A SIMPLE CORRECTION TO REMOVE THE BIAS OF THE GINI COEFFICIENT DUE TO GROUPING

Tom Van Ourti and Philip Clarke*

Abstract—We propose a first-order bias correction term for the Gini index to reduce the bias due to grouping. It depends on only the number of individuals in each group and is derived from a measurement error framework. We also provide a formula for the remaining second-order bias. Both Monte Carlo and EU and U.S. empirical evidence show that the first-order correction reduces a considerable share of the bias, but that some remaining second-order bias is increasing in the variance. We propose a procedure that addresses the remaining second-order bias by using additional information.

I. Introduction

THE Gini index is the most commonly applied inequality measure in the literature, probably because of its link with Lorenz curves, which give an intuitive and graphical representation of inequality. Its main application has been in the measurement of inequalities in income and wealth, but it also has a long history in other areas. For example, it has appeared as an inequality measure of health indicators (among others, Le Grand, 1987; Pradhan, Sahn, & Younger, 2003), educational attainment (among others, Sheret, 1988; Lin, 2007), business concentration (among others, Hart, 1971; Buzzacchi & Valletti, 2006), scientific publications and citations (among others Allison & Stewart, 1974), legislative malapportionment (Alker, 1965), astronomy (Abraham, van den Bergh, & Nair, 2003), and many others.

A long-standing problem in calculating the Gini index is how to deal with data grouped by categories or into ranges (Gastwirth, 1972; Abounoori & McCloughan, 2003). This issue commonly arises with income or tax statistics that are often grouped for confidentiality reasons. Grouped data are also the main source of information on income distributions provided through the POVCALNET interactive computational tool of the World Bank (World Bank, 2008) and the UNU-WIDER World Income Inequality Database (UNU-WIDER, 2008). Recent publications on regional inequality (Guest & Swift, 2008), global income inequality (Milanovic, 2002; 2005; Sala-i-Martin, 2006), and global wealth inequality (Davies et al., 2010) have also used grouped data.[1] Previous empirical research suggests that the grouping of income into relatively small number of categories imparts a nonnegligible downward bias. For example, using the 1984 U.S. Current Population Survey and the 1979–1980 Israeli Family Expenditure Survey, Lerman and Yitzhaki (1989) show that the bias from using grouped data with ten and five income categories is about 2.5% and 7% of the Gini as calculated from microdata. Davies and Shorrocks (1989) report biases of similar magnitude from grouping Canada's 1984 Survey of Consumer Finance.

Two solutions have been proposed to cope with the dependence of the Gini index on the number of groups. First, a common approach, when average incomes of each income group are known, is to reduce the bias due to grouping by fitting parametric functions that satisfy the properties of a theoretical Lorenz curve. The estimated parameters are then used to estimate the Gini coefficient (Kakwani, 1980a, 1986; Villaseñor and Arnold, 1989; Basmann et al., 1990, Ryu & Slottje, 1996). This approach is popular among applied researchers (Datt & Ravallion, 1992; Bigsten & Shimeles, 2007) and has been implemented in the POVCAL software of the World Bank (2008). A second approach is to define nonparametric bounds on the Gini index (Gastwirth, 1972; Mehran, 1975; Murray, 1978; Fuller, 1979; Kakwani, 1980a; Ogwang, 2003, 2006), which has the advantage that, compared to parametric functions, it does not make any assumption on the shape of the underlying Lorenz curve, but requires information on the lower and upper limit of each group. The lower bound of the Gini corresponds to the situation where all individuals within a group are supposed to have the same mean amount of this group, while the upper bound reflects a situation where inequality is maximal in each of the groups.

Deltas (2003) has attempted to address the related issue of small-sample bias. Here the bias arises not because of grouping, but is due to having only a few observations such

Received for publication October 8, 2008. Revision accepted for publication March 24, 2010.

