{i,2} = x_{i,1}$ and that $x_{j,1} = x_{j,2}$. This is an identification assumption that will allow us to make inference in this model. The second identification assumption is that the parameters are constant over all the individuals within the same period.

Let us now discuss the joint distribution of the four error terms which are $(\tilde{u}_{j,1}, u_{i,1}, u_{j,2}, \tilde{u}_{i,2})$. We assume that it is Gaussian with zero mean and variance-covariance matrix Σ. This matrix has a particular structure which results from the following very simple assumptions. We have assumed that within each period, the two error terms (corresponding to observed and to latent variables) have the same variance for identification reasons. We assume now that the individuals are not correlated, which means that there is no spatial correlation, simply because our data are not informative on that dimension. The important parameter to specify is the correlation ρ between two income observations for the same individual over the two periods. If we translate these assumptions into mathematical terms, we have:

$$\begin{aligned} \mathrm{Corr}(\tilde{u}_{j,1}, u_{i,1}) = 0, \mathrm{Corr}(\tilde{u}_{j,1}, u_{j,2}) = \rho, \mathrm{Corr}(\tilde{u}_{j,1}, \tilde{u}_{i,2}) = 0, \\ \mathrm{Corr}(u_{i,1}, u_{j,2}) = 0, \mathrm{Corr}(u_{i,1}, \tilde{u}_{i,2}) = \rho, \\ \mathrm{Corr}(u_{j,2}, \tilde{u}_{i,2}) = 0, \end{aligned}$$

which leads to the following variance-covariance matrix:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \rho\sigma_1\sigma_2 & 0 \\ 0 & \sigma_1^2 & 0 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & 0 & \sigma_2^2 & 0 \\ 0 & \rho\sigma_1\sigma_2 & 0 & \sigma_2^2 \end{pmatrix} \tag{13}$$

So even if (13) has a block diagonal structure, this does not mean that we have imposed restrictions that could be testable.

Dang et al. (2014) have chosen to simulate only one of the two latent variables, considering one of the two distributions, $\tilde{y}_{j,1}|x_{j,2}$ or $\tilde{y}_{i,2}|x_{i,1}$, treating thus the two periods independently. For instance, for the first period, they are using:

$$\tilde{y}_{j,1}|x_{j,2} \sim N\left(x_{j,2}\beta_1, \sigma_1^2\right). \tag{14}$$

By doing so, they loose a part of the available information. We prefer first to simulate both $\tilde{y}_{j,1}$ and $\tilde{y}_{i,2}$, and second to condition on all the other variables, so as to take into account the between periods correlation. Due to the particular structure of matrix (13), we have however the simplification $p(\tilde{y}_{j,1}|x_{j,2}, y_{i,1}, y_{j,2}, \tilde{y}_{j,2}) = p(\tilde{y}_{j,1}|x_{j,2}, y_{j,2})$:

$$\tilde{y}_{j,1}|x_{j,2}, y_{j,2} \sim N\left(x_{j,2}\beta_1 + \rho\frac{\sigma_1}{\sigma_2}(y_{j,2} - x_{j,2}\beta_2), \sigma_1^2(1-\rho^2)\right), \tag{15}$$

which means that it is necessary to condition only on the observed variables when simulating the latent variables. Conversely, for simulating $\tilde{y}_{i,2}$, we have:

$$\tilde{y}_{i,2}|x_{i,1}, y_{i,1} \sim N\left(x_{i,1}\beta_2 + \rho\frac{\sigma_2}{\sigma_1}(y_{i,1} - x_{i,1}\beta_1), \sigma_2^2(1-\rho^2)\right). \tag{16}$$

Equations (15) and (16) will be our basic tools to simulate the missing observations and thus build a completed panel. So despite the apparent restrictive structure of (13), the simulation of the latent variables takes into account as much information as possible.

A model written as an incomplete data problem leads logically to a Gibbs sampler. Missing observations are generated conditionally on values of the parameters which are then reevaluated conditionally on the simulated missing variables. However in our case, things are not as simple as that. The identifying assumptions $x_{i,2} = x_{i,1}$ and $x_{j,1} = x_{j,2}$ provide information for both the regression coefficients β_i and the latent variables. For the present we have no specific source of information for the correlation coefficient ρ while it is needed for simulating the latent variables. We need to provide an extra source of information and this will be the object of Section 4.6. It is important to remember that Dang et al. (2014) provide only bounds for ρ.

4.5. Inference on the Regression Parameters

Conditional on simulated values for the latent variables, our two period regression model with variance-covariance matrix (13) corresponds to a simple SURE (Seemingly Unrelated Regression) model. In Bauwens et al. (1999, Chap. 9), it is shown how Bayesian inference in a SURE model can be done using a Gibbs sampler for estimating jointly the regression parameters and the correlation structure. We can also decompose inference into two stages: the regression parameters on one hand and the correlation structure on the other hand. For that purpose, we decompose Σ as given in (13) into the product of two diagonal matrices S containing σ_1 and σ_2, sandwiching a correlation matrix R, so that $\Sigma = S\,R\,S$, following a suggestion made in Barnard et al. (2000). In our case, these matrices are:

$$S = \begin{pmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_1 & 0 & 0 \\ 0 & 0 & \sigma_2 & 0 \\ 0 & 0 & 0 & \sigma_2 \end{pmatrix}, R = \begin{pmatrix} 1 & 0 & \rho & 0 \\ 0 & 1 & 0 & \rho \\ \rho & 0 & 1 & 0 \\ 0 & \rho & 0 & 1 \end{pmatrix}.$$

Let us now factorise the product $S\,R\,S = P\,P'$ using a Choleski decomposition. We then form the matrix $L = S\,P^{-1}$ which is equal to:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -\rho\sigma_2/(\sigma_1\sqrt{1-\rho^2}) & 0 & 1/\sqrt{1-\rho^2} & 0 \\ 0 & -\rho\sigma_2/(\sigma_1\sqrt{1-\rho^2}) & 0 & 1/\sqrt{1-\rho^2}) \end{pmatrix}.$$

If we pre-multiply our system of four equations by L, the matrix of variance covariance of the error terms will be transformed into a diagonal matrix:

$$L\,\Sigma\,L' = \begin{pmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_1^2 & 0 & 0 \\ 0 & 0 & \sigma_2^2 & 0 \\ 0 & 0 & 0 & \sigma_2^2 \end{pmatrix} = S^2,$$

meaning that we can make inference on the two blocks separately. Let us pre-multiply our model (11) and (12) by L. Due to the particular structure of L, the model for the first period is left unchanged:

$$\begin{cases} \tilde{y}_{j,1} &= x_{j,2}\beta_1 + v_{j,1}, \\ y_{i,1} &= x_{i,1}\beta_1 + v_{i,1}, \end{cases} \tag{17}$$

with $v_{j,1} = \tilde{u}_{j,1}$, $v_{i,1} = u_{i,1}$. The model for the second period is modified into a conditional model with:

$$\begin{cases} \frac{1}{\sqrt{1-\rho^2}}(y_{j,2} - \rho\omega\tilde{y}_{j,1}) &= \frac{1}{\sqrt{1-\rho^2}}(x_{j,2}\beta_2 - \rho\omega x_{j,2}\beta_1) + v_{j,2}, \\ \frac{1}{\sqrt{1-\rho^2}}(\tilde{y}_{i,2} - \rho\omega y_{i,1}) &= \frac{1}{\sqrt{1-\rho^2}}(x_{i,1}\beta_2 - \rho\omega x_{i,1}\beta_1) + v_{i,2}, \end{cases} \tag{18}$$

with $\omega = \sigma_2/\sigma_1$ and

$$v_{j,2} = (u_{j,2} - \rho\omega\tilde{u}_{j,1})/\sqrt{1-\rho^2}, \qquad v_{i,2} = (\tilde{u}_{i,2} - \rho\omega u_{i,1})/\sqrt{1-\rho^2}.$$

Let us now define the following matrices with $n = n_1 + n_2$ rows so as to write our model in a matrix form:

$$ys_1 = \begin{pmatrix} \tilde{y}_1 \\ y_1 \end{pmatrix}, \qquad ys_2 = \begin{pmatrix} y_2 \\ \tilde{y}_2 \end{pmatrix}, \qquad X = \begin{pmatrix} X_2 \\ X_1 \end{pmatrix},$$

where $\tilde{y}_1$, y_1 are respectively two vectors of dimension n_2 and n_1 (y_2, $\tilde{y}_2$ are of dimension n_2 and n_1) and X_2, X_1 are two matrices with corresponding number of rows n_2 and n_1. Our model for the first period can be written as:

$$ys_1 = X\beta_1 + v_1, \tag{19}$$

and Bayesian inference on its regression parameters can be done separately from inference on the other parameters. Because the model is symmetric between the two periods, we can adopt the reverse ordering for the Choleski decomposition, obtain another matrix L and consider for period two the symmetric model:

$$ys_2 = X\beta_1 + v_2. \tag{20}$$

So conditional on the simulated values for the latent variables, the two regression models (19) and (20) can be analysed separately. Under a non-informative prior, the posterior densities of $\beta_1|\sigma_1^2$ and $\beta_2|\sigma_2^2$ are two conditional normal densities while those of σ_1^2 and σ_2^2 are two inverse gamma2 densities with:[5]

[5] See Bauwens et al. (1999, Appendix A) for a definition of the densities used.

$$
\begin{aligned}
p(\beta_1|\sigma_1^2, ys_1) &= f_N(\beta_1|\beta_{*1}, \sigma_1^2 M_*^{-1}), && (21)\\
p(\sigma_1^2|ys_1) &= f_{i\gamma}(\sigma_1^2|s_{*1}, \nu_*), && (22)\\
p(\beta_2|\sigma_2^2, ys_2) &= f_N(\beta_2|\beta_{*2}, \sigma_2^2 M_*^{-1}) && (23)\\
p(\sigma_2^2|ys_2) &= f_{i\gamma}(\sigma_2^2|s_{*2}, \nu_*), && (24)
\end{aligned}
$$

where $\nu_* = n$, $M_* = X'X$ and:

$$
\begin{aligned}
\beta_{*1} &= M_*^{-1} X' ys_1,\\
\beta_{*2} &= M_*^{-1} X' ys_2,\\
s_{*1} &= ys_1' ys_1 - \beta_{*1}' M_* \beta_{*1},\\
s_{*2} &= ys_2' ys_2 - \beta_{*2}' M_* \beta_{*2}.
\end{aligned}
$$

We have the needed formulae for conducting Bayesian inference on the regression parameters, conditional on the simulated values for the latent variables. We must now detail how to make inference on the correlation parameter ρ.

4.6. Inference on Correlation

Returning to the remaining equations in the system (17) and (18) where ρ appears explicitly, let us multiply them both by $\sqrt{1-\rho^2}$, rearrange the terms and adopt a similar matrix notation as before. We get:

$$ys_2 - X\beta_2 = \rho\omega[ys_1 - X\beta_1] + v_2\sqrt{1-\rho^2}, \quad (25)$$

where v_2 is defined in (20) with zero mean and variance-covariance matrix $\sigma_2 I_n$. From the previous step, we have knowledge of β_1, β_2, ω and σ_2^2. So in theory we could recover information on ρ using this regression which is nothing but a regression of the residuals from the period 2 model on the residuals from period 1 model. However, things are not as simple as that, because in fact no new information is brought in by this autoregressive model which only compares observations to their simulated counterpart ($\tilde{y}_{i,2}$ to $y_{i,1}$ and $y_{j,2}$ to $\tilde{y}_{j,1}$). We have to find a source of extra information.

Verbeek and Vella (2005) and Verbeek (2008) have proposed to use the grouping technique of Deaton (1985) for making inference in autoregressive pseudo panel models, using the fact that taking group averages is equivalent to IV estimation with the group dummies as instruments. We shall use time invariant information such as birth date and gender as instruments to determine $c = 1, \ldots, C$ cohorts and then compute the cell means of each vector or matrix of observations. We collect these C cell means into new vectors noted $y_{2,c}$, $y_{1,c}$ having each C rows and two new matrices $X_{2,c}$, $X_{1,c}$ with also C rows. They correspond respectively to the C cell means of y_2, y_1 and of X_2, X_1. Let us now define the following vectors and matrices:

$$yc_2 = \begin{pmatrix} y_{2,c} \\ y_{2,c} \end{pmatrix}, \qquad yc_1 = \begin{pmatrix} y_{1,c} \\ y_{1,c} \end{pmatrix}, \qquad Xc = \begin{pmatrix} X_{2,c} \\ X_{1,c} \end{pmatrix}, \quad (26)$$

which all have $2 \times C$ rows. Let us rearrange (25) into:

$$ys_2 = \rho\omega ys_1 + X[\beta_2 - \rho\omega\beta_1] + v_2\sqrt{1-\rho^2}. \quad (27)$$

We then replace in (27) ys_2 by yc_2, ys_1 by yc_1 and X by Xc so as to get:

$$yc_2 = \rho yc_1\omega + Xc\beta_3 + u_c. \quad (28)$$

When comparing (28) to (27), we see that we have replaced y_2 by $y_{2,c}$, but also $\tilde{y}_2$ by the same $y_{2,c}$. A similar replacement was done for the endogeneous variable of the first period. This means that we have replaced latent variables by observed cell means.[6] This model can be simplified if we consider the deviation matrix $M_X = I_C - Xc(Xc'Xc)^{-1}Xc'$. Let us pre-multiply (28) by M_X to get:

$$M_X\, yc_2 = \rho M_X\, yc_1\omega + M_X\, u_c. \tag{29}$$

We shall use (29) to get information on ρ, assuming that ω was determined in the previous step.[7] With (29), we have an autoregressive regression model where inference can be conducted in a simple way, conditionally on ω. We assume that the error term $M_X\, u_c$ is of zero mean and variance-covariance matrix $\sigma_3^2 I_{2C}$. It is interesting to introduce the possibility of an informative natural conjugate prior for ρ. We consider the following prior information:

$$\mathrm{E}(\rho) = r_0, \qquad \mathrm{Var}(\rho|\sigma_3^2) = \sigma_3^2/m_0, \qquad \mathrm{E}(\sigma_3^2) = s_0/(\nu_0 - 2).$$

The conditional posterior density of $\rho|\sigma_3^2$ is a normal density and that of σ_3^2 is an inverse gamma2 density with:

$$p(\rho|\sigma_3^2, yc_2, \omega) = f_N(\rho|\rho_*, \sigma_3^2/r_*), \tag{30}$$
$$p(\sigma_3^2|yc_2, \omega) = f_{i\gamma}(\sigma_3^2)|s_*, nc_*), \tag{31}$$

where:

$$\begin{aligned} r_* &= yc_1' M_X yc_1 \omega^2 + m_0, \\ \rho_* &= r_*^{-1}(yc_1' M_X yc_2 \omega + m_0 r_0), \\ s_* &= s_0 + r_0^2 m_0 + yc_2' M_X yc_2 - \rho_*^2/r_*, \\ nc_* &= 2C + \nu_0. \end{aligned}$$

The parameter ρ has to be constrained to a specific range. It has to be strictly less than 1 for stability reasons and Dang et al. (2014) have shown that it has to be positive. So we have to constraint $\rho \in [0, 1]$. Consequently, we shall not draw ρ directly from its conditional posterior density $p(\rho|\sigma_3^2, yc_2, \omega)$, but from this posterior density truncated between 0 and 1.

4.7. A Separate Dynamic Equation for Implementing a Diff-In-Diff Strategy

The last equation we have to discuss corresponds to (3) where we introduce the effect of the wall in the second period. We have to suppose that the households impacted by the wall do not change over the two periods, which means that the wall is a time invariant variable (even if its effect is diluted over time) and that the households do not migrate (we shall document that point later). Let us call W the vector containing the index of the impacted households by the wall in period 1 and in period 2. Let us define a random dummy variable d_1 which indicates for period 1 if a household was in a state of poverty or not:

$$d_1 = \mathbb{1}(ys_1 < 0). \tag{32}$$

This variable is a function of all the parameters because ys_1 contains simulated values of a latent variable. We can then specify our last equation to estimate as being:

$$ys_2 = d_1 W\gamma_1 + (1 - d_1)W\gamma_2 + X\beta_3 + u, \tag{33}$$

---

[6] As a matter of fact, it would be quite difficult to define cell means for the simulated values of the latent variables.
[7] Expressed in a totaly different way, we could find a similar idea in Dang and Lanjouw (2013).

with u being Gaussian with zero mean and variance-covariance matrix $\sigma_u^2 I_n$. It can be put in a matrix form if we define:

$$Z = [d_1 W, (1-d_1)W, X], \qquad \delta' = (\gamma_1, \gamma_2, \beta_3').$$

The posterior density of δ in the regression $ys_2 = Z\delta + u$ has the usual form under a non-informative prior:

$$p(\delta|\sigma_u^2, ys_2) = f_N(\delta|\delta_*, \sigma_u^2(Z'Z)^{-1}), \tag{34}$$
$$p(\sigma_u^2|ys_2) = f_{i\gamma}(\sigma_u^2|s_{*u}, \nu_*), \tag{35}$$

with $\nu_* = n$ and

$$\delta_* = (Z'Z)^{-1}Z'ys_2, \qquad s_{*u} = ys_2' ys_2 - \delta_*' Z'Z\,\delta_*.$$

We propose to measure the effect of the wall on poverty dynamics by a diff-in-diff strategy. Poverty follows its own trend which can be measured by a transition matrix, as we were able to simulate the income-to-needs ratios, ys_1 and ys_2, for the two periods, using (15) and (16). The corresponding matrix of marginal poverty transition is defined as:

$$P = \begin{pmatrix} \Pr(ys_1 < 0 \cap ys_2 < 0) & \Pr(ys_1 < 0 \cap ys_2 > 0) \\ \Pr(ys_1 > 0 \cap ys_2 < 0) & \Pr(ys_1 > 0 \cap ys_2 > 0) \end{pmatrix},$$

with rows summing to one and generic element p_{ij}. p_{11} is the probability of staying in a state of poverty between periods 1 and 2 while p_{21} is the probability of entering poverty in period 2. For each draw l of the parameters, this matrix can be evaluated as:

$$\begin{aligned} p_{11} &= \sum d_1^{(l)} \times d_2^{(l)} / \sum d_1^{(l)}, \\ p_{12} &= \sum d_1^{(l)} \times (1 - d_2^{(l)}) / \sum d_1^{(l)}, \\ p_{21} &= \sum (1 - d_1^{(l)}) \times d_2^{(l)} / \sum (1 - d_1^{(l)}), \\ p_{22} &= \sum (1 - d_1^{(l)}) \times (1 - d_2^{(l)}) / \sum (1 - d_1^{(l)}), \end{aligned}$$

where $d_1^{(l)}$ is defined in (32) with ys_1 being replaced by $ys_1^{(l)}$ and $d_2^{(l)}$ is defined accordingly for period 2. These dummy variables indicate for the two periods and for each draw if an household is in a state of poverty or not. As they are computed for each MCMC draw of the parameters, p_{11} and p_{21} are estimated by averaging over the MCMC output.

These marginal probabilities are modified when we take into account the wall effect by means of the dynamic conditional model (33). With the Gibbs output for γ_1 and γ_2, we can evaluate the vectors of poverty persistence s_2 and poverty entry e_2. More precisely, for poverty persistence, we have:

$$s_2 = \Phi_2\left(-\frac{W\gamma_1 + X\beta_3}{\sigma_u}, -\frac{X\beta_1}{\sigma_1}; \rho\right) \Big/ \Phi\left(-\frac{X\beta_1}{\sigma_1}\right), \tag{36}$$

which corresponds to the ratio between a joint probability and a marginal probability, Φ_2 being the bivariate Gaussian cumulative distribution. While for poverty entry, we get:

$$e_2 = \Phi_2\left(-\frac{W\gamma_2 + X\beta_3}{\sigma_u}, \frac{X\beta_1}{\sigma_1}; -\rho\right) \Big/ \Phi\left(\frac{X\beta_1}{\sigma_1}\right). \tag{37}$$

If we impose the restriction $\gamma_1 = 0$, then $p_{11} = s_2$, while the restriction $\gamma_2 = 0$ would imply $p_{21} = e_2$. Consequently the net effect of the wall is measured by the influence of γ_1 and γ_2 and is obtained by computing the posterior density of the two differences: $s_2 - p_{11}$ for the effect of the wall on poverty persistence; $e_2 - p_{21}$ for the effect of the wall on poverty entry.

4.8. Summarised Gibbs Sampler Algorithm

Inference on β_1, σ_1, β_2, σ_2, ρ and $\delta' = (\gamma_1, \gamma_2, \beta_3')$ is provided by a Gibbs sampler for which we have found all the conditional posterior densities. We summarise the whole process in the following Gibbs sampler algorithm:

Define starting values for:

1. β_1, σ_1 obtained using OLS on $y_1 = X_1\beta_1 + v_1$.
2. β_2, σ_2 obtained using OLS on $y_2 = X_2\beta_2 + v_2$.
3. compute $\omega = \sigma_2/\sigma_1$.
4. determine cohort cells means for y_1, y_2 and X_1, X_2.
5. compute yc_1, yc_2, X_c and M_X.
6. ρ obtained using OLS on $M_X yc_2 = \rho\, M_X yc_1 \omega + u_c$.

Start iterations for the Gibbs sampler over $l = 1, \ldots, m$:

1. simulate the latent variables $\tilde{y}_{j,1}$ and $\tilde{y}_{i,2}$, given $\rho^{(l-1)}$, $\beta_1^{(l-1)}$, $\beta_2^{(l-1)}$, $\omega^{(l-1)}$ using (15) and (16) so as to form $ys_1^{(l)}$ and $ys_2^{(l)}$.
2. draw $\beta_1^{(l)}$, $\beta_2^{(l)}$ and $\sigma_1^{(l)}$, $\sigma_2^{(l)}$ using the posterior densities (21)–(24), corresponding to the full marginal models for periods 1 and 2, compute $\omega^{(l)} = \sigma_2^{(l)}/\sigma_1^{(l)}$.
3. draw $\rho^{(l)}$ and $\sigma_3^{2(l)}$ using the normal posterior density (30) truncated on [0, 1] and the inverse-gamma2 posterior density (31), based on the auxiliary model $M_X yc_2 = \rho\, M_X yc_1 \omega^{(l-1)} + u$.
4. compute $d_1^{(l)}$ from $ys_1^{(l)}$ and $d_2^{(l)}$ from $ys_2^{(l)}$.
5. deduce $p_{11}^{(l)}$ and $p_{21}^{(l)}$ from $d_1^{(l)}$ and $d_2^{(l)}$.
6. draw $\delta^{(l)}$ using the posterior densities (34) and (35).
7. evaluate the probabilities $s_2^{(l)}$ and $e_2^{(l)}$ defined in (36) and (37), using bivariate cumulative Gaussian routines and compute the differences $s_2^{(l)} - p_{11}^{(l)}$ and $e_2^{(l)} - p_{21}^{(l)}$.
8. store the draws.
9. redo the loop.

From this MCMC output, we can produce graphs of the posterior density of poverty entry and poverty persistence. All computations were done in `R`, using the packages `truncnorm` for drawing from a truncated normal and `pbivnorm` for the bivariate Gaussian cumulative.

5. Empirical Results

The general setting is as follows. We have chosen gender of the household head, age, age squared (divided by 10), urban, camp and Jerusalem as time invariant explanatory variables to generate the pseudo panel. We have excluded the level of education from this list as there are missing observations in the second period for this variable. We have used the year of birth and gender to define the cohorts in the model for ρ. We used 25 classes for the year of birth of male household heads and only 10 classes for female heads who are smaller in number. Year of birth classes are determined using quantile intervals. This made 35 classes and 70 observations for the double regression (28). We ran a specification search for this regression, retaining as explanatory variables sex, age, age^2, urban, refugee camp. We used an informative Gaussian prior for $\rho|\sigma_3^2$ for which we have chosen $\mathrm{E}(\rho) = 0.50$ and $SD(\rho) = 0.50$. The corresponding prior on the variance of the error term is an inverse-gamma2 for which we have chosen $\nu_0 = 10$ and $s_0 = 2$. This is a rather mild prior, but we shall provide a sensitivity analysis. We used non-informative priors for the remaining parameters of the model.

5.1. Marginal Models and Poverty Dynamics

With 10,000 draws for the Gibbs sampler and 500 additional draws for warming up the chain, the posterior moments of ρ are 0.399 for the mean and 0.017 for the standard deviation, using our informative prior. If we had used a non-informative prior on ρ, the posterior mean would have been 0.297 and the posterior standard error 0.013. The graph of the posterior density given in Figure 1 clearly indicates the sensitivity of this parameter to the given prior information, even when the latter is mild. However, we shall see that the final result (entry in poverty and poverty persistence) is much less sensitive to the prior specification.

