WORKING PAPER 1921

Rank correction: a new approach to differential nonresponse in wealth survey data

Rafael Wildauer and Jakob Kapeller

November 2019

POST-KEYNESIAN ECONOMICS SOCIETY

Rank Correction: A New Approach to Differential Nonresponse in Wealth Survey Data

Rafael Wildauer[1] and Jakob Kapeller[2]

[1]University of Greenwich
[2]University Duisburg-Essen and Johannes Kepler University Linz

November, 2019

Abstract

This paper is concerned with the problem of modelling the tail of the wealth distribution with survey data in the context of differential nonresponse. In order to deal with the problem post data collection, it is standard practice to combine wealth survey data with observations from rich lists and then fit a Pareto tail. In contrast, our approach does not require information about individual wealth holdings from rich lists and is thus applicable in situations where such information is not available. Applying the procedure to wealth survey data (HFCS, SCF, WAS) yields estimates of top wealth shares, which are closely in line with estimates from the World Inequality Database and thus represent a likely improvement over the raw survey data.

Keywords: differential nonresponse, Pareto tail, post data collection, survey data, wealth distribution

JEL Classification: D31, C46, C81

The usual disclaimer applies. We thank Rob Calvert Jump, Ines Heck, Karsten Köhler and Stephan Steinerberger for useful comments and discussions.

1 Introduction

The last decade saw the publication of several novel data sources suitable for studying the distribution of wealth. Among these are the World Inequality Database (`www.wid.world`), the Household Finance and Consumption Survey (HFCS) carried out under the auspices of the ECB, the UK's Wealth and Asset Survey (WAS) as well as efforts to use data leaks on offshore wealth holdings Alstadsæter, Johannesen, and Zucman (2019). For the United States the Survey of Consumer Finances (SCF) has been conducted regularly and consistently since 1989 and is considered as the most reliable source for assessing the distribution of private wealth. In Europe, three waves of the HFCS (2011, 2014, 2017) have already been conducted; however, currently only the first two waves are available to the research community and provide information on the distribution of wealth for 15 (wave 1) respectively 20 (wave 2) EU countries. For many of these countries this data source is a true novelty as reliable alternative data sources for assessing the distribution of private wealth have not been available before. In addition, recent works by Piketty, Saez, and Zucman (2016) and Saez and Zucman (2016) are of special relevance as they move towards producing data on wealth that are consistent with micro (Survey of Consumer Finances, SCF) as well as macro (Financial Accounts) sources.

High quality data on the distribution of wealth is crucial not only for better understanding the dynamics of growth and accumulation, but also to allow for a better informed public debate on distributional issues and to assess the role of private wealth in terms of tax policy. Traditionally, two forms of micro-data have been used to obtain information on the distribution of wealth: data from surveys such as the HFCS or the SCF as well as data from administrative sources, especially tax authorities. While traditionally conceived as rival approaches, which both come with their own limitations (Piketty, 2014), Saez and Zucman (2016, p. 569) point out that survey and administrative data can be used as complements in order to derive a more detailed and fine-grained assessment of the distribution of wealth. Indeed the fruitful effort of constructing Distributional National Accounts (Piketty et al., 2016) relies on a combination of survey, tax and national accounts data to arrive at a picture of the distribution of wealth that is as accurate as possible. Nevertheless, an obvious backdrop of this strategy is that administrative (tax) data is often not available for research purposes. In these cases, surveys are the key means to collect information about the distribution of household wealth.

Collecting and applying survey data comes with challenges. For one, estimates derived from surveys that come without a suitable oversampling strategy suffer from median-bias[1] and thereby typically underestimate the share of wealth held in the tail of the distribution (Eckerstorfer et al., 2016). For another, the probability of participating in such surveys is negatively correlated with household wealth itself, a phenomenon known as differential nonresponse. The evidence for differential nonresponse is compelling and can be illustrated with reference to the SCF, where tax data on capital incomes are used to identify affluent households prior to data collection. While the response rate in the stratified random sample is about 70%, it sharply decreases for the so-called list sample of affluent households, which are ex ante identified based on tax records. Here, even the poorest stratum has a response probability of only 50%, which further decreases to 12% for the stratum of the wealthiest households (Bricker, Henriques, Krimmel, & Sabelhaus, 2016, p. 282). Similarly, D'Alessio and Faiella (2002) report a response rate of 26% for the lowest wealth group which declines to 9% in the highest wealth group when in 1998 anonymized data from a commercial bank was used to identify affluent individuals in an oversampling effort for the Italian wealth survey. For the HFCS Osier (2016) emphasizes that nonresponse rates are not random and that additional data especially on income or wealth would be desirable to improve sample designs.

In this paper we focus on the issue of differential nonresponse. In practical terms there are three approaches to this problem: The first, is to do nothing, use the data as it is and to assume that the efforts made by the administrators of the survey were sufficient to deal with the problem. Most importantly this would require a strong over-sampling strategy where information which is available already prior to data collection, is used to identify wealthy households and include a disproportionate amount of them in the gross sample to ensure enough responses despite a lower

[1] Median bias refers to a situation where the median of the sampling distribution of the Pareto alpha estimator is different from the population parameter which the researcher aims to estimate.

response probability. The second approach is to fit a Pareto distribution to the tail of the survey data and use the estimated distribution to describe the tail instead of the tail observations (Jayadev, 2008; Eckerstorfer et al., 2016). The third approach extends the second by adding journalists' rich lists like the "Forbes 400" for the US or the "Manager Magazin" for Germany to the original survey data. The resulting data set is then used to estimate a Pareto distribution (Vermeulen, 2018) and subsequently the fitted distribution is applied to describe the tail of the wealth distribution.

This paper aims to add a fourth approach to this list, the rank correction approach. It relies on much less external information compared to the rich list approach and yields better results than the first two. This means the rank correction approach can be used when the rich list approach is not feasible due to the lack or poor quality (Capehart, 2014; Kopczuk, 2015) of rich list data. This advantage will become apparent when applying the method to actual data (section 5). In addition the rank correction approach can be regarded as a substitute and robustness check for the rich list approach with the main advantage that all modelling assumptions are made in a transparent and explicit way. This is in contrast to the reliance on rich lists where key methodological aspects as well as differences across countries are left in the dark.

The core idea of the rank correction approach is to correct the ranks of the sample observations (i.e. the cumulative sum of the survey weights) in order to take into account that the most affluent households are much less likely to be included in the sample. This preserves the linearity of the relationship between logarithms of household wealth and rank (cumulative weights) underlying the Pareto distribution, which is exploited when fitting the distribution to the data. We will demonstrate that this simple adjustment is able to substantially reduce the bias from differential nonresponse when fitting a Pareto distribution to the tail of the available survey data.

Applying such an approach to the second wave of the HFCS data shows that the average estimate of the Pareto tail index declines from 2.4 obtained from a baseline regression without rank correction to 1.9 after implementing the rank correction procedure. Using these Pareto tails to replace the tail from the survey leads to an average increase of aggregate wealth by 5%. Correspondingly, the average top 1% wealth share increases from 16.8% to 20% and the average top 0.1% share from 5% to 8% . Comparing these results with exogenous sources on the distribution of wealth indicates that the rank correction procedure improves upon raw survey measures. For example the WID provides top 1% wealth shares for the US, France and the UK (37%, 23.4% and 19.9%) which compare very well with the rank correction approach (37.6%, 22.6% and 17.3%). This outcome represents a clear improvement relative to the top wealth shares obtained from the raw survey data (35.4%, 18.7% and 15.1%).

The rest of the paper is organised as follows. Section 2 introduces the rank correction approach. Section 3 analyses the performance of the rank correction procedure by means of Monte Carlo simulations. In Section 4 we introduce two rules of thumb which are suitable to guide the implementation of the rank correction approach in practice. Section 5 contains an application to data from the HFCS, SCF and WAS. Section 6 contains a summary and concludes.

2 The Rank Correction Approach

2.1 Fitting Pareto tails to wealth survey data: the standard approach

The standard approach of fitting a Pareto tail to wealth survey data is to fit the complementary cumulative distribution function (CCDF) of the Pareto distribution to the empirical CCDF derived from the available sample. The theoretical CCDF for a random variable X following a type I Pareto distribution is defined as follows [2]:

$$CCDF_T(x_i) = Pr(X > x_i) = \left(\frac{x_m}{x_i}\right)^{\alpha} \tag{1}$$

Let's assume a sample of households with net wealth $x = (x_1, \ldots, x_n)$ and corresponding survey weights $w = (w_1, \ldots, w_n)$, where the number of households represented by the available sample is defined as $N = \sum_{i=1}^{n} w_i$. Arranging the data in descending order (i.e., from the most to the least

[2] Throughout the paper we refer to type I Pareto distributions when we talk about Pareto distributions.

affluent observation) yields a data vector denoted as $x_d = (x_{(1)}, \ldots, x_{(n)})$ with the corresponding vector of weights $w_d = (w_{(1)}, \ldots, w_{(n)})$. Then the empirical CCDF is defined as:

$$CCDF(x_{(i)}) = \frac{\sum\limits_{1 \leq j \leq i} w_{(j)}}{N} \tag{2}$$

Combining the theoretical and empirical CCDFs provides the basis for a linear regression:

$$\ln\left(\sum_{1 \leq j \leq i} w_{(j)}\right) = c_1 - \alpha \ln(x_{(i)}) + \epsilon_i \tag{3}$$

where $c_1 = \ln(N) + \alpha \ln(x_m)$. Equation (3) is then estimated by OLS. The estimated Pareto tail index α is then used to describe household wealth above x_m. Vermeulen (2018) goes one step further and incorporates Gabaix and Ibragimov's (2011) bias correction into this standard estimator leading to:

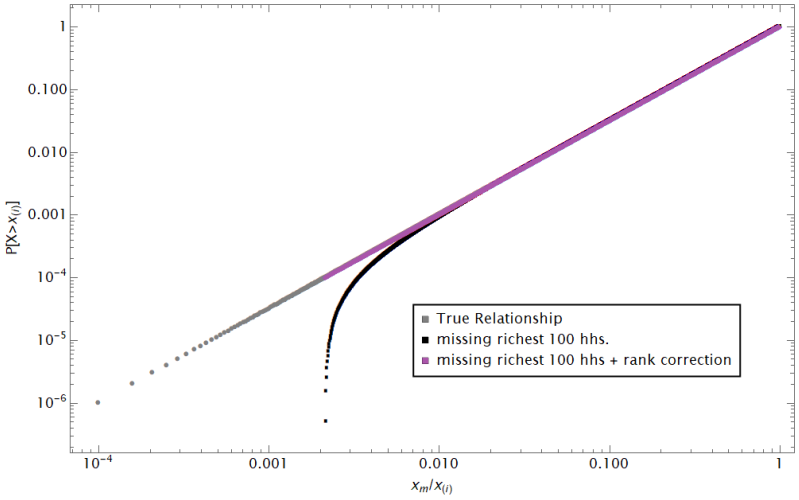
$$\ln\left((i - 0.5)\frac{\bar{N}_i}{\bar{N}}\right) = c_2 - \alpha \ln(w_{(i)}) + \epsilon_i \tag{4}$$

where $\bar{N}_i = \frac{1}{n}\sum_{k=1}^{i} w_{(k)}$ and $\bar{N}$ is the average weight. See Vermeulen's (2018) online appendix for the detailed derivation and Wildauer and Kapeller (2019) for a simpler alternative and a general discussion of the appropriate definition of the empirical CCDF.