* Van Ourti: Erasmus University Rotterdam, Tinbergen Institute, NETSPAR; Clarke: University of Sydney, Australian National University.

This research was partially supported from the NETSPAR project Income, Health and Work across the Life Cycle, and from the project The Dynamics of Income, Health and Inequality over the Lifecycle (known as the ECuity III Project), which is funded in part by the European Commission's Quality of Life and Management of Living Resources program (contract QLK6-CT-2002-02297). We thank EUROSTAT for access to the European Community Household Panel Users' Database, version of December 2003, and the Netherlands Central Bureau of Statistics for access to the linked data sets used for this research ("Own calculations of Erasmus University Rotterdam using data files made available by the Netherlands Central Bureau of Statistics on the regional income distribution of persons and households derived from the Tax Administration"). Part of this research was undertaken while T.V.O. was a postdoctoral fellow of the Netherlands Organisation for Scientific Research, Innovational Research Incentives Scheme—Veni. P.C. receives support from a Senior Research Fellowship from the University of Sydney. We thank Hans van Kippersluis for helpful comments and excellent research assistance. The study has benefited from the comments and suggestions of Teresa Bago d'Uva, Bob Breunig, George Deltas, John Einmahl, Andrew Leigh, an anonymous referee, as well as participants at seminars given at Australian National University, Tilburg University, and Erasmus University Rotterdam. We finally thank Xander Koolman and Cristina Hernández-Quevedo for assisting with STATA code. The usual caveats apply, and all remaining errors are our responsibility.

[1] One referee noted "In effect, for reasons of tractability and also because not all countries provide micro data ... global income inequality calculations would probably have to be done with grouped data until such time when a single world-wide survey is conducted."

The Review of Economics and Statistics, August 2011, 93(3): 982–994

as might occur when calculating the Gini of subpopulations using small (sub-)samples or due to few firms in an industry when studying business concentration. Deltas (2003) addresses the small-sample bias with a first-order correction term that depends on only the number of observations.[2] The main advantage of this correction term is its relative simplicity and transparency in application, but it neglects that the small-sample bias of the Gini is distribution specific. Nevertheless, Monte Carlo simulations show that his correction term manages to reduce the small-sample bias.

Inspired by Deltas (2003), we develop a simple first-order correction term to deal with the bias of the Gini due to grouping by treating grouping as a form of measurement error. Our first-order correction, which at its simplest form involves multiplying the Gini by $K^2/(K^2-1)$, where K is the number of groups, differs from the methods based on fitting parametric functions and the nonparametric bounds in that it can be applied without information on the average incomes or income ranges of each income group, that is, it needs only information on the number of individuals in each income group or range. This has unrivaled advantages when one only has access to estimates of the Gini index based on grouped data without observing the underlying average incomes or income ranges, as is, for example, the case for the majority of countries in the UNU-WIDER World Income Inequality Database (UNU-WIDER, 2008). Also in case the underlying average incomes or ranges are observed, our correction method has the advantage of being simple and transparent. However, because it is not exploiting the information on average incomes or income ranges, its performance will depend on the shape of the underlying unobserved income distribution. In other words, the bias in the Gini due to grouping is distribution specific, and a second-order bias might remain after applying the first-order correction. While the latter second-order bias is zero for some specific distributions, Monte Carlo evidence shows that it is in general low but increasing in the variance of the underlying distribution. We confirm this Monte Carlo evidence in an empirical illustration: our first-order correction term reduces a large share of the bias due to grouping when applied to the income distributions of fifteen European countries and the United States. We also develop a procedure that addresses the remaining second-order bias by imposing additional information. Our results show that this procedure could be used as an alternative to existing correction methods involving fitting conventional parametric forms to the data.

The remainder of this paper contains four sections. We start by illustrating the usefulness of OLS in obtaining an estimate of the Gini. The next section derives our first-order correction and applies Monte Carlo simulations to increase the understanding of the remaining second-order bias. We then illustrate our methods on data for fifteen European countries and the United States in the fourth section. The final section contains the conclusions.