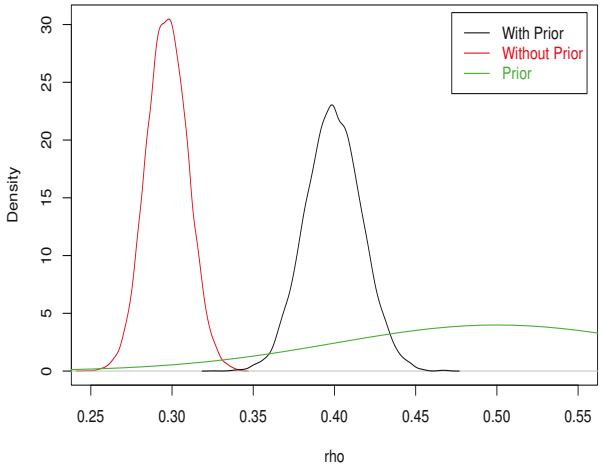


Figure 1. Prior and posterior densities for ρ. The prior standard deviation was divided by 5 in order to keep a reasonable scale on the graph. The actual prior is much less informative.

The posterior moments of the parameters of the two marginal models that are used to generate the pseudo-panel are displayed in Table 6. Inference results change between the two periods. In particular the influence of the variables *sex*, *age* and *urban* is weakly determined for the second period. Using a non-informative prior on ρ would change slightly the posterior expectation of ρ as indicated above, but would not change strongly the other regression parameters. We must note that a usual regression model for explaining the income-to-needs ratio would include many other covariates such as household size, dependency ratio, sources of income. However, these are not time invariant and because we are in a pseudo panel context, we cannot use these variables.

From this model, we can estimate the matrix of marginal probability transitions between the two states of poverty and non-poverty in the absence of a specific modelled effect of the wall. This result, as displayed in Table 7, is in a way comparable to what could be computed from the approach of Dang et al. (2014), except that we have normalised the transition matrix when they have not and that we have estimated ρ while they have given only bounds. The probability of poverty entry is equal to 0.109 while the probability of poverty persistence is 0.319 which are both rather low values. This might be due to omitted variables. Of course the wall effect is not present in this marginal model, but are also missing the impact of other restrictive measures set up by the Israeli government, some examples of which were reported in the introduction. However, we are interested in measuring the difference in poverty dynamics between an hypothetical situation described by our marginal models and a situation where the wall is introduced. So the variables which are omitted both in Tables 6 and 8 should not impact too much the difference in poverty dynamics measured by γ_1 and γ_2.

Table 6. Marginal models over the two periods.

	2004			2011		
Regressors	**Mean**	**SD**	**Ratio**	**Mean**	**SD**	**Ratio**
Intercept	0.682	0.131	5.194	0.634	0.117	5.421
Sex	0.151	0.044	3.399	0.041	0.033	1.219
Age	−0.013	0.005	−2.423	−0.007	0.005	−1.461
Age2	0.014	0.005	2.600	0.009	0.004	2.098
Urban	0.080	0.027	2.961	0.037	0.024	1.511
Camp	−0.087	0.036	−2.401	−0.172	0.031	−5.601
Jerusalem	0.712	0.035	20.253	0.711	0.037	19.373
σ	0.519	0.008	61.670	0.542	0.007	76.531
ρ	0.399	0.0174	23.00			

These results were obtained with the informative prior $\mathrm{E}(\rho) = 0.50$ and $SD(\rho) = 0.50$ and a non-informative prior on the remaining parameters. Mean is the sample average of the draws, SD means standard deviation of the draws and Ratio is the mean divided by the SD.

Table 7. Implied marginal transition matrix.

	Poor	**Non Poor**	**S.d.**	**S.d.**
Poor	0.319	0.681	0.017	0.017
Non Poor	0.109	0.891	0.005	0.005

The lines represent transition probabilities from state i to state j, from period 1 to period 2. The second panel indicates the standard deviations.

Table 8. Conditional transition model with state dependence (wall effect) and an informative prior.

Regressors	**Mean**	**S.d.**	**Ratio**
Intercept	0.609	0.119	5.129
γ_1	−0.254	0.049	−5.132
γ_2	0.109	0.024	4.492
Sex	0.035	0.033	1.036
Age	−0.006	0.005	−1.291
Age^2	0.008	0.004	1.910
Urban	0.023	0.025	0.922
Camp	−0.164	0.031	−5.271
Jerusalem	0.727	0.038	19.298
σ	0.537	0.007	74.066

These results were obtained with the informative prior $\mathrm{E}(\rho) = 0.50$ and $SD(\rho) = 0.50$ and a non-informative prior on the remaining parameters. Mean is the sample average of the draws, S.d. means standard deviation of the draws and Ratio is the mean divided by the S.d.

5.2. State Dependence and Wall Effect

Let us now examine inference results for the transition model, which takes into account the influence of the wall and of state dependence. Results are displayed in Table 8. The state dependence effect, which corresponds to the wall effect while being in a poverty state in period 1, has a marked negative effect. The value of γ_2 is positive and strongly significant. Both are going to alter the probabilities of poverty persistence and poverty entry. The main determinants of poverty dynamics are being in a camp and living in Jerusalem, which are too opposed situations.

There are various types of endogeneity problems that could affect inference on γ_1 and γ_2, on top of the effect of omitted variables. First, the construction of the wall might have had a general equilibrium effect. However, a Keynesian effect focusing on the demand side seems more realistic for these areas that have almost no production and where consumption is mostly dependent on imports of Israeli goods. Production in areas located beside the wall is mostly for self-consumption. Second, a selection

bias could occur if the prospect of the construction of the wall had induced potentially affected people to move. However, the West Bank is a small area (around 5300 square km) which makes internal migration more costly than daily transportation, especially because we are taking into account only localities located in the inside of the West Bank side of the wall. Moreover, PCBS (2010) gives evidence that internal migration is minimal in the West Bank; only 24.4% of Palestinians have changed their place of residence in the previous 6 years (2004–2010). Among those who migrated only 3% had done that due to the wall's construction or due to measures of the Israeli occupation forces.[8]

In Table 9, we recall in the top panel the marginal probabilities of poverty entry and persistence as obtained from the two marginal models. The second panel provides the same quantities obtained using (36) and (37) using the estimated values of γ_1 and γ_2. In the bottom panel of Table 9, we compute the difference between the two types of probabilities which provides a measure of the impact of the Wall on poverty dynamics. Taking into account the wall has a large effect on poverty dynamics. For those who were already poor in period 1, the wall increases their probability of staying poor by 58 percentage points. For those who were not in poverty, the probability of entering into poverty during the second period is increased by 18 percentages points.

We have reproduced in Figure 2 the posterior densities of these probabilities, using plain lines. We compare these probabilities to those obtained under a non-informative prior on ρ, using dashed lines. With a non-informative prior on ρ, the differential in probability of poverty entry is slightly increased while the differential in poverty persistence is slightly decreased. However, these differences are mild. So the prior information we gave had a sizable influence on the posterior density of ρ, but not on the posterior density of poverty entry and persistence differentials.

Table 9. Wall effect on poverty dynamics.

Effects	Mean	S.d.	Ratio
Marginal Persistence	0.320	0.017	19.38
Marginal Entry	0.109	0.005	21.11
Conditional Persistence	0.897	0.026	34.91
Conditional Entry	0.287	0.007	43.10
Diff. in persistence	0.578	0.029	20.17
Diff. in entry	0.178	0.007	24.80

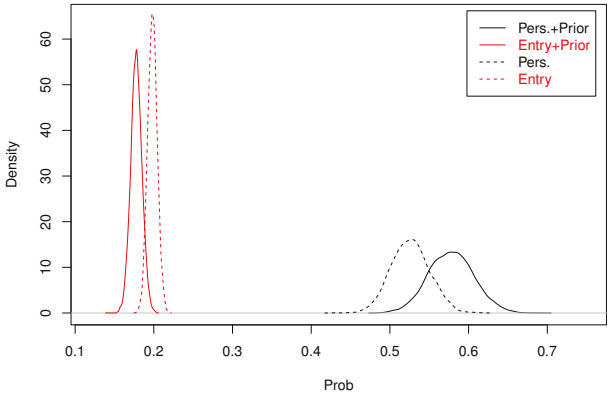


Figure 2. Posterior density of poverty persistence and poverty entry differentials due to the Wall.

[8] We are grateful to a referee for pointing out the potential problems of endogeneity.

5.3. Convergence Checks

Convergence of the Gibbs algorithm was checked for our evaluation of poverty persistence and poverty entry after the wall. We used normalised CUMSUM plots, used for instance in Bauwens and Lubrano (1998). Figure 3 shows that with only 2000 draws and 500 draws for warming up of the chain, we get already a good convergence for the final estimators, well within the 20% confidence band. In addition, for 10,000 as we used, convergence is more than satisfactory.

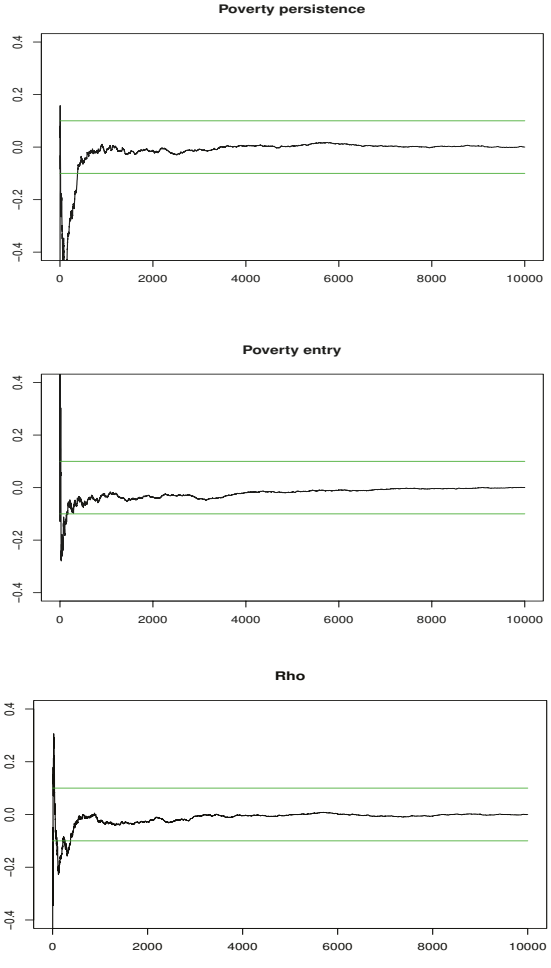


Figure 3. CUMSUM graphs with 20% confidence bands for assessing Gibbs convergence using an informative prior.

6. Conclusions

Governments have always been tempted to build walls to solve problems confronting them. The first historical example is the Great Wall of China, the building of which started three centuries B.C., soon followed by the Hadrian Wall in Scotland built by the Roman Emperor Hadrian. These walls were designed to keep others out of the frontiers. In the Twentieth century, the meaning of the Wall in Berlin was different, the aim being to keep people in. The destruction of this wall was synonymous

of freedom regained. Following Cohen (2006), the aim of the wall built by the Israeli government is to keep people both in and out. We can note that this wall is by no mean an exception in the world, witness the long list detailed in Vallet and David (2012a, 2012b) showing their exponential increase since the fall of Berlin wall. However, no other wall has been subject of so many emotional reactions as underlined by Cohen (2006). It is certainly because its route "*harms Palestinians to a disproportionate degree.*" However, the same author concludes his paper by the following sentence: "*Perhaps then, in the context of a negotiated settlement between Israelis and Palestinians, it can join other famous walls in becoming obsolete.*" This statement finds some grounds in the results of Longo et al. (2014) who exploited a natural experiment based on a 2009 policy toward the "easement" of checkpoints in the West Bank. They found, using a diff-in-diff approach that "easement" made the treated populations less likely to support violence.

We have shown empirically that the wall had a large impact on poverty dynamics in the West Bank. This result was obtained using pseudo panels and a Bayesian approach. The method we propose provides a coherent way of simulating unobserved values, stating the question in the framework of incomplete data problems. It resulted in a Gibbs sampler from which we could provide inference for two different types of probabilities concerning poverty entry and poverty persistence. These different types come from the distinction between a marginal model and a conditional model taking into account state dependence, or phrased differently taking or not into account the previous state of poverty. In our model, we allow the wall to have a different effect, depending if the household was poor or not during the previous period. We identified a clear different effect which allows us to evaluate how the dynamics of poverty was impacted by the wall. We have documented a rather weak effect of 1.50% using a simple diff-in-diff approach when examining stylised facts and no panel structure. With a much more elaborated model using pseudo panels, we could extract more precise information from our sample which led to measuring an increase of 18 percentage points for poverty entry and of 58 percentage points for poverty persistence. In addition, these results were quite robust to the specification of the prior used for ρ. However, we must note that these results are dependent of important identification assumptions which are not testable. It would be interesting to have a true panel in order to be less constrained in the specification of the model in order to check the accuracy of our evaluation of the impact of the wall on poverty dynamics.

Author Contributions: Both authors contributed equally to the paper.

Funding: This work was supported by the A*MIDEX project (No ANR-11-IDEX-0001-02) funded by the "Investissements d'Avenir" French Government program, managed by the French National Research Agency (ANR) and later by the ANR project (No. ANR-17-CE39-0009-01) funded by the French National Research Agency (ANR), within the framework of the TMENA2 projects.

Acknowledgments: We would like to thank Luc Bauwens for penetrating discussions as well as Mohammad Abu-Zaineh, Thibault Gajdos and Stephen Bazen. This paper was presented in various places: Marseille, workshop on Multidimensional Inequalities, May 2015; Orleans, French Econometric Conference, December 2015; Jiangxi University of Economics and Finance, April 2016; Rimini, 10th Annual RCEA Bayesian Econometric Workshop, May 2016. We would like also to thank two anonymous referees for their useful comments and suggestions. Of course, remaining errors and shortcomings are solely ours. This paper was written at the occasion of the visit of Tareq Sadeq in Marseille during the year 2014–2015 financed by a European mobility grant.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Variables in the Data Bases

The data base provided by the Palestinian Central Bureau of Statistics contains a lot of variables, among which we have use those depicted in Table A1.

Table A1. Variables used for estimation.

Name	Variable 2004	Variable 2011
Age of H	age	d5
Sex of H	sex	d4-1
Year school HH	yerschol1	d15
Rural	loctype=2	loc_type=2
Camp	loctype=3	loc_type=3
H size adults	adults	no_adults
H size	ir04_mal+ir04_fem	IR04_male + IR04_female
Wall impact	id13=2	id10=2
No Wall impact	id13=3	id10=3
Jerusalem	id13=0	id10=0
Cons adj H size	consadj	consadj
Consumption	cons	TOT_Cons
Public water access	h12a=1	h9a=1 + h9a=2
Own agricultural land	i18=1	h24=1

References

Adnan, Wifag. 2015. Who gets to cross the border? The impact of mobility restrictions on labor flows in the West Bank. *Labour Economics* 34: 86–99. [CrossRef]

Aliber, Michael. 2003. Chronic poverty in South Africa: Incidence, causes and policies. *World Development* 31: 473–90. [CrossRef]

Atran, Scott, Robert Axelrod, and Richard Davis. 2007. Sacred barriers to conflict resolution. *Science, American Association for the Advancement of Science* 317: 1039–40. [CrossRef] [PubMed]

Bane, Mary Jo, and David T. Ellwood. 1986. Slipping into and out of poverty: The dynamics of spells. *Journal of Human Ressources* 21: 1–23. [CrossRef]

Barnard, John, Robert McCulloch, and Xiao-Li Meng. 2000. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* 10: 1281–311.

Bauwens, Luc, and Michel Lubrano. 1998. Bayesian inference on GARCH models using the Gibbs sampler. *The Econometrics Journal* 1: 23–46. [CrossRef]

Bauwens, Luc, Michel Lubrano, and Jean-François Richard. 1999. *Bayesian Inference in Dynamic Econometric Models*. Oxford: Oxford University Press.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics* 119: 249–75. [CrossRef]

Bornstein, Avram S. 2001. Border enforcement in daily life: Palestinian day labourers and entrepreneurs crossing the green line. *Human Organization* 60: 298–307. [CrossRef]

Cappellari, Lorenzo, and Stephen P. Jenkins. 2004. Modelling low income transitions. *Journal of Applied Econometrics* 19: 593–610. [CrossRef]

Carter, Michael R., and Christopher B. Barrett. 2006. The economics of poverty traps and persistent poverty: An asset-based approach. *Journal of Development Studies* 42: 178–99. [CrossRef]

Carter, Michael R., and Julian May. 2001. One kind of freedom: Poverty dynamics in post-apartheid South Africa. *World Development* 29: 1987–2006. [CrossRef]

Cleaver, Kevin M., and Götz A. Schreiber. 1994. *Reversing the Spiral: The Population, Agriculture, and Environment Nexus in Sub-Saharan Africa*. Technical Report. Washington: The World Bank.

Cohen, Shaul E. 2006. Israel's West Bank barrier: An impediment to peace? *Geographical Review* 96: 682–95. [CrossRef]

Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. 2011. *Using Repeated Cross-Sections to Explore Movements into and out of Poverty.* Policy Research Working Paper 5550. Washington: World Bank.

Dang, Hai-Anh, and Peter F. Lanjouw. 2013. *Measuring Poverty Dynamics with Synthetic Panels Based on Cross-Sections.* Working Paper WPS6504. Washington: Development Research Group, Poverty and Inequality Team, The World Bank.

Dang, Hai-Anh, Peter Lanjouw, Jill Luoto, and David McKenzie. 2014. Using repeated cross-sections to explore movements into and out of poverty. *Journal of Development Economics* 107: 112–28. [CrossRef]

Deaton, Angus. 1985. Panel data from time series of cross-sections. *Journal of Econometrics* 30: 109–26. [CrossRef]

Dercon, Stefan. 1998. Wealth, risk and activity choice: Cattle in western Tanzania. *Journal of Development Economics* 55: 1–42. [CrossRef]

Dreze, Jean, and P. V. Srinivasan. 1997. Widowhood and poverty in rural india: Some inferences from household survey data. *Journal of Development Economics* 54: 217–34. [CrossRef]

Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. 2003. Micro-level estimation of poverty and inequality. *Econometrica* 71: 355–64. [CrossRef]

Leach, Melissa, Tim Forsyth, and Ian Scoones. 1998. *Poverty and Environment: Priorities for Research and Policy*. Technical Report. Prepared for the United Nations Development Programme and European Commission. Brighton: Institute of Development Studies.

Fuwa, Nobuhiko. 2007. Pathways out of rural poverty: A case study in socioeconomic mobility in the rural Philippines. *Cambridge Journal of Economics* 31: 123–44. [CrossRef]

Hareuveni, Eyal. 2012. *Arrested Development: The Long Term Impact of Israel's Seperation Barrier in the West Bank*. Technical Report. Jerusalem: BeTSELEM—The Israeli Information Center for Human Rights in the Occupied Territories. ISBN 978-965-7613-00-9.

Hassan, Zaha. 2005. Building walls and burning bridges: Legal obligations of the United States with respect to Israel's construction of the wall of separation in occupied Palestinian territory. *Willamette Journal of International Law and Dispute Resolution* 13: 197–244.

Ibañez, Ana María, and Andrés Moya. 2010. Do conflicts create poverty traps? Asset losses and recovery for displaced households in Colombia. In *The Economics of Crime: Lessons for and from Latin America*. Cambridge: National Bureau of Economic Research, Inc., pp. 137–72.

Jalan, Jyotsna, and Martin Ravallion. 2000. Is transient poverty different? Evidence from rural China. *Journal of Development Studies*, 36: 82–99. [CrossRef]

Jayne, Thomas S., Takashi Yamano, Michael T. Weber, David Tschirley, Rui Benfica, Antony Chapoto, and Ballard Zulu. 2003. Small holder income and land distribution in Africa: Implications for poverty reduction strategies. *Food Policy* 28: 253–75. [CrossRef]

Kattan, Victor. 2007. The legality of the West Bank wall: Israel's High Court of Justice v. the International Court of Justice. *Vanderbilt Journal of Transnational Law* 40: 1425–1522.

Lanjouw, Peter, and Martin Ravallion. 1995. Poverty and household size. *The Economic Journal* 105: 1415–34. [CrossRef]

Leuenberger, Christine. 2016. Maps as politics: Mapping the West Bank barrier. *Journal of Borderlands Studies* 31: 339–64. [CrossRef]

Lillard, Lee A., and Robert J. Willis. 1978. Dynamic aspects of earning mobility. *Econometrica* 46: 985–1012. [CrossRef]

Longo, Matthew, Daphna Canetti, and Nancy Hite-Rubin. 2014. A checkpoint effect? Evidence from a natural experiment on travel restrictions in the West Bank. *American Journal of Political Science* 58: 1006–23. [CrossRef]

Malone, Andrew R. 2004. Water now: The impact of Israel's security fence on Palestinian water rights and agriculture in the West Bank. *Case Western Reserve Journal of International Law* 36: 639–72.

PCBS. 2000. *Palestinian Labor Force Survey 1999*. Technical Report. Ramallah: Palestinian Central Bureau of Statistics.

PCBS. 2005. *Palestinian Labor Force Survey 2004*. Technical Report. Ramallah: Palestinian Central Bureau of Statistics.

PCBS. 2010. *Migration's Survey in the Palestinian Territory*. Technical Report. Ramallah: Palestinian Central Bureau of Statistics.

Reynolds, Kyra. 2015. Palestinian agriculture and the Israeli separation barrier: The mismatch of biopolitics and chronopolitics with the environment and human survival. *International Journal of Environmental Studies* 72: 237–55. [CrossRef]

Rodgers, Joan R., and John L. Rodgers. 1993. Chronic poverty in the United States. *The Journal of Human Resources* 28: 25–54. [CrossRef]

Roy, Sara. 2000. Palestinian society and economy: The continued denial of possibility. *Journal of Palestine Studies* 30: 5–20. [CrossRef]

Scherr, Sara J. 2000. A downward spiral? Research evidence on the relationship between poverty and natural resource degradation. *Food Policy* 25: 479–98. [CrossRef]

Usher, Graham. 2006. The wall and the dismemberment of Palestine. *Race and Class* 47: 9–30. [CrossRef]
Vallet, Élisabeth, and Charles-Philippe David. 2012a. Du retour des murs frontaliers en relations internationales. *Etudes Internationales* 43: 5–25. [CrossRef]
Vallet, Élisabeth, and Charles-Philippe David. 2012b. Introduction: The (re)building of the wall in international relations. *Journal of Borderlands Studies* 27: 111–19. [CrossRef]
Verbeek, Marno. 2008. Pseudo-panels and repeated cross-sections. In *The Econometrics of Panel Data*. Edited by Laszlo Matyas and Patrick Sevestre. Berlin/Heidelberg: Springer, pp. 369–83.
Verbeek, Marno, and Francis Vella. 2005. Estimating dynamic models from repeated cross-sections. *Journal of Econometrics* 127: 83–102. [CrossRef]
World Bank. 2008. *West Bank and Gaza: The Economic Effects of Restricted Access to Land in the West Bank*. Technical Report 47323. Washington: Social and Economic Development Group, Middle East and North Africa Region, The World Bank.
Zimmerman, Frederick J., and Michael R. Carter. 2003. Asset smoothing, consumption smoothing and the reproduction of inequality under risk and subsistence constraints. *Journal of Development Economics* 71: 233–60. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Income Inequality, Cohesiveness and Commonality in the Euro Area: A Semi-Parametric Boundary-Free Analysis

Gordon Anderson [1], Maria Grazia Pittau [2], Roberto Zelli [2,*] and Jasmin Thomas [1]

[1] Department of Economics, University of Toronto, Toronto, ON M5S, Canada; anderson@chass.utoronto.ca (G.A.); jasmin.thomas@mail.utoronto.ca (J.T.)
[2] Department of Statistical Sciences, Sapienza University of Rome, 00185 Roma RM, Italy; grazia.pittau@uniroma1.it
* Correspondence: roberto.zelli@uniroma1.it; Tel.: +39-06-49910782

Received: 13 December 2017; Accepted: 8 March 2018; Published: 21 March 2018

Abstract: The cohesiveness of constituent nations in a confederation such as the Eurozone depends on their equally shared experiences. In terms of household incomes, commonality of distribution across those constituent nations with that of the Eurozone as an entity in itself is of the essence. Generally, income classification has proceeded by employing "hard", somewhat arbitrary and contentious boundaries. Here, in an analysis of Eurozone household income distributions over the period 2006–2015, mixture distribution techniques are used to determine the number and size of groups or classes endogenously without resort to such hard boundaries. In so doing, some new indices of polarization, segmentation and commonality of distribution are developed in the context of a decomposition of the Gini coefficient and the roles of, and relationships between, these groups in societal income inequality, poverty, polarization and societal segmentation are examined. What emerges for the Eurozone as an entity is a four-class, increasingly unequal polarizing structure with income growth in all four classes. With regard to individual constituent nation class membership, some advanced, some fell back, with most exhibiting significant polarizing behaviour. However, in the face of increasing overall Eurozone inequality, constituent nations were becoming increasingly similar in distribution, which can be construed as characteristic of a more cohesive society.

Keywords: income distribution; inequality; mixtures; Gini; EU-SILC

JEL Classification: C14; D31; I32

1. Introduction

As Milanovic (2011) observes, growing inequalities between states in federations such as the Eurozone can be seen as a catalyst for the deterioration of social cohesion and support for the Union's institutions amongst its citizens. Thus, measurements of aspects of wellbeing of the Eurozone as an entity in itself and of its constituent nations are regarded as basic information for evaluating the progress of the Eurozone toward greater social cohesion within and between its various constituencies (Brandolini 2007) and such measurements have become important in core European institutional documents and debates (Filauro 2017).