2.2 A graphical motivation of the rank correction approach

Fitting Pareto distributions to wealth survey data without rich lists can be used to improve estimates for total wealth or the amount of wealth held in the tail of the distribution (Eckerstorfer et al., 2016; Vermeulen, 2018). However, while fitting a Pareto distribution to survey data which suffers from differential unit nonresponse improves the estimate of the tail wealth compared to an estimate which is purely based on the survey data, the Pareto model still underestimates the actual tail wealth between 17% and 4% (Vermeulen, 2018, p. 377). The reason for the bad performance of the OLS estimator of the Pareto model in a situation of differential unit nonresponse is that the log-linear relationship between the empirical and theoretical CCDF breaks down. The rank correction and the rich list approach aim to restore that linear relationship with the crucial difference that the rank correction approach requires much less additional information. We will proceed by illustrating the rank correction approach by means of a simple example.

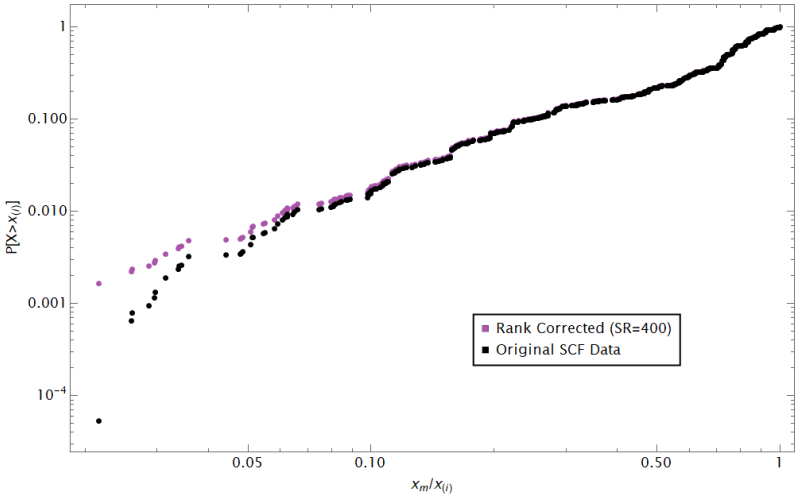
Figure 1: Motivating the rank correction approach



Let's assume we are concerned with a hypothetical tail population of 1 million households (N_{T1}) which are described by the Pareto distribution with minimum level of wealth of €1 million (x_m) and a shape parameter equal to 1.5 ($\alpha = 1.5$). In Figure 1 we plot the empirical CCDF based on equation 2 against the ratio $\frac{x_m}{x_{(i)}}$ representing the theoretical CCDF from equation 1 on a log-log scale. In grey we have the log linear population relationship between $\frac{x_m}{x_{(i)}}$ and $CCDF(x_{(i)})$

where the slope of that line represents the shape parameter α. The black dots represent the same relationship with the additional assumption that the most affluent 100 households are not observed and hence are not included in the computation of the empirical CCDF. This illustration shows that wealth survey data, which does not account for the very top of the wealth distribution gives an inaccurate representation of the wealth distribution in log-linear terms. In contrast, the purple dots take the fact that the most affluent 100 observations are not observed into account by correcting the weights accordingly and, hence, retain the log-linear relationship between $\frac{x_m}{x_{(i)}}$ and $CCDF(x_{(i)})$. On a theoretical level, the rich list approach assumes that the researcher has access to a rich list which adequately represents the richest observations not included in the survey. That means by relying on rich lists the researcher assumes (part of) the problem away. In contrast, the rank correction approach (purple) does not require any information about the missing households other than how many of them are missing.

Figure 2: Rank correction applied to the SCF



The example in Figure 1 is a highly stylized illustration of why differential nonresponse leads to the breakdown of the log-linear relationship between the theoretical and empirical CCDF. In contrast, Figure 2 provides an illustration using data from the Survey of Consumer Finances (2016 wave). The SCF's design is tailored to protect the privacy of its participants and, hence, explicitly excludes individuals from the Forbes 400 List. Plotting the relationship between $\frac{x_m}{x_{(i)}}$ and $CCDF(x_{(i)})$ for the observations representing the richest 250,000 US households based on the original SCF data reveals the breakdown of the log-linear relationship due to the exclusion of these richest 400 households. However, after adjusting the weights for the omission of the top 400 household, the purple dots again conform to a log-linear relationship. Therefore correcting the survey weights by taking the missing observations at the very top into account, will lead to improved estimates of the Pareto shape parameter.

In practice surveys might not only suffer from the explicit exclusion of super rich households from the target population due to privacy concerns but also from more general forms of differential nonresponse throughout the upper tail of the distribution. The Monte Carlo simulations in section (3) will address this general problem in greater detail.

2.3 Deriving the rank correction estimator

The rank correction estimator is derived in three steps. First, we incorporate Gabaix and Ibragimov's (2011) (G&I from here on) argument that it has long been known that OLS estimation of equation (3) yields a biased estimate of the shape parameter α (e.g. Aigner & Goldberger, 1970). G&I show that subtracting the value 1/2 from $\sum_{j=1}^{i} w_{(j)}$ will eliminate this bias. However, G&I also assume that $w_i = w_j = 1$ so that the expression $\sum_{j=1}^{i} w_{(j)}$ is equivalent to the rank of observation i as represented by the index number $(i) = (1), \ldots, (n)$. Therefore, extending the bias correction proposed by G&I to complex survey weights corresponds to computing the empirical CCDF in the

following way:

$$CCDF(x_{(i)})_{G\&I} = \frac{\left(\sum_{1\leq j\leq i} w_{(j)}\right) - 0.5w_{(j)}}{N} \tag{5}$$

We can reformulate $CCDF(x_{(i)})_{G\&I}$ after defining $w_{(0)} = 0$:

$$CCDF(x_{(i)})_{AV} = \frac{\left(\sum_{1\leq j\leq i} w_{(j)}\right) - 0.5w_{(j)}}{N} = \frac{2\left(\sum_{1\leq j\leq i} w_{(j)}\right) - w_{(j)}}{2N} = \frac{\sum_{1\leq j\leq i} w_{(j-1)} + \sum_{1\leq j\leq i} w_{(j)}}{2N} \tag{6}$$

Wildauer and Kapeller (2019) show that $CCDF(x_{(i)})_{AV}$ is the average between the empirical CCDF based on a data vector in descending order and the empirical CCDF based on a data vector in ascending order. This provides a simple interpretation of G&I's bias correction in the context of complex survey weights. The interested reader is referred to that paper.

The second step is to account for missing observations at the top of the wealth distribution. This means the total population represented by the survey $N = \sum_{j=1}^{n} w_{(j)}$ does not include these missing observations and therefore each individual weight is scaled down proportionally. For this purpose we define a vector of adjusted weights $w'_d = (w'_{(1)}, \ldots, w'_{(n)})$ where

$$w'_{(i)} = w_{(i)}\left(1 - \frac{u}{N}\right) \tag{7}$$

and u represents the correction factor used to account for the number of super wealthy households which are excluded from the sample due to privacy concerns on the one hand, as well as for more general forms of differential nonresponse on the other. This adjustment ensures that the number of households represented by the sample is unchanged.

The third and final step is to use the correction factor u to correct the ranks of the observations at hand. This is achieved by shifting all ranks up by the according amount. So the rank corrected empirical CCDF is defined as:

$$CCDF(x_{(i)})_{RC} = \frac{\left[\sum_{j=0}^{i-1} w'_{(j)} + \sum_{j=1}^{i} w'_{(j)}\right] + 2u}{2N} \tag{8}$$

Then we can combine equation 8 with the theoretical CCDF (equation 1) and obtain the rank correction regression equation which can be estimated by OLS:

$$\ln\left(\left[\sum_{j=0}^{i-1} w'_{(j)} + \sum_{j=1}^{i} w'_{(j)}\right] + 2u\right) = c_2 - \alpha \ln(x_{(i)}) + \epsilon_i \tag{9}$$

where $c_2 = \alpha \ln(x_m) + \ln(2N)$ and $w'_{(0)} = 0$.

3 A Simulation Study

The crucial question emerging from this argument is how to choose an appropriate rank correction factor (u) to adjust the weights before estimating the shape parameter of the Pareto distribution? In this context, two major aspects are relevant for determining the rank correction factor: first, the rank correction factor should account for a possible ex ante exclusion of the richest households (e.g. due to privacy concerns). Secondly, the correction factor should address the problem that disproportionately many affluent households are missing from the sample due to differential nonresponse problems[3]. Both problems can be tackled by choosing an adequate correction factor u which consists of two parts. The first part is the number of super rich households which are excluded due

[3] In the terminology of Little and Rubin (2019) observations are missing not at random (MNAR).

to privacy concerns (SR) and the second part is a correction factor due to other forms of differential nonresponse (DNR) and thus:

$$u = SR + DNR \tag{10}$$

The purpose of this section is to show that the correction factor u can be modelled as the sum of two independent correction factors (SR and DNR) and to demonstrate which characteristics of the available data should be taken into account when determining these factors. Choosing adequate correction factors allows for substantial improvements over naive estimations of the shape parameter which ignore the privacy and the differential nonresponse problems. A rule of thumb for applications to actual survey data is presented in the next section.

3.1 Rank correction and privacy restrictions on the sample design

Vermeulen (2018) simulates a tail population of 1 million households, roughly in line with French and German samples in the second wave of the HFCS: There are 1.24 million millionaire households out of a total sample of 39.7 million housheolds in Germany and 930,000 millionaire households out of 29 million households in France. We will use N_T to denote tail populations and N_C to denote the population of the country as a whole.

Thus for our Monte Carlos simulations we follow Eckerstorfer et al. (2016) and Vermeulen (2018) and assume a country population of 40 million households ($N_C = 40 \cdot 10^6$) and a tail population of 1 million households ($N_{T1} = 10^6$), which follows a Pareto distribution with the scale parameter $x_{min} = 1,000,000$ and shape parameter $\alpha = 1.5$. We simulate net sample sizes ranging from 0.2‰ to 6‰ of the tail population which corresponds to a range of net sample sizes from approximately 200 to approximately 6000 observations. The first response mechanism we analyse only incorporates the exclusion of super rich households due to privacy concerns by setting the response probability of these households on rich lists to zero[4]. Thus we define the response mechanism as:

$$R_1(x_i) = \begin{cases} 0.4, & \text{for } x_{min} \leq x_i < x_{\text{SR}} \\ 0, & \text{for } x_{\text{SR}} \leq x_i \leq x_{max} \end{cases} \tag{11}$$

where $R_1(x_i)$ is the response probability of household i depending on its wealth (x_i), such that the general response rate is 40% for all households[5], while super-rich households show a response probability of 0%. Here x_{SR} denotes the level of wealth of the poorest household which is excluded from the sampling frame because of privacy concerns due to being listed on a rich list. x_{max} denotes the wealth of the most affluent household in the population N_{T1}.