[2] It involves multiplying the Gini estimated from a small sample with the inverse of its potential maximum: $n/(n-1)$.

II. Estimation of the Gini index

The Gini can be estimated using several equivalent formulas. For our purposes the following one is the most useful (Pyatt, Chen, & Fei, 1980),

$$G_n = \frac{2\sum_{i=1}^{n} y_i R_i}{n\bar{y}} - 1 = \frac{2\mathrm{cov}(y_i, R_i)}{\bar{y}}, \tag{1}$$

where y_i is the income of individual $i = 1,\ldots,n$ with individuals ranked from poor to rich, that is, $y_1 \leq y_2 \leq \cdots \leq y_n$, $R_i = n^{-1}(i - 1/2)$ is the fractional income rank (Lerman & Yitzhaki, 1989), and $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ denotes average income.[3] A simple transformation of equation (1) shows that the Gini can also be calculated as the OLS estimate of β (Kakwani, Wagstaff, & van Doorslaer, 1997),

$$2\sigma_R^2 \frac{y_i}{\bar{y}} = \alpha + \beta R_i + \varepsilon_i, \tag{2}$$

where $\sigma_R^2 = (n^2 - 1)(12n^2)^{-1}$ is the variance of R_i (Milanovic, 1997), ε_i is an error term with zero mean, and α, β are parameters. It is important to note that the equality between equations (1) and (2) holds under the properties of OLS as arithmetic tool and that no additional assumptions need be made.

III. The Bias of the Gini due to Grouping and a First-Order Correction Term

In this section, we present an exact expression for the bias of the Gini due to grouping and derive and discuss the properties of a first-order correction term to address this bias. We start with the easier case of groups of equal size and next generalize to groups of unequal size. Our approach proceeds as follows. First, we compare equation (2) for n observations and for a situation where one constructs K groups from these n observations.[4] In other words, we assume that an estimate of the Gini based on grouped data is available and next analyze how this estimate differs from the one that would be obtained from the underlying individual data. Second, we derive an exact expression for the difference between both estimators by drawing a parallel with the econometric literature on measurement error models (for example, Cameron & Trivedi, 2005, chapter 26).

[3] We discuss the Gini of income, but obviously everything also holds for any variable whose distribution is analyzed.

[4] Note the similarity with the difference between the OLS and between estimator for panel models (Cameron & Trivedi, 2005).

Third, an intuitive first-order correction term to address the bias in the grouped data estimator results from this exact expression. It is termed first-order since, in contrast to the existing methods based on fitting parametric functions and the nonparametric bounds, it does not need information on average incomes per income group or the income ranges.

A. Groups of Equal Size

In order to understand the bias of the Gini that results from grouping n observations into K groups of equal size, it is helpful to see that equation (2) reduces to

$$2\sigma^2_{R^K}\frac{y_g}{\overline{y}} = \alpha^K + \beta^K R_g + \varepsilon_g, \tag{3}$$

where we have added K superscripts to refer to the grouped data case, $R_g = K^{-1}(g - 1/2)$ is the fractional income rank of group $g = 1,\ldots,K$, $\sigma^2_{R^K} = (K^2 - 1)(12K^2)^{-1}$ is the variance of R_g, and y_g is the average income within group g. The OLS estimate of β^K equals the Gini index calculated from the K groups and is a downwardly biased estimator of the Gini calculated from n observations due to the convexity of the underlying Lorenz curves,[5]

$$\begin{aligned}\beta^K = G_n^K &= \frac{2\sum_{g=1}^{K} y_g R_g}{K\overline{y}} - 1 \\ &= \frac{2\mathrm{cov}(y_g, R_g)}{\overline{y}} \le G_n = \beta.\end{aligned} \tag{4}$$

