Economics Bulletin

Volume 40, Issue 3

Gender-specific top incomes: are they Pareto distributed?

Terhi Ravaska
VATT Institute for Economic Research

Abstract

Estimating the Pareto alpha parameters from the top income distribution separately for women and men can provide information on gender inequality at the top (Atkinson et al. 2018). However, whether the top incomes for women and men are really Pareto distributed has not been tested. I fit a Pareto type I distribution to the Finnish total population administrative income data for the years 1995-2014 and find that for women, a Pareto distributional assumption is only plausible for certain years. Also, I find that Pareto models apply for very high incomes which are above the commonly assumed income thresholds for Pareto distribution.

This work was supported by the Academy of Finland Strategic Research Council project Work Inequality and Public Policy (number 293120). I thank the seminar participants and the referee for their helpful feedback.
Citation: Terhi Ravaska, (2020) "Gender-specific top incomes: are they Pareto distributed?", *Economics Bulletin*, Volume 40, Issue 3, pages 1994-2004
Contact: Terhi Ravaska - terhi.ravaska@vatt.fi.
Submitted: June 08, 2020. **Published:** August 08, 2020.

1 Introduction

A common distributional assumption for top incomes and wealth is the Pareto type I distribution (Atkinson et al., 2011; Cowell, 2011; Vermeulen, 2018; Alvaredo et al., 2018; Davies and Di Matteo, 2020). Pareto type I distribution is a parametric distribution with shape parameter alpha, which, as stated by Atkinson (2017, p. 140), is a "convenient first summary of the extent of income concentration". For this Pareto distribution there exist simple formulae for calculating inequality indices with the parameter estimate of alpha making it a compelling model to study the evolution of top income inequality[1]. There is also a far-reaching view that this Pareto model fits the top income data well even though many goodness-of-fit tests used do not reliably distinguish Pareto distributions from other heavy-tailed distributions (Cirillo, 2013; Jenkins, 2017).

Recently, the Pareto type I model has also been used to study gender differences at the top of the income distribution[2]. For various countries, Atkinson et al. (2018) estimate Pareto coefficients separately for women and men at the top 1% or top 5% of the joint income distribution. A ratio of these Pareto coefficients is used as an indicator of glass ceiling and specifically it tells how fast women are disappearing from the top of the distribution compared to men. They conclude that for all countries the concentration of the male top incomes is stronger and for some countries has become stronger in recent years. However, they do not test whether the top incomes are in fact Pareto-distributed.

The purpose of this note is to test whether a Pareto type I distribution is a plausible assumption for the top of the gender-specific income distribution in Finland. I use Finnish administrative data (without top-coding) including the entire Finnish population for the years 1995-2014. Administrative data reduce the estimation bias present in survey data often used in the top income literature. Also the administrative data include a comprehensive income concept so we can study the true top income distribution.

A common method to estimate shape parameter alpha is ordinary least squares (OLS) regression for a predetermined top income group (Atkinson et al., 2011; Cowell, 2011;

[1]Besides measuring inequality, the shape of the income distribution is important for the analysis of redistribution policies. Pareto alpha parameters are applied for example in calculations of optimal tax rates for top incomes, see Piketty et al. (2014) and Bac. There is also extensive literature on the models generating the Pareto tail for income and wealth distributions, for surveys see Benhabib and Bisin (2018), Gabaix (2009) and Jones (2015).

[2]Pareto coefficients have also been estimated by age in Badel et al. (2018). They find that in the countries studied (the US, Canada, Sweden and Denmark), the Pareto coefficients decrease (inequality increase) with age.

Atkinson, 2017). In the OLS regression a log of empirical survivor function is regressed on the log of income and a constant. The slope of the regression line is the shape parameter alpha. Previous literature has discussed the problems of OLS in fitting Pareto distributions (Goldstein et al., 2004; Clauset et al., 2009) so instead I follow the statistical techniques proposed in Clauset et al. (2009) and combine maximum likelihood fitting methods with a Kolmogorov–Smirnov statistic as a goodness-of-fit test to determine the lower income threshold and the corresponding shape parameter.