Table 1: Size of rich lists relative to underlying survey populations

country	(1) rich list	(2) year	(3) entries	(4) hhds	(5) population (hhds)	(6) relative size
Austria	Trend	2014	100	300*	3.9 million	0.077‰
Germany	Manager Magazin	2014	517	1490	39.7 million	0.039‰
Spain	El Mundo	2012	118	309	17.4 million	0.020‰
France	Challenges	2015	500	1500*	29.0 million	0.052‰
Poland	Wprost	2019	100	300*	13.5 million	0.022‰
					average	0.042‰

The star indicates that no information on the number of households was available and hhds = 3·entries was used. Highly exhaustive rich lists relative to their country size, like in the case of Belgium, the Netherlands or the UK were ignored, in order to obtain conservative estimates.

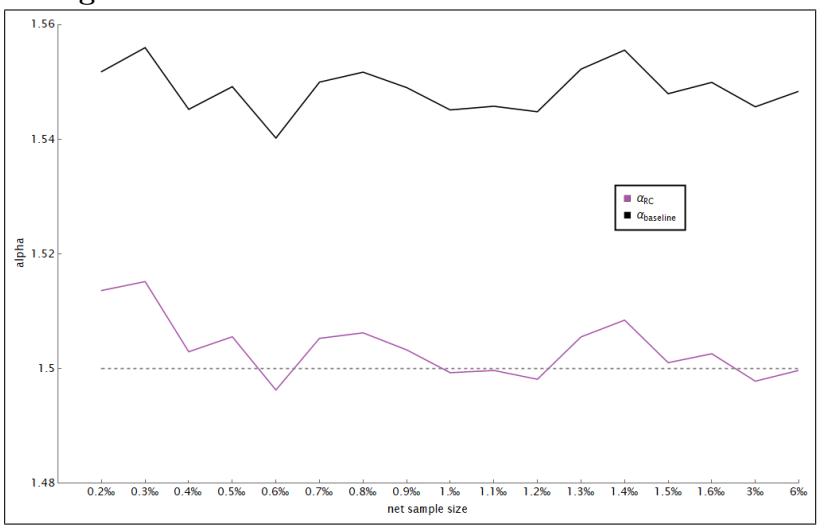
Table 1 compares the size of rich lists relative to the total population for 5 European countries. It is important to note that the number of entries on rich lists in column (3) is not equivalent to the number of households (column 4) on these rich lists. The reason is that the lists published by journalists often treat entire families as one observation, although these families often comprise

[4]While it is not true that these super rich households exhibit a zero response probability, if they are excluded from the sample by design, setting their response rate to zero is an equivalent and convenient way of incorporating such a mechanism into our simulations. This is in line with SCF practice (Kennickell & Woodburn, 1997, p. 5).

[5]This is roughly in line with available data. For example Bricker et al. (2016) report response rates for the richest strata in the SCF between 50% and 12%.

several households. For example, ownership of Volkswagen and Porsche is spread across several families in Austria and Germany and similarly the ownership of Aldi, a large German retailer, is spread across several households. In order to ensure comparability with the HFCS and the SCF, for which the household is the unit of measurement, we split these families into the number of households they represent, given the necessary information is available. If this information was not available, we assumed that each entry represents three households, which is the average number that prevails for Germany and Spain. Based on the average rich list size of 0.042‰ in Table 1, we will assume that the most affluent 0.04‰ of all households are excluded from the sample due to privacy concerns. With a total population of 40 million that is equivalent to 1600 households (i.e. $SR = 1600$) which is roughly the number of households which are on the German rich list from the Manager Magazin.

Figure 3: Simulation results: rank correction vs baseline



Rank correction estimator based on equation (9) with correction factor $u = SR = 1600$ compared to Vermeulen's (2018) baseline estimator based on equation (4) for a population of $N_{T1} = 10^6$ following a Pareto Distribution with $x_{min} = 10^6$ and $\alpha = 1.5$ and response mechanism $R_1(x_i)$, based on means over 200 draws per sample size.

Figure 3 presents simulation results based on 200 draws for each sample size from population N_{T1} with the uniform response mechanism $R_1(x_i)$, which excludes the 1600 most affluent households due to privacy concerns. We compare the performance of the rank correction estimator $\hat{\alpha}_{RC}$, based on equation (9) with correction factor $u = SR = 1600$ against Vermeulen's (2018) baseline estimator $\hat{\alpha}_{\text{baseline}}$, based on equation (4). The key result is that the rank correction estimator outperforms the baseline estimator substantially and yields estimates of the shape parameter much closer to the true population parameter value of $\alpha = 1.5$. Thus, not taking into account the exclusion of super rich households due to privacy concerns induces a substantial bias in the estimation of the shape parameter even if no other differential nonresponse problem plagues the data.

3.2 Rank correction and differential nonresponse

To illustrate how the rank correction procedure can provide more reliable estimates of the top tail of the wealth distribution in the context of general forms of differential nonresponse, we conduct a Monte Carlo simulation which makes use of Vermeulen's (2018) response mechanism. This response mechanism models the response probability as a declining function of household wealth[6].

$$R_2(x_i) = 0.903 - 0.036594 \ln(x_i) \tag{12}$$

where R_2 is the response probability of household i and x_i is that household's net wealth. This response mechanism $R_2(x_i)$ represents a situation where the sampling procedure suffers from differential nonresponse but no households are excluded due to privacy concerns as in the previous

[6]Vermeulen (2018) defines the nonresponse probability (NR) instead of the response probability (R). The formulations are equivalent since NR = 1-R

section. Separating these two arguments in our analytical exercises enables us to investigate the two problems under consideration – exclusion of potential observations due to privacy concerns and general forms of differential nonresponse - separately.

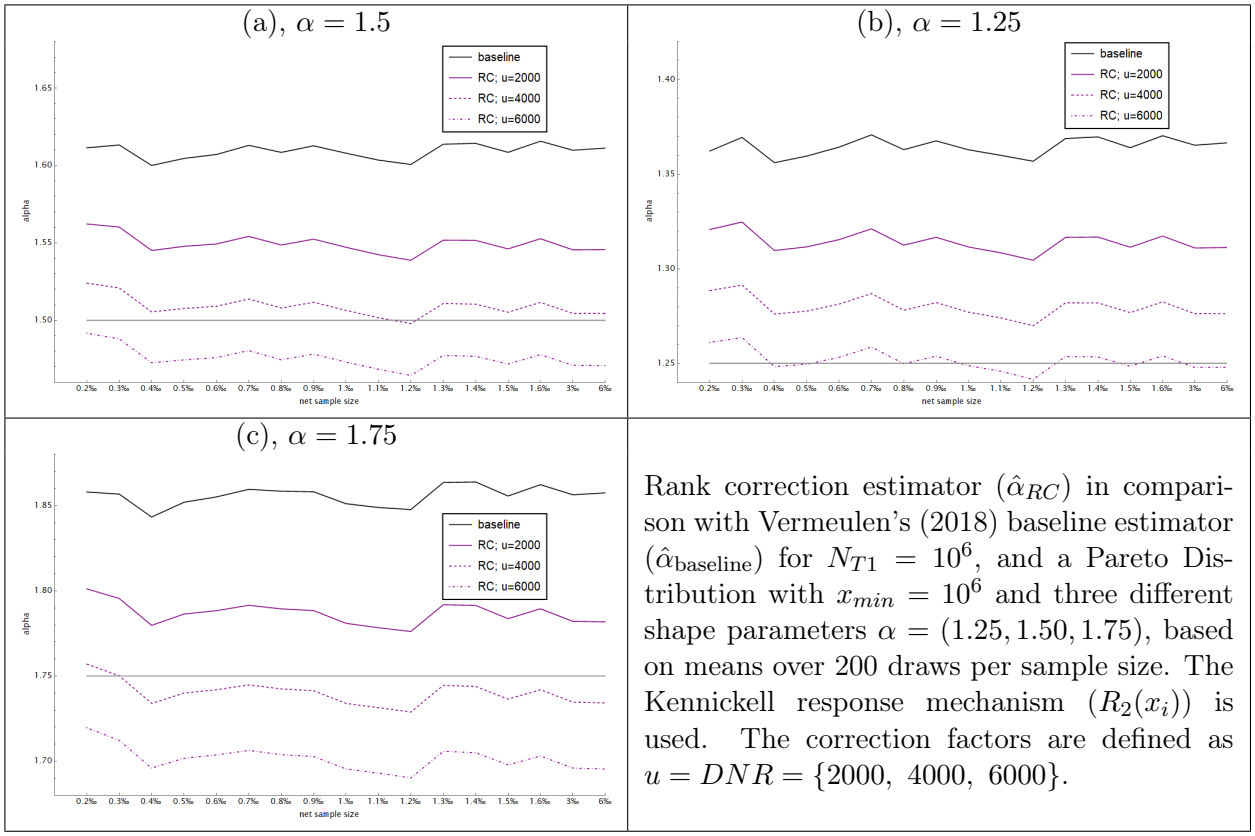
Table 2: Response probabilities for response mechanism $R_2(x_i) = 0.903 - 0.036594$

	Max	Mean	top 2000	top 4000	top 6000	top 100	top 10
α=1.25	40%	37%	19%	21%	22%	10%	4%
α=1.50	40%	37%	22%	24%	25%	15%	10%
α=1.75	40%	38%	25%	26%	27%	18%	14%

Response probabilities based on response mechanism $R_2(x_i) = 0.903 - 0.036594$ and population $N_{T1} = 10^6$ and different α's. Maximum and mean over entire population. Then response rates for the most affluent 2000, 4000, 6000, 100 and 10 households in the population.

Vermeulen's response mechanism produces response probabilities as presented in Table 2 for populations of 1 million households when assuming different shape parameters for the Pareto distribution. The main feature of R_2 is that since the response probability is a logarithmic function of a variable exhibiting a Pareto tail, it falls off rather slowly as can be seen by the fact that the maximum response probability (which corresponds to the poorest household in the population) is very close to the average across the entire population. However, for a very small number of households at the very top such as the most affluent 100 or 10 households, response rates are markedly different. The average response rates of the most affluent households are presented in steps of 2000 households. The pattern is that the 2000 most affluent households, exhibit an average response rate between 19% and 25%. These differences stem from the fact that different shape parameters also affect the wealth of super-rich households and thus their response rates. This response mechanism was estimated by Vermeulen (2018) based on data from Kennickell and Woodburn (1997), who exploited the fact that the SCF uses high quality tax data to design their sample which allows for a comparison of ex ante information on wealth with ex post data on response rates. As a consequence, this mechanism has a solid empirical basis[7].