Next, we establish an exact relationship between G_n and G_n^K in equations (2) and (3). Comparing the latter equations reveals that both the right-hand side and the left-hand side differ. The difference in the right-hand side can be interpreted as a measurement error problem; we observe the rank of income at the level of the groups rather than one at the level of the n observations. More exactly, we add an equation that describes the measurement error problem:

$$R_i^g = R_i + \delta_i^g, \tag{5}$$

where δ_i^g is the measurement error and R_i^g is the fractional income rank of group g defined at the individual level, that is, every individual in group g gets the fractional income rank of group g—R_g. Due to the properties of the fractional income rank, this measurement error is uniformly distributed and has zero mean. Substituting equation (5) into equation (2) gives

$$2\sigma^2_R\frac{y_i}{\overline{y}} = \alpha + \beta R_i^g + (\varepsilon_i - \beta\delta_i^g). \tag{6}$$

[5] A downward bias occurs if there is income variation within at least one of the K groups; there is no bias if there is no income variation in each of the K groups.

It is impossible to estimate β from equation (6) using OLS (as an arithmetic tool) since we do not observe $(\varepsilon_i - \beta\delta_i^g)$.[6] Instead, we can only estimate

$$2\sigma^2_R\frac{y_i}{\overline{y}} = \alpha^{MER} + \beta^{MER} R_i^g + \eta_i, \tag{7}$$

where η_i is a zero-mean error term and the superscript *MER* refers to measurement error. Using some algebra and exploiting the fact both δ_i^g and R_i^g and ε_i and R_i are uncorrelated (which holds due to using OLS as an arithmetic tool only), it is easy to show that the OLS estimate of β^{MER} in equation (7) and the OLS estimate of $\beta = G_n$ in equation (2) are related:[7]

$$\beta^{MER} = G_n + \frac{\frac{1}{n}\sum_{i=1}^{n}\delta_i^g\varepsilon_i}{\sigma^2_{R^K}}. \tag{8}$$

In order to derive an expression relating G_n and G_n^K, we need to establish one additional relationship that addresses the difference between the left-hand side of equations (2) and (3). After some algebra, one can establish that

$$\begin{aligned}\beta^{MER} &= G_n^K\left(\frac{\sigma^2_R}{\sigma^2_{R^K}}\right) \\ &= G_n^K\left[\frac{K^2(n^2-1)}{n^2(K^2-1)}\right],\end{aligned} \tag{9}$$

which shows that β^{MER} is related to G_n^K by the ratio of the variances of the actual fractional income rank and that of the fractional income rank of group g.

Combining equations (8) and (9) allows us to come up with a useful equation that expresses the Gini estimated from n observations as a function of, among others, the Gini estimated from a grouping of these n observations:

$$\begin{aligned}G_n &= G_n^K\left(\frac{\sigma^2_R}{\sigma^2_{R^K}}\right) - \frac{\frac{1}{n}\sum_{i=1}^{n}\delta_i^g\varepsilon_i}{\sigma^2_{R^K}} \\ &= G_n^K\left[\frac{K^2(n^2-1)}{n^2(K^2-1)}\right] - \left[\frac{12K^2}{K^2-1}\right]\left[\frac{1}{n}\sum_{i=1}^{n}\delta_i^g\varepsilon_i\right].\end{aligned} \tag{10}$$

[6] We do not observe $(\varepsilon_i - \beta\delta_i^g)$ since we consider the hypothetical situation where the actual income levels y_i are observed but the corresponding actual fractional income ranks R_i are not. This assumption makes sense since equations (5) to (8) focus on the difference between the right-hand sides of equations (2) and (3), that is, interpreting the difference in the right-hand side as a measurement error problem; without addressing the difference in the left-hand side (or in other words, the fact that actual income levels are observed). The difference in the left-hand side is addressed in equation (9). Therefore, the assumption of observing actual income levels, but not their corresponding fractional ranks, is auxiliary, and not needed to sustain equation (10), which gives an exact expression for the difference between equations (2) and (3).