I find that a Pareto type I distribution is a plausible assumption for men throughout the observation period. For women, however, the Pareto assumption is plausible only for certain years and especially at the end of the observation period. For these years, the income threshold for Pareto distribution is lower for women. I also observe that the income threshold where a Pareto distribution can be assumed is much higher (99.9th percentile and above) than that commonly used when fitting Pareto distribution to top incomes. These observations suggest that women's top income distribution has become more similar to that of men but the top of the joint income distribution is still dominated by men. As Finland is at the top of several international rankings[3] of gender-equality, these results raise the question of whether women's top incomes in other countries can be assumed to be Pareto distributed a-priori either.

2 Data and methodology

The data come from Statistics Finland income registers and include comprehensive information on annual income formation including taxes paid and income transfers received. The data cover the years 1995-2014 and include all individuals living in Finland (approximately 5.2 million people). The variable of interest is the annual individual gross income, which includes wage income, self-employment income, cash transfers from the government and capital incomes excluding realized capital gains. Items that are taxed at the source (e.g. interest income from bank deposits) or capital income that is tax-exempt (e.g imputed net rent from owner-occupied housing) are not included in the data. There is no top-coding in the data. The income variables are deflated to 2014 prices using the Finnish consumer price index. Table 1 summarizes the data on the top income groups.

[3]For example World Economic Forum ranked Finland in the 4th place in the global gender-equality index for 2020.

Table 1: Summary information on the data, euros

	top 10%				top 1%		
year	observations	threshold, €	median, €	mean, €	threshold, €	median, €	mean, €
1995	501 671	49 687	61 890	72 457	100 142	123 969	151 056
1996	501 748	50 675	63 080	73 741	101 949	125 804	153 086
1997	502 997	51 390	64 167	75 918	104 822	130 541	164 332
1998	503 915	52 264	65 519	78 619	108 267	136 186	177 940
1999	504 959	52 848	66 520	83 507	112 422	143 568	215 421
2000	505 669	52 090	65 708	83 956	112 443	144 640	226 156
2001	507 129	51 811	65 467	81 446	112 036	143 915	203 724
2002	508 129	51 759	65 447	81 519	112 586	144 119	203 897
2003	509 363	52 598	66 624	82 587	115 183	148 412	203 055
2004	510 796	54 507	69 193	87 220	121 116	157 168	224 240
2005	512 827	55 233	70 138	87 396	122 594	158 509	216 894
2006	514 731	55 426	70 386	88 185	122 991	159 434	222 822
2007	516 952	55 571	70 956	89 357	126 294	164 201	227 418
2008	519 312	53 800	68 556	85 371	121 140	156 316	210 631
2009	521 575	54 103	68 971	84 477	120 455	153 781	199 198
2010	523 796	54 463	69 359	85 546	121 531	156 151	205 855
2011	525 549	52 709	67 280	83 624	118 558	153 038	205 944
2012	528 137	51 344	65 415	80 144	114 530	146 244	188 878
2013	530 449	50 645	64 585	79 200	112 721	143 851	187 773
2014	532 383	50 178	63 960	78 267	111 522	142 713	184 737

The cumulative distribution function (CDF) of a Pareto type I distribution[4] is

$$F(y) = 1 - (\frac{y}{y_{min}})^{-\alpha}, \text{ when } y > y_{min}. \tag{1}$$

In the above notation, y denotes income, $y_{min} > 0$ is the threshold where the Pareto assumption is valid, and the shape parameter, alpha, $\alpha > 1$ describes the heaviness of the tail distribution. The smaller the α is, the greater the heaviness and the larger the top income inequality.

The α parameters are estimated from the gender-specific income distributions with the estimation method of maximum likelihood (ML). The maximum likelihood estimator is

$$\hat{\alpha} = n \left[\sum_{i=1}^{n} ln \frac{x_i}{x_{min}} \right]^{-1}, \tag{2}$$

where n is the number of observations. The method of maximum likelihood gives consistent parameter estimates in the limit of a large sample size. The asymptotic standard error for the estimated alpha parameter is approximated by $\sqrt{\hat{\alpha}^2/n}$.

[4]More information on the properties of Pareto distributions is found in Arnold (2015) and on statistical methods for distributional analysis in Cowell and Flachaire (2015).

As seen in equations 1 and 2 the Pareto Type 1 distribution is valid only above a certain threshold. If the threshold is wrong, the model is misspecified. There is a bias-variance trade-off where too low a threshold leads to a biased estimate and too high a threshold increases the variance. To find the optimal threshold, where the bias is minimized, I estimate the α parameters for varying thresholds and use graphical tools and statistical testing to determine the optimal value.