The sense of cohesion amongst constituent nations hinges on notions of belonging, an absence of alienation from each other, or, in the presence of such alienation, that the process going forward is toward a less alienated state (OECD 2011). This has much to do with concepts of inequality, segmentation and polarization within and between groups and the sense in which they are directionally dynamic processes. When constituent nations are equally unequal with relatively similar income

levels, there is a commonality of situation which promotes cohesiveness, whereas, when such inequality and income levels are not equally shared, the situation is somewhat more divisive and alienating. Since notions of what constitutes poorness and wellness in income terms may vary across constituencies, these possibilities are perhaps best understood in the context of general income groupings or classifications in the Eurozone. Here, in analyzing the class structure anatomy of an overall euro area distribution and the relationship of its components to its constituent nations, these issue are addressed in terms of household incomes in the euro area and 18 of its constituent nations for the period 2006–2015.

For the longest time in income classification and measurement, income grouping have been identified by employing "hard", somewhat arbitrary and contentious income cut-offs or boundaries (see, for example, Anderson 2010; Atkinson and Brandolini 2013; Ravallion 2010, 2012). The main problem being that analysis is overly influenced by boundary choice, especially when intertemporal comparisons are being made.[1] Aside from measurement error or data contamination issues (see Deaton 2010), categorizing poorness and wellness in such an arbitrary fashion can prejudice other aspects of analysis. For example, defining classes by quantiles fixes class sizes over time precluding analysis of poverty reduction strategies. Tying class boundaries to some proportion of a location measure ties movement of classes to movements in the overall distribution and assumes away the possibility of independent class variation (incidentally contravening the focus axiom frequently invoked in poverty analysis).

Here, by employing mixture distribution techniques in a general euro area distribution, the number and size of groups or classes is determined by the commonalities in their income patterns and processes without resort to such hard boundaries.[2] This facilitates analysis of individual nation membership of income groupings and the progress of those nations through the overarching Eurozone income class structure. In so doing, some new indices of polarization and segmentation are developed in the context of a decomposition of the Gini coefficient and the roles of, and relationships between, these groups in societal income inequality, poverty, polarization and societal segmentation are examined. In addition, since it may be prudent to work with a simple ordinal classification that does not impute cardinal measure to wellbeing, a measure, the Utopia Index, that provides a complete cardinal ordering of wellbeing, although the basis of comparison (class membership) is only an ordinal classification, has been developed and implemented. These ideas are applied to an analysis of the Eurozone income distribution over the decade spanning 2006–2015. Implications for the individual constituent nations of the collective are explored. What emerged in the generic Eurozone distribution is a four-class, increasingly unequal polarizing structure with income growth in all four classes. With regard to individual nation results over the sample period, six nations are seen to be progressing through the class structure with twelve regressing. In terms of class transitions, thirteen nations are seen to exhibit polarizing patterns with four exhibiting converging behavior over the period. With regard to the Utopia Index, nation rankings appeared to be fairly stable over time exceptions were Finland and France who made significant advances and Greece which declined. In the following, Section 2 discusses the algebraic relationship between income classes, the Gini coefficient and measures of inequality, polarization and segmentation of subgroups. In Section 3 income data for the European countries are presented and discussed. Section 4 outlines the details of mixture distribution estimation and reports the main empirical results of the empirical analysis. Conclusions are drawn in Section 5.

---

[1] Witness The World Bank, 2017 GNI per capita ($ US equivalent) thresholds used for classifying nation income status. These were established in 1989—based upon previously established operational criteria–and inflation updated each year, or the United Nations $1 a day or the subsequent changes in the United Nations Development goals $1 a day poverty measure.

[2] Mixture distributions have also been used to deal with measurement error/data contamination problems, see Alvarez-Esteban et al. (2016). On the usefulness of mixture models for distributional analysis, see Cowell and Flachaire (2015).

2. Income Classes and the Gini Coefficient: Inequality, Polarization and Segmentation of Subgroups

2.1. Mixture of Distribution to Identify Income Classes

In each year, income data are interpreted as a sample from a mixture of K components in unknown proportions $w_1, \cdots, w_K$. Each component represents the income distribution of a homogeneous group of households, that is a household belonging to group k faces income opportunities described by the distribution f_k. Given some assumptions regarding the nature of the f_ks, these components can be specified to belong to some parametric family (normality or log normality are popular specifications that can be theoretically rationalized).

If the components are assumed to belong to the normal family, the mixture density can be written as:[3]

$$f\left(x;\mathbf{\Psi}\right)=\sum_{k=1}^{K}w_k f_k\left(x;\mu_k,\sigma_k^2\right), \tag{1}$$

where $f_k\left(x;\mu_k,\sigma_k\right)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$ and w_k, the mixing weight, represent the proportion of the population in class k. The vector $\mathbf{\Psi}=\left(w_1,\cdots,w_{K-1},\boldsymbol{\xi}'\right)'$ contains all the unknown parameters of the mixture model; in this case, $\boldsymbol{\xi}=\left(\mu_1,\cdots,\mu_K,\sigma_1^2,\cdots,\sigma_K^2\right)$.

Estimation of the triple w_k,μ_k,σ_k^2 for $k=1,\cdots,K$ yields much information about the structure of the society. For convenience and without loss of generality, suppose the types are ordered $(k>j\Rightarrow\mu_k\geq\mu_j)$ then μ_1 corresponds to the average income of the poor type and w_1 represents the proportion of the poor type in society, i.e., the relative poverty rate. This does mean that the class membership of a household with income x cannot be determined with certainty. However, such an analysis has many advantages, e.g., classes are determined without resort to arbitrary chosen boundaries, hence they are allowed to vary independently over time in terms of their size, location and scale.

Knowledge of $f_k(x|\mu_k,\sigma_k^2)$ also facilitates within class inequality measurement and between class polarization and segmentation measurement facilitating the study of such concepts in the context of an overall distribution.

It is also of interest to see how the individual nation states that make up the community have fared in terms of the income classifications. This may be examined by generating class membership probabilities for each member state over the period. Once the parameters of the components and the values of the class shares are estimated, posterior or conditional probabilities τ_i that household i with income x^* is in the kth group can be computed since:

$$\tau_{ik}=Pr\left(x^*\in k'\text{th class}\right)=\frac{w_k f_k(x^*)}{\sum_{k=1}^{K}w_k f_k(x^*)},\ \text{for } k=1\cdots K. \tag{2}$$

Effectively, this provides K group membership indices for each agent in the population. Note that it is possible for the group distributions to overlap, that is for an household with income x^* to potentially be a member of more than one group. To the extent that these distributions do not overlap (perfect segmentation in the terminology of Yitzhaki 1994), knowing the household income will completely determine the household's group and all of the households in a group. To the extent that they do overlap, household income will only partially define its group membership in the sense that its probability of being in a particular group is all that can be obtained.

Given the estimated ex-post probabilities of each household i belonging to a specific class k, τ_{ik}, the unbiased probability of class membership of each constituent nation h, based upon the average probability of membership in a particular class in a given nation, can be derived as follows. For country

[3] These ideas are readily generalized to multidimensional environments (see Anderson et al. 2017a).

h (where $h = 1, \cdots, H$) with n_h observations x_i, $i = 1 \cdots, n_h$, the probability that country h is in class k is given by:

$$\theta_{hk} = \frac{1}{n_h} \sum_{i=1}^{n_h} \tau_{ik} \text{ for } k = 1 \cdots, K. \tag{3}$$

Alternatively, following a traditional "hard" classification, we can assign households to components according to their maximal conditional probability:

$$\mathbb{I}_{i,k} = \begin{cases} 1 & \text{if } \tau_{ik} = \max(\tau_{i1}, \cdots, \tau_{iK}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbb{I}_{i,k}$ is the indicator variable of the i-th household to belong to class k. Then, we can calculate the proportion of households resident in country h simply as:

$$\theta_{hk}^{(a)} = \frac{1}{n_h} \sum_{i=1}^{n_h} \mathbb{I}_{i,k} \text{ for } k = 1 \cdots, K. \tag{5}$$

2.2. The Gini Coefficient and Segmentation of Subgroups

An inherent problem with the Gini coefficient, highlighted in Bourguignon (1979), is that it is not generally subgroup decomposable. Following Mookherjee and Shorrocks (1982) and Anderson and Thomas (2017), the Gini coefficient may be written as the sum of three components as follows:[4]

$$\text{GINI} = \sum_{k=1}^{K} w_k^2 \frac{\mu_k}{\mu} G_k + \frac{1}{\mu} \sum_{k=2}^{K} \sum_{j=1}^{k} w_k w_j |\mu_k - \mu_j| + \text{NSF}, \tag{6}$$

where G_k is the Gini associated with the kth subgroup and $\sum_{k=1}^{K} w_k = 1$ so that $\mathrm{E}_{f(x)}(x) = \mu = \sum_{k=1}^{K} w_k \mu_k$ and NSF may be thought of as the "non-segmentation" factor. The NFS may be written as:

$$\text{NSF} = \frac{2}{\mu} \sum_{k=2}^{K} \sum_{j=1}^{k-1} w_k w_j \int_0^{\infty} f_k(y) \int_y^{\infty} f_j(x)(x - y) dx dy. \tag{7}$$

The Gini is thus a weighted sum of subgroup Ginis plus a weighted sum of subgroup "dominating mean differences" divided by the overall mean (in essence a between group Gini coefficient BGINI) plus a component which is the weighted sum of the extent to which there are individuals in lower group j who overlap with, i.e., have greater incomes than, individuals in upper group k weighted by the extent to which they have more. In essence, the Gini is a linear function of within and between group Gini coefficients plus a term measuring the extent to which subgroups are not segmented.

Considering NSF, first note that, when subgroups k and j are perfectly segmented (so that $f_k(x) = 0$ for all x such that $f_j(x) > 0$ and $f_j(x) = 0$ for all x such that $f_k(x) > 0$), the corresponding term in the component vanishes. In the particular case where this is true for all $j \neq k$, the Gini is sub-group decomposable (Mookherjee and Shorrocks 1982).

Noting that in general all three components of GINI are non-negative and that $0 \leq \text{NSF} \leq \text{GINI}$, then SI, a segmentation index, may be written as:

$$\text{SI} = 1 - \frac{\text{NSF}}{\text{GINI}}, \quad 0 \leq \text{SI} \leq 1. \tag{8}$$

[4] This decomposition is readily extended to the Absolute Gini (Hey and Lambert 1980; Weymark 2003) by multiplying these equations by the overall mean from whence it may be seen that the overall Absolute Gini is a weighted sum of subgroup Absolute Ginis, the between group Absolute Gini and the Absolute Non Segmentation factor. Results in Giles (2004) facilitate inference. Derivation of the decomposition in the context of continuous distributions is shown in Appendix A.1.

SI provides a measure of the degree to which constituent groups are segmented or do not overlap. The analysis can be done with respect to particular groups, so the extent to which the poor or the rich are segmented from the rest of society may be readily analyzed.

Consider a generic group $g \in (1, \cdots, K)$. The specific non-segmentation factor of group g can be obtained as:

$$\text{NSF}_g = \frac{2}{\mu} \sum_{\substack{k=1 \\ k \neq g}}^{K} w_k w_g \int_0^\infty f_g(y) \int_y^\infty f_k(x)(x-y)dxdy. \tag{9}$$

This is twice a weighted sum of the (expected) average value of the excess of incomes of people in group g over those of people in the other groups normalized by average income which is of interest in contemplating the "isolation" of the group.

Clearly, NSF_g could be inserted in place of NSF in (8) to obtain an index of the segmentation of the specific group.

2.3. Comparing Constituent Distributions: Polarization, Transvariation and Utopia-Dystopia Index

2.3.1. Polarization

Conceptually, polarization is based upon notions of between group alienation and within group association. Duclos et al. (2004) captured this in an axiomatically developed general polarization index covering many, possibly latent, groups which may be written as:

$$P_\alpha(f) = S \int_0^\infty f(x) \int_0^\infty f(y)^{1+\alpha} |y-x| dydx \tag{10}$$

Here, S is a standardizing factor and α is the polarization sensitivity factor confined to $[0.25, 1]$. Note that (10) is not unlike a Gini coefficient (if α is set to 0 and S is suitably chosen, (10) becomes the continuous distribution version of Gini). $P_\alpha(f)$ can be interpreted as the scaled expected value of all possible rectangles formed under the distribution with height $f(y)^{1+\alpha}$ and base $|y-x|$ where $f(y)$ reflects the association component (larger $f(y)$ reflects more association) and $|y-x|$ reflects the alienation factor.

The index was developed by contemplating "sliding" and "squeezing" translations of basic constituent densities which respectively increased distances between constituent groups or intensified concentration around group means.[5] Slides change the relative locations of groups reflected in BGINI and, under certain circumstances, reduce NSF (i.e., increase the chance of segmentation), squeezes on the other hand simply reduce NSF without affecting BGINI. Note that, while polarizing slides can be associated with increasing between group inequalities, polarizing squeezes cannot, so a sufficient condition for establishing polarization (convergence) between groups is a combination of increased (decreased) between group inequality, BGINI, and segmentation, SI, suggesting a Gini-based polarization index PG of the form:

$$PG = \left(\frac{\text{BGINI}}{\text{GINI}}\right)^\alpha \text{SI}^{1-\alpha} \text{ for } 0 < \alpha < 1. \tag{11}$$

A group-specific polarization index for each class can also be obtained as:

$$\text{Pol}_g = w_g^\alpha \left(1 - \frac{\text{NSF}_g}{\text{GINI}}\right)^{1-\alpha} \text{ for } 0 < \alpha < 1. \tag{12}$$

[5] In the context of mixture distributions, these ideas can be explored by considering the component distributions to be the basic densities.

2.3.2. Transvariation

When all subgroups are perfectly segmented, the strongest form of stochastic dominance between them prevails, and there will be a strict complete first order dominance relationship between all constituent groups consistent with any monotonic non decreasing well-being measure function of income. All such measures would be unambiguous. Perfect segmentation is a sufficient (but not necessary) condition for such an ordering. However, normality of constituent distributions precludes such segmentation (all distributions overlap somewhere on the real line).

It is therefore useful to introduce a j-th index of transvariation (Anderson et al. 2017b) able to capture the degree of overlapping between K continuous distributions:

$$TR_j = \int_0^\infty \left[max\left(F_1^j(x), F_2^j(x), \cdots, F_K^j(x)\right) - min\left(F_1^j(x), F_2^j(x), \cdots, F_K^j(x)\right)\right] dx \tag{13}$$

where $F^j(x) = \int_0^\infty F^{j-1}(z)dz$, with $F^0(x) = f(x)$ and $j = 0, 1, 2, \cdots$.

Zero order transvariation ($j = 0$) is the many distribution version of Gini's classic two distribution transvariation (Gini 1916; Pittau and Zelli 2017). Under perfect segmentation, $TR_{(j=0)} = K$ and, when all distributions are identical, $TR_{(j=0)} = 0$ so that $TR_{(j=0)}/K$ provides a good index of the degree of segmentation.

2.3.3. Utopia-Dystopia

When the K classes C_k, $k = 1, \cdots, K$ have only an ordinal ranking (so that $C_i \preceq C_j$ for $i < j$ where the operator $\preceq$ indicates C_j is at least as good as C_i), a countries relative wellbeing measure can still be obtained from the discrete cumulative density functions $F_h(C_k)$, $k = 1, \cdots, K$, for countries $h = 1, \cdots, H$ across the K latent classes since:

$$\begin{aligned} &F_{h_m}(C_k) \leq F_{h_n}(C_k), \text{ for all } k = 1, \cdots, K-1 \text{ with strict inequality somewhere} \\ &\Rightarrow \mathbb{E}_{h_m}(C) \geq \mathbb{E}_{h_n}(C). \end{aligned}$$

That is to say, when country m's class membership density first order dominates that of country n, m will have a higher expected class membership than country n.[6] Inference for the comparisons can be conducted using the maximum modulus distributions for multiple simultaneous comparisons (Stoline and Ury 1979).

A relative Utopia-Dystopia measure for country h, $UI(h)$ $h = 1, \cdots, H$ (Anderson and Leo 2017; Anderson et al. 2017c) can be developed as:

$$UI(h) = \frac{\sum_{k=1}^{K} \left(\max_h F_h(C_k) - F_h(C_k)\right)}{\sum_{k=1}^{K} \left(\max_h F_h(C_k) - \min_h F_h(C_k)\right)}, \tag{14}$$

where $\max_h F_h(C_k)(\min_h F_h(C_k))$ corresponds to the maximum (minimum) value of $F_h(C_k)$ over all h. The index can be shown to reside in [0, 1] with 1 representing unequivocal "Utopia" (the best of all nations in that its distribution first order dominates all others) and 0 corresponding to an unequivocal "Dystopia" (such a nation's distribution is first order dominated by all others) and have many desirable properties of a wellbeing index.

3. Data Issues

Monitoring income inequality as well as other indicators related to personal income distribution within European countries relies on comparable and internationally harmonized estimates for the

[6] Note that only First Order Dominance comparisons can be made here since the ordering is not endowed with cardinal measure.

member states. The European Union Survey on Income and Living Conditions (EU-SILC) is the harmonized household-level survey that is widely used for these purposes (see, e.g., Longford 2014). The cross-sectional component of EU-SILC is a collection of annual national surveys of socio-economic conditions of individuals and households in the EU countries. All national surveys in EU-SILC have standard questionnaires and procedures for data processing and yield ex-ante harmonized micro-data that allow homogeneous inter-country comparisons using a uniform protocol. The EU-SILC project is carried out under European Union legislation (council regulation No. 1177/2003) and it was formally launched in 2004 for the EU15. In 2006 EU-SILC covered the EU25 Member States as well as Norway and Iceland.

To analyze the evolution of the Euro area income distribution over time, four temporally equi-spaced waves, namely 2006, 2009, 2012 and 2015, were chosen and the Euro area defined as those countries that are currently using the euro. Since data for Malta are only available from the 2008 wave, this country is excluded from analysis leaving 18 Euro zone countries.[7]

The income reference period refers to the previous year, consequently analysis with EU-SILC files actually refers to 2005–2014. The income is the total household net disposable income (variable HY020 in the SILC mnemonics), obtained by aggregation of all income sources from all household members net of direct taxes and social contributions. All observations are weighted by cross-sectional weights (variable DB090).

The EU-SILC income definition does not include capital gains, leading to a potential under-estimation of household income, especially top-incomes. Other sources of potential bias in the upper tail of the income distribution derived from EU-SILC are discussed in Törmälehto (2017). However, an increasing number of countries implementing EU-SILC combine interview-based data with register data on incomes. This strategy is expected to mitigate the well-documented low accuracy of sample surveys in estimating top-incomes.

Assuming cohabitation generates economies of scale in consumption and therefore needs do not grow proportionally with members, incomes are age and size-adjusted using the so called modified-OECD equivalence scale.This scale assigns a value of 1 to the household head, of 0.5 to each additional adult member aged 14 and over and of 0.3 to each child aged under 14.

Even if countries share the same currency, the question arises as to whether purchasing power parities should be used to compare incomes from different countries by eliminating the differences in price levels between them. Given significant disparities in the cost of living between countries, it adjust nominal incomes, the PPP index for the household final consumption expenditure is used.[8] Households whose income is less than zero were excluded from the sample. Thus, the final distribution considered is the real disposable size-adjusted income distribution of a weighted sample of households resident in the euro area.

4. Empirical Results

4.1. Number of Classes and Estimation of Mixture Parameters in the Community Income Distribution

We assume that the overall income distribution in the Eurozone can be described by a mixture of normal distributions. To ensure comparability of inequality measures of the distributions over time, the same origin at zero is taken by excluding negative incomes. Therefore, the component densities were taken to be truncated normal, with the number of components to be established. The assumption of normality may be too restrictive, since in principle any functional form can be taken into account. The choice of normality stems from a twofold motivation. Firstly, mixture of normal distributions form a much more general class. In fact, any absolutely continuous distribution can be approximated by

[7] Namely: Austria, Belgium, Cyprus, Estonia, Finland, France, Germany, Greece, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Portugal, Slovakia, Slovenia, and Spain.

[8] For a discussion on the use of PPPs in the EU income distribution see Brandolini (2007). For some recent results on the EU-wide and Eurozone income inequality using EU-SILC data, see Filauro (2017).

a finite mixture of normals with arbitrary precision (Marron and Wand 1992). Secondly, a mixture model of normals seems to capture better than other functional forms the idea of a polarized economy where relatively homogeneous groups of households are clustered around their expected incomes. The assumption of normality, in fact, results from additive shocks to the expected income of each strata.

The unknown mixture parameters (means, variances and proportions of each component) are estimated by maximum likelihood (ML) via the expectation-maximization (EM) algorithm (Dempster et al. 1977). Starting from a given number of components and an initial parameter $\Psi^{(0)}$, the first stage of the algorithm (E-step) is to assign to each data point its current posterior probabilities $\hat{\tau}_{ik}$ given by (2). The second stage (M-step), comprises estimation of sample means and variances of the normal densities and mixing weights $\hat{w}_k$ estimated as the means of the probabilities $\hat{\tau}_{ik}$. The estimates of the parameters are used to re-attribute a set of improved probabilities of group membership and the sequence of alternate E and M steps continues until a satisfactory degree of convergence occurs to the ML estimates. It is well known that the likelihood function of normal mixtures is unbounded and the global maximizer does not exist (McLachlan and Peel 2000). Therefore, the maximum likelihood estimator of Ψ should be the root of the likelihood equation corresponding to the largest of the local maxima located. The solution usually adopted is to apply a range of starting solutions for the iterations. The model was fitted repeatedly using a variety of initial values. Deterministic starting values based on separate models for the outcome based on K means (Kaufman and Rousseeuw 1990) were employed and the model fitted. The model was then fitted 10 additional times based on random jittering of the starting solution, and then another 10 times based on random jittering of the estimates at convergence of the previous runs. The results in the empirical application were fairly stable with respect to the starting solution in the sense that the same maximum for the likelihood or a value very close to it was invariably obtained.

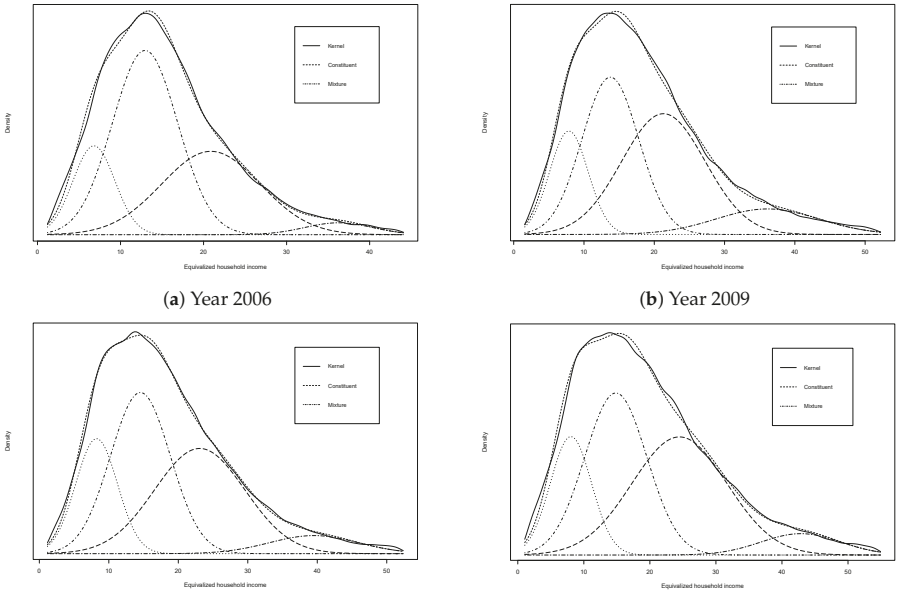
The number of components has been assessed by using the Bayesian's information criterion (BIC). Although regularity conditions do not hold for mixture models, Keribin (2000) showed that BIC is consistent for choosing the number of components in a mixture. In addition, we calculated the Akaike's information criterion (AIC), the consistent Akaike's information criterion (CAIC) and the AIC with a parameter penalty factor of three (AIC3), which is proved to perform well in a mixture context (Andrews and Currim 2003). Given sample size of between 141,000 and 154,000 observations per year, all the criteria yield similar results and picked a four or five-component mixture as the 'best' parsimonious model for all the years (see Table 1). Specifically, a four-component mixture is selected in the year 2006, while for the remaining years a five-component mixture seems to be preferable according to the criteria. However, the difference of the values taken by all the criteria between four and five component is marginal. Therefore, although the best fitting with five components, we decided to stay with four-components. In fact, adding a fifth component yields a negligible improvement in fit, leaving the first three components unchanged and splitting the fourth component into two classes. Moreover, the fifth component accounts for a very limited proportion of the whole population (from 0.2% to 0.7%) and it largely overlaps the fourth component due to its high variance. The four-component mixture instead is always characterized by distinct means, relatively modest dispersion and non-negligible size. There are no bizarre situations in the model fits such as clusters with very small variance or very flat components with large dispersion and very small probabilities. There is also an absence of components with similar means but different shape due to their disparate variances, etc., i.e., components that can play a role in improving the fit of the whole distribution but may be unacceptable in terms of economic interpretability. The four components can be interpreted as "low" (L), "lower-middle" (LM), "upper-middle" (UM) and "high" (H) income groups.