Table 3: Simulation results for the averaged rank correction estimator



Rank correction estimator ($\hat{\alpha}_{RC}$) in comparison with Vermeulen's (2018) baseline estimator ($\hat{\alpha}_{\text{baseline}}$) for $N_{T1} = 10^6$, and a Pareto Distribution with $x_{min} = 10^6$ and three different shape parameters $\alpha = (1.25, 1.50, 1.75)$, based on means over 200 draws per sample size. The Kennickell response mechanism ($R_2(x_i)$) is used. The correction factors are defined as $u = DNR = \{2000,\ 4000,\ 6000\}$.

[7]However, concerns regarding the extent to which this mechanism applies universally across countries and time remain. This is an important area for further research and depends crucially on central banks' access to individual tax data which would enable them to implement high quality oversampling strategies.

Table 3 presents simulation results based on 200 draws for each sample size from population N_{T1} with the Vermeulen-Kennickell response mechanism ($R_2(x_i)$), without excluding any households due to privacy concerns. Results are reported for three different correction factors $u = DNR = \{2000,\ 4000,\ 6000\}$. Panels (a)-(c) of Table 3 confirm that $\hat{\alpha}_{\text{baseline}}$ exhibits a substantial bias and overestimates the shape parameter independent of the sample size or the value of the shape parameter itself. The novelty is that the rank correction estimator $\hat{\alpha}_{RC}$ exhibits a much smaller bias across all three levels of the chosen correction factor u, independent of the specific sample size or the true underlying shape parameter. The rank correction approach with $u = 4000$ yields results which are very close to the true population parameter even for different shape parameters.

If the underlying population exhibits an especially thick tail (meaning low shape parameters as in panel (b) with $\alpha = 1.25$), a choice of $u = 4000$ proves to be a conservative choice which yields an estimate of the shape parameter slightly above the true value. For populations with less thick tails (i.e. panel (c) with $\alpha = 1.75$) the rank correction approach based on $u = 4000$ only slightly overestimates the tail (i.e. underestimates α). Nevertheless in both cases, the rank correction approach produces results which exhibit a much smaller bias compared to the baseline estimator which ignores the nonresponse problem altogether. Furthermore, the simulation exercise confirms that choosing $u = 4000$ is indeed a conservative choice as it hardly over-estimates the tail thickness while improving the baseline estimate in all cases. Choosing $u = 2000$ represents an extremely conservative level of correction, which never overestimates the tail thickness.

Based on Table 3 we conclude that the rank correction approach works. It works well across a plausible set of sample sizes and shape parameters of the Pareto distribution. By choosing adequate upper and lower bounds for the correction factor u, it is possible to obtain estimates of the shape parameter which are substantially closer to the true population parameter. The remainder of the section moves on to analyse the privacy and differential nonresponse problem in combination and then adds two robustness checks. The first of these addresses the sensitivity of the results to changes in the tail population size and the second check addresses the sensitivity of the results to changes in the response mechanism.

3.3 Rank correction: privacy restrictions and differential nonresponse

While the previous two sections investigated the restrictions imposed by privacy concerns and differential nonresponse in isolation, this section demonstrates that we can choose a correction factor which addresses both problems jointly and is based on equation (10): $u = SR+DNR$. For doing so, we conduct another set of simulations for a population of 1 million households ($N_{T1} = 10^6$) following a Pareto Distribution with $x_{min} = 10^6$ and three different shape parameters $\alpha = \{1.25,\ 1.50,\ 1.75\}$.

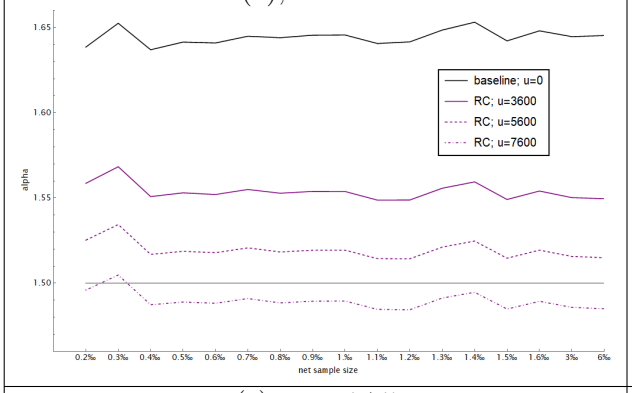
The response mechanism is a combination of $R_1(x_i)$ and $R_2(x_i)$:

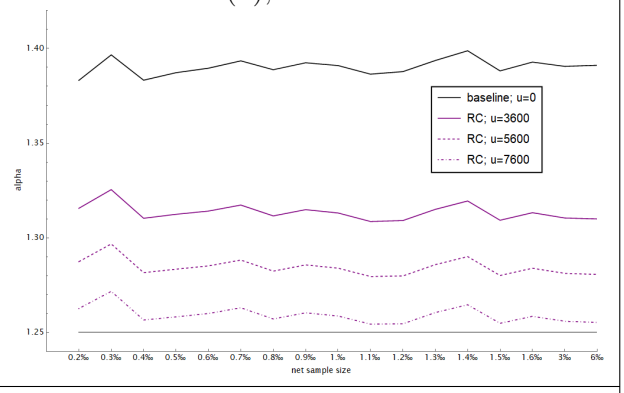
$$R_3(x_i) = \begin{cases} 0.903 - 0.036594 \ln(x_i), & \text{for } x_{min} \leq x_i < x_{\text{SR}} \\ 0, & \text{for } x_{\text{SR}} \leq x_i \leq x_{N_1} \end{cases} \tag{13}$$

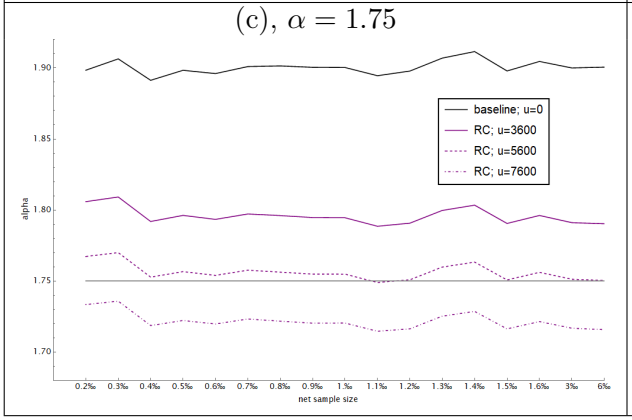
Instead of the fixed 40% response rate for all households who are not on the rich list (as is the case with $R_1(x_i)$), the new response mechanism $R_3(x_i)$, models the response rate for these households as a declining function of wealth in the same way $R_2(x_i)$ did. Results are presented in Table 4 for choosing u as $u = SR + DNR$ with $SR = 1600$ and $DNR = \{2000,\ 4000,\ 6000\}$ yielding $u = \{3600,\ 5600,\ 7600\}$.

Table 4 confirms that the problem of exclusion of super rich households due to privacy concerns and differential unit nonresponse can be tackled by specifying a correction factor which is the sum of the privacy (SR) and general differential nonresponse (DNR) correction factors. Panel (a) confirms that for a shape parameter $\alpha = 1.5$ the rank correction approach based on $u = 5600$ eliminates the bias from which baseline OLS regressions are suffering almost completely across all sample sizes. The conservative lower bound of $u = 3600$ exhibits a substantially lower bias compared to the baseline, but overestimates the shape parameter for all sample sizes. A more aggressive correction factor like $u = 7600$, slightly underestimates the shape parameter and thus overestimates the tail thickness. Panels (b) and (c) confirm that the conservative correction factor of $u = 3600$ always represents a substantial improvement over the baseline, and even an aggressive correction based on $u = 7600$ will only slightly overestimate the tail thickness in the case of relatively high shape parameters such as 1.75 but performs very well for lower shape parameters.

Table 4: Simulation results for the averaged rank correction estimator







Rank correction estimator ($\hat{\alpha}_{RC}$) in comparison with Vermeulen's (2018) baseline estimator ($\hat{\alpha}_{baseline}$) for $N_{T1} = 10^6$, and a Pareto Distribution with $x_{min} = 10^6$ and three different shape parameters $\alpha = \{1.25,\ 1.50,\ 1.75\}$, based on means over 200 draws per sample size and response mechanism ($R_3(x_i)$). The correction factors are defined as $u = DNR + SR = \{3600,\ 5600,\ 7600\}$.

.

3.4 Robustness check: varying the tail population size

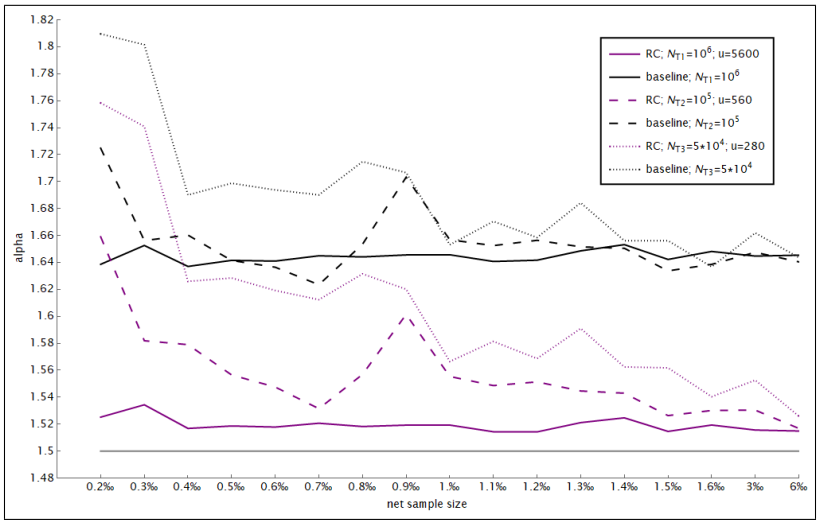
After demonstrating that the rank correction approach can be used to address the problem of privacy restrictions as well as the problem of differential nonresponse jointly, this section investigates to what extent choosing the correction factor is sensitive to the size of the tail population itself. For this purpose we conduct simulations for two populations: $N_{T2} = 10^5$ and $N_{T3} = 5 \cdot 10^4$ together with the response mechanism from the previous section ($R_3(x_i)$) which combines the problem of privacy restrictions and general differential nonresponse.

Results are presented in Figure 4. Based on the results from the previous section the moderate correction factor $u = 5600$ was chosen for the baseline specification based on the standard sample $N_{T1} = 10^6$. For the two smaller samples $N_{T2} = 10^5$ and $N_{T3} = 5 \cdot 10^4$ the correction factor was scaled proportional to the change in the population size. Thus for N_2 we used $u = 560$ and for N_3 we used $u = 280$.