To determine the right threshold y_{min}, a typical first step is to plot α's against different thresholds and choose the optimal threshold as the minimum income level beyond which the plot is horizontal[5]. However, as convincingly discussed in Clauset et al. (2009) and Cirillo (2013), the graphical tools are not a sufficient way to determine Pareto distribution. I complement the analysis with statistical testing and utilize the Kolmogorov-Smirnov (KS) statistic as a goodness-of-fit test[6]. This statistic measures the distance between the cumulative distribution functions of the fitted and the empirical distribution, that is,

$$D = \max_{y \geq y_{min}} |F(y) - P(y)| \tag{3}$$

where $F(y)$ is the CDF of the empirical data with lower threshold at y_{min}, and $P(y)$ is the CDF of the Pareto model with the best fit over the same range. The optimal threshold is the value that minimizes this distance D. I report the p-values of this test statistic in table 2.

3 Results

Figure 1 describes the gender composition and income shares at the top of the joint income distribution. During the years 1995-2014, the share of women in the top 10% of the income distribution was 28.6% on average. The share of women decreases rapidly toward the very top of the income distribution. In 1995, the share of women in the top 1% was 15.4%, and this has increased over time to 21.2% in 2014 [7]. This under-representation of women translates to different gender-specific tail distributions.

Figure 2 presents the estimated α's against different thresholds for selected years[8]. Except for 1995, the α's estimated from the men's distribution become approximately horizontal

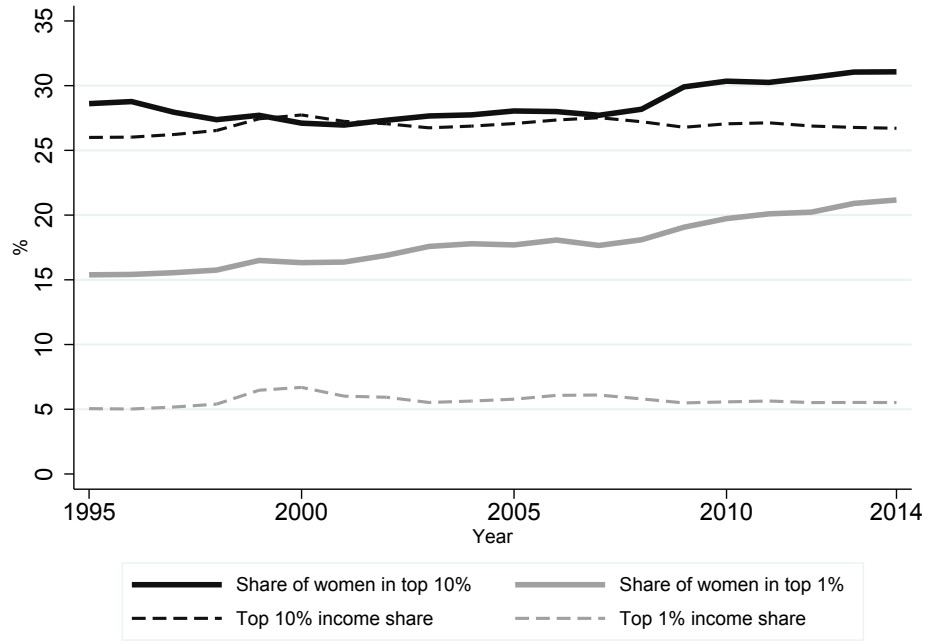
[5]The plot should be horizontal as the α parameter is constant in Pareto distribution above the right threshold.

[6]The estimation was run with Stata 16 using software by Jenkins and Van Kerm (2015) and the KS statistic was calculated with Stata package smirnov.

[7]In the top 0.1% the share of women was 15.0% in 1995 and 16.9% in 2014.

[8]Figures for other years are available from the author.