Figure 1 visually compares the fitted four component mixtures for all the years of the analysis with the corresponding estimated kernel density.[9]

[9] For the purpose of comparison, the variance of each component population was inflated by a factor of $1 + h^2/\sigma_i^2$ to match that of the kernel density, where h is the estimated bandwidth of the kernel.

Table 1. The choice of the number of components according to BIC, CAIC, Akaike and Akaike3.

N. of Components	Loglik	BIC	AIC	CAIC	AIC3
			2006		
1	−470,056	940,136	940,114	940,138	940,138
2	−466,189	932,437	932,383	932,442	932,393
3	−465,139	930,373	930,286	930,381	930,302
4	−464,693	929,516	929,397	929,527	929,419
5	−464,695	929,556	929,404	929,570	929,432
			2009		
1	−510,880	1,021,784	1,021,762	1,021,786	1,021,766
2	−500,152	1,000,363	1,000,309	1,000,368	1,000,319
3	−498,302	996,699	996,612	996,707	996,628
4	−497,977	996,084	995,965	996,095	995,987
5	−497,864	995,894	995,742	995,908	995,770
			2012		
1	−528,245	1,056,514	1,056,492	1,056,516	1,056,496
2	−517,724	1,035,507	1,035,453	1,035,512	1,035,463
3	−516,046	1,032187	1,032,100	1,032,195	1,032,116
4	−515,655	1,031,441	1,031,321	1,031,452	1,031,343
5	−515,636	1,031,438	1,031,286	1,031,452	1,031,314
			2015		
1	−562,189	1,124,402	1,124,380	1,124,404	1,124,384
2	−552,080	1,104,220	1,104,165	1,104,225	1,104,175
3	−550,175	1,100,445	1,100,358	1,100,453	1,100,374
4	−549,761	1,099,653	1,099,533	1,099,664	1,099,555
5	−549,701	1,099,569	1,099,416	1,099,583	1,099,444



(**a**) Year 2006 (**b**) Year 2009

(**c**) Year 2012 (**d**) Year 2015

Figure 1. Eurozone income distribution: Four component mixtures with the corresponding estimated kernel density for the years 2006–2015.

4.2. Inequality, Polarization and Segmentation in an Income Class Decomposition of the Eurozone Income Distribution

Table 2 reports the salient statistics compatible with the mixtures shown in Figure 1 for all years: the estimated mean (μ) and standard deviation (σ) of each truncated normal component along with its corresponding mixing proportion (w). As the mixing proportions (w) indicate, no single group overwhelms (no group ever accounts for more than half the population).

Table 2. Estimated parameters of the components of the mixtures.

Year	**2006**			**2009**			**2012**			**2015**		
Class	μ	σ	w	μ	σ	w	μ	σ	w	μ	σ	w
Low (L)	6.77	2.50	0.15	7.84	2.68	0.16	8.21	2.98	0.19	7.99	3.14	0.18
Lower-Middle (LM)	12.95	3.90	0.49	13.75	3.99	0.35	14.54	4.26	0.38	14.83	4.66	0.36
Upper-Middle (UM)	20.88	5.92	0.33	21.30	5.90	0.39	23.08	6.47	0.37	24.47	7.17	0.40
High (H)	36.27	4.01	0.03	36.11	7.29	0.10	39.52	6.36	0.06	42.85	5.98	0.06

Note: μ and σ are expressed in PPA-adjusted thousand Euros. w are the mixing proportions.

As can be seen mean, incomes have generally grown for all groups over the period with growth rates of 1.80% 1.45%, 1.72% and 1.80%, respectively, with a slight downturn for the lowest income group in 2015. Income variation has grown steadily over the period for the Low, Lower-Middle and Upper-Middle groups with a Kuznets curve-like inverted U shaped profile for the richest group. The Low and High income groups are the smallest in size with the former in the range of 15 to 19% of the population and the latter between 3% and 10% of the population. The size of the poorest group has grown over the period as has the size of the upper middle and high income groups, but the lower middle income group has diminished substantially, suggestive of some polarization in the community. To examine this, an analog of the Duclos et al. (2004) polarization measure for mixtures of K normal distributions[10] can be calculated by the the ERP index, an adaptation of Esteban and Ray's discrete version of (10) computed at distribution modes and their respective abscissa scaled by the respective relative subgroup size where:

$$ERP = \sum_{k=1}^{K} \sum_{j=1}^{K} |\mu_k - \mu_j| \left(\frac{w_k}{\sigma_k} + \left(\frac{w_j}{\sigma_j} \right)^{1+\alpha} \right) \frac{1}{\sqrt{2\pi}}. \tag{15}$$

Here, α corresponds to the polarization intensity parameter (chosen by the investigator) should be in the range (0.25, 1]. This corresponds to an aggregation over all possible pairs in many groups of the two group trapezoidal polarization measure proposed in Anderson (2010) and follows the interpretation of Duclos et al. (2004) polarization measures as the expected value of all possible trapezoids that can be formed from modal differences and their appropriately scaled abscissa. Here, it is offered as an alternative to the segmentation based polarization measure proposed in (11) above. Table 3 presents the statistics for a range of polarization intensity parameters and confirms, with a slight hiatus in 2012, that there appears to be ongoing income polarization throughout the period.

Table 3. Polarization coefficients.

α	**2006**	**2009**	**2012**	**2015**
0.25	1.070	1.049	0.976	1.133
0.5	0.869	0.871	0.796	0.943
1	0.746	0.768	0.693	0.838

[10] See Pittau et al. (2010) for a discussion on polarization measurements within a normal mixture framework.

Turning to the Gini decomposition analysis, Table 4 presents various subgroup Gini coefficients and Table 5 records the decomposition results.[11] As measured by the Gini coefficient, income inequality in the euro area had increased by the end of the period overall and in all income subgroups. It dipped in 2009 for all but the high income group (whose inequality peaked at this time) perhaps due to the economic exigencies of the time (recall 2009 refers to the year 2008, the year of the economic crash). Although the four groups appear to be well segmented (there appears to be little group overlap), the degree of segmentation has not changed much over the period. The within group inequality component diminished over the period while between group inequality increased. The overall segmentation-based polarization index coheres with the results in Table 3, but the decomposition results suggest that it is predominantly polarization of the poor and rich that is underlying these effects.

A better measure of inequality of distribution between the four classes that captures the lack of commonality in the sub-distributions is the zero order normalized transvariation (TR_0), as presented in Section 2.3. It was 0.1403 in 2006 and 0.1398, 0.1401, and 0.1454, respectively, in the subsequent observation years, suggesting very little change in the inequality of subgroup distributions over the 2006–2012 period but a substantial increase in subgroup distributional differences in 2015.

Table 4. Gini coefficients: overall and by subgroups.

	Overall	Low	Lower-Middle	Upper-Middle	High	Non-Lower	Non-High
2006	0.385	0.283	0.239	0.226	0.088	0.330	0.343
2009	0.400	0.267	0.231	0.221	0.158	0.342	0.335
2012	0.404	0.281	0.233	0.224	0.125	0.338	0.349
2015	0.421	0.298	0.249	0.233	0.104	0.352	0.362

Note: Non-Lower stands for all subgroups excluding the lowest. Non-High stands for all subgroups excluding the highest. Overall Gini coefficient standard errors (Giles 2004) are always of an order less than 0.002.

Table 5. Gini decomposition, segmentation and polarization of subgroups measured by the within and between components, the non-segmentation factor (NSF), the segmentation index (SI), the Gini-based polarization index (PG), the polarization index of the low income group (Pol_{poor}) and the polarization index of the high income group (Pol_{rich}). The latest are obtained fixing $\alpha = 0.5$.

Year	Within Gini	Between Gini	NSF	SI	PG	Pol_{poor}	Pol_{rich}
2006	0.084	0.202	0.099	0.743	0.624	0.370	0.172
2009	0.068	0.223	0.109	0.726	0.636	0.375	0.297
2012	0.072	0.223	0.110	0.729	0.634	0.408	0.236
2015	0.078	0.231	0.112	0.734	0.635	0.396	0.232

4.3. The Progress of Individual Constituent Nations

Table 6 presents the class membership probabilities for each country in each year of the analysis calculated according to Equation (3).[12] Estonia, Lithuania, Latvia and Slovakia all had over 50% membership of the Low income group at the beginning of the period whereas, of that group, only Latvia was in that position (along with Greece, who had well under 50% membership in 2006) at the end of the period. Indeed, the variability of nation poor group membership experience diminished substantially over period suggesting a greater sharing of poor group membership amongst the constituent nations. At the other end of the income spectrum, five nations (Austria, Germany, Ireland, Luxembourg and Netherlands) enjoyed over 10% of their population in the high income group,

[11] Inferential comparison of Gini coefficients was implemented using Giles's (2004) simple regression technique. As Modarres and Gastwirth (2006) and Davidson (2009) both indicate, Giles (2004) overstates the magnitude of the standard error so it can be considered an upper bound. Since it turns out to be very small relative to observed differences in the Gini coefficients rendering differences significant, further more sophisticated computations were deemed to be unwarranted.

[12] Similar results have been obtained adopting the alternative estimate of country membership as in Equation (5).

while Luxembourg's outstanding 26.7% membership of the high income group at the beginning of the period was somewhat diluted by the end where they were joined by France and Finland as the only members in the over 10% group.

Table 6. Membership probabilities for each country: years 2006–2015. Income groups are labeled Low (L), Lower-Middle (LM), Upper-Middle (UM) and High (H).

	2006				2009				2012				2015			
Nation	**L**	**LM**	**UM**	**H**	**L**	**LM**	**UM**	**H**	**L**	**LM**	**UM**	**H**	**L**	**LM**	**UM**	**H**
Austria	0.051	0.360	0.483	0.105	0.101	0.355	0.436	0.108	0.086	0.294	0.463	0.157	0.074	0.329	0.518	0.078
Belgium	0.077	0.413	0.433	0.077	0.139	0.395	0.394	0.072	0.118	0.365	0.426	0.092	0.100	0.386	0.463	0.052
Cyprus	0.113	0.400	0.394	0.093	0.157	0.360	0.375	0.108	0.143	0.355	0.375	0.126	0.223	0.410	0.329	0.038
Germany	0.069	0.375	0.452	0.105	0.115	0.357	0.421	0.107	0.099	0.311	0.447	0.143	0.107	0.355	0.467	0.071
Estonia	0.531	0.369	0.095	0.004	0.525	0.348	0.119	0.008	0.487	0.352	0.149	0.013	0.409	0.384	0.194	0.013
Greece	0.243	0.452	0.267	0.038	0.334	0.409	0.228	0.028	0.447	0.384	0.158	0.012	0.528	0.359	0.109	0.004
Spain	0.196	0.440	0.312	0.052	0.248	0.388	0.298	0.066	0.256	0.375	0.304	0.066	0.333	0.411	0.239	0.017
Finland	0.089	0.423	0.413	0.074	0.101	0.344	0.433	0.122	0.077	0.290	0.468	0.165	0.048	0.264	0.548	0.140
France	0.079	0.420	0.419	0.082	0.075	0.320	0.447	0.158	0.071	0.295	0.470	0.164	0.038	0.257	0.572	0.132
Ireland	0.109	0.424	0.365	0.102	0.160	0.392	0.355	0.094	0.179	0.388	0.350	0.083	0.173	0.404	0.377	0.046
Italy	0.126	0.427	0.376	0.071	0.181	0.387	0.351	0.081	0.160	0.351	0.388	0.102	0.171	0.397	0.387	0.044
Lithuania	0.579	0.333	0.085	0.003	0.545	0.326	0.118	0.011	0.533	0.329	0.128	0.009	0.492	0.358	0.141	0.009
Luxembourg	0.032	0.228	0.474	0.267	0.041	0.244	0.457	0.258	0.035	0.214	0.467	0.285	0.042	0.249	0.549	0.160
Latvia	0.604	0.311	0.080	0.005	0.587	0.297	0.105	0.011	0.594	0.291	0.106	0.009	0.537	0.332	0.126	0.006
Netherlands	0.024	0.336	0.523	0.118	0.043	0.320	0.500	0.137	0.049	0.314	0.502	0.135	0.057	0.361	0.522	0.060
Portugal	0.343	0.426	0.191	0.040	0.436	0.360	0.171	0.033	0.413	0.365	0.186	0.035	0.411	0.392	0.183	0.014
Slovenia	0.127	0.517	0.327	0.029	0.179	0.465	0.322	0.033	0.181	0.435	0.345	0.039	0.191	0.480	0.313	0.016
Slovakia	0.619	0.318	0.061	0.002	0.524	0.365	0.106	0.004	0.336	0.442	0.208	0.014	0.376	0.474	0.148	0.002

The extent to which income class structures vary across the constituent nations can be assessed by considering the discrete many distribution transvariation (TRMD), analogue of Gini's Transvariation of class membership distributions (Anderson et al. 2017b) across the 18 nations in Table 6. This statistic, a number between 0 and 1, measures the extent to which a collection of distributions differ. If all distributions are identical, it will take on the value 0, while, if nations are perfectly segmented or different in the extreme (each is in one class and no other and there is at least one nation in each class), the statistic will take on the value 1. Letting p_{hk} be an element of the 18×4 matrix P whose (h, k) element is the probability that nation h is in class k, $h = 1, \cdots, 18$, $k = 1, \cdots, 4$ and $P_{.k}$ corresponds to the kth column of P, then

$$\text{TRMD} = \frac{\sum_{k=1}^{4} \left(\max(P_{.k}) - \min(P_{.k})\right)}{4}.$$

The TRMD class membership distribution measures for the observation years from 2006 onwards were 0.403, 0.354, 0.365 and 0.338, respectively, indicative of some convergence over the period, i.e., nations were becoming more alike in their income class membership distributions.

To assess a nation's progress in terms of class membership, an ordinal comparison of income class structure can be performed.[13] In essence a first order dominance comparison over discrete ordered states is performed wherein there is no attribution of cardinal measure class differences, all that is asserted is that higher classes are preferred to lower classes.

Let $F_{h,2006}(C(k)) = \sum_{j=1}^{k} p_{hj,2006}$ and $F_{h,2015}(C(k)) = \sum_{j=1}^{k} p_{hj,2015}$ and consider the difference $D_h(C(k)) = F_{h,2006}(C(k)) - F_{h,2015}(C(k))$, for $k = 1, \cdots, K-1$ and $h = 1, \cdots, H$. If $D_h(C(k)) \leq 0$ for all k, then 2006 dominates 2015 marking an unequivocal downward transition or deterioration in income class. If $D_h(C(k)) \geq 0$ for all k, then 2015 dominates 2006 marking an unequivocal upward transition or improvement in income classes. The studentized maximum modulus distribution (Stoline and Ury 1979) indicates an asymptotic 0.01 critical value of 2.934 for the "t" statistic in three way multiple comparisons. Based upon this, Table 7 reports the First Order Dominance comparison of 2006 and 2015. Perhaps most clearly seen in the 2006-2015 comparison diagrams in the Appendix A.2

[13] This can be thought of as avoiding data contamination issues that would be present in ordinal comparison.

twelve countries[14] (Austria, Belgium, Cyprus, Germany, Greece, Spain, Ireland, Italy, Luxembourg, Netherland, Portugal and Slovenia) suffered a deterioration in their fortune in that 2006 stochastically dominated 2015, whereas the remaining six countries (Estonia, Finland, France, Lithuania, Latvia and Slovakia) enjoyed advances in their income classification fortunes wherein 2015 dominated 2006. It is of interest to note that with the exception of France and Finland all of these countries joined the union recently in 2004 whereas with the exception of Slovenia the countries that suffered a deterioration in class status all joined the union much earlier in the process and were relatively long time members.

Table 7. First order class comparisons (ordinal comparisons).

Country	$D\_h(C(1))$	$D\_h(C(2))$	$D\_h(C(3))$	t Values			
Austria	−0.023	0.008	−0.027	5.14	0.88	5.05	2006 dominates 2015
Belgium	−0.023	0.004	−0.026	4.36	0.43	5.71	2006 dominates 2015
Cyprus	−0.110	−0.120	−0.055	13.22	10.71	9.63	2006 dominates 2015
Germany	−0.038	−0.018	−0.033	10.75	2.90	9.41	2006 dominates 2015
Estonia	0.122	0.107	0.008	12.99	15.86	4.48	2015 dominates 2006
Greece	−0.285	−0.192	−0.034	39.81	28.49	12.98	2006 dominates 2015
Spain	−0.137	−0.108	−0.035	24.37	18.20	14.88	2006 dominates 2015
Finland	0.041	0.200	0.065	11.77	29.93	15.16	2015 dominates 2006
France	0.041	0.204	0.051	12.44	30.38	11.85	2015 dominates 2006
Ireland	−0.064	−0.044	−0.056	9.63	4.63	11.31	2006 dominates 2015
Italy	−0.045	−0.015	−0.026	12.23	2.93	10.91	2006 dominates 2015
Lithuania	0.087	0.062	0.006	8.43	9.28	3.76	2015 dominates 2006
Luxembourg	−0.010	−0.031	−0.106	2.16	2.83	10.65	2006 dominates 2015
Latvia	0.067	0.046	0.000	6.71	7.47	0.00	2015 dominates 2006
Netherlands	−0.033	−0.058	−0.057	11.46	8.06	13.54	2006 dominates 2015
Portugal	−0.068	−0.034	−0.026	7.56	4.39	7.98	2006 dominates 2015
Slovenia	−0.064	−0.027	−0.013	11.76	3.82	5.93	2006 dominates 2015
Slovakia	0.243	0.087	0.000	25.84	14.83	0.00	2015 dominates 2006

A nations propensity for intertemporal polarization or convergence can be assessed using a j period polarization statistic, $PS_h(j)$, which reflects the extent to which the nations incomes are moving from the center to the tails over j periods. Defining the probability of being in the tails of the distribution (i.e., in low or high income class) in period t as $P_{lh,t}$ and the probability of being in the middle of the distribution (i.e., in the lower or upper middle) in period t as $P_{mh,t} = (1 - P_{lh,t})$,

$$PS_h(j) = 0.5 + (P_{lh,t} - P_{mh,t}) - (P_{lh,t-j} - P_{mh,t-j}) = 0.5 + 2(P_{lh,t} - P_{lh,t-j}).$$

Based on the null hypothesis of no change in polarization, the polarization statistic would equal 0.5 and would be asymptotically $N(0.5, 0.25((1/n_t) + (1/n_{t-j})))$ where n is sample size. Values of $PS(j)$ greater than 0.5 imply polarization, while values less than 0.5 imply convergence. These are presented in Table 8. Over the 2006–2015 period, Estonia, Lithuania and Slovakia converged significantly (Latvia did not move significantly), whereas all other nations diverged or polarized significantly with Greece making the most striking movement.

Finally, Table 9 presents the results of the Utopia/Dystopia index for the euro area countries across years along with the corresponding ranking of the eighteen countries. As can be seen, rankings are fairly stable over time with Luxembourg enjoying Utopian status in two years and Latvia and Slovakia each suffering Dystopian status in one year. Significant movers are Greece, which dropped from 13th to 18th place over the period, and Finland and France who each displayed considerable improvement in ranking over the period.

[14] Note that for Austria and Belgium the second class component is estimated positive, however not significantly so, that is to say one could not reject the hypothesis that the component was negative, thus taken with the significant 1st and 3rd components one could not reject the joint hypothesis that 2006 dominates 2015 for these two countries.

Table 8. Nation polarization statistics and standard errors.

	2006–2009		2006–2012		2009–2012		2006–2015		2009–2015		2012–2015	
Country	$PS_h(j)$	**s.e.**	$PS_h(j)$	**s.e.**	$PS_h(j)$	**s.e.**	$PS_h(j)$	**s.e.**	$PS_h(j)$	**s.e.**	$PS_h(j)$	**s.e.**
Austria	0.605	0.009	0.673	0.009	0.568	0.009	0.492	0.009	0.387	0.009	0.319	0.009
Belgium	0.614	0.009	0.611	0.009	0.497	0.009	0.495	0.009	0.381	0.009	0.384	0.009
Cyprus	0.618	0.012	0.627	0.011	0.509	0.012	0.610	0.011	0.492	0.012	0.483	0.011
Germany	0.597	0.006	0.637	0.006	0.540	0.006	0.509	0.006	0.412	0.006	0.372	0.006
Estonia	0.495	0.010	0.428	0.010	0.433	0.010	0.273	0.010	0.278	0.010	0.345	0.010
Greece	0.663	0.009	0.855	0.010	0.692	0.009	1.002	0.008	0.839	0.007	0.647	0.008
Spain	0.632	0.006	0.647	0.006	0.515	0.006	0.704	0.007	0.572	0.006	0.557	0.006
Finland	0.619	0.007	0.657	0.007	0.538	0.007	0.549	0.007	0.430	0.007	0.392	0.007
France	0.644	0.007	0.648	0.007	0.504	0.007	0.519	0.007	0.375	0.007	0.371	0.007
Ireland	0.585	0.010	0.602	0.010	0.517	0.010	0.516	0.010	0.431	0.010	0.414	0.010
Italy	0.630	0.005	0.629	0.005	0.499	0.005	0.537	0.005	0.407	0.005	0.408	0.005
Lithuania	0.448	0.010	0.421	0.010	0.473	0.010	0.338	0.010	0.390	0.010	0.417	0.010
Luxembourg	0.501	0.012	0.542	0.011	0.541	0.010	0.307	0.012	0.306	0.012	0.265	0.011
Latvia	0.478	0.010	0.488	0.010	0.510	0.009	0.367	0.010	0.389	0.009	0.379	0.009
Netherlands	0.577	0.007	0.585	0.007	0.508	0.007	0.451	0.007	0.374	0.007	0.366	0.007
Portugal	0.672	0.011	0.631	0.010	0.459	0.010	0.584	0.009	0.412	0.009	0.453	0.008
Slovenia	0.613	0.007	0.628	0.007	0.515	0.007	0.602	0.007	0.489	0.008	0.474	0.008
Slovakia	0.315	0.010	−0.042	0.010	0.143	0.010	0.014	0.010	0.199	0.010	0.556	0.010

Table 9. Utopia Index and Rank for each country: years 2006–2015.

	2006		2009		2012		2015	
Country	**UI**	**Rank**	**UI**	**Rank**	**UI**	**Rank**	**UI**	**Rank**
Austria	0.78	3	0.73	5	0.79	5.00	0.82	4
Belgium	0.69	5	0.62	8	0.65	7.00	0.71	7
Cyprus	0.66	8	0.65	7	0.65	8.00	0.48	10
Germany	0.75	4	0.71	6	0.75	6.00	0.74	6
Estonia	0.08	15	0.06	15	0.11	16.00	0.19	13
Greece	0.43	13	0.30	13	0.14	15.00	0.01	18
Spain	0.50	12	0.47	12	0.44	12.00	0.29	12
Finland	0.67	7	0.75	4	0.81	4.00	0.96	3
France	0.69	6	0.83	3	0.81	2.00	0.98	2
Ireland	0.66	9	0.61	9	0.55	10.00	0.57	9
Italy	0.62	10	0.57	10	0.61	9.00	0.58	8
Lithuania	0.04	16	0.05	16	0.06	17.00	0.07	16
Luxembourg	0.99	1	1.00	1	1.00	1.00	1.00	1
Latvia	0.03	17	0.01	18	0.00	18.00	0.02	17
Netherlands	0.84	2	0.86	2	0.81	3.00	0.80	5
Portugal	0.31	14	0.19	14	0.21	14.00	0.18	14
Slovenia	0.53	11	0.49	11	0.48	11.00	0.46	11
Slovakia	0.00	18	0.04	17	0.25	13.00	0.16	15

5. Concluding Remarks

By employing mixture distribution techniques to determine the number and size of groups or classes in an income distribution, a four class model of the income distribution of the Eurozone countries over the decade spanning 2006–2015 has been developed without resort to hard class boundaries. Some new indices of polarization and segmentation are developed in the context of a decomposition of the Gini coefficient and the roles of, and relationships between, these groups in societal income inequality, poverty, polarization and societal segmentation are examined. Implications for the individual constituent nations of the collective are explored.

When viewed as an entity in itself, what emerged was a four-class, increasingly unequal polarizing structure with income growth in all four classes for the Eurozone. With regard to individual constituent nation class membership results over the sample period, six nations were seen to be advancing (Estonia, Finland, France, Lithuania, Latvia, and Slovenia) and twelve falling back (Austria, Belgium,

Cyprus, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Slovenia and Spain), all of whom, with the exception of Slovenia, were longer time members of the EU. In terms of an ordinally ranked class structure, with the exception of Estonia, Lithuania and Slovakia, who converged significantly, and Latvia, who did not move, all nations exhibited significant polarizing behavior, in the form of significantly divergent behaviour in class transitions over the period. Thus, in the face of increasing overall inequality in the Eurozone, indicated in Tables 2–4, and the within country increasing inequality/polarization, indicated in Table 8, the increasing commonality of class membership distributions across its constituent nations heralded by the across nation Transvariation statistics can be construed as characteristic of a more cohesive society, in essence constituent nations are becoming more alike through their individual increasing variation engendering increasingly overlapping income distributions between nations. Thus, while the Eurozone is experiencing increasing income class inequality and polarization, its constituent nations are experiencing increasing similarities in their respective income distribution.