The important result from Figure 4 is that smaller population sizes require larger samples in order to achieve a comparable reduction in the bias due to privacy concerns and differential nonresponse, compared to the baseline Vermeulen (2018) specification. This pattern is not restricted to the rank correction estimator but is also present in the baseline estimator and the scale of the problem is more pronounced for small samples. For example the mean over 200 draws for the point estimate of the shape parameter (α) is 1.64 for the baseline estimator applied to samples of size 0.2‰ and based on a population of $N_{T1} = 10^6$. Using the same sample size but reducing the population sizes to $N = 10^5$ and $N = 5 \cdot 10^5$ yields baseline estimates of 1.72 and 1.81 respectively. The rank correction procedure improves upon these baseline results by producing alpha estimates of 1.53, 1.66 and 1.76 respectively. These differences are highly persistent as sample sizes increase. However for the largest samples of around 6‰ these population-size induced differences disappear for practical purposes: the averaged baseline estimates we obtain are 1.65, 1.64 and 1.64 for the three population sizes of 10^6, 10^5 and $5 \cdot 10^4$. The rank correction procedure improves upon these baseline results by producing shape parameter estimates of 1.51, 1.52 and 1.53, respectively (rounded to two comma digits).

The implication of these results is twofold: First, scaling the correction factor proportional to

Figure 4: Simulation results for different population sizes



Population sizes of $N_{T1} = 10^6$, $N_{T2} = 10^5$ and $N_{T3} = 5 \cdot 10^4$ are used together with accordingly scaled correction factors: $u = 5600$, $u = 560$ and $u = 280$.

the population size is a good strategy to adapt the rank correction approach for different population sizes. Second, estimates based on small populations suffer from substantially larger biases, especially for smaller samples. This second result calls for an additional adjustment when dealing with small samples from small populations. However in the case of the ECB's HFCS there is a strong tendency that smaller countries rely on much larger samples relative to the population compared to larger countries which partially compensates for this problem. For example, the smallest countries in the second wave of the HFCS are Malta and Luxembourg with net sample sizes of 6.2‰ and 7.6‰ .

3.5 Robustness check: varying the response mechanism

The last robustness check is to assess the sensitivity of the required correction factor (u) to changes in the response mechanisms. The sensitivity is assessed by running simulations with two alternative response mechanisms. The first alternative, $R_{A1}(x_i)$, represents a downward shift of the intercept compared to the baseline mechanism $R_3(x_i)$. The second alternative, $R_{A2}(x_i)$, represents a decrease in the slope parameter compared to the baseline mechanism $R_3(x_i)$:

$$R_{A1}(x_i) = \begin{cases} 0.85 - 0.036594 \cdot \ln(x_i), & \text{for } x_{min} \leq x_i < x_{\text{topX}} \\ 0, & \text{for } x_{\text{topX}} \leq x_i \leq x_N \end{cases} \tag{14}$$

$$R_{A2}(x_i) = \begin{cases} 0.903 - 0.042 \cdot \ln(x_i), & \text{for } x_{min} \leq x_i < x_{\text{topX}} \\ 0, & \text{for } x_{\text{topX}} \leq x_i \leq x_N \end{cases} \tag{15}$$

Table 5 outlines the differences between these two alternative response mechanisms and the Vermeulen-Kennickell mechanism ($R_3(x_i)$) used so far. The first mechanism represents a downward intercept shift which means that the response probability of each household is uniformly decreased by about 5%. The second mechanism reduces the average response rate by about 8 percentage points.

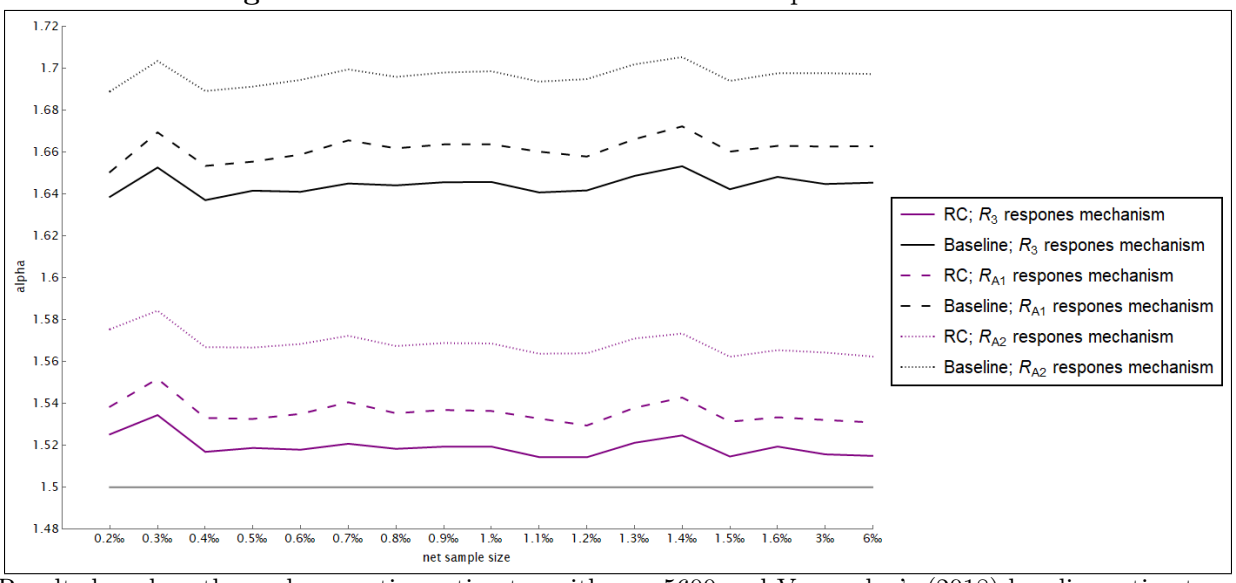
Table 5: Response probabilities for alternative response mechanisms

	Max	Mean	top 1600	1600^{th} to 3600^{th} richest household
$R_3(x_i)$	40%	37%	0%	25%
$R_{A1}(x_i)$	34%	32%	0%	20%
$R_{A2}(x_i)$	32%	29%	0%	16%

Response probabilities based on different response mechanisms and population $N_{T1} = 10^6$ and $\alpha = 1.5$. Maximum and mean over entire population. Then average response rate for the most affluent 1600 households. Then average response rates for 2000 next richest observations.

Figure 4 presents estimation results based on the three different response mechanisms. Compared to the baseline both alternative mechanisms increase the bias the estimates of the shape parameter exhibit both in the case of the standard estimator and after applying a rank correction estimator with the moderate correction factor $u = 5600$. Decreasing the slope parameter increases the bias more than the intercept shift in the response mechanism. This is due to the fact that decreasing the slope parameter widens the gap in response rates between affluent and less affluent households and thus contributes to an underestimation of the tail thickness (overestimation of the shape parameter). Nevertheless, the rank correction procedure provides results which are substantially closer to the true underlying population parameter compared to the standard approach which ignores the differential nonresponse problem.

Figure 5: Simulation results for different response mechanisms



Results based on the rank correction estimator with $u = 5600$ and Vermeulen's (2018) baseline estimator.

4 A Rule of Thumb for Real Data Applications

The key challenge in applied work is to choose an adequate value for the correction factor u. The first component of u depends directly on the number of households which are excluded from the target population (SR) because of privacy concerns. In some cases, as in the case of the Survey of Consumer Finances (SCF), the amount of excluded households is openly communicated. In other cases the number of entries on publicly available rich lists can provide guidance. In general we propose to choose SR as 0.004% of the underlying population of the country (N_C). The second component of u is the correction factor for addressing general problems of differential nonresponse throughout the sample (DNR). Based on the above simulation results we conclude that there are three factors which need to be taken into account when determining DNR:

First, DNR is proportional to the size of the tail population (N_T) as can be seen in Figure 4. That means when starting out from a moderate correction factor like $DNR = 4000$ which was established for a tail population size of 10^6 households, this factor needs to be scaled accordingly for different tail populations. Second, even if an oversampling strategy is part of the sampling process, not all oversampling strategies are equal in terms of quality and efficacy. Oversampling regions or entire states which are known to be more wealthy will be much less useful in oversampling the tail of the wealth distribution compared to oversampling schemes which rely on high quality administrative data such as tax forms[8]. The simulations were conducted under the assumption that no oversampling takes place and thus if oversampling does take place, the moderate correction factor $DNR = 4000$ as the point of departure has to be adjusted accordingly. Third, choosing DNR is insensitive to the

[8]In the second wave of the HFCS oversampling rates of millionaires are 506% in Spain, 409% in Slovakia and 329% in France. All three countries use tax data in some form. The corresponding rates are 173% for Germany, 72% for Belgium and 58% for Poland (regional income information). The picture is more dire for countries without any oversampling strategy like Italy (9%).

sample size, especially for large tail populations around 10^6 households (see Tables 3 and 4). Only for very small tail population sizes, the sample size starts to play a role. Hence, when considering small populations researchers should take the sample size into account and potentially adjust the chosen correction factor. In addition, the simulations have shown that DNR is relatively insensitive to different values of the population shape parameter of the Pareto distribution (see Tables 3 and 4). By relatively insensitive we mean that the rank correction approach always outperformed the baseline specification in our simulations and thus provides estimates which are closer to the population parameter. Lastly, the rank correction approach is relatively insensitive to changes in the response mechanism. In Figure 5 we can see that the rank correction approach consistently outperforms the baseline approach which just ignores the differential nonresponse problem at hand.

Based on these factors we propose a rule of thumb for setting DNR. The point of departure is $DNR = 4000$ which is then adjusted based on the three factors which proved to have an important impact on the results of the Monte Carlo simulations: The population size (a_P), the sample size (a_S) if populations are small and the oversampling strategy in place (a_O).

$$DNR = 4000 \cdot a_P \cdot a_S \cdot a_O \tag{16}$$

The population adjustment factor takes the size of the tail population relative to a tail population of $N_{T1} = 10^6$ into account. Setting $a_P = 1$ would refer to a tail population of 10^6 and $a_P = 0.6$ to a tail population of $6 \cdot 10^5$ and so on. The sample adjustment factor takes into account the larger differential nonresponse bias for small samples from small populations. We will discuss how to choose a_S in the next section. The oversampling adjustment factor takes into account whether an oversampling strategy was applied and if so, the quality of that scheme. For cases with no oversampling a_0 should be set to 1. For increasingly successful oversampling strategies a_0 should be adjusted downwards towards 0.

Overall our rule of thumb allows the researcher to derive a correction factor which systematically takes the characteristics of the survey at hand into account. While we demonstrated that the rank correction procedure is an improvement over simply ignoring the likely problem of differential nonresponse, we are aware that this approach cannot fully resolve the fundamental problem of taking an unknown response mechanism into account. Nevertheless, in the next section we will demonstrate the usefulness of this rule of thumb by applying it to actual household survey data.