Figure 1: Share of women and income shares at the top of the joint income distribution



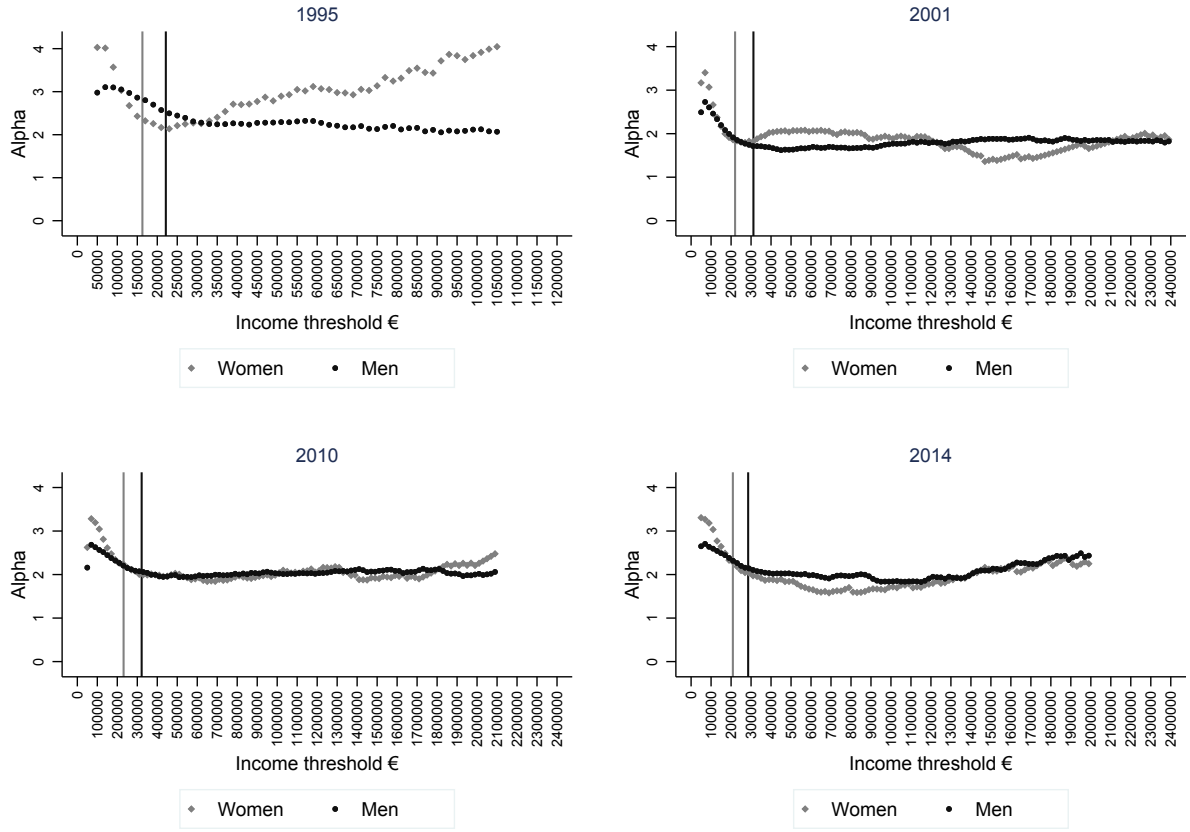
Notes: Income concept individual gross income excluding realized capital gains.

in the limit of the top 0.1% (black vertical line) male earners, which is much higher than the 95th or 99th percentiles, which are commonly assumed as the Pareto threshold. For 1995, the plausible lower threshold for the men's distribution is even greater than the limit of the top 0.1%.

For men, graphically, the Pareto type I distribution provides a plausible fit throughout the observation period. From figure 2 we can conclude that for women, the Pareto assumption is clearly rejected for 1995, as the estimated α's are increasing against the thresholds. For other years, it is more difficult to reject the Pareto assumption. Even though the α's show a slight increase at the very top, in this part, there are also very few observations. Based on the visual inspection, I reject the Pareto assumption for all of the years (in the period of 1995-2014) in which the plots for α's are not clearly horizontally aligned. For women, a Pareto type I distribution is a plausible fit for the years 2001, 2004-2006 and 2010-2014. For these years the KS goodness-of-fit statistic is calculated.

Table 2 shows the results from the ML estimation. I follow Clauset et al. (2009) and conclude that the Pareto distribution assumption should not be rejected when the p-value for the KS-statistic is greater than 0.1. I also require that the p-values be above this threshold for the subsequent income thresholds to avoid choosing too low of a threshold due to statistical fluctuations.

Figure 2: Alpha parameter estimates by threshold



Notes: Vertical lines represent the top 0.1% thresholds, grey line for women's income distribution and black for men. The α parameters are estimated with Maximum Likelihood. Note that for year 1995 the horizontal axis is different than for other years as the income distribution was less dispersed in that year.