While this paper has reached some conclusions on the anatomy of the income distribution in the Euro area, several opportunities for extending its scope remain. For example, the European Union Survey on Income and Living Conditions provides a rich set of socioeconomic covariates that can be used to model the class membership of the households, along the lines of Anderson et al. (2016). Furthermore, the possibility of using concomitant variables at different levels of aggregation—households nested within EU countries (Konte 2016)—can contribute to explain variation between and within countries for each income class.

Author Contributions: All authors contributed equally to the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Gini Segmentation

To derive the segmentation version of GINI in the context of continuous distributions note that:

$$\begin{aligned}
\text{GINI} &= \frac{1}{\text{E}(x)}\int_0^\infty\int_0^\infty f(y)f(x)|x-y|dx\,dy = \\
&= \frac{1}{\sum_{k=1}^K w_k\mu_k}\int_0^\infty\int_0^\infty \sum_{k=1}^K w_k f_k(y)\sum_{k=1}^K w_k f_k(x)|x-y|dx\,dy = \\
&= \frac{1}{\mu}\int_0^\infty\int_0^\infty\left[\sum_{k=1}^K w_k^2 f_k(y)f_k(x)|x-y| + \sum_{k=2}^K w_k f_k(y)\sum_{j=2}^{k-1} w_j f_j(x)|x-y|\right]dx\,dy = \\
&= \frac{1}{\mu}\int_0^\infty\int_0^\infty\left[\sum_{k=1}^K w_k^2\frac{\mu_k}{\mu_k}f_k(y)f_k(x)|x-y|\right]dxdy + \\
&+ \frac{1}{\mu}\int_0^\infty\int_0^\infty\left[\sum_{k=2}^K w_k f_k(y)\sum_{j=2}^{k-1} w_j f_j(x)|x-y|\right]dxdy = \\
&= \sum_{k=1}^K w_k^2\frac{\mu_k}{\mu}\text{GINI}_k + \frac{2}{\mu}\int_0^\infty\int_0^\infty\left[\sum_{k=2}^K w_k f_k(y)\sum_{j=2}^{k-1} w_j f_j(x)|x-y|\right]dxdy
\end{aligned}$$

where

$$\text{GINI}_k = \frac{1}{\mu_k}\int_0^\infty\int_0^\infty f_k(y)f_k(x)|x-y|dx\,dy.$$

From the second component, consider a typical term:

$$\begin{aligned}
&\int_0^\infty \int_0^\infty \left[w_k f_k(y) w_j f_j(x) |x-y| \right] dx, dy = \\
&= w_k w_j \int_0^\infty f_k(y) \int_0^\infty f_j(x) |x-y| dx, dy = \\
&= w_k w_j \int_0^\infty f_k(y) \left[-\int_0^y f_j(x)(x-y)dx + 2\int_y^\infty f_j(x)(x-y)dx \right] = \\
&= w_k w_j \int_0^\infty f_k(y) \left[-(\mu_j - y) + 2\int_y^\infty f_j(x)(x-y)dx \right] = \\
&= w_k w_j \left[|\mu_j - \mu_k| + 2\int_0^\infty f_k(y) \int_y^\infty f_j(x)(x-y)dxdy \right].
\end{aligned}$$

Note $\int_0^\infty f_k(y) \int_y^\infty f_j(x)(x-y)dxdy > 0$, so that:

$$\begin{aligned}
\text{GINI} = &\sum_{k=1}^{K} w_k^2 \frac{\mu_k}{\mu} G_k + \frac{1}{\mu} \sum_{k=2}^{K} \sum_{j=1}^{k} w_k w_j |\mu_k - \mu_j| + \\
&+ \frac{2}{\mu} \sum_{k=2}^{K} \sum_{j=1}^{k-1} w_k w_j \int_0^\infty f_k(y) \int_y^\infty f_j(x)(x-y)dxdy.
\end{aligned}$$

Appendix A.2. National Class Membership Cumulative Density

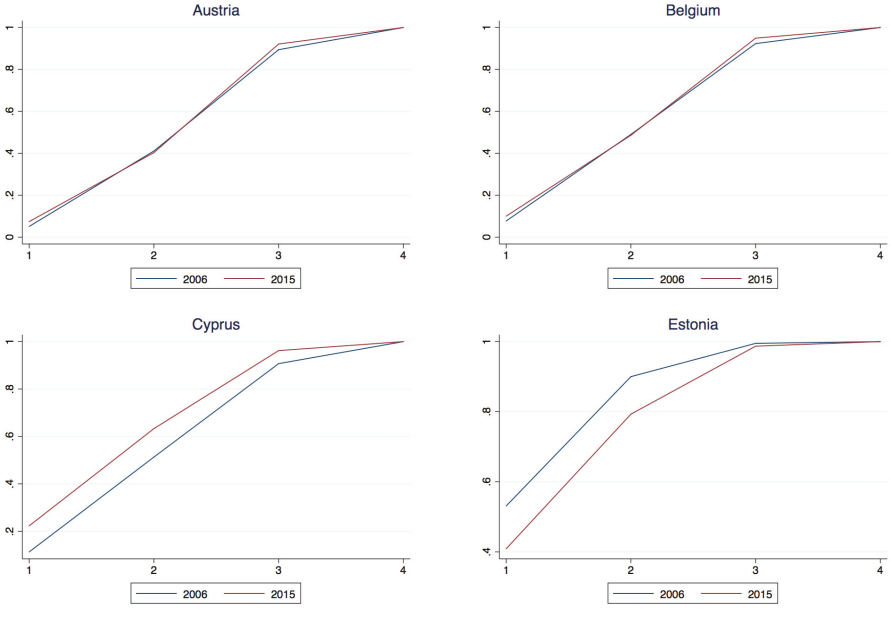


Figure A1. *Cont.*

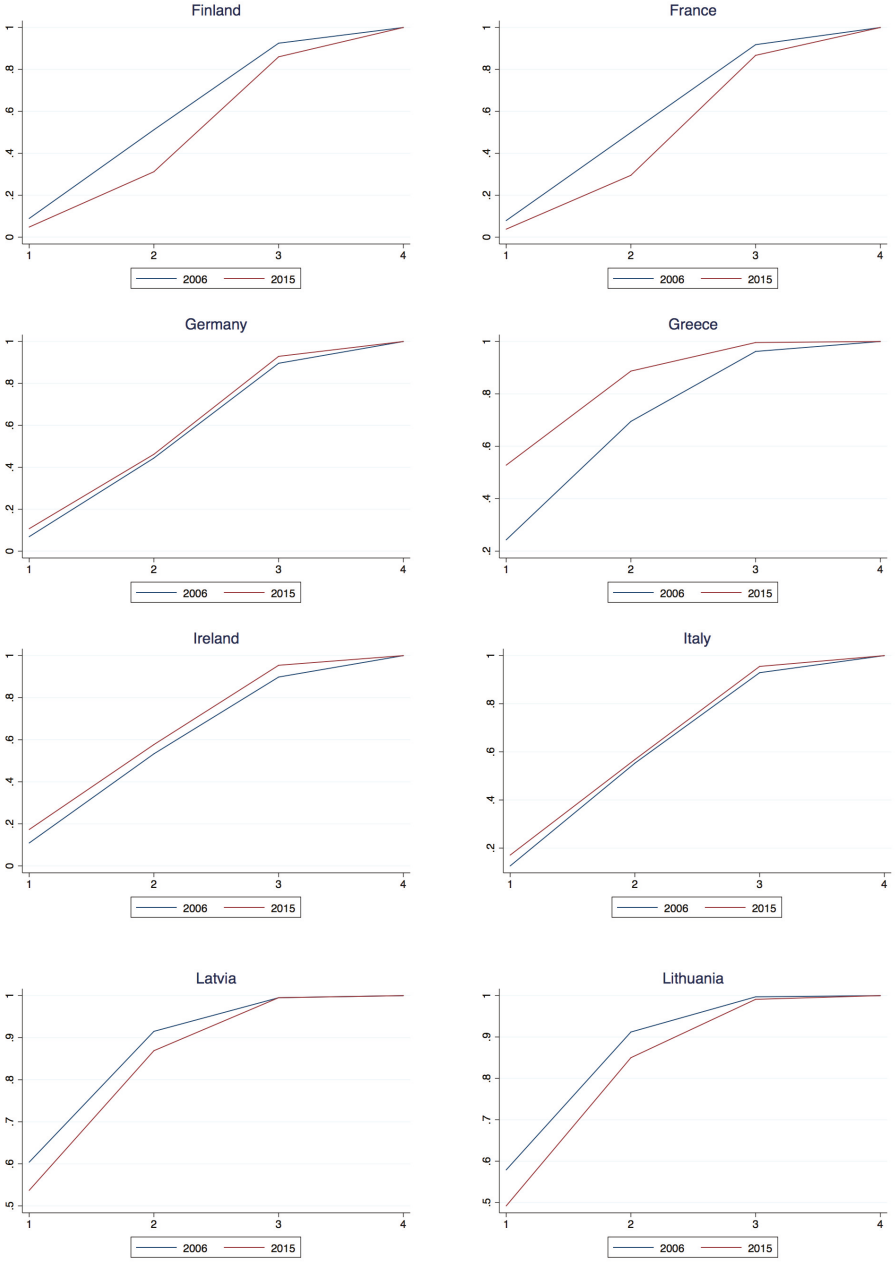


Figure A1. *Cont.*

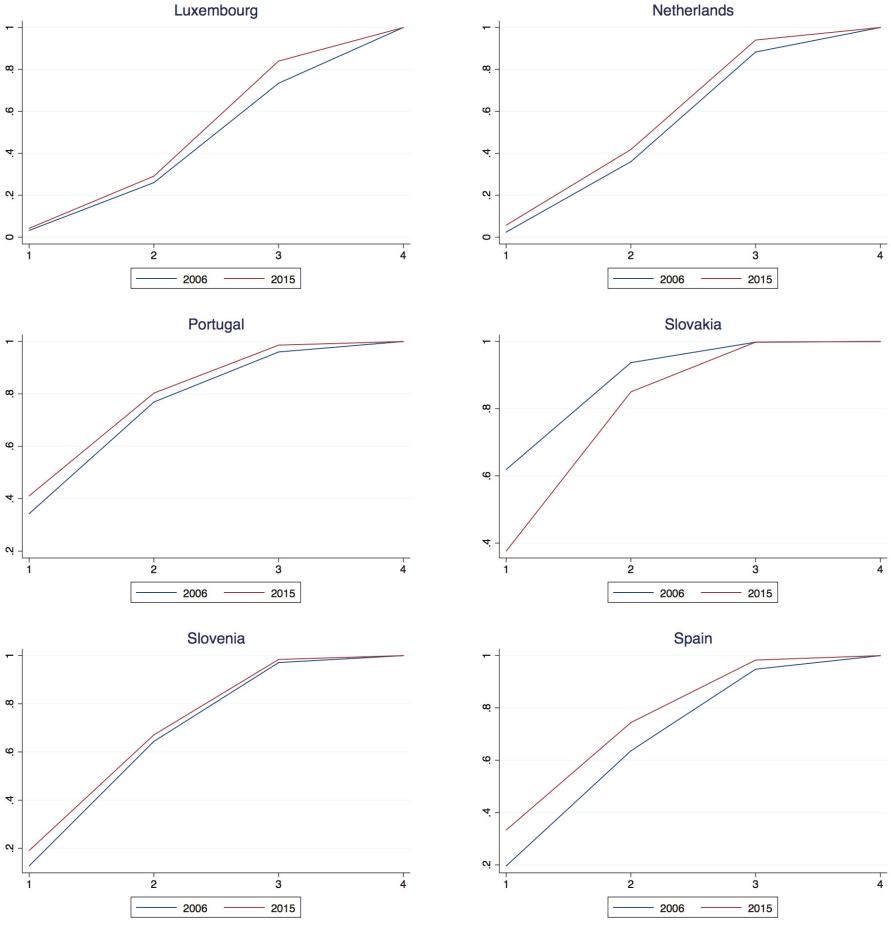


Figure A1. Class membership cumulative density, years 2006 and 2015.

References

Anderson, Gordon J. 2010. Polarization of the Poor: Multivariate Relative Poverty Measurement Sans Frontiers. *Review of Income and Wealth* 56: 84–101.

Anderson, Gordon J., Alessio Farcomeni, Maria Grazia Pittau, and Roberto Zelli. 2016. A new approach to measuring and studying the characteristics of class membership: The progress of poverty, inequality and polarization of income classes in urban China. *Journal of Econometrics* 191: 348–59.

Anderson, Gordon J., Alessio Farcomeni, Maria Grazia Pittau, and Roberto Zelli. 2017a. More equal yet less similar: Human development and the progress of nation wellbeing since 1990. A multidimensional mixture distribution analysis. Working paper. Mimeo, University of Toronto, Toronto, Canada.

Anderson, Gordon J., Oliver Linton, and Jasmin Thomas. 2017b. Similarity, dissimilarity and exceptionality: Generalizing Gini's transvariation to measure 'differentness' in many distributions. *Metron* 75: 161–80.

Anderson, Gordon J., and Teng Wah Leo. 2017. On Providing a Complete Ordering of Non-Combinable Alternative Prospects. Working paper. Mimeo, University of Toronto, Toronto, Canada.

Anderson, Gordon J., Thierry Post, and Yoon-Jae Whang. 2017c. Ranking Incomparable Prospects: The Utopia Index. Working paper. Mimeo, University of Toronto, Toronto, Canada.

Anderson, Gordon J., and Jasmin Thomas. 2017. More Unequal Yet Increasingly Similar Incomes: Polarization, Segmentation and Ambiguity in the Changing Anatomy of Constituent Canadian Income Distributions in the 21st Century. Working Paper 587, Department of Economics, University of Toronto, Toronto, Canada.

Andrews, Rick L., and Imran S. Currim. 2003. A comparison of segment retention criteria for finite mixture logit's models. *Journal of Marketing Research* 40: 235–43.

Atkinson, Anthony B., and Andrea Brandolini. 2013. On the identification of the middle class. In *Income Inequality*. Edited by Anet C. Gornick and Markus Jäntti. Stanford: Stanford University Press, pp. 77–100.

Álvarez-Esteban, Pedro C., Eustasio del Barrio, Juan A. Cuesta-Albertos, and Carlos Matrán. 2016. A contamination model for the stochastic order. *Test* 25: 751–74.

Bourguignon, François. 1979. Decomposable Income Inequality Measures. *Econometrica* 47: 901–20.

Brandolini, Andrea. 2007. Measurement of income distribution in supranational entities: The case of the European Union. In *Inequality and Poverty Re-Examined*. Edited by Stephen P. Jenkins and John Micklewright. Oxford: Oxford University Press, pp. 62–83.

Cowell, Frank A., and Emmanuel Flachaire. 2015. Statistical methods for distributional analysis. In *Handbook of Income Distribution*. Edited by Anthony B. Atkinson and François Bourguignon. North Holland: Elsevier, vol. 2A, chp. 6, pp. 359–465.

Davidson, Russell. 2009. Reliable inference for the Gini index. *Journal of Econometrics* 150: 30–40.

Deaton, Angus. 2010. Price Indexes, Inequality, and the Measurement of World Poverty. *American Economic Review* 100: 5–34.

Dempster, Arthur P., Nard M Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society* 69: 1–38.

Duclos, Jean-Yves, Joan Esteban, and Debraj Ray. 2004. Polarization: Concepts, Measurement, Estimation. *Econometrica* 72: 1737–72.

Filauro, Stefano. 2017. *European Incomes, National Advantages: EU-Wide Inequality and Its Decomposition by Country and Region*. EERI Research Papers Series No 05/2017. Economics and Econometrics Research Institute (EERI), Brussels, Belgium.

Giles, David E. A. 2004. Calculating a Standard Error for the Gini Coefficient: Some Further Results. *Oxford Bulletin of Economics and Statistics* 66: 425–33.

Gini, Corrado. 1916. Il concetto di transvariazione e le sue prime applicazioni. *Giornale degli Economisti e Rivista di Statistica* 52: 13–43.

Hey, John D., and Peter J. Lambert. 1980. Relative Deprivation and the Gini Coefficient: Comment. *Quarterly Journal of Economics* 95: 567–73.

Kaufman, Leonard, and Peter J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Keribin, Christine. 2000. Consistent estimation of the order of mixture model. *Sankhya* 62: 49–66.

Konte, Maty. 2016. The effects of remittances on support for democracy in Africa: Are remittances a curse or a blessing? *Journal of Comparative Economics* 44: 1002–22.

Longford, Nicholas T. 2014. *Statistical Studies of Income, Poverty and Inequality in Europe: Computing and Graphics in R Using EU-SILC*. Boca Raton: Chapman & Hall/CRC Press.

Marron, J. S., and M. P. Wand. 1992. Exact mean integrated squared error. *The Annals of Statistics* 20: 712–36.

McLachlan, Geoffret, and David Peel. 2000. *Finite Mixture Models*. New York: Wiley.

Milanovic, Branko. 2011. *The Haves and the Have-Nots: A Brief and Idiosyncratic History of Global Inequality*. New York: Basic Books.

Modarres, Reza, and Joseph L. Gastwirth. 2006. A cautionary note on estimating the standard error of the Gini index of inequality. *Oxford Bulletin of Economics and Statistics* 68: 385–90.

Mookherjee, Dilip, and Anthony Shorrocks. 1982. A decomposition analysis of the trend in UK income inequality. *Economic Journal* 92: 886–902.

OECD. 2011. *Perspectives on Global Development 2012: Social Cohesion in a Shifting World*. Paris: OECD Publishing.

Pittau, Maria Grazia, and Roberto Zelli. 2017. At the roots of Gini's transvariation: Extracts from 'Il concetto di transvariazione e le sue prime applicazioni'. *Metron* 75: 127–40.

Pittau, Maria Grazia, Roberto Zelli, and Paul A. Johnson. 2010. Mixture Models, Convergence Clubs and Polarization. *Review of Income and Wealth* 56: 102–22.

Ravallion, Martin. 2010. Mashup Indices of Development. Policy Research Working Paper. No. 5432. World Bank, Washington, DC, USA.

Ravallion, Martin. 2012. Why Don't We See Poverty Convergence? *Economic Review* 102: 504–23.

Stoline, Michael R., and Hans K. Ury. 1979. Tables of the Studentized Maximum Modulus Distribution and an Application to Multiple Comparisons among Means. *Technometrics* 21: 87–93.

Törmälehto, V. -M. 2017. High Income and Affluence: Evidence From the European Union Statistics on Income and Living Conditions (EU-SILC). Statistical Working Papers, Eurostat, Luxembourg.

Yitzhaki, Shlomo. 1994. Economic Distance and overlapping of distributions. *Journal of Econometrics* 61: 147–59.

Weymark, John. 2003. Generalized Gini Indices of Equality of Opportunity. *Journal of Economic Inequality* 1: 5–24.

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

econometrics

MDPI

Article

Foreign Workers and the Wage Distribution: What Does the Influence Function Reveal?

Chung Choe [1] and Philippe Van Kerm [2,*]

[1] Department of Economics, Hanyang University, 55 Hanyangdaehak-ro, Sangnok-gu, Ansan, Gyeonggi-do 426-791, Korea; choechung@gmail.com
[2] Luxembourg Institute of Socio-Economic Research and University of Luxembourg, Maison des Sciences Humaines, 11 Porte des Sciences, L-4366 Esch/Alzette, Luxembourg
* Correspondence: philippe.vankerm@liser.lu; Tel.: +352-585855-607

Received: 31 January 2018; Accepted:31 August 2018; Published: 7 September 2018

Abstract: This paper draws upon influence function regression methods to determine where foreign workers stand in the distribution of private sector wages in Luxembourg, and assess whether and how much their wages contribute to wage inequality. This is quantified by measuring the effect that a marginal increase in the proportion of foreign workers—foreign residents or cross-border workers—would have on selected quantiles and measures of inequality. Analysis of the 2006 Structure of Earnings Survey reveals that foreign workers have generally lower wages than natives and therefore tend to haul the overall wage distribution downwards. Yet, their influence on wage inequality reveals small and negative. All impacts are further muted when accounting for human capital and, especially, job characteristics. Not observing any large positive inequality contribution on the Luxembourg labour market is a striking result given the sheer size of the foreign workforce and its polarization at both ends of the skill distribution.

Keywords: immigrant wages; wage inequality; cross-border workers; influence function; RIF regression; Luxembourg

JEL Classification: J15; J31; J61

1. Introduction

While abundant research has documented the 'nativity wage gap'—the wage difference between foreign and native workers; see, among many others, Chiswick (1978); Borjas (1985, 1995); Adsera and Chiswick (2007)—much less evidence is available on how foreign workers actually 'fit in' and contribute to the shape of the wage distributions in host countries. The objective of this paper is to illustrate methods to examine the matter in question and describe how much foreign workers' wages contribute to private sector wage inequality in Luxembourg.

To be clear, our approach is descriptive or 'static': we are not attempting to identify causal or general equilibrium impacts of immigrants on the total wage distribution (or on native workers wages); see Blau and Kahn (2012); Card (2009) for reviews of this contentious literature.[1] Instead,

[1] The equilibrium impacts of immigration on the distribution of native workers wages remains a debated topic. They crucially depend on the degree of complementarity or substitutability between foreign and native labour—and this may vary by occupation and skill groups—so net impacts are not unambiguous. In the United States for instance, while Grossman (1982), Card (1990) or, more recently, Card (2009) and Ottaviano and Peri (2012) show that the impact of immigration on native wages is small or negligible, Borjas (1999, 2003) find that immigration lowers the wage of competing native workers. Manacorda et al. (2012) and Dustmann et al. (2013) show that the impact of immigration on wages in the UK is heterogeneous across the distribution: the overall effect on native wages is positive as a combination of a negative effect at lower percentiles

we document where foreign workers' earnings stand in the earnings distribution and quantify their contribution to wage inequality. As we explain shortly, we do so by calculating the effect that a notional marginal substitution of native workers for foreign workers would have on the shape of the earnings distribution—on wage inequality indicators in particular. Checking if substituting native workers for (observationally equivalent) foreign workers has any impact on distributional statistics turns out to be an indirect but informative way to apprehend how foreign workers' wages fit in and contribute to the shape of the overall wage distribution.

The analysis builds upon (recentered) influence function (RIF) regression methods proposed in Firpo et al. (2009) which capture how marginal changes in the distribution of covariates impact on distributive statistics of interest. These statistics have typically been quantiles in recent applications, but the methods can be applied to any statistic summarizing particular features of a distribution. We consider both a series of quantiles (to analyse the overall shape of the distribution) and dispersion measures, namely the variance, the Gini coefficient and three percentile ratios (to examine wage inequality). Two key advantages of the RIF regression approach over, say, conventional inequality decomposition methods (Shorrocks 1984) is that RIF regressions (i) apply generally to any conventional statistic of interest and (ii) allow us to assess the distributive impact of foreign workers both unconditionally and conditionally, that is, holding covariates (such as human capital and job characteristics) that may account for wage differentials between native and foreign workers constant, as we describe in more details in Section 2.[2]

With a share of foreign workers in total employment above 70 percent, Luxembourg is the European Union country relying most on foreign labour to fuel its domestic economic activity. These foreign workers are composed of both immigrants (that is, foreign workers residing in Luxembourg) and cross-border workers (that is, foreign workers residing in neighbouring countries—Belgium, France or Germany—but who are employed and work in Luxembourg). As per official statistics from Statec, in the first quarter of 2013, Luxembourg natives represented just 29.2 percent of total employment, immigrants represented 26.8 percent and cross-border workers represented 43.9 percent of total employment.[3] In the fourth quarter of 2006—the period covered by our analysis, see *supra*—the respective proportions were at 31.2 percent, 26.4 percent and 42.3 percent. This atypical situation arises from the small size of the domestic population, comparatively high wages and the supply of labour from neighbouring regions (see, e.g., Annaert 2004; OECD 2012).

Foreign labour in Luxembourg is largely European but is nonetheless heterogeneous in skill and human capital and polarized on both low-skill and high-skill positions (Amétépé and Hartmann-Hirsch 2011; Fusco et al. 2014). Such polarization is not uncommon. Similar concentration on both tails has been documented for example in Switzerland (Müller and Ramirez 2009) or the United Kingdom (Dustmann et al. 2013). Because of this concentration in both ends of the occupation ladder and their sheer number, foreign workers are typically perceived as pushing earnings dispersion upwards. This paper confronts such a claim to data derived from a matched employer-employee dataset, the 2006 Luxembourg Structure of Earnings survey. We separate out the potential contributions of immigrants and cross-border workers as they differ in their characteristics—the latter being generally younger, better educated and with more recent and weaker attachment to the Luxembourg labour market. Cross-border workers are also less strongly polarized in skills and occupations than immigrants, exhibit lower within-group wage inequality and may therefore be expected to have smaller influence on overall wage dispersion.

---

of the distribution but a positive effect at higher percentiles. No or small positive impacts have been identified in Spain and Israel (Carrasco et al. 2008; Friedberg 2001).

[2] See Van Kerm et al. (2017) for a study of the anatomy of wage differentials between natives and foreign workers using alternative approaches.

[3] See http://www.statistiques.public.lu/stat/TableViewer/tableView.aspx?ReportId=12916 (accessed 2018-08-30).