5 Application to Wealth Survey Data

In this section the rank correction approach is applied to the second wave of the Household Finance and Consumption Survey (HFCS), the 2013 wave of the Survey of Consumer Finanes (SCF) and the fourth wave (2012-2014) of the UK's Wealth and Asset Survey (WAS). We use the aggregate netwealth measures from the HFCS (variable DN3001) and the SCF (variable networth; SCF summary dataset). The WAS exhibits an important difference compared to the HFCS and the SCF in that it also includes model based estimates of pension wealth. That means it estimates the current value of future pension claims. To make our wealth measure comparable across surveys, we exclude pension wealth in the WAS[9].

5.1 Choosing the correction factor (u) in practice

Applying the rank correction approach outlined above to actual wealth survey data, requires the researcher to determine the correction factor u. First, let's begin with the problem of privacy concerns. We argued that SR should be chosen either proportional to the population under investigation (i.e. setting $SR = 0.004\% \cdot N_C$) or with reference to the number of entries in publicly available rich lists. Table 6 presents values for SR based on both approaches contingent on whether rich lists are available (columns 2 to 4). Column (5) reports the value we implement in the subsequent analysis.

The decision whether SR is determined proportional to the population or based on rich list information is based on two principles. First if we have reliable information on the number of

[9]Thus we define netwealth = TotWlthW4 − TOTPENw4_aggr. The amount of wealth the WAS adds in form of claims on future pensions is substantial. For example the number of millionaires in wave 4 based on TotWlthW4 amounts to 2.75 million while there are 887,209 millionaires based on the netwealth variable which excludes pension wealth.

Table 6: Determining SR

country	(1) $0.004\% \cdot N_C$	(2) richlist	(3) entries	(4) households	(5) SR
Austria	155	Trend	100	300*	155
Belgium	192	De Rijkste Belgen	637	1,911*	1,911
Germany	1,587	Manager Magazin	517	1,490	1,490
Spain	697	El Mundo	118	309	309
Finland	105				105
France	1,161	Challenges	500	1,500*	1,161
Greece	171				171
Italy	988				988
Netherlands	304	Quote 500	500	1,500*	1,500
Portugal	161				161
Poland	540	Wprost	100	300*	300
US_{2013}	4,901	Forbes 400	400	1,200*	400
$UK_{2012-2014}$	1,024	Sunday Times	1,000	3,000*	3,000

The star in column (4) indicates no information on the number of households was available and hhds $= 3 \cdot$ entries was used.

households on rich lists, this information is used (Germany, Spain), otherwise we use the proportional measure except for Belgium, Netherlands, the UK and Poland. For the first three countries an unusually exhaustive rich list relative to the size of the country is available. In this context the proportional values seem to understate the extent of privacy concerns. In the case of Poland the proportional result (540) is almost twice as large as the estimated number of households (300) on the rich list and thus we choose the more conservative correction factor of 300. Finally for the US we only choose a correction factor of 400 because of the exceptionally high quality oversampling strategy in place and the explicitly communicated exclusion of the Forbes 400.

For determining DNR our proposed rule of thumb was $DNR = 4000 \cdot a_P \cdot a_S \cdot a_O$ where a_P, a_S and a_O are population, sample and oversampling adjustment factors. Columns (1) to (3) of Table 7 present three important characteristics of the country datasets in order to determine these adjustment factors: the number of observations (*Obs.*) with net wealth above the threshold of €1 million ($ 15 million in the case of the US), the number of households these observations represent (i.e. the tail population N_T) and the sample size (s) in per mille with respect to this tail population (i.e. $s = Obs/N_T \cdot 1000$). Then, the population adjustment factor a_P is defined as the ratio of the country specific tail population (N_T) to the reference tail population of 1 million households used in the Monte Carlo simulations in section 5:

$$a_P = N_T/10^6 \tag{17}$$

For three countries we deviate from this rule because we have strong concerns about how reliably these surveys are measuring the size of the tail (i.e. households with net wealth in excess of €1 million). These three countries are Greece, the Netherlands and Poland. Our suspicion stems from the fact that these three samples only contain a handful of millionaire observations as can be seen in column (1) of Table 7. Therefore we match these three countries with similar countries and assign similar tail sizes. For example we pair Greece with Spain and define the Greek tail size to be 50% of the Spanish tail, relative to the country size:

$$a_{P,GR} = \frac{N_{C,GR}}{N_{C,ES}} \cdot a_{P,ES} \cdot 0.5 \tag{18}$$

The 50% adjustment is a precautionary measure. In equivalent manner we define, the Dutch tail proportional to the German tail:

$$a_{P,NL} = \frac{N_{C,NL}}{N_{C,DE}} \cdot a_{P,DE} \tag{19}$$

as well as the Polish tail proportional to the German tail (due to a lack of a better matching country), with a 33% precautionary adjustment:

$$a_{P,PL} = \frac{N_{C,PL}}{N_{C,DE}} \cdot a_{P,DE} \cdot 0.33 \tag{20}$$

The sample adjustment factor a_S is supposed to take into account the increasing severity of the differential nonresponse bias when estimating the Pareto tail index based on small samples for small populations (see Figure 4). In order to determine a_S for the countries listed in Table 7, we used a stepwise search procedure to find the optimal correction factor u_{opt} in the simulations based on smaller tail populations $N_{T2} = 10^5$ and $N_{T3} = 5 \cdot 10^4$. This search procdure is outlined in the Appendix. We then determined the relationship between the sample size (s, in per mille) and the optimal correction factors with two regressions of the form:

$$u_{opt}/u_{base} = \beta_0 + \beta_1 \ln(s) + \epsilon \tag{21}$$

where $u_{base} = 560$ was used in the regression for the case of a tail population of $N_{T2} = 10^5$ and $u_{base} = 280$ for the case of a tail population of $N_{T3} = 5 \cdot 10^4$. These are the standard correction factors scaled proportional to the population size (i.e. instead of $u = 5600$ for $N_{T1} = 10^6$, $u = 560$ for $N_{T2} = 10^5$). We then used the fitted values from these two regressions ($\frac{\hat{u_{opt}}}{u_{base}}$), as proxies for a_S, for countries with tail population sizes close to 10^5 and $5 \cdot 10^4$. The details and regression results are outlined in the Appendix. For example in the case of Austria we have a tail population of $N_T = 129,309$ households and a sample size of $s = 0.67‰$. Based on regression (21) we obtain a fitted value and thus correction factor of 1.7 for a population of 100,000 and a sample of $s = 0.7‰$ and thus we set $a_S = 1.7$. We chose an adjustment factor of 1 for those countries which either exhibit a large tail population close to or over 1 million households (i.e. Germany, Spain, France, Italy, USA) or which exhibit sample sizes in excess of 6‰(i.e. Finland).

Table 7: Choosing DNR for HFCS and SCF data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$Obs.$	N_T	s	a_P	a_S	a_O	DNR	SR	u
Austria	86	129,304	0.665	0.13	1.7	0.8	703	155	858
Belgium	209	258,007	0.811	0.26	1.3	0.7	939	1911	2,850
Germany	379	1,243,129	0.305	1.24	1	0.6	2,984	1,490	4,474
Spain	1,272	600,989	2.117	0.6	1	0.1	240	309	549
Finland	445	53,655	8.294	0.05	1	0.1	21	105	126
France	1,638	930,511	1.76	0.93	1	0.05	186	1,161	1,347
Greece	12	14,280	0.871	0.07	7.5	0.9	1,986	171	2,157
Italy	255	717,846	0.355	0.72	1	1	2,871	988	3,859
Netherlands	31	101,740	0.305	0.24	2	1	1,903	1500	3,403
Portugal	209	78,195	2.723	0.08	1.2	0.7	263	161	424
Poland	19	45,625	0.412	0.14	3.5	0.9	1,776	300	2,076
USA_{2013}	502	482,202	1.041	0.50	1	0.01	19	400	419
$UK_{2012-2014}$	1,190	887,209	1.341	0.9	1	1	3,549	3,000	6,549

N_T is defined as the tail population of households above €1 million in the HFCS, above £1 million in the WAS and above $15 million in the SCF.

The adjustment factor a_O is supposed to take into account the varying degrees to which different countries implement oversampling strategies and the quality of exogenous data on which these oversampling strategies are based. For countries which do not implement any oversampling strategy an adjustment factor of 1 should be chosen (Italy and Netherlands). For those countries which implement oversampling strategies, the adjustment factor should be adjusted towards 0 with increasing quality of the oversampling strategy. As a consequence, countries that rely on precise individual tax data for oversampling purposes (US, France, Spain, Finland) were given adjustment factors very close to 0. The next category of countries are those which use regional income data (Germany, Belgium, Greece, Poland) or dwelling floor space (Portugal, Poland) and the fourth group are countries which only use geographic information to oversample (Austria). When it comes to judging the quality of the oversampling strategies applied and the severity of the remaining differential nonresponse bias, the number of tail observations in column (1) provides crucial information. From there it becomes clear that despite the fact that Greece or Poland use regional income and real estate price information for their oversampling design, the small number of observations above the €1 million threshold, indicates the poor performance of these oversampling strategies. Accordingly these two countries have a much higher adjustment factor of 0.9 compared to other countries relying

on similar information for their oversampling strategies. For the UK we chose an adjustment factor of 1 because the oversampling strategy is focussed solely on the top decile and based on a single stratum. Since the Pareto tail we are fitting falls entirely into this one stratum, within this stratum there is no oversampling.

Combining the three adjustment factors (columns 4 to 6) yields DNR, reported in column (7) which together with the previously discussed values for SR yield the final correction factor u, reported in column (9). It is these values for u which haven been used to produce the results which are presented in the next section.

5.2 Estimation results

Table 8 presents the results after implementing the rank correction procedure based on the correction factor (u) as defined in the previous section. Column (1) reports the Pareto tail index which is obtained from fitting a Pareto distribution to the survey data above the threshold (€1 million and £1 million for the HFCS and WAS data respectively, and $15 million for the US data) based on Vermeulen's (2018) baseline estimator $\hat{\alpha}_{baseline}$ (equation 4). Column (2) reports the results from the rank correction estimator $\hat{\alpha}_{RC}$ (equation 9), relying on the correction factors reported in column (3).

Table 8 contains two important results: First, the Pareto tail parameter based on the rank correction approach is smaller than the baseline result for all countries. This result is strongly in line with our Monte Carlo simulations. Secondly, there are substantial differences across countries with respect to how much the rank correction approach impacts the estimated Pareto alpha. This second result is also expected as the country specific correction factors take differences in size and the likely quality of the implemented oversampling strategy into account. We see the most pronounced corrections and the highest values for the tail parameter for those countries where we deemed the oversampling scheme to be of the lowest quality (and which come with the smallest number of observations in the tail). For example for Greece the estimated tail index declines from 3.62 to 1.92 and for the Netherlands from 4.5 to 3.58.