The second column of the table 2 shows the optimal threshold percentile from the gender-specific income distribution. The threshold percentiles vary between 99.86th and 99.98th and the estimated α (fourth column) range from 1.31 to 2.31. The economic significance of the alpha estimates are also be presented with Gini coefficients for the top incomes, which for Pareto type I distribution are calculated with formulae $1/(2\alpha - 1)$ (Arnold, 2015). This shows that the variation in α translates to wide range of Gini coefficients. The tail of the distribution is heavier (more inequality) for men for most years (excluding the years 2005, 2011 and 2014).

The evolution of top income inequality (for men) at the beginning of the observation period follows the inequality evolution for total population. At the turn of the century, the Finnish economy grew rapidly at the same time increasing the overall Gini from 0.26 to 0.31 (Jäntti et al., 2010). Also the top income shares increased rapidly. For example

for the very top, in 2000, the share of the (factor) income share of the top 1% was 9.13%, whereas in 1995, it was 5.86% . This economic growth spurt especially benefited men, and their top income inequality was highest in this period. For the period after 2000 the overall Gini has remained stable while at the top the Gini evolution has been more volatile.

The central result from table 2 is that the Pareto assumption can hold for women only for certain years and only for very high incomes. However, the increased representation of women at the top over time has translated to women's top income distribution to have fatter tails similar to men. Nevertheless, for the whole period or for larger top income groups the direct comparison of the alpha parameters between genders is not possible and thus we need more ample measures of the evolution of glass-ceiling at the top of the income distribution. Ravaska (2018) shows that other indicators, such as women's income composition and representation in high-paying occupations, have changed over this period in Finland explaining the evolution in top incomes. Women have gained ground at the top and the glass-ceiling in this sense has got thinner.

Table 2: Maximum likelihood estimation of Pareto alphas and optimal threshold

Table 2.A Women

Year	Percentile for threshold	α (s.e.)	KS p-value	Gini
2001	99.96	1.932 (0.0782)	0.292	0.349
2004	99.90	1.715 (0.0453)	0.653	0.412
2005	99.91	1.874 (0.0516)	0.483	0.364
2006	99.90	1.919 (0.0513)	0.344	0.352
2010	99.95	2.046 (0.0708)	0.464	0.323
2011	99.94	1.974 (0.0666)	0.454	0.339
2012	99.94	2.173 (0.0671)	0.116	0.299
2013	99.94	2.127 (0.0655)	0.222	0.307
2014	99.94	2.092 (0.0654)	0.150	0.314

Table 2.B Men

Year	Percentile for threshold	α (s.e.)	KS p-value	Gini
1995	99.95	2.310 (0.0558)	0.535	0.276
1996	99.97	2.220 (0.0633)	0.363	0.291
1997	99.95	2.060 (0.0462)	0.385	0.321
1998	99.93	1.885 (0.0370)	0.688	0.361
1999	99.98	1.317 (0.0466)	0.283	0.612
2000	99.98	1.324 (0.0453)	0.161	0.607
2001	99.95	1.622 (0.0373)	0.154	0.446
2002	99.89	1.793 (0.0282)	0.213	0.387
2003	99.97	1.733 (0.0553)	0.150	0.405
2004	99.96	1.627 (0.0451)	0.586	0.444
2005	99.87	1.983 (0.0283)	0.345	0.337
2006	99.90	1.787 (0.0294)	0.451	0.388
2007	99.86	1.905 (0.0259)	0.202	0.356
2008	99.91	1.988 (0.0347)	0.231	0.336
2009	99.89	2.153 (0.0345)	0.229	0.302
2010	99.94	1.975 (0.0416)	0.571	0.339
2011	99.89	1.985 (0.0306)	0.240	0.337
2012	99.95	2.010 (0.0480)	0.992	0.331
2013	99.94	1.993 (0.0409)	0.423	0.335
2014	99.91	2.114 (0.0371)	0.125	0.310

The estimation was run separately for income thresholds above 50 000 euros with 20 000 euros steps. Table presents parameter values for the optimal fit to the gender specific top income distribution.