Our first baseline finding is that, with only a few exceptions, foreign workers tend to drive the wage distribution down: most quantiles of the wage distribution would be reduced by an increase in the share of foreign workers. This is a direct implication of the lower wages generally paid to foreign workers compared to natives (Van Kerm et al. 2017). However, we also find that their impact on wage inequality is most often insignificant or even negative. All effects are further muted when human capital and job characteristics are taken into account (that is, when one considers substitutions of native workers for foreign workers with similar characteristics).

Section 2 formally defines the parameters of interest and details the RIF regression methodology. Section 3 describes the data used in the analysis. Section 4 presents our results and Section 5 concludes.

2. Methods

Our analysis of the distributive footprint of foreign workers rests on a simple extension of the influence function regression methods developed in Firpo et al. (2009). We assess the contribution of foreign workers to the overall distribution of wages by measuring the impact that a (marginal) substitution of native workers by foreign workers would have on several distributional statistics (holding the wages of different types of workers unchanged). This impact depends on the locations of wage distributions of native and foreign workers relative to each other and reflect the implications of this configuration for different distributional statistics of interest.

2.1. Gâteaux Derivatives for Changes in Covariate Distributions

The central concept is the *directional derivative* of statistical functionals known as the Gâteaux derivative (Dugger and Lambert 2014; Gâteaux 1913). Let F be the cumulative distribution function of the random variable studied—here, wages in Luxembourg—and $v(F)$ denote a functional of interest, such as a mean, median, variance, or some more complex measure such as the Gini coefficient. Denote by G some alternative distribution function (which we will further specify below). The Gâteaux derivative of v at F *in the direction of* G captures how v responds to an infinitesimal modification of F obtained by mixing the two distributions

$$\nabla v_{F\to G} := \lim_{\epsilon \downarrow 0} \frac{v(H_\epsilon^{F,G}) - v(F)}{\epsilon} \tag{1}$$

where

$$H_\epsilon^{F,G}(y) = (1-\epsilon)F(y) + \epsilon G(y)$$

(see Hampel et al. 1986; Huber 1981).

To study the contribution of foreign workers, let us construct G as follows. First, write the distribution F as a combination of wage distributions for different workers types weighted by their respective employment shares:

$$F(y) = \sum_{x\in\Omega} s_x \, F_{Y|X=x}(y)$$

where Ω denotes a set of K workers types, s_x is the proportion of workers of type x in total employment, and $F_{Y|X=x}$ denotes the wage distribution among workers of type x. We will here distinguish workers by their nationality and country of residence and form a partition across three types $\Omega = \{\text{National resident}, \text{Foreign resident}, \text{Cross-border worker}\}$ (see below for detailed definitions). Let us now define a distribution obtained by increasing the share of type k workers and reducing the share of type r workers (the 'reference' type) by the same amount, but holding the conditional distributions $F_{Y|X=x}$ constant:

$$G_k^1(y) := (s_k + t)\, F_{Y|X=k}(y) \;+\; (s_r - t)\, F_{Y|X=r}(y) \;+\; \sum_{x\in\Omega\setminus\{k,r\}} s_x \, F_{Y|X=x}(y).$$

The distributions F and G_k^1 are obtained by 'swapping shares' and differ only in the relative proportions of workers k and r. With this definition, the Gâteaux derivative of $v(F)$ in the direction of G_k^1, $\nabla v_{F\to G_k^1}$, captures how v would respond to an exchange of type r workers for type k workers. We label this measure the *unconditional effect* (UE) of workers type k on v

$$\text{UE}(v(F),k) := \nabla v_{F\to G_k^1}.$$

The sign and magnitude of $\text{UE}(v(F),k)$ depend on differences in the conditional distributions of the different types of workers, their shares, and on the nature of v. For example, imagine type k workers were mainly found in both tails of the wage distribution (that is, if $F_{Y|X=k}$ had a bimodal distribution with modes at high and low wages) and type r workers were mainly located in the middle of the wage distribution. One would then expect $\text{UE}(v(F),k)$ for central tendency measures (such as the mean or the median) to be about zero, while $\text{UE}(v(F),k)$ for measures of inequality or dispersion would be positive since we would have more workers with wages in the tails of the distributions.

Our definition of G_k^1 is a special case of the more general form of counterfactual distribution G^* examined in Firpo et al. (2009) which is obtained by changing the (joint) distribution of (potentially many) conditioning variables from F_X to G_X but holding the conditional distributions of wages $F_{Y|X=x}$ constant,

$$G^*(y) := \int_{\Omega_X} F_{Y|x}(y)dG_X(x).$$

Firpo et al. (2009, Theorem 1) demonstrate that for such a G^*, the Gâteaux derivative is obtained by integrating the conditional expectation of the *influence function* of $v(F)$ with respect to the conditioning variables X,[4]

$$\nabla v_{F\to G^*} = \int_{\Omega_X} \text{E}\left[\text{IF}(y;v,F)|X=x\right]\, d(G_X^* - F_X)(x). \tag{2}$$

(The shape of the influence function $\text{IF}(y;v,F)$ is described in more detail below.) Firpo et al. (2009, Corollary 1) further show that in case distribution G_X^* is obtained by applying a 'location shift' to a continuous variable from X_j to $X_j + t$ then $\nabla v_{F\to G^*}$ is equal to the average partial derivative

$$\nabla v_{F\to G^*} = \int \frac{\partial \text{E}\left[\text{IF}(y;v,F)|X=x\right]}{\partial x} dF_X(x). \tag{3}$$

This motivates Firpo et al.'s (2009) application of (recentered) influence function regression to estimate $\nabla v_{F\to G^*}$ (see below).

A 'location shift' is inadequate for covariates following a multinomial distribution—such as our workers type variable—since the distribution of $X + t$ is undefined when X can only take on a set of fixed, discrete values. An analogous expression can be however derived for G^* based on 'shares swaps' rather than 'location shifts'. Consider first the case of a single covariate. The 'shares swap' described above to construct G_k^1 consists in exchanging a fixed fraction of type k data mass for type r data mass. We show in Appendix A that $\nabla v_{F\to G_k^1}$ based on such a 'shares swap' can be expressed as

$$\nabla v_{F\to G_k^1} \quad = \quad \left(\text{E}\left[\text{IF}(y;v,F)|X=k\right] - \text{E}\left[\text{IF}(y;v,F)|X=r\right]\right) \times t \tag{4}$$

where $\text{E}[\text{IF}(y;v,F)|X=x]$ is the expected value of $\text{IF}(y;v,F)$ among workers of type x. This result is a extension of Firpo et al. (2007, Corollary 3) to the case of multinomial variables and is a discrete data variant of Equation (3). Note however the presence of the scaling factor t. While Firpo et al. (2009) set $t = 1$ for the 'location shift' definition of G^*, this would imply implausible subgroup shares $(s_r - t) \leq 1$

[4] Please note that Theorem 1 in Firpo et al. (2009) integrates the *recentered* influence function defined as $\text{RIF}(y;v,F) = v(F) + \text{IF}(y;v,F)$. Our expression in terms of the influence function is equivalent since $v(F)$ in the recentered influence function expression of the theorem can be differenced away.

and $(s_k + t) \geq 1$ in our definition of G^*. We adopt instead $t = 0.10$ to keep counterfactual subgroup shares between 0 and 1.

Consider now the case of multiple covariates. One limitation of G_k^1 is that changing the shares of workers k and r may imply changing the distribution of other relevant characteristics that determine wages and are correlated with workers types. In this case, it would be unclear if $\text{UE}(v(F), k)$ captures the effect of changes in workers type directly or through the implied change in other characteristics. Imagine foreign workers are low-skilled and primarily work in low-paid occupations. The UE on, say, mean wages of an increase in the proportion of foreign workers is then likely negative, since it increases disproportionately low-paid workers overall. In the presence of multiple conditioning variables, there is interest in constructing a 'shares swap' which transforms the distribution of variable X but leaves the distribution of other observable conditioning variables Z unchanged, as Firpo et al. (2009) do with location shift counterfactuals. We therefore consider a substitution between native and immigrant workers done conditionally on a set of human capital and/or job characteristics Z. A proportion t of workers are assumed substituted within all configurations of Z so as to marginally change the proportion of foreign workers but preserve the overall distribution of workers' characteristics Z in the population. This is obtained by defining

$$\begin{aligned} G_k^2(y) \quad := \quad & \int_{\Omega_Z} \left((s_{k|Z=z} + t)\, F_{Y|X=k,Z=z}(y) \;+\; (s_{r|Z=z} - t)\, F_{Y|X=r,Z=z}(y) \right. \\ & \left. + \sum_{x \in \Omega \setminus \{k,r\}} s_{x|Z=z}\, F_{Y|X=x,Z=z}(y) \right) f_Z(z) dz \end{aligned}$$

and $F_{Y|X=x,Z=z}$ and $s_{x|Z=z}$ now denote respectively the conditional distribution of wage given worker type x and characteristics z and the share of workers of type x among workers with characteristics z. This substitution leaves the distribution of covariates Z unchanged. Appendix A shows that, by an immediate extension of Firpo et al. (2007, Corollary 3), $\nabla v_{F \to G_k^2}$ also takes the form of a discrete average partial effect

$$\begin{aligned} \text{UPE}(v(F), k) \quad := \quad & \nabla v_{F \to G_k^2} \\ = \quad & \left(\int_{\Omega_Z} \text{E}[\text{IF}(y; v, F) | X = k, Z = z] - \text{E}[\text{IF}(y; v, F) | X = r, Z = z]\; f_Z(z) dz \right) t. \end{aligned} \tag{5}$$

This defines the second quantity of interest that we examine in the paper and that we label the 'unconditional *partial* effect' (UPE) of workers type k on v, $\text{UPE}(v(F), k)$, as Firpo et al. (2009) do.[5]

2.2. Influence Functions

The influence function of a functional reflects its sensitivity to different areas of the distribution. Formally, the influence function is the Gâteaux derivative of F in the direction of a Dirac distribution Δ_y that has point mass at y:

$$\text{IF}(y; v, F) := \nabla v_{F \to \Delta_y} = \lim_{\epsilon \downarrow 0} \frac{v((1-\epsilon)F + \epsilon \Delta_y) - v(F)}{\epsilon}$$

and $\Delta_y(s) = 0$ if $s < y$ and 1 otherwise (Hampel 1974). Expressions for influence functions have been derived for most functionals commonly used in distributive analysis (see, e.g., Essama-Nssah and Lambert 2012).

[5] The unconditional partial effect is labelled a 'policy effect' in Rothe (2010) or a 'counterfactual effect' in Chernozhukov et al. (2013).

In our empirical application, we examine a variety of distributional functionals: (i) 19 ventiles are examined to show the general contribution of foreign workers to the overall wage distribution, and (ii) several dispersion measures are examined to capture their influence on wage inequality, namely the variance, the Gini coefficient, and the percentile ratios P90/P10, P90/P50 and P50/P10.[6] These functionals have well-known influence function. The IF for quantile τ is

$$\mathrm{IF}(y; q_\tau, F) = \frac{\tau - \mathbf{1}\left[y \leq q_\tau\right]}{f(q_\tau)}$$

where $f(q_\tau)$ is the density function at the quantile τ (Firpo et al. 2009). The IF for ratios of quantiles is obtained by applying the derivation rules described in Deville (1999):

$$\mathrm{IF}\left(y; \mathrm{R}(q_h, q_l), F\right) = \frac{1}{q_l}\left(\mathrm{IF}(y; q_h, F) - \frac{q_h}{q_l}\mathrm{IF}(y; q_l, F)\right)$$

where $\mathrm{R}(q_h, q_l) = \frac{q_h}{q_l}$ denotes the ratio of two quantiles, corresponding here to the percentiles ratios P90/P10, P90/P50 and P50/P10. Finally, the IF for the variance and the Gini coefficient are, respectively,

$$\mathrm{IF}(y; \mathrm{Var}, F) = \mathrm{Var}(F) + (y - \mu(F))^2$$

and

$$\mathrm{IF}(y; \mathrm{GINI}, F) = -\frac{\mu(F) + y}{\mu(F)}\mathrm{GINI}(F) + 1 - \frac{y}{\mu(F)} + \frac{2}{\mu(F)}\int_0^y F(x)dx$$

(see, e.g., Essama-Nssah and Lambert 2012).

2.3. Estimation by Influence Function Regression

$\mathrm{UE}(v(F), k)$ and $\mathrm{UPE}(v(F), k)$ are functions of conditional expectations of influence functions evaluated at different values of covariates. Assuming a linear and additive relationship between x, z and $\mathrm{IF}(y; v, F)$ leads to an estimator for UPE or UPE called the RIF-OLS estimator by Firpo et al. (2009):

$$\mathrm{E}[\mathrm{IF}(y; v, F)|X = x, Z = z] = \alpha + z\gamma + \sum_{g \in \Omega \setminus \{r\}} \beta_g \mathbf{1}\left[g = x\right] \tag{6}$$

where $\mathbf{1}\left[g = x\right]$ is a dummy for worker type x and z is a vector of potential additional covariates. Inserting Equation (6) in Equations (4) and (5) shows that, under this specification, $t\beta_k$ equals $\mathrm{UE}(v(F), k)$ in the absence of any additional covariates z_i and equals $\mathrm{UPE}(v(F), k)$ if covariates z_i are included in the model. (Notice that the dummy for natives ($\mathbf{1}\left[g = r\right]$) is taken as omitted category.). An influence function regression therefore provides a straightforward way to estimate UE and UPE.

A more flexible specification can allow for interactions between X and Z in the estimation of UPE and still provide straightforward estimation

$$\mathrm{E}[\mathrm{IF}(y; v, F)|X = x, Z = z] = \alpha + z\gamma + \sum_{g \in \Omega \setminus \{r\}} \left(\beta_g + z\gamma_g\right) \mathbf{1}\left[g = x\right]. \tag{7}$$

Inserting this specification in Equation (5) gives

$$\mathrm{UPE}(v(F), k) = t\left(\beta_k + \int_{\Omega_Z} z\gamma_k \, f_Z(z)dz\right), \tag{8}$$

[6] Estimates based on several other relative inequality measures were also examined (quantile group shares ratios, the standard deviation of log wage, generalized entropy measures) and lead to similar conclusions. They are not reported here but are available on request.

that is, UPE$(v(F), k)$ is the discrete partial effect of the workers type dummy variables on IF$(y; v, F)$ averaged over all additional covariates distributions.

Note that the dependent variable in the RIF-OLS regressions (IF$(y_i; v, F)$) is a function of F which is unknown but itself derived from one's sample. First, the value of IF$(y_i; v, \hat{F})$ is computed for all sample observations i for the statistic of interest v. Second, IF$(y_i; v, \hat{F})$ is regressed by OLS on worker type dummies x_i and, for UPE$(v(F), k)$, on additional covariates z_i. This two-stage procedure therefore results in complex sample dependence between observations which can be taken into account by resorting to bootstrap resampling for inference (Firpo et al. 2009).

3. Data

Our analysis exploits data from the 2006 Luxembourg Structure of Earnings Survey. The survey is collected in all European Union countries on the basis of common variable definitions and sampling design defined in European Community regulations. It aims to provide detailed information on earnings in the European Union. The Luxembourg SES is collected by STATEC—Institut national de la statistique et des études économiques, the national statistical institute.

The SES is a nationally representative matched employer-employee survey covering, in 2006, non-profit and private sector firms (NACE C–K and M–O) employing at least 10 workers. This sampling frame covers 79 percent of salaried workers in Luxembourg at the time of the survey (STATEC 2009).[7] The distinctive feature of the SES in the context of Luxembourg is that—since it is based on a sampling frame of firms—it collects information on both resident and cross-border workers.

The survey has a two-stage design. A sample of firms (stratified by firm size) was drawn in a first stage. A sample of workers from the selected firms was drawn in a second stage. In total, the 2006 Luxembourg SES dataset covers 1856 firms and 31,329 workers (STATEC 2009). Information is available on both employers (sector of activity, size, collective agreement coverage) and employees (earnings plus basic demographic information (including educational achievements) and occupation and job characteristics).

Table 1 describes the distribution of gross hourly wage in our sample for three distinct groups of workers: Luxembourg nationals, immigrants (foreign residents) and cross-border worker. Hourly wage is calculated as the earnings received in the month of reference of the survey (October 2006) divided by the number of paid hours worked in the month. We limit the sample to workers aged 18 to 65. Luxembourg workers have much higher wages than foreign workers. Mean wage of Luxembourg workers is 29% higher than mean wage of immigrant workers (€23.23 versus €18.02) and 30% higher than mean wage for cross-border workers (€23.23 versus €17.83). Median wages differ across the groups in even stronger proportions with Luxembourg workers having 46% and 36% higher median wage than immigrant workers and cross-border workers respectively. While cross-border workers and immigrant workers have similar mean and median wage, wage dispersion—whether measured by the Gini coefficient, the variance of wages or percentile ratios—is much higher among immigrant workers. In turn, native workers exhibit about the same degree of wage inequality as immigrant workers, but at much higher levels of wage. Differences in the distribution of wages across the three groups are illustrated further in Figure 1 which shows the density function of the overall wage distribution (on a logarithmic scale) along with the distributions for the three subgroups scaled by their employment shares. The three groups clearly exhibit different distributions of wage both in location and in spread. Given this complex configuration, the contribution of foreign workers to overall inequality is far from obvious and is difficult to deduce from subgroup summary indicators.

[7] Most noticeably, civil servants and agricultural sector workers are excluded from the sampling frame. These sectors employ only few foreign workers (in particular cross-border workers).

Table 1. Employment share and hourly wage distribution statistics by worker type (Luxembourg nationals, immigrants and cross-border workers).

	Luxembourg Nationals	Immigrant Workers	Cross-Border Workers
Employment share	0.25	0.27	0.49
Mean and selected percentiles (€)			
Mean	23.2	18.0	17.8
10th percentile (P10)	10.7	9.1	10.2
25th percentile (P25)	14.5	10.9	12.0
Median (P50)	20.3	13.9	14.9
75th percentile (P75)	27.9	20.1	20.4
90th percentile (P90)	37.0	31.6	28.8
Measures of dispersion and inequality			
Standard deviation	15.3	13.2	10.3
Gini coefficient	0.284	0.303	0.251
P90/P10 ratio	3.5	3.5	2.8
P50/P10 ratio	1.9	1.5	1.5
P90/P50 ratio	1.8	2.3	1.9

Notes: Based on the 2006 Luxembourg Structure of Earnings Survey. Sample weights applied.

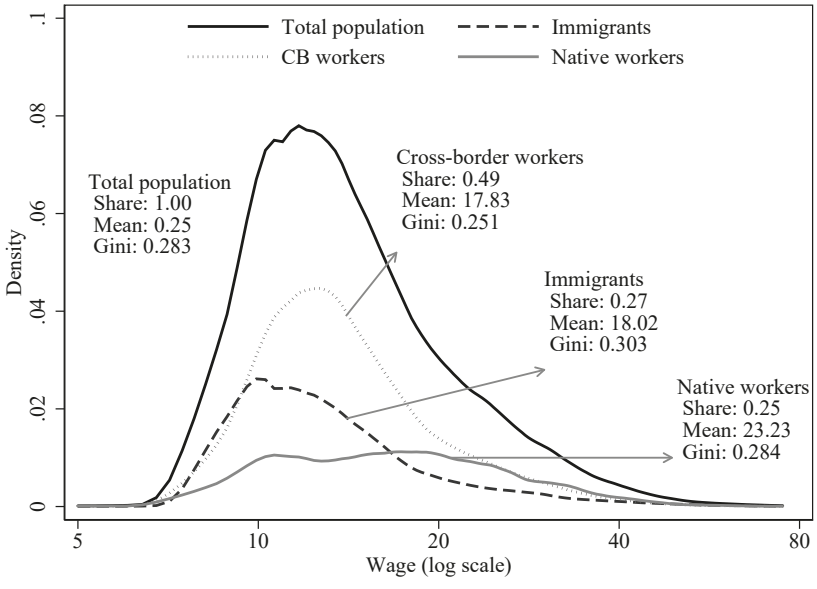


Figure 1. Hourly wage distribution: density function estimates for total population and by worker type. Note: Subgroup densities are scaled by subgroup employment share. Source: Structure of Earnings Survey for Luxembourg 2006.

Additionally, the three subgroups differ a lot in job and productivity-related characteristics; see Table 2. This partly accounts for the wage differences between natives and foreign workers. For instance, native workers are more likely to work in large firms, have much longer job tenure or more likely to hold supervisory positions. Immigrants are more likely to work in the construction sector while cross-border workers are more likely to work in the real estate sector. Worth noting is also the polarized distribution of educational achievements of immigrants—with a higher fraction of both primary and tertiary education workers. Cross-border workers generally have higher educational

achievements, but are also younger and have the lowest job tenure. Careful qualification of the contribution of foreign workers to wage inequality ought to account for these differences.

Table 2. Detailed nationality, human capital and job characteristics by worker type (Luxembourg nationals, immigrants and cross-border workers).

	Luxembourg Nationals	Immigrant Workers	Cross-Border Workers
Luxembourg	1.00	–	–
Belgian	–	0.08	0.22
French	–	0.13	0.50
German	–	0.05	0.22
Portuguese	–	0.47	0.01
Other EU	–	0.18	0.04
Non-EU	–	0.10	0.01
Female	0.39	0.38	0.32
Age	39.90	37.63	37.20
Primary educ. or less (ref.)	0.11	0.24	0.08
Secondary education	0.80	0.62	0.80
Tertiary education	0.08	0.14	0.12
Years at current employer	11.82	6.32	5.59
Manager	0.17	0.14	0.14
10–49 employees in firm	0.24	0.32	0.27
50–249 employees in firm	0.24	0.30	0.35
250–499 employees in firm (ref.)	0.11	0.13	0.14
500–999 employees in firm	0.08	0.11	0.11
1000+ employees in firm	0.33	0.14	0.13
Part time contract	0.18	0.15	0.13
Industry/Manufacture	0.17	0.10	0.18
Construction	0.05	0.21	0.14
Wholesale	0.12	0.10	0.13
Hotel/Restaurant	0.01	0.06	0.03
Trans/Comm	0.16	0.07	0.09
Finance	0.17	0.16	0.17
Real estate	0.08	0.18	0.19
Education, Health & Other not-for-profit (ref.)	0.24	0.11	0.08
Managerial	0.07	0.06	0.04
Professional	0.10	0.09	0.12
Associate professional	0.23	0.13	0.18
Clerk	0.23	0.11	0.15
Service worker	0.09	0.10	0.11
Craft and trade worker	0.13	0.21	0.20
Manufacturers	0.08	0.09	0.13
Low skilled and laborer (ref.)	0.08	0.20	0.07
Number of observations	7537	8367	15105

Notes: Based on the 2006 Luxembourg Structure of Earnings Survey. Sample weights applied.

4. Results

How do foreign workers contribute to the shape of the private sector wage distributions in Luxembourg?

4.1. Unconditional Impacts: UE Estimates

Figure 2 shows our first baseline results: the UEs on a set of 19 quantiles from the 5th percentile to the 95th percentile for both immigrants and cross-border workers.[8] The quantile UEs are negative for both groups of workers: given the configuration of each group's wage distribution relative to each other, a marginal increase in the share of foreign workers would haul the overall wage distribution downwards. The quantile UEs for immigrants are lower than for cross-border workers until the 70th percentile. This is a reflection of immigrants' comparatively lower wages and their relatively larger concentration in the bottom part of the wage distribution. Beyond the 70th percentile, the quantile UEs for cross-border workers continue to decline while they start increasing for immigrants. This finding brings empirical support to claims that immigrants in Luxembourg include both top earners that contribute to high wages as much as natives do (yet not *more*) and low skill migrants that drive bottom quantiles down (Amétépé and Hartmann-Hirsch 2011). The pattern does not hold true however for cross-border workers.

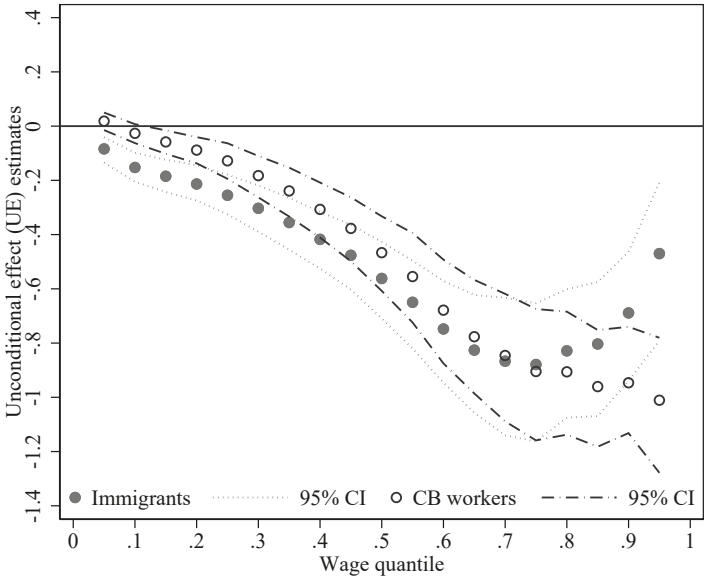


Figure 2. Estimates of unconditional effects (UE) on quantiles of the wage distribution for immigrant and cross-border workers.