Table 8: Baseline and rank correction shape parameter (α) estimates

	(1)	(2)	(3)
	α with u=0	α with u	u
Austria	1.426	1.331	858
Belgium	2.198	1.901	2,850
Germany	1.588	1.470	4,474
Spain	1.756	1.609	549
Finland	2.085	1.900	126
France	1.631	1.567	1,347
Greece	3.626	1.923	2,157
Italy	2.416	2.156	3,859
Netherlands	4.497	3.577	3,403
Portugal	2.157	1.901	424
Poland	2.347	1.780	2,076
US_{2013}	1.830	1.650	419
$UK_{2012-2014}$	1.973	1.772	6,549

Comparing baseline and rank correction shape parameter (α) estimates for HFCS, SCF and WAS data.

The estimated alpha coefficient provides a first intuition on how the rank corrected fit of the Pareto tail will impact on the overall picture of household wealth. Table 9 provides a detailed comparison between the total amount of household wealth measured by the original survey data (column 1) and after applying the rank correction approach and replacing all households above the threshold with the estimated Pareto tail[10]. Column (3) presents the ratio of the corrected data over the original survey data. Comparing Tables 8 and 9 reveals that there is a clear pattern that countries with the smallest tail indices also exhibit the strongest corrections. This is expected since

[10]We are following Vermeulen (2018) for ease of comparability.

smaller alphas represent thicker tails. The important caveat is that despite the fact that countries with weak or non-existing oversampling strategies (Netherlands, Poland, Italy, Greece) exhibited a substantial correction of their estimated tail indices in Table 8, due to the still relatively thin tail, the corrections of aggregate wealth are moderate. The likely reason for this outcome is that fitting a Pareto tail relies on the fact that some information about the tail is captured by the survey. In situations with very limited information about the tail, fitting the Pareto distribution becomes increasingly difficult. Overall the recorded corrections are moderate for most countries, between 2% (Poland, Italy, Finland) and almost 7% (Germany) with the exception of Austria (16%) and the Netherlands (0.4%).

Table 9: Aggregate household net wealth

	(1) original	(2) rank correction (RC)	(3) RC/original
Austria	998	1,162	1.164
Belgium	1,584	1,644	1.037
Germany	8,500	9,073	1.067
Spain	4,768	5,016	1.052
Finland	512	526	1.027
France	7,033	7,439	1.058
Greece	445	465	1.046
Italy	5,590	5,721	1.023
Netherlands	1,147	1,151	1.004
Portugal	627	659	1.051
Poland	1,301	1,324	1.018
US_{2013}	66,762	69,150	1.036
$UK_{2012-2014}$	6,599	6,847	1.038

Aggregate household net wealth based on original survey data and rank corrected data in billion Euro. For the US and UK in billion USD and billion GBP, respectively.

Table 10 provides a clearer picture about the impact the rank correction approach has on different measures of wealth concentration and wealth inequality. A striking but expected result is that the top 0.1% and top 1% wealth shares increase substantially when applying the rank correction approach, even in countries where the correction of aggregate wealth is modest. The reason for this phenomenon lies in the fact that the thresholds of €1 million, £1 million and $15 million is between the top 1% and top 10% cut-off in most countries.

Table 10: Household net wealth shares

	(1) original top 0.1%	(2) RC top 0.1%	(3) original top 1%	(4) RC top 1%	(5) original top 10%	(6) RC top 10%
Austria	10.9	19.4	25.4	34.2	55.5	61.4
Belgium	1.3	4.4	12.	13.9	42.5	43.5
Germany	6.3	14.	23.6	29.6	59.8	62.3
Spain	6.4	8.2	16.3	19.8	45.6	48.3
Finland	4.	5.1	13.3	15.3	45.2	46.7
France	7.3	9.8	18.7	22.6	50.8	53.4
Greece	1.5	4.9	9.2	12.7	42.5	44.9
Italy	2.6	3.8	11.7	13.2	42.9	44.1
Netherlands	0.7	1.9	9.8	10.	43.6	43.6
Portugal	4.1	6.1	14.4	18.2	52.1	54.4
Poland	1.9	4.7	11.7	13.3	41.8	42.9
USA_{2013}	13.1	15.4	35.4	37.6	75.0	75.9
$UK_{2012-2014}$	5.6	6.3	15.1	17.3	45.7	47.6

Household net wealth shares based on original survey data and rank corrected data expressed in % of total household wealth.

5.3 Reconciliation with other data sources

Another crucial question is, how well the results based on the rank correction approach align with other existing information on the distribution of wealth. This section compares the results provided by the rank correction approach with other, unrelated sources of information on the distribution and extent of private household net wealth. The two crucial sources of exogenous information against which we compare our results are first, the World Inequality Database (WID) and second, journalists' rich lists for individual countries.

The methods used to construct the WID series for the US are discussed in Piketty et al. (2016) and the accompanying data appendix (Tables II-E1 to E13 contain the wealth share estimates). The country specific details for applying this methodology to France are discussed in Garbinti, Goupille-Lebret, and Piketty (2016) and in the accompanying appendices. The methods used for the UK series are discussed in Alvaredo, Atkinson, and Morelli (2018), its working paper version and the online appendix. The most important difference between the WID concentration measures and the survey based concentration measures is that the former are based on net personal wealth, which means that the unit of analysis is the individual instead of the household. One of the first steps of the WID methodology is to split married couples in survey or tax data into two observations with equal net wealth shares. This means some differences in the results stem from these methodological differences.

Table 11: Top wealth shares: WID vs rank correction

	country	data and method	(1) top 0.1%	(2) top 1%	(3) top 10%
(1)	France	World Inequality Database	8.2	23.4	55.3
(2)	France	rank correction estimator	9.8	22.6	53.4
(3)	France	uncorrected survey data (HFCS)	7.3	18.7	50.8
(4)	USA	World Inequality Database	20.3	37.0	73.2
(5)	USA	rank correction estimator	15.4	37.6	75.9
(6)	USA	uncorrected survey data (SCF)	13.1	35.4	75.0
(7)	UK	World Inequality Database		19.9	51.9
(8)	UK	rank correction estimator	6.3	17.3	47.6
(9)	UK	uncorrected survey data (WAS)	5.6	15.1	45.7

Source: Authors' computations based on data from the Household Finance and Consumption Survey (HFCS), Survey of Consumer Finances (SCF), Wealth and Asset Survey (WAS) and the World Inequality Database (WID). Comparison of French, US and UK wealth shares for the years 2014 and 2013 and 2012-2014 respectively.

Against that background, Table 11 compares wealth concentration ratios from the WID (rows 1, 4 and 7) with the results from the rank correction approach (rows 2, 5 and 8) and raw survey based measures (rows 3, 6 and 9). The rank correction results for France, the US and the UK clearly represent an improvement over the raw survey data and are closer to the WID measures than the raw counterparts. In the case of France the RC measures are below the WID values except for the top 0.1% share and thus can be regarded as moderate concentration measures against the WID background which represents the most precise effort in the literature to produce concentration measures based on a combination of tax, survey and national accounts data. For the US case, the rank correction based top shares are slightly higher than the WID results except for the top 0.1%. For the UK, the WID does not provide an entry for the top 0.1% share in 2012. Overall the RC based measures help close the gap between the WID and the raw survey measures which we interpret as support for the rank correction approach.

Another exogenously available source of information about the top tail of the wealth distribution are journalists' rich lists. Table 12 column (1) lists the number of billionaires in the population according to the raw survey data (i.e. billionaire observations times their weight) which indicates that no country except the US with the SCF has an oversampling strategy in place which is suitable to capture billionaires[11]. Column (3) reports the number of billionaire entries on publicly available rich lists and since these entries often comprise large families, column (4) reports an estimate trying to disentangle these families into individual households. In Germany this leads to a substantial fall

[11] Another factor is the larger size of the US.

Table 12: Number of billionaire households

country	(1) survey	(2) rich list	(3) bn. entries	(4) bn. households	(5) RC data
Austria	0	Trend	33	35	18
Belgium	0	Forbes	3		0
Germany	0	Manager Magazin	134	77	49
Spain	0	El Mundo	27	11	9
Finland	0	Forbes	4		0
France	0	Challenges	72	74	19
Greece	0	Forbes	3		0
Italy	0	Forbes	41		0
Netherlands	0	Forbes	8		0
Portugal	0	Forbes	3		0
Poland	0	Forbes	5		0
USA	43	Forbes	491		474
UK	0	Sunday Times	40		4

Note that Forbes does not refer to the Forbes 400 list but to the Forbes list of billionaires worldwide.

in the number of billionaire households from 134 to 77 because a series of billionaire families are divided into less than billionaire households. Finally, column (5) reports the number of billionaires according to the estimated Pareto tail after applying the rank correction procedure. It can be seen that for many of the smaller countries, no billionaires are expected. At the same time for those countries where billionaire lists are available the rank correction procedure substantially improves upon the raw data numbers and in general yields plausible results which are close to the number of billionaire households found on rich lists. In all cases the rank correction approach produces fewer billionaires than are reported on rich lists. We interpret this result as general support for the claim that the rank correction approach is a rather conservative tool for addressing the nonresponse bias in survey data as the results it provides probably still underestimate the actual degree of concentration. The probable underestimation of the rank correction approach becomes apparent for countries like Italy, the Netherlands and the UK, all three of which do not have an oversampling strategy in place[12].

6 Summary and Conclusion

Against the background that for many countries household surveys are the only available source of data on the distribution of wealth, this paper presents a new approach for tackling the problem that surveys tend to underestimate household wealth due to differential nonresponse. The core idea of this rank correction procedure is to fit a Pareto distribution to the tail of the data, by means of a log-rank-log-wealth regression (Kratz & Resnick, 1996), after adjusting the accumulated survey weights (i.e. the rank) first, for the number of super rich households which are not included in the sample design due to their appearance on rich lists and resulting privacy concerns (SR) and second, for the under-representation of rich households at the top due to general forms of differential nonresponse (DNR). This procedure yields an overall correction factor $u = SR + DNR$.

For determining the two components (SR, DNR) of the correction factor u, we propose two rules of thumb. The first heuristic is for determining the number of households excluded at the top due to privacy concerns. If explicit information about the exclusion of super rich households is available from the data provider (as is the case with the Survey of Consumer Finances) that should be used. Alternatively we propose to choose SR proportional to the total population based on the average size of publicly available rich lists across Austria, Germany, Spain, France and Poland: $SR = 0.004\% \cdot N_C$.

The second heuristic is concerned with the determination of DNR and we propose to start from a correction factor of 4000, which is optimal for a tail population size of 1 million households and

[12]The WAS for the UK only oversamples the top decile of the wealth distribution and thus for the purpose of fitting a tail which starts within the top decile, there is no oversampling.

the Vermeulen-Kennickell response mechanism (Vermeulen, 2018). This correction factor can then be tailored to the characteristics of the data set at hand by applying three adjustment factors: $DNR = 4000 \cdot a_P \cdot a_S \cdot a_O$. The adjustment factor a_P takes the size of the tail population into account (relative to the 1 million baseline), a_S scales the correction factor to account for small sample sizes from small tail populations and a_O is an adjustment factor which takes the quality and degree of oversampling at the top into account.