4 Conclusion

In this note, I have shown that a Pareto type I distribution is a reasonable assumption for men's annual income distribution in Finland for the years 1995-2014, but for women, the Pareto assumption is valid only for certain years. I also showed that a Pareto type I distribution is only a plausible assumption for very high incomes exceeding the 99.9th percentile. The adequacy of the Pareto distribution was first assessed visually and then using maximum likelihood methods and Kolmogorov-Smirnov goodness-of-fit tests.

While the use of Pareto type I model is wide-spread due to its simplicity in economics and particularly in studies on top incomes, some other power laws might provide improved fit to the top income data. It was beyond the scope of this note to test alternative distributional assumptions. Jenkins (2017) has shown that for UK data, a Pareto type II distribution fits the top income data better, and this distribution might also provide useful information about gender-specific distributions.

References

Alvaredo, F., Atkinson, A. B., and Morelli, S. (2018) "Top wealth shares in the UK over more than a century" *Journal of Public Economics* **162**, 26 – 47.

Arnold, B. (2015) *Pareto distributions,* Taylor & Francis group 2nd edition.

Atkinson, A. (2017) "Pareto and the Upper Tail of the Income Distribution in the UK: 1799 to the Present" *Economica* **84**(**334**), 129–156.

Atkinson, A. B., Casarico, A., and Voitchovsky, S. (2018) "Top Incomes and the Gender Divide" *Journal of Economic Inequality* **16**, 225–256.

Atkinson, A. B., Piketty, T., and Saez, E. (2011) "Top Incomes in the Long Run of History" *Journal of Economic Literature* **49**(**1**), 3–71.

Badel, A., Daly, M., Huggett, M., and Nybom, M. (2018) "Top Earners: Cross-Country Facts" *Federal Reserve Bank of St. Louis Review, Third Quarter 2018,* **100**(**3**), 237–57.

Benhabib, J. and Bisin, A. (2018) "Skewed Wealth Distributions: Theory and Empirics" *Journal of Economic Literature* **56**(**4**), 1261–91.

Cirillo, P. (2013) "Are your data really Pareto distributed?" *Physica A: Statistical Mechanics and its Applications* **392**(**23**), 5947–5962.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009) "Power-Law Distributions in Empirical Data" *SIAM Review* **51**(**4**), 661–703.

Cowell, F. (2011) *Measuring Inequality,* Oxford University Press 3rd edition.

Cowell, F. A. and Flachaire, E. (2015) "Statistical Methods for Distributional Analysis" In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution* volume 2 of *Handbook of Income Distribution* pages 359 – 465 Elsevier.

Davies, J. and Di Matteo, L. (2020) "Long Run Canadian Wealth Inequality in International Context" *The Review of Income and Wealth* forthcoming.

Gabaix, X. (2009) "Power Laws in Economics and Finance" *Annual Review of Economics* **1**(**1**), 255–294.

Goldstein, M., Morris, S., and Yen, G. (2004) "Problems with fitting to the power-law distribution" *The European Physical Journal B* **41**, 255–258.

Jäntti, M., Riihelä, M., Sullström, R., and Tuomala, M. (2010) "The trends in top income shares in Finland 1966–2007" In Atkinson, A. B. and Piketty, T., editors, *Top Incomes: Global Perspective* chapter 8, Oxford University Press.

Jenkins, S. (2017) "Pareto Models, Top Incomes and Recent Trends in UK Income Inequality" *Economica* **84**, 261–289.

Jenkins, S. and Van Kerm, P. (2015) Paretofit: Stata module to fit a type 1 Pareto distribution Retrieved from https://EconPapers.repec.org/RePEc:boc:bocode:s456832 (accessed March 3, 2020).

Jones, C. I. (2015) "Pareto and Piketty: The Macroeconomics of Top Income and Wealth Inequality" *Journal of Economic Perspectives* **29**(**1**), 29–46.

Piketty, T., Saez, E., and Stantcheva, S. (2014) "Optimal Taxation of Top Labor Incomes: A Tale of Three Elasticities" *American Economic Journal: Economic Policy* **6**(**1**), 230–71.

Ravaska, T. (2018) "Top incomes and income dynamics from a gender perspective: Evidence from Finland 1995–2012" Labour Institute for Economic Research, Working Papers 321.

Vermeulen, P. (2018) "How Fat is the Top Tail of the Wealth Distribution?" *The Review of Income and Wealth* **64**(**2**), 357–387.

World Economic Forum "Global Gender Gap Report 2020" Retrieved from https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality, (accessed March 3, 2020).