What these patterns imply for wage dispersion and inequality can be read from estimates of UEs on inequality measures reported in Table 3 under Model 1. All estimates (but one) for both immigrant and cross-border workers are negative. The only exception is the contribution to the P90/P50 ratio of immigrant workers. Overall, wages of foreign workers tend to *reduce* inequality; moving towards a distribution composed of 10 percentage point less Luxembourg workers and 10 percentage points more foreign workers (holding the wage distributions within each group constant) would lead to a

[8] RIF regression calculations were done with the statistical software Stata (version 14.2) (StataCorp 2015) and the user-written package for Stata `rifreg` available from Nicole Fortin at http://faculty.arts.ubc.ca/nfortin/datahead.html. Bootstrap confidence intervals for the UE and UPE estimates were constructed on the basis of 1,000 replications from a repeated half-sample bootstrap resampling scheme (Saigo et al. 2001) and account for the two-stage design of the survey (see Section 3). We use the *rhsbsample* Stata user-written package for generating the replication weights (Van Kerm 2013). Pointwise confidence intervals are based on the bias-corrected percentile method (Efron 1981).

reduction of inequality. Magnitudes are very small however. They are close to zero and not statistically different thereof for immigrant workers. UE estimates are different from zero for cross-border workers, but they remain small. For example, using these estimates to approximate the change in the Gini suggests a reduction of the Gini of only 0.005 against a baseline Gini of 0.283.[9] This finding contrasts what could be conjectured from the polarization of foreign workers in both tails of the skill and pay distributions and the evidence on the quantile UEs from Figure 2. The wage distribution of foreign workers does not appear so polarized (compared to the distribution of native workers) as to drive inequality upwards.

The only exception to the negative UEs is the contribution of immigrant workers to the P90/P50 ratio. This ratio differs somewhat from the other indices in that it captures 'upper half inequality'. The rebound of the quantile UEs for immigrant workers at the top of the distribution shown in Figure 2 explains this exception. All quantiles would be dragged down by "moving towards" more immigrant workers but the decline is stronger at the median than at the top of the distribution. This suggests that a share of foreign workers are located at high wages and do contribute to upper-end inequality.

Table 3. Estimates of UEs (Model 1) and UPEs (Model 2 and Model 3) on the variance of hourly wages and indicators of relative inequality.

	UE	**UPE**			
	Model 1	**Model 2a**	**Model 2b**	**Model 3a**	**Model 3b**
			Variance		
Immigrant worker	−7.58	−6.97	−7.92	−1.91	−1.16
	(6.75)	(7.85)	(8.59)	(4.83)	(3.53)
Cross-border worker	−14.17	−13.52	−14.57	−6.76	−5.31
	(6.20)	(7.29)	(7.85)	(4.30)	(2.59)
			Gini coefficient		
Immigrant worker	−0.0003	−0.0004	−0.0007	0.0002	0.0006
	(0.0016)	(0.0017)	(0.0019)	(0.0012)	(0.0011)
Cross-border worker	−0.0054	−0.0053	−0.0053	−0.0030	−0.0021
	(0.0015)	(0.0015)	(0.0017)	(0.0010)	(0.0008)
			Percentile ratio P90/P10		
Immigrant worker	−0.022	−0.005	−0.009	0.003	0.010
	(0.014)	(0.013)	(0.013)	(0.011)	(0.012)
Cross-border worker	−0.091	−0.071	−0.068	−0.044	−0.026
	(0.013)	(0.011)	(0.011)	(0.010)	(0.010)
			Percentile ratio P50/P10		
Immigrant worker	−0.032	−0.023	−0.024	−0.012	−0.012
	(0.005)	(0.005)	(0.004)	(0.003)	(0.003)
Cross-border worker	−0.043	−0.033	−0.032	−0.022	−0.020
	(0.005)	(0.005)	(0.004)	(0.004)	(0.003)

[9] Formally, the approximation is the leading term of a von Mises expansion of functional differences (Fernholz 1983; Hampel 1974):

$$\begin{aligned} v(G) - v(F) &= \int_0^1 \mathrm{IF}(y; v, F) dG(y) + r(v, F, G) \\ &\approx \int_0^1 \mathrm{IF}(y; v, F) dG(y). \end{aligned}$$

The accuracy of the approximation depends on how close G is to F. The connection to the Gâteaux derivative becomes clearer if one sets $\epsilon = 1$ in the definition Equation (1): the von Mises approximation consists in linearly projecting the Gâteaux derivative (defined for infinitesimal mixing $\epsilon \to 0$) all the way from F to G (corresponding to $\epsilon = 1$). This is unlikely to be a good approximation for F and G far apart and therefore provides another argument for setting t relatively small in the defnition of G.

Table 3. *Cont.*

	UE	UPE			
	Model 1	**Model 2a**	**Model 2b**	**Model 3a**	**Model 3b**
			Percentile ratio P90/P50		
Immigrant worker	0.028	0.027	0.026	0.017	0.021
	(0.011)	(0.009)	(0.009)	(0.007)	(0.007)
Cross-border worker	−0.002	−0.002	−0.001	0.001	0.009
	(0.011)	(0.010)	(0.009)	(0.007)	(0.007)

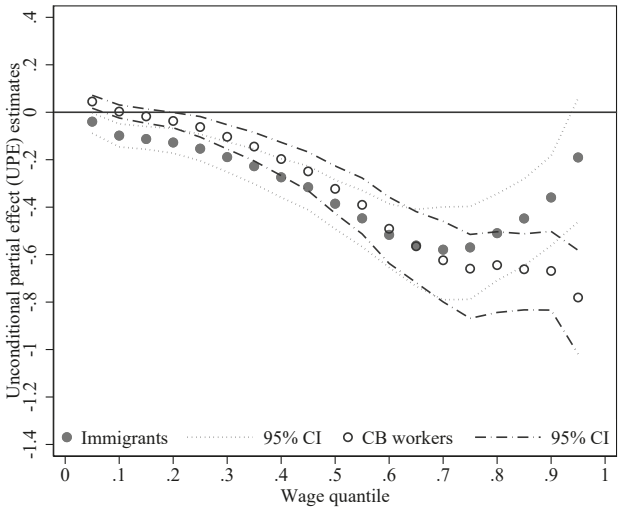
Notes: Based on the 2006 Luxembourg Structure of Earnings Survey data. Model 1 does not include covariates (and therefore estimates UEs), Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Models 2a and 3a are based on a simple additive RIF-OLS regression models, models 2b and 3b allow interaction between nationality groups and all covariates in the RIF-OLS regression. Bootstrap standard errors in brackets.

4.2. Accounting for Human Capital and Job Characteristics: UPE Estimates

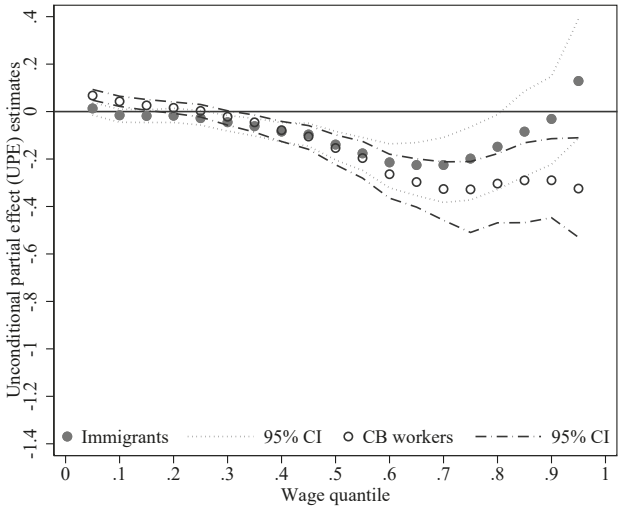
Cross-border workers and immigrants have markedly different characteristics from native workers. It is therefore useful to consider UPEs that capture the locations of each subgroup's wage distributions relative to each other conditionally on covariates. This is where the interest of the methodology materializes since traditional inequality decomposition analysis does not lean itself to examining conditional wage distributions with a rich set of covariates. UPE captures the direction of a change in quantiles and inequality measures for a change in the share of foreign and national workers that would keep constant the overall distribution of human capital and job characteristics.

UPEs on 19 quantiles are shown in Figure 3: the top panel accounts for individual characteristics only (age, gender and level of education), the bottom panel accounts for both individual and job characteristics. (Full regression coefficient estimates for three selected quantiles are reported in Appendix B.) Adjusting for individual characteristics reduces only moderately the absolute impact of foreign workers. However, further adjusting for job characteristics markedly reduces this impact. Quantile UPEs remain generally negative (except at the very bottom for cross-border workers and at the very top for immigrants), but they are much smaller in absolute value and globally significantly below zero only between the 35th and 80th percentiles. UPEs are, overall, very small throughout the bottom half of the distribution. UPEs for cross-border workers become more markedly negative for quantiles above the median and UPEs for immigrants display again a U-shape with declining values until the 65th percentile and an increase up to about zero for the highest quantile.

The impacts of foreign workers on inequality measures are further muted. Table 3 under Model 2 and Model 3 reports estimates of inequality UPEs. Estimates under Model 2 control for human capital characteristics only; estimates under Model 3 also control for job characteristics. Two sets of estimates are reported. Estimates under Model 2a and 3a are based on the basic additive IF-OLS model of Equation (6)—also used in Figure 3. Estimates under Model 2b and 3b are based on the more flexible specification (7) allowing for interactions. All UPE estimates have the same sign as the UE estimates. They remain negative and mostly significantly different from zero for cross-border workers, while immigrant workers' impacts on wage dispersion and inequality remain negative, small and not significantly different from zero, with the exception of the P90/P50 ratio. The size of the coefficients in Model 2 is comparable to UE estimates (Model 1). The coefficients fall in (absolute) size in Model 3 as part of the wage differences is driven by job characteristics, yet the coefficients tend to remain significantly different from zero where the UEs were already significant.



(**a**)



(**b**)

Figure 3. Estimates of unconditional partial effects (UPE) on quantiles of the wage distribution for immigrant and cross-border (CB) workers. (**a**) Conditional on age, education level and gender; (**b**) Conditional on age, education level, gender and job characteristics.

4.3. Impacts by Disaggregate Nationality Groups

Disaggregation of foreign workers types by less coarsely defined nationality groups leads to more homogeneous subgroups and sheds some light on the patterns just observed. Cross-border workers from France, Belgium and Germany form a relatively homogeneous labour force in terms of skill composition. There is much more heterogeneity across immigrant groups. Portuguese immigrants with generally low educational achievements form the largest share of immigrants (47 percent in our

sample). Belgian, French and German immigrants taken together represent 26 percent of our sample, other EU immigrants, 18 percent and non-EU immigrants, 10 percent.

Figures 4 and 5 show disaggregated quantile UEs and UPEs by country of residence (for cross-border workers) and by broad nationality groups (for immigrants). The coefficients are for marginal impacts of substituting workers from one of these groups against native residents. Unsurprisingly, the impacts of the three cross-border groups are similar, with the largest negative impact attributed to Belgian residents. There is much more heterogeneity across immigrant groups. Portuguese immigrants are consistently found to depress all quantiles: they are paid relatively low wages. However the impact largely disappears (even at high quantiles) after controlling for individual and job characteristics. At the other end of the spectrum, Belgian, German and French residents appear to have positive UEs, especially at top quantiles: they are paid higher wages than natives and drive up the top quantiles.[10] Non-EU and other EU immigrants have parallel profiles although at different levels. UEs and UPEs for both groups exhibit a markedly U-shape; this suggests that a fraction of the population in these groups tend to receive low pay (and therefore quickly reduce quantiles in the bottom of the distribution) while a fraction of the groups is highly paid and increases top quantiles. This pattern is particularly strong for non-EU immigrants that—after accounting for human capital and job characteristics—is the group that depresses most bottom quantiles and increases most top quantiles.

Table 4 shows disaggregated UPE estimates for the Gini and the variance and Table 5 shows UPE estimates for the percentile ratios. The P90/P50 ratio apart, the only group that appears to contribute positively and significantly to inequality is the non-EU immigrants. By contrast, wages of French and Belgian cross-border workers are consistently found to significantly reduce inequality (with similar orders of magnitude). All other groups generally have no systematically significant impacts after controlling for human capital and job characteristics. The UPEs of Portuguese residents is negative and significant on the variance but the effects disappear on the Gini coefficient.

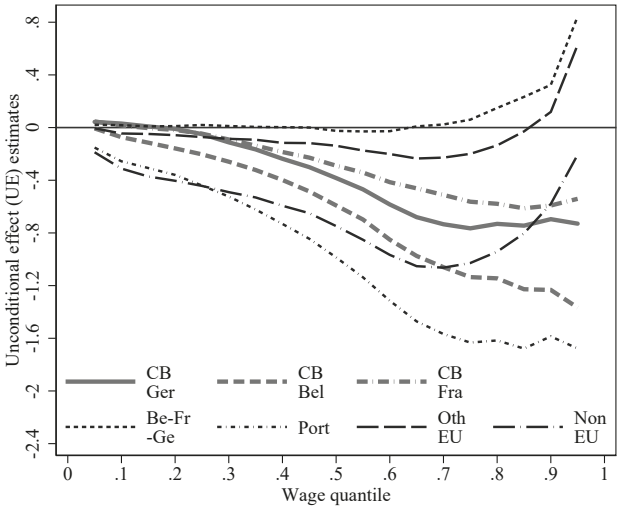
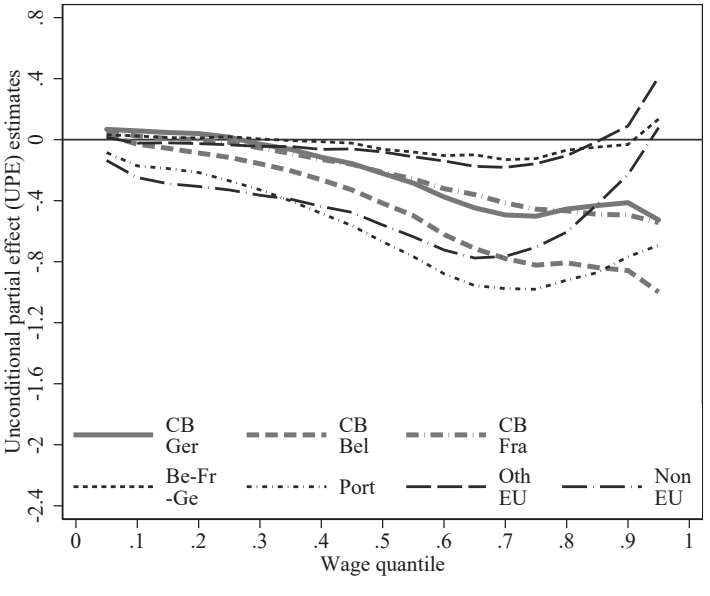
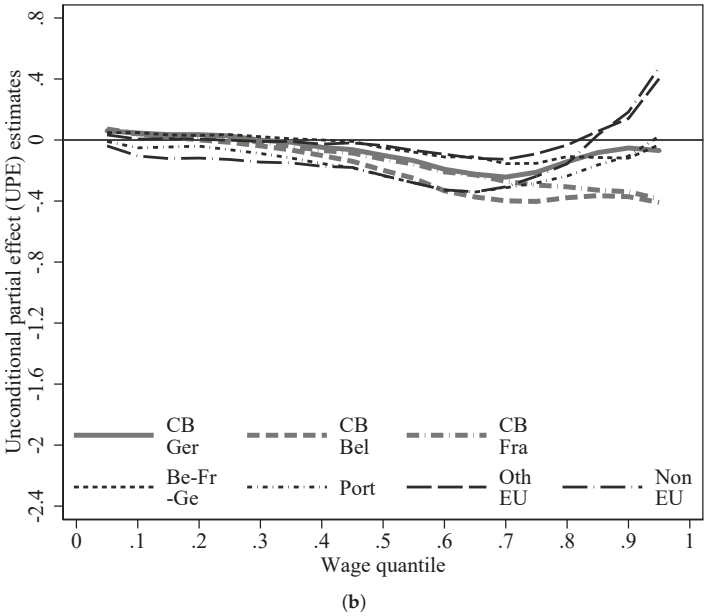


Figure 4. Estimates of unconditional effects (UE) on quantiles of the wage distribution for immigrant and cross-border workers disaggregated by nationality groups.

[10] Endogenous selection is likely at play here with high wage workers from Belgium, France or Germany affording the potential costs of migrating into Luxembourg.



(**a**)



(**b**)

Figure 5. Estimates of unconditional partial effects (UPE) on quantiles of the wage distribution for immigrant and cross-border workers disaggregated by nationality groups. (**a**) Conditional on age, education level and gender; (**b**) Conditional on age, education level, gender and job characteristics.

Examination of percentile ratios shows again how the U-shape of quantile UEs imply that foreign workers tend to increase dispersion in the upper half of the distribution and reduce dispersion in the bottom half. This is mostly marked for non-EU residents.

Table 4. Estimates of UEs (Model 1) and UPEs (Model 2 and Model 3) on the variance of hourly wages and the Gini coefficient for disaggregated nationality groups.

	UE	**UPE**			
	Model 1	**Model 2a**	**Model 2b**	**Model 3a**	**Model 3b**
			Variance		
Be-Fr-Ge resident	3.91	−6.84	−3.91	−6.37	−4.74
	(9.31)	(11.77)	(10.57)	(9.33)	(5.34)
Portuguese resident	−19.42	−10.56	−20.31	−1.41	−13.45
	(7.13)	(6.95)	(8.83)	(3.03)	(3.79)
Other EU resident	−0.32	−4.95	−7.33	−2.66	−4.42
	(8.02)	(9.42)	(9.25)	(6.52)	(4.23)
Non-EU resident	5.42	5.98	0.63	11.51	3.58
	(5.33)	(5.09)	(6.85)	(5.60)	(8.46)
German CB	−10.72	−10.41	−11.22	−3.62	−1.77
	(6.63)	(7.76)	(8.29)	(5.40)	(4.47)
French CB	−17.66	−15.51	−17.82	−7.27	−8.71
	(6.62)	(7.56)	(8.34)	(4.05)	(2.95)
Belgian CB	−10.20	−11.92	−12.07	−8.23	−4.32
	(5.24)	(6.57)	(6.68)	(4.08)	(2.22)
			Gini coefficient		
Be-Fr-Ge resident	0.0029	−0.0019	−0.0006	−0.0026	−0.0016
	(0.0019)	(0.0023)	(0.0021)	(0.0018)	(0.0012)
Portuguese resident	−0.0046	−0.0012	−0.0063	0.0007	−0.0061
	(0.0019)	(0.0018)	(0.0019)	(0.0010)	(0.0010)
Other EU resident	0.0027	0.0006	−0.0003	0.0007	−0.0003
	(0.0022)	(0.0024)	(0.0022)	(0.0018)	(0.0013)
Non-EU resident	0.0060	0.0058	0.0047	0.0057	0.0050
	(0.0021)	(0.0020)	(0.0026)	(0.0018)	(0.0028)
German CB	−0.0044	−0.0043	−0.0037	−0.0013	−0.0001
	(0.0016)	(0.0017)	(0.0019)	(0.0014)	(0.0016)
French CB	−0.0066	−0.0059	−0.0067	−0.0034	−0.0039
	(0.0016)	(0.0016)	(0.0018)	(0.0010)	(0.0008)
Belgian CB	−0.0040	−0.0047	−0.0046	−0.0036	−0.0019
	(0.0014)	(0.0015)	(0.0015)	(0.0010)	(0.0007)

Notes: Based on the 2006 Luxembourg Structure of Earnings Survey data. Model 1 does not include covariates (and therefore estimates UEs), Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Models 2a and 3a are based on a simple additive RIF-OLS regression models, models 2b and 3b allow interaction between nationality groups and all covariates in the RIF-OLS regression. Bootstrap standard errors in brackets.

Table 5. Estimates of UEs (Model 1) and UPEs (Model 2 and Model 3) on three percentile ratios (P90/P10, P50/P10 and P90/P50) for disaggregated nationality groups.

	UE	UPE			
	Model 1	**Model 2a**	**Model 2b**	**Model 3a**	**Model 3b**
		Percentile ratio P90/P10			
Be-Fr-Ge resident	0.028	−0.011	0.001	−0.029	−0.011
	(0.016)	(0.013)	(0.013)	(0.013)	(0.012)
Portuguese resident	−0.082	−0.023	−0.091	0.007	−0.082
	(0.019)	(0.018)	(0.017)	(0.013)	(0.015)
Other EU resident	0.027	0.017	0.013	0.013	0.009
	(0.021)	(0.019)	(0.018)	(0.018)	(0.016)
Non-EU resident	0.042	0.058	0.042	0.055	0.067
	(0.026)	(0.023)	(0.028)	(0.020)	(0.028)
German CB	−0.083	−0.063	−0.046	−0.021	−0.004
	(0.015)	(0.013)	(0.017)	(0.014)	(0.021)
French CB	−0.106	−0.081	−0.084	−0.052	−0.048
	(0.014)	(0.012)	(0.012)	(0.010)	(0.011)
Belgian CB	−0.068	−0.059	−0.059	−0.048	−0.026
	(0.013)	(0.011)	(0.011)	(0.011)	(0.010)
		Percentile ratio P50/P10			
Be-Fr-Ge resident	−0.005	−0.010	−0.009	−0.013	−0.009
	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)
Portuguese resident	−0.059	−0.040	−0.048	−0.015	−0.023
	(0.008)	(0.007)	(0.008)	(0.005)	(0.007)
Other EU resident	−0.007	−0.005	−0.004	−0.005	−0.002
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
Non-EU resident	−0.026	−0.017	−0.024	−0.007	−0.011
	(0.009)	(0.008)	(0.008)	(0.007)	(0.006)
German CB	−0.044	−0.032	−0.030	−0.018	−0.022
	(0.007)	(0.006)	(0.006)	(0.004)	(0.005)
French CB	−0.048	−0.038	−0.036	−0.027	−0.022
	(0.006)	(0.005)	(0.005)	(0.004)	(0.004)
Belgian CB	−0.032	−0.025	−0.027	−0.019	−0.021
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
		Percentile ratio P90/P50			
Be-Fr-Ge resident	0.025	0.006	0.012	−0.001	0.004
	(0.009)	(0.008)	(0.008)	(0.008)	(0.008)
Portuguese resident	0.024	0.037	0.004	0.025	−0.023
	(0.017)	(0.014)	(0.015)	(0.009)	(0.012)
Other EU resident	0.026	0.017	0.013	0.014	0.008
	(0.012)	(0.011)	(0.010)	(0.011)	(0.010)
Non-EU resident	0.060	0.059	0.058	0.043	0.057
	(0.014)	(0.012)	(0.015)	(0.011)	(0.016)
German CB	0.004	0.002	0.011	0.010	0.026
	(0.012)	(0.010)	(0.012)	(0.008)	(0.014)
French CB	−0.004	−0.002	−0.006	0.002	−0.002
	(0.013)	(0.011)	(0.010)	(0.008)	(0.008)
Belgian CB	−0.001	−0.004	−0.002	−0.006	0.010
	(0.009)	(0.009)	(0.009)	(0.007)	(0.007)

Notes: Based on the 2006 Luxembourg Structure of Earnings Survey data. Model 1 does not include covariates (and therefore estimates UEs), Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Models 2a and 3a are based on a simple additive RIF-OLS regression models, models 2b and 3b allow interaction between nationality groups and all covariates in the RIF-OLS regression. Bootstrap standard errors in brackets.

5. Summary and Conclusions

This paper assesses how foreign workers fit in and contribute to the shape of the wage distribution in Luxembourg, a high immigration country. We empirically confront claims that foreign workers wages inflate overall wage inequality. We do this indirectly. Following Firpo et al. (2009) and Rothe (2010), we define distributional parameters of interest that quantify how various distributional statistics would respond to a notional marginal substitution of native workers for foreign workers. This is estimated both unconditionally and conditionally on workers' human capital and job characteristics using influence function regression.

We remark first that quantiles of the wage distribution are generally driven *down* by foreign workers. Exceptions are only found for top quantiles which are driven up by immigrants from neighbouring countries and other EU countries (Portugal excluded) if we do *not* condition on covariates, and by non-EU immigrants and EU immigrants (Portugal and neighbouring countries excluded) if we condition on human capital and job characteristics. This is consistent with the fact that foreign workers are polarized and 'sandwich' the distribution at both high skill and low skill positions. The implication of these patterns for wage inequality is very limited however. There is hardly any indication that immigrants wages inflate the variance, the Gini coefficient or percentile ratios. The only significant exception is for non-EU immigrants—not more than 10 percent of immigrants—that appear to contribute positively to wage dispersion. All other immigrants affect inequality downwards if at all, while cross-border workers significantly drive inequality down.

Influence function regression reveals well-suited to the type of analysis conducted in this paper, and it could easily be expanded to additional distributional statistics, such as measures of earnings polarization or low pay. Of course, resulting estimates of marginal impacts must not be mis-interpreted as long-term general equilibrium effects of migration. We do not estimate longer-term equilibrium effects of such a change in employment composition or, to put it differently, we assume wages of employed workers to be unaffected by a marginal increase in foreign workers in total employment.[11] Our results instead provide descriptive evidence on how the structure of wages of foreign workers contribute to the overall wage distribution and in particular to wage inequality.