By means of Monte Carlo simulations, we show that choosing an appropriate correction factor yields a substantial reduction in the bias of the estimated shape parameter of the Pareto distribution (α) in a situation of differential nonresponse. Applying the rank correction procedure to data from the SCF and the HFCS results in significant corrections of top wealth shares. Our results are closer in line with other existing top wealth share estimates than are the raw survey estimates. For example the WID provides top 1% wealth shares for the US, France and the UK (37%, 23.4% and 19.9%) which compare very well with the rank correction approach (37.6%, 22.6% and 17.3%) and represent a clear improvement over raw survey estimates (35.4%, 18.7% and 15.1%).

The key advantage of the rank correction procedure over similar existing methods such as Vermeulen's (2018) rich list approach, is that it requires much less exogenous information. Especially in situations where rich list data is not available or of poor quality, the rank correction approach is a substantial improvement over standard OLS fitted Pareto tails. The rank correction approach enables the researcher to take differential nonresponse problems into account even if rich list data is not available. In addition, the rank correction approach forces the researcher to be explicit and transparent about the required modelling assumptions. In contrast the rich list approach assumes that rich lists are measured correctly and thus implicitly incorporates all assumptions and judgements made by the journalists compiling these rich lists and the problems which come with it (Capehart, 2014; Kopczuk, 2015). In this sense the rank correction approach can serve as an alternative and robustness check to the rich list approach.

Lastly, we want to reiterate that the rank correction approach, at its core, is a heuristic. Nevertheless, we do think it is an important improvement over simply ignoring differential nonresponse problems. That said, it is of the utmost importance to address the root cause of the problem and improve oversampling strategies in existing wealth surveys. From a European perspective, the introduction of the HFCS was a massive step forward. The next step would be to streamline the oversampling strategies in the HFCS across countries and base them on individual tax data. Granting access to tax information is a sensitive issue but the success of the SCF demonstrates that it is by all means feasible. In the meantime researchers have to rely on heuristics and are forced to make difficult judgements about how to deal with this problem. While this situation is not ideal, we think that simply ignoring the problem actually makes matters worse and amplifies the underlying problem. The alternative presented in this paper might not be as elegant or beautiful as we would like it to be but at the end of the day it is presumably better to be approximately right than precisely wrong.

Appendix: Choosing the sample adjustment (a_S) factor

Optimal correction factor search procedure

In order to find optimal correction factors u_{opt} for different population sizes (N_{T2} and N_{T3}), we adopt the following search procedure: The first step is to estimate a regression based on equation (9) with a correction factor $u_1 = 560$ for N_{T2} ($u_2 = 280$ for N_{T3}) and calculate the deviation of the obtained estimate from the population value: $d_1 = \hat{\alpha}_1 - 1.5$. If this deviation is positive we continue with the second step which is to increase the correction factor by by 50 $u_2 = u_1 + 50$ and re-estimate equation (9) with this new correction factor. Based on the resulting estimate of the shape parameter we can calculate $d_2 = \hat{\alpha}_2 - 1.5$. We continue these steps until the difference turns negative. This search algorithm is applied to all 200 samples for each sample size. The optimal correction factor for a given sample size is the average across the 200 samples. The table below contains the results for both tail populations:

Table 13: Optimal correction factors

	$N_{T2} = 10^5$		$N_{T3} = 5 \cdot 10^4$	
netsample (in ‰)	u_{opt}	α	u_{opt}	α
0.2	2260	1.502	2060	1.504
0.3	1310	1.500	1720	1.504
0.4	1210	1.503	940	1.503
0.5	1010	1.501	900	1.505
0.6	910	1.503	840	1.502
0.7	760	1.505	780	1.504
0.8	960	1.504	860	1.503
0.9	1310	1.505	780	1.504
1.0	960	1.501	540	1.502
1.1	910	1.500	600	1.503
1.2	910	1.503	540	1.504
1.3	860	1.502	640	1.502
1.4	860	1.500	520	1.501
1.5	710	1.504	500	1.504
1.6	760	1.501	420	1.502
3.0	760	1.500	460	1.502
6.0	660	1.501	360	1.502

Source: Authors' calculations.

Sample adjustment factor (a_S) regressions

We regressed the vector of net sample sizes s=(0.4 ‰, ... , 1.6 ‰) on the ratio of the optimal correction factors from Table (13) to the standard correction factors scaled according to the size of the tail population (i.e. $u = 560$ and $u = 260$):

$$u_{opt}/560 = \beta_0 + \beta_1 \ln(s) + \epsilon$$

$$u_{opt}/280 = \beta_0 + \beta_1 \ln(s) + \epsilon$$

The values 560 and 280 represent the standard correction factor $u = 5600$, scaled to tail population sizes of $N_{T2} = 10^5$ and $N_{T3} = 5 \cdot 10^4$. The standard correction factor $u = 5600$ was chosen for $N_{T1} = 10^6$ and thus these two alternative tail populations represent 10% and 5% of N_{T1} and the correction factors 560 and 280 are rescaled in the same proportion.

The regression results are reported in the table below. We excluded extremely small samples (0.2‰ and 0.3‰) as well as very large samples (3‰ and 6‰) to increase the stability of the regression. The fitted values from these regressions can be interpreted as the sample size adjustment factor (a_S).

Table 14: Sample adjustment factor regressions

	Estimate	Standard Error	t-Statistic	P-Value
Regresion 1 for N_{T2}: $u_{opt}/560 = \beta_0 + \beta_1 \ln(s) + \epsilon$				
β_0	-1.08098	1.23219	-0.877283	0.40
β_1	-0.389043	0.17419	-2.23345	0.047
Regresion 2 for N_{T3}: $u_{opt}/280 = \beta_0 + \beta_1 \ln(s) + \epsilon$				
β_0	-6.91819	1.168	-5.9231	0.00
β_1	-1.32464	0.165145	-8.02112	0.00

Source: Authors' calculations. The regression excludes the smallest samples of 0.2‰and 0.3‰and the two largest samples of 3‰and 6‰.

Table 15 reports fitted values for our core range of net sample sizes. These fitted values can be used as sample adjustment factors (a_S) for tail populations of similar sizes. For example in the case of Austria we have a tail population of $N_T = 129,309$ households and a sample size of $s = 0.67‰$. Based on the above regression, we obtain a fitted value and thus correction factor of 1.7 for a population of 100,000 and a sample of $s = 0.7‰$ and thus we set $a_S = 1.7$.

Table 15: Sample adjustment factors (a_S)

netsample size ‰	for N_{T2}	for N_{T3}
0.4 ‰	2.0	3.4
0.5 ‰	1.9	3.2
0.6 ‰	1.8	2.9
0.7 ‰	1.7	2.7
0.8 ‰	1.7	2.5
0.9 ‰	1.6	2.4
1.0 ‰	1.6	2.2
1.1 ‰	1.6	2.1
1.2 ‰	1.5	2.0
1.3 ‰	1.5	1.9
1.4 ‰	1.5	1.8
1.5 ‰	1.4	1.7
1.6 ‰	1.4	1.6

Source: Authors' calculations.
$N_{T2} = 10^5$ and $N_{T3} = 5 \cdot 10^4$.

References

Aigner, D. J., & Goldberger, A. S. (1970). Estimation of pareto's law from grouped observations. *Journal of the American Statistical Association*, *65*(330), 712–723.

Alstadsæter, A., Johannesen, N., & Zucman, G. (2019, jun). Tax evasion and inequality. *American Economic Review*, *109*(6), 2073–2103. doi: 10.1257/aer.20172043

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

Bricker, J., Henriques, A., Krimmel, J., & Sabelhaus, J. (2016). Measuring income and wealth at the top using administrative and survey data. *Brookings Papers on Economic Activity*, *2016*(1), 261–331.

Capehart, K. W. (2014). Is the wealth of the world's billionaires not paretian? *Physica A: Statistical Mechanics and its Applications*, *395*, 255–260. doi: 10.1016/j.physa.2013.09.026

D'Alessio, G., & Faiella, I. (2002). Non-response behaviour in the bank of italy's survey of household income and wealth. *Temi di discussione (Bank of Italy Economic working papers)*, *2002*(462).

Eckerstorfer, P., Halak, J., Kapeller, J., Schütz, B., Springholz, F., & Wildauer, R. (2016). Correcting for the missing rich: An application to wealth survey data. *Review of Income and Wealth*, *62*(4), 605–627. doi: 10.1111/roiw.12188

Gabaix, X., & Ibragimov, R. (2011). Rank - 1/2: A simple way to improve the ols estimation of tail exponents. *Journal of Business & Economic Statistics*, *29*(1), 24–39. doi: 10.1198/jbes.2009.06157

Garbinti, B., Goupille-Lebret, J., & Piketty, T. (2016). Accounting for wealth inequality dynamics: Methods, estimates and simulations for france (1800-2014). *WID Working Paper Series*, *2016/5*.

Jayadev, A. (2008). A power law tail in india's wealth distribution: Evidence from survey data. *Physica A: Statistical Mechanics and its Applications*, *387*(1), 270–276. doi: 10.1016/j.physa.2007.08.049

Kennickell, A. B., & Woodburn, R. L. (1997). *CONSISTENT WEIGHT DESIGN FOR THE 1989, 1992 AND 1995 SCFs, AND THE DISTRIBUTION OF WEALTH* (Tech. Rep.). Federal Reserve Board Survey of Consumer Finances Working Papers.

Kopczuk, W. (2015). What do we know about the evolution of top wealth shares in the united states? *Journal of Economic Perspectives*, *29*(1), 47–66. doi: 10.1257/jep.29.1.47

Kratz, M., & Resnick, S. I. (1996, jan). The qq-estimator and heavy tails. *Communications in Statistics. Stochastic Models*, *12*(4), 699–724. doi: 10.1080/15326349608807407

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data.* John Wiley and Sons Ltd.

Osier, G. (2016). Unit non-response in household wealth surveys: Experience from the eurosystem's household finance and consumption survey. *European Central Bank Statistics Paper Series*, *2016*(15).

Piketty, T. (2014). *Capital in the twenty-first century.* Harvard University Press.

Piketty, T., Saez, E., & Zucman, G. (2016). Distributional national accounts: Methods and estimates for the united states. *NBER Working paper*, *22945*. doi: 10.3386/w22945

Saez, E., & Zucman, G. (2016). Wealth inequality in the united states since 1913: Evidence from capitalized income tax data. *The Quarterly Journal of Economics*, *131*(2), 519–578. doi: 10.1093/qje/qjw004

Vermeulen, P. (2018). How fat is the top tail of the wealth distribution? *Review of Income and Wealth*, *64*(2), 357–387.

Wildauer, R., & Kapeller, J. (2019). A comment on fitting pareto tails to complex survey data.