[7] δ_i^g and R_i^g are uncorrelated since R_i^g equals the average R_i of group g, that is, $\sum_{i\in g}\delta_i^g R_i^g = \sum_{i\in g}(R_i^g - R_i)R_i^g = 0$, and hence $\sum_{i=1}^{n}\delta_i^g R_i^g = \sum_{g=1}^{K}\left(\sum_{i\in g}\delta_i^g R_i^g\right) = 0$.

Assuming that $n \to +\infty$ and $K < +\infty$ (that is, the number of groups in the population and their relative size is fixed) results in

$$G_\infty = \frac{K^2}{K^2 - 1}[G_\infty^K - 12\text{cov}(\delta_i^g, \varepsilon_i)]. \quad (11)$$

Equation (11) reveals some interesting insights. First, we have used the properties of OLS only as an arithmetic tool and the properties of the fractional rank to come up with equation (11). Second, a first-order correction term to address the bias of the grouped data estimator of the Gini and an expression for the remaining second-order bias result self-evidently from equation (11). The first-order correction $(K^2 - 1)^{-1} K^2$ does depend on only the number of income groups (hence "first-order"). Therefore, it can, in contrast to existing methods, also be used to correct estimates of the Gini index based on grouped data without observing the underlying average incomes or income ranges. The performance of the first-order correction term can be inferred from the remaining second-order bias, $12\ \text{cov}(\delta_i^g, \varepsilon_i)K^2(K^2 - 1)^{-1}$ and will depend on the shape of the underlying unobserved income distribution. In other words, the expression for the remaining second-order bias reflects that the bias in the Gini due to grouping is distribution specific. Third, the first-order correction term has two intuitive interpretations: it equals a "grouped data" adjustment of the variance of the fractional rank, which turns out to be identical to the so-called attenuation bias in the classical measurement error model (for example, Cameron & Trivedi, 2005, section 26.2.3), and it is related to the inverse of the covariance between the grouped and actual fractional rank, $(K^2 - 1)^{-1} K^2 = [12\ \text{cov}(R_i^g, R_i)]^{-1}$, which implies a low/(high) first-order correction term for a high/(low) covariance. The second-order bias also has an intuitive interpretation as it is a function of the covariance between the measurement error and the error term from equation (2).

A few things can be said about this covariance. It will be smaller the higher the number of groups K, which is easily inferred from the equality $\text{cov}(\delta_i^g, \varepsilon_i) = \text{cov}(R_i^g, \varepsilon_i)$. In addition, its value and sign are unknown since, although one knows that δ_i^g is uniformly distributed with zero mean, the error term ε_i is unobservable without the underlying individual-level data. Nevertheless, it is straightforward to get an idea on its sign and magnitude if one has an idea on the shape of the underlying unobservable distribution function of y_i.

First, if the unobserved y_i is uniformly distributed (or income levels are linearly related to the fractional income rank), the covariance term will be 0 since the variance of ε_i equals 0 and no second-order bias will remain after applying the first-order correction term. While this is mainly informative for uniformly distributed attributes, it also involves an interesting reference case for nonuniformly distributed attributes, such as income distributions. Second, the covariance term might also equal 0 for some nonuniform distributions. Since the requirement $\text{cov}(\delta_i^g, \varepsilon_i) = 0$ might hold for an infinite number of distributions, we cannot enumerate all cases here. An interesting case is the distribution determined by $y_i = R_i^2$. Here the covariance term equals 0 since ε_i is symmetrically distributed around the median fractional rank $R_i = 0, 5$. Another example is the beta distribution with parameters 0.5 and 1. However, a distribution where income is not linearly related to the income rank will not generally lead to a 0 covariance term. In the latter case, the covariance term might be negative (implying an undercorrection after applying the first-order correction term) or positive (implying an overcorrection).

In order to increase the understanding of the performance of the first-order correction term under different distributional assumptions, and consequently the sign and magnitude