Author Contributions: Both authors contributed equally to the paper.

Acknowledgments: This research was part of the InWIn project (*Information and wage inequality: Evidence on wage differences between natives, immigrants and cross-border workers in Luxembourg*) supported by the Luxembourg 'Fonds National de la Recherche' (contract C10/LM/785657). The 2006 Luxembourg Structure of Earnings Survey data were provided by STATEC—Institut national de la statistique et des études économiques. Comments by Michela Bia, Jacques Brosius, Alessio Fusco, Stephen Jenkins, Jean Ries and participants to the BCL Household Finance and Consumption Workshop, ECINEQ 2017 and seminar participants at ISER (University of Essex) are gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Assume for clarity, and without loss of generality, that the multinomial variable of workers type X takes on values $0, 1, \ldots, K$ and its cumulative distribution F_X can be written

$$F_X(x) = \sum_{j=0}^{K} s_j \mathbf{1}\left[x \geq j\right]$$

[11] Please note that estimates of general equilibrium effects of immigration available for other countries are in fact generally small (Blau and Kahn 2012; Card 2009), although of course these findings may not necessarily apply to the Luxembourg case.

(where **1** [cond] is 1 if cond is true and 0 otherwise). The distribution of workers type after shifting shares from the reference worker type r to worker type k, $G_X^{1,k}$, is then

$$G_X^{1,k}(x) = \sum_{j=0}^{K} s_j^* \mathbf{1}[x \geq j]$$

where $s_r^* = s_r - t$, $s_k^* = s_k + t$, and $s_j^* = s_j$ otherwise. The implied difference in the probability distributions $d(G_X^{1,k} - F_X)$ is

$$\begin{aligned} d(G_X^{1,k} - F_X)(x) &= t\,(\mathbf{1}[x = k] - \mathbf{1}[x = r]) \\ &= \begin{cases} t & \text{if } x = k \\ -t & \text{if } x = r \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Using Theorem 1 from Firpo et al. (2009) and inserting $d(G_X^{1,k} - F_X)(x)$ in Equation (2) we have our resulting expression

$$\mathrm{UE}(v(F), k) = (\mathrm{E}[\mathrm{IF}(y; v, F)|X = k] - \mathrm{E}[\mathrm{IF}(y; v, F)|X = r])\, t.$$

Derivation of the expression for $\mathrm{UPE}(v(F), k)$ follows the same logic, after conditioning on covariates Z. Let us first write the joint distribution of covariates X and Z

$$\begin{aligned} F_{X,Z}(x, z) &= F_{X|Z\leq z}(x)\, F_Z(z) \\ &= \left(\sum_{j=0}^{K} s_{j|Z\leq z} \mathbf{1}[x \geq j]\right) F_Z(z) \end{aligned}$$

and $G_{X,Z}^{2,k}$ as

$$G_{X,Z}^{2,k}(x, z) = \left(\sum_{j=0}^{K} s_{j|Z\leq z}^* \mathbf{1}[x \geq j]\right) F_Z(z)$$

where the shares of workers types conditional on Z are $s_{r|Z\leq z}^* = s_{r|Z\leq z} - t$, $s_{k|Z\leq z}^* = s_{k|Z\leq z} + t$, and $s_{j|Z\leq z}^* = s_{j|Z\leq z}$ otherwise. We now have

$$d(G_{X,Z}^{2,k} - F_{X,Z})(x, z) = \begin{cases} t\, dF_Z(z) & \text{if } x = k \\ -t\, dF_Z(z) & \text{if } x = r \\ 0 & \text{otherwise} \end{cases}$$

Using Theorem 1 from Firpo et al. (2009) again, now expanded to the full set of covariates X and Z, namely

$$\nabla v_{F\to G^*} = \int_{\Omega_Z} \int_{\Omega_X} \mathrm{E}[\mathrm{IF}(y; v, F)|X = x, Z = z]\; d(G_{X,Z}^* - F_{X,Z})(x, z) \tag{A1}$$

and substituting $d(G_{X,Z}^{2,k} - F_{X,Z})(x, z)$ in Equation (A1) yields our second main expression

$$\begin{aligned} \mathrm{UPE}(v(F), k) &= \nabla v_{F\to G_k^2} \\ &= \left(\int_{\Omega_Z} (\mathrm{E}[\mathrm{IF}(y; v, F)|X = k, Z = z] - \mathrm{E}[\mathrm{IF}(y; v, F)|X = r, Z = z])\; dF_Z(z)\right) t. \end{aligned}$$

Appendix B. Detailed Influence Function Regression Results

Table A1. Coefficient estimates of influence function regressions - P10.

	Aggregate Nationality Groups			Disaggregate Nationality Groups		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Immigrant worker	−0.154 *	−0.102 *	−0.018 †			
Cross-border worker	−0.027 ‡	0.002	0.042 *			
Be-Fr-Ge resident				0.018 ‡	0.024 †	0.049 *
Portuguese resident				−0.258 *	−0.178 *	−0.056 *
Other EU resident				−0.045 *	−0.023 ‡	0.006
Non-EU resident				−0.314 *	−0.250 *	−0.107 *
German CB				0.031 †	0.059 *	0.047 *
French CB				−0.073 *	−0.031 †	0.039 *
Belgian CB				0.016	0.021 †	0.039 *
Female		−0.187 *	−0.121 *		−0.185 *	−0.119 *
Age		0.434 *	0.385 *		0.425 *	0.384 *
Age squared/100		−0.005 *	−0.004 *		−0.005 *	−0.004 *
Secondary education		0.130 *	0.008		0.100 *	0.002
Tertiary education		0.304 *	0.066 †		0.245 *	0.053 ‡
Job tenure		0.186 *	0.148 *		0.191 *	0.152 *
Job tenure squared/100		−0.004 *	−0.003 *		−0.004 *	−0.003 *
Manager			0.008			0.006
10–49 employees in firm			0.033 ‡			0.035 ‡
50–249 employees in firm			0.041 †			0.042 †
500–999 employees in firm			−0.070 ‡			−0.069 ‡
1000+ employees in firm			0.021			0.023
Part time contract			−0.074 *			−0.074 *
Industry/Manufacture			−0.086 *			−0.087 *
Construction			−0.069 *			−0.060 †
Wholesale			−0.268 *			−0.267 *
Hotel/Restaurant			−0.374 *			−0.372 *
Trans/Comm			−0.050 ‡			−0.050 ‡
Finance			−0.064 *			−0.067 *
Real estate			−0.159 *			−0.157 *
Managerial			0.452 *			0.435 *
Professional			0.488 *			0.471 *
Associate professional			0.505 *			0.488 *
Clerk			0.493 *			0.477 *
Service worker			0.301 *			0.289 *
Craft and trade worker			0.380 *			0.371 *
Manufacturers			0.386 *			0.376 *
Constant	1.042 *	0.005	−0.132	1.042 *	0.057	−0.111

Notes: Based on 2006 Luxembourg Structure of Earnings Survey data. *, † and ‡ indicate statistical significance at 1, 5 and 10 percent levels respectively (repeated half-sample percentile bootstrap confidence intervals not covering zero at the corresponding confidence levels). Model 1 does not include covariates, Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Binary or multinomial covariates scaled by $t = 0.1$.

Table A2. Coefficient estimates of influence function regressions - P50 (median).

	Aggregate Nationality Groups			Disaggregate Nationality Groups		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Immigrant worker	−0.564 *	−0.390 *	−0.142 *			
Cross-border worker	−0.468 *	−0.325 *	−0.155 *			
Be-Fr-Ge resident				−0.025	−0.064 †	−0.052 †
Portuguese resident				−0.989 *	−0.677 *	−0.241 *
Other EU resident				−0.138 *	−0.081 †	−0.037
Non-EU resident				−0.751 *	−0.562 *	−0.234 *
German CB				−0.384 *	−0.222 *	−0.100 *
French CB				−0.594 *	−0.419 *	−0.201 *
Belgian CB				−0.290 *	−0.219 *	−0.128 *
Female		−0.020	−0.134 *		−0.018	−0.129 *
Age		0.851 *	0.564 *		0.820 *	0.559 *
Age squared/100		−0.010 *	−0.006 *		−0.010 *	−0.006 *
Secondary education		0.429 *	0.033		0.331 *	0.016
Tertiary education		1.285 *	0.194 *		1.092 *	0.165 *
Job tenure		0.415 *	0.286 *		0.427 *	0.295 *
Job tenure squared/100		−0.004 *	−0.004 *		−0.005 *	−0.004 *
Manager			0.216 *			0.212 *
10-49 employees in firm			−0.054			−0.054
50-249 employees in firm			−0.026			−0.026
500-999 employees in firm			−0.001			0.003
1000+ employees in firm			0.160 ‡			0.165 ‡
Part time contract			0.071 *			0.069 *
Industry/Manufacture			−0.223 *			−0.222 *
Construction			−0.522 *			−0.502 *
Wholesale			−0.527 *			−0.521 *
Hotel/Restaurant			−0.519 *			−0.509 *
Trans/Comm			−0.137 ‡			−0.141 ‡
Finance			0.031			0.024
Real estate			−0.324 *			−0.310 *
Managerial			1.017 *			0.972 *
Professional			1.153 *			1.112 *
Associate professional			1.033 *			0.992 *
Clerk			0.644 *			0.607 *
Service worker			0.258 *			0.233 *
Craft and trade worker			0.328 *			0.309 *
Manufacturers			0.265 *			0.239 *
Constant	1.939 *	−0.589 *	−0.066	1.939 *	−0.417 *	−0.009

Notes: Based on 2006 Luxembourg Structure of Earnings Survey data. *, † and ‡ indicate statistical significance at 1, 5 and 10 percent levels respectively (repeated half-sample percentile bootstrap confidence intervals not covering zero at the corresponding confidence levels). Model 1 does not include covariates, Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Binary or multinomial covariates scaled by $t = 0.1$.

Table A3. Coefficient estimates of influence function regressions - P90.

	Aggregate Nationality Groups			Disaggregate Nationality Groups		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Immigrant worker	−0.718 *	−0.378 *	−0.031			
Cross-border worker	−0.986 *	−0.699 *	−0.301 *			
Be-Fr-Ge resident				0.338 †	−0.034	−0.123
Portuguese resident				−1.652 *	−0.809 *	−0.108
Other EU resident				0.124	0.093	0.147
Non-EU resident				−0.604 *	−0.239	0.193 ‡
German CB				−0.725 *	−0.430 *	−0.055
French CB				−1.284 *	−0.895 *	−0.386 *
Belgian CB				−0.616 *	−0.519 *	−0.352 *
Female		−0.268 *	−0.380 *		−0.262 *	−0.377 *
Age		0.543 *	−0.108		0.485 *	−0.130
Age squared/100		−0.000	0.006 *		0.000	0.006 *
Secondary education		1.029 *	0.297 *		0.890 *	0.284 *
Tertiary education		4.034 *	1.220 *		3.768 *	1.202 *
Job tenure		0.654 *	0.302 *		0.679 *	0.325 *
Job tenure squared/100		−0.006	0.002		−0.007	0.002
Manager			0.787 *			0.793 *
10-49 employees in firm			−0.147			−0.153 ‡
50-249 employees in firm			−0.034			−0.030
500-999 employees in firm			0.213 ‡			0.227 ‡
1000+ employees in firm			0.063			0.081
Part time contract			0.507 *			0.503 *
Industry/Manufacture			−0.888 *			−0.881 *
Construction			−0.896 *			−0.883 *
Wholesale			−0.704 *			−0.685 *
Hotel/Restaurant			−0.826 *			−0.809 *
Trans/Comm			−0.278			−0.301
Finance			−0.228			−0.231
Real estate			−1.087 *			−1.051 *
Managerial			4.910 *			4.891 *
Professional			2.221 *			2.208 *
Associate professional			1.024 *			1.001 *
Clerk			0.054			0.035
Service worker			0.165 †			0.158 †
Craft and trade worker			0.129			0.119
Manufacturers			−0.046			−0.059
Constant	3.879 *	0.116	2.136 *	3.879 *	0.384	2.195 *

Notes: Based on 2006 Luxembourg Structure of Earnings Survey data. *, † and ‡ indicate statistical significance at 1, 5 and 10 percent levels respectively (repeated half-sample percentile bootstrap confidence intervals not covering zero at the corresponding confidence levels). Model 1 does not include covariates, Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Binary or multinomial covariates scaled by $t = 0.1$.

Table A4. Coefficient estimates of influence function regressions - Variance.

	Aggregate Nationality Groups			Disaggregate Nationality Groups		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Immigrant worker	−7.58	−6.97	−1.91			
Cross-border worker	−14.17 *	−13.52 *	−6.76 †			
Be-Fr-Ge resident				3.91	−6.84	−6.37
Portuguese resident				−19.42 *	−10.56 *	−1.41
Other EU resident				−0.32	−4.95	−2.66
Non-EU resident				5.42	5.98	11.51 †
German CB				−10.72 *	−10.41 †	−3.62
French CB				−17.66 *	−15.51 *	−7.27 *
Belgian CB				−10.20 *	−11.92 *	−8.23 *
Female		−5.25 *	−10.38 ‡		−5.20 *	−10.42 ‡
Age		−36.81 *	−42.57 *		−37.43 *	−42.97 *
Age squared/100		0.60 *	0.61 *		0.60 *	0.62 *
Secondary education		10.50 *	2.55 ‡		9.65 *	2.67 †
Tertiary education		63.31 *	31.23 *		61.75 *	31.52 *
Job tenure		10.92 *	7.59 †		11.27 *	7.80 †
Job tenure squared/100		−0.42 *	−0.25 ‡		−0.42 *	−0.25 †
Manager			2.43			2.67
10–49 employees in firm			0.31			0.13
50–249 employees in firm			12.25 †			12.34 †
500–999 employees in firm			1.31			1.49
1000+ employees in firm			−1.79			−1.61
Part time contract			20.26 *			20.21 *
Industry/Manufacture			−22.51 †			−22.34 †
Construction			−23.64 *			−23.79 *
Wholesale			−14.48			−14.22
Hotel/Restaurant			−16.43 †			−16.51 †
Trans/Comm			−15.91			−16.27
Finance			−21.29			−21.07
Real estate			−25.67 †			−25.33 †
Managerial			103.07 *			104.00 *
Professional			18.75 ‡			19.77 ‡
Associate professional			7.13			7.93
Clerk			1.31			2.10
Service worker			−0.39			0.11
Craft and trade worker			2.85			3.29
Manufacturers			0.59			1.16
Constant	41.24 *	72.65 *	101.35 *	41.24 *	74.77 *	101.11 *

Notes: Based on 2006 Luxembourg Structure of Earnings Survey data. *, † and ‡ indicate statistical significance at 1, 5 and 10 percent levels respectively (repeated half-sample percentile bootstrap confidence intervals not covering zero at the corresponding confidence levels). Model 1 does not include covariates, Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Binary or multinomial covariates scaled by $t = 0.1$.

Table A5. Coefficient estimates of influence function regressions - Gini.

	Aggregate Nationality Groups			Disaggregate Nationality Groups		
	Model 1	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
Immigrant worker	−0.0003	−0.0004	0.0002			
Cross-border worker	−0.0054 *	−0.0053 *	−0.0030 *			
Be-Fr-Ge resident				0.0029 ‡	−0.0019	−0.0026 ‡
Portuguese resident				−0.0046 *	−0.0012	0.0007
Other EU resident				0.0027 ‡	0.0006	0.0007
Non-EU resident				0.0060 *	0.0058 *	0.0057 *
German CB				−0.0044 *	−0.0043 *	−0.0013
French CB				−0.0066 *	−0.0059 *	−0.0034 *
Belgian CB				−0.0040 *	−0.0047 *	−0.0036 *
Female		−0.0006	−0.0023 †		−0.0006	−0.0023 †
Age		−0.0192 *	−0.0213 *		−0.0194 *	−0.0215 *
Age squared/100		0.0003 *	0.0003 *		0.0003 *	0.0003 *
Secondary education		0.0033 *	0.0012 ‡		0.0032 *	0.0013 ‡
Tertiary education		0.0254 *	0.0126 *		0.0253 *	0.0129 *
Job tenure		0.0012 ‡	−0.0002		0.0014 ‡	−0.0001
Job tenure squared/100		−0.0001 ‡	0.0000		−0.0001 ‡	0.0000
Manager			0.0028 *			0.0029 *
10–49 employees in firm			−0.0003			−0.0004
50–249 employees in firm			0.0022			0.0022
500–999 employees in firm			0.0017			0.0018 ‡
1000+ employees in firm			−0.0009			−0.0008
Part time contract			0.0067 *			0.0067 *
Industry/Manufacture			−0.0060 †			−0.0060 †
Construction			−0.0056 †			−0.0058 †
Wholesale			−0.0003			−0.0002
Hotel/Restaurant			−0.0007			−0.0007
Trans/Comm			−0.0028			−0.0029
Finance			−0.0043			−0.0042
Real estate			−0.0073 *			−0.0071 *
Managerial			0.0398 *			0.0403 *
Professional			0.0018			0.0023
Associate professional			−0.0051 †			−0.0047 †
Clerk			−0.0085 *			−0.0081 *
Service worker			−0.0034 *			−0.0031 *
Craft and trade worker			−0.0054 *			−0.0052 *
Manufacturers			−0.0072 *			−0.0069 *
Constant	0.0311 *	0.0534 *	0.0679 *	0.0311 *	0.0538 *	0.0677 *

Notes: Based on 2006 Luxembourg Structure of Earnings Survey data. *, † and ‡ indicate statistical significance at 1, 5 and 10 percent levels respectively (repeated half-sample percentile bootstrap confidence intervals not covering zero at the corresponding confidence levels). Model 1 does not include covariates, Model 2 includes human capital covariates, Model 3 includes both human capital and job characteristics. Binary or multinomial covariates scaled by $t = 0.1$.

References

Adsera, Alicia, and Barry Chiswick. 2007. Are there gender and country of origin differences in immigrant labor market outcomes across European destinations? *Journal of Population Economics* 20: 495–526. [CrossRef]

Amétépé, Fofo, and Claudia Hartmann-Hirsch. 2011. An outstanding positioning of migrants and nationals: The case of Luxembourg. *Population Review* 50: 195–217.

Annaert, Jean-Luc. 2004. A bright spot in the heart of Europe: What can we learn from the Luxembourg success story? *ECFIN Country Focus* 1: 15.

Blau, Francine D., and Lawrence M. Kahn. 2012. Immigration and the distribution of incomes. NBER Working Paper 18515, National Bureau of Economic Research, Cambridge, MA, USA.

Borjas, George J. 1985. Assimilation, changes in cohort quality, and the earnings of immigrants. *Journal of Labor Economics* 3: 463–89. [CrossRef]

Borjas, George J. 1995. Assimilation and changes in cohort quality revisited: What happened to immigration earnings in the 1980s? *Journal of Labor Economics* 13: 201–45. [CrossRef] [PubMed]

Borjas, George J. 1999. *Heaven's Door*. Princeton: Princeton University Press.

Borjas, George J. 2003. The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market. NBER Working Paper 9755, National Bureau of Economic Research, Cambridge, MA, USA.

Card, David. 1990. The impact of the mariel boatlift on the Miami labor market. *Industrial and Labor Relations Review* 43: 245–57. [CrossRef]

Card, David. 2009. Immigration and inequality. *American Economic Review* 99: 1–21. [CrossRef]

Carrasco, Raquel, Juan F. Jimeno, and A. Carolina Ortega. 2008. The effect of immigration on the labor market performance of native-born workers: Some evidence for Spain. *Journal of Population Economics* 21: 627–48. [CrossRef]

Chernozhukov, Victor, Ivan Fernandez-Val, and Blaise Melly. 2013. Inference on counterfactual distributions. *Econometrica* 81: 2205–68. [CrossRef]

Chiswick, Barry R. 1978. The effect of americanization on the earnings of foreign-born men. *Journal of Political Economy* 86: 897–921. [CrossRef]

Deville, Jean-Claude. 1999. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* 25: 193–203.

Dugger, Daniel, and Peter J. Lambert. 2014. The 1913 paper of René Gâteaux, upon which the modern-day influence function is based. *Journal of Economic Inequality* 12: 149–52. [CrossRef]

Dustmann, Christian, Tommaso Frattini, and Ian P. Preston. 2013. The effect of immigration along the distribution of wages. *Review of Economic Studies* 80: 145–73. [CrossRef]

Efron, Bradley. 1981. Nonparametric standard errors and confidence intervals. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 9: 139–58. [CrossRef]

Essama-Nssah, Boniface, and Peter J. Lambert. 2012. Influence functions for policy impact analysis. In *Inequality, Mobility and Segregation: Essays in Honor of Jacques Silber (Research on Economic Inequality, Volume 20)*. Edited by John A. Bishop and Rafael Salas. Bingley: Emerald Group Publishing, chp. 6, pp. 135–59.

Fernholz, Luisa Turrin. 1983. *von Mises Calculus For Statistical Functionals*. Number 19 in Lecture Notes in Statistics. New York: Springer-Verlag.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2007. Unconditional quantile regressions. Technical Working Paper 339, National Bureau of Economic Research, Cambridge, MA, USA.

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. 2009. Unconditional quantile regressions. *Econometrica* 77: 953–73.

Friedberg, Rachel M. 2001. The impact of mass migration on the Israeli labor market. *Quarterly Journal of Economics* 116: 1373–408. [CrossRef]

Fusco, Alessio, Philippe Van Kerm, Aigul Alieva, Luna Bellani, Fanny Etienne-Robert, Anne-Catherine Guio, Iryna Kyzyma, Kristell Leduc, Philippe Liégeois, Maria Noel Pi Alperin, et al. 2014. Luxembourg: Has inequality grown enough to matter? In *Changing Inequalities and Societal Impacts in Rich Countries: Thirty Countries Experiences*. Edited by Brian Nolan, Wiemer Salverda, Daniele Checchi, Ive Marx, Abigail McKnight, István György Tóth and Herman van de Werfhorst. Oxford: Oxford University Press.

Gâteaux, René. 1913. Sur les fonctionnelles continues et les fonctionnelles analytiques. *Comptes Rendus de l'Académie des Sciences de Paris* 157: 325–27. Translation published as (2014) 'A note on continuous functionals and analytic functions'. "Rediscovered classics" series. *Journal of Economic Inequality* 12: 153–55. [CrossRef]

Grossman, Jean Baldwin. 1982. The substitutability of natives and immigrants in production. *Review of Economics and Statistics* 64: 596–603. [CrossRef]

Hampel, Frank R. 1974. The influence curve and its role in robust estimation. *Journal of the American Statistical Association* 69: 383–93. [CrossRef]

Hampel, Frank R., Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley series in Probability and Statistics. Hoboken: John Wiley and Sons, Inc.

Huber, Peter J. 1981. *Robust Statistics*. Wiley series in Probability and Statistics. Hoboken: John Wiley and Sons, Inc.

Manacorda, Marco, Alan Manning, and Jonathan Wadsworth. 2012. The impact of immigration on the structure of wages: Theory and evidence from Britain. *Journal of the European Economic Association* 10: 120–51. [CrossRef]

Müller, Tobias, and José Ramirez. 2009. Wage inequality and segregation between native and immigrant workers in Switzerland: Evidence using matched employee-employer data. In *Occupational and Residential Segregation (Research on Economic Inequality, Volume 17)*. Edited by Yves Flückiger, Sean F. Reardon, and Jacques Silber. Bingley: Emerald Group Publishing, pp. 205–43.

OECD. 2012. *OECD Economic Surveys: Luxembourg 2012*. Paris: OECD Publishing.

Ottaviano, Gianmarco I. P., and Giovanni Peri. 2012. Rethinking the effect of immigration on wages. *Journal of the European Economic Association* 10: 152–97. [CrossRef]

Rothe, Christoph. 2010. Nonparametric estimation of distributional policy effects. *Journal of Econometrics* 155: 56–70. [CrossRef]

Saigo, Hiroshi, Jun Shao, and Randy R. Sitter. 2001. A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. *Survey Methodology* 27: 189–96.

Shorrocks, Anthony F. 1984. Inequality decomposition by population subgroups. *Econometrica* 52: 1369–85. [CrossRef]

StataCorp. 2015. *Stata Statistical Software: Release 14*. College Station: StataCorp LP.

STATEC. 2009. La structure des salaires en 2006. Bulletin du STATEC 1–2009, STATEC Institut national de la statistique et des études économiques, Luxembourg. Available online: http://www.statistiques.public.lu/catalogue-publications/bulletin-Statec/2009/PDF-Bulletin-1-2009.pdf (accessed on 3 September 2018).

Van Kerm, Philippe. 2013. Repeated half-sample bootstrap resampling. Paper presented at United Kingdom Stata Users' Group Meetings 2013, London, UK, 12–13 September 2013. Available online: https://ideas.repec.org/p/boc/usug13/10.html (accessed on 3 September 2018).

Van Kerm, Philippe, Seunghee Yu, and Chung Choe. 2017. Decomposing quantile wage gaps: A conditional likelihood approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 65: 507–27. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

Econometrics Editorial Office
E-mail: econometrics@mdpi.com
www.mdpi.com/journal/econometrics

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland

Tel: +41 61 683 77 34
Fax: +41 61 302 89 18

www.mdpi.com

ISBN 978-3-03897-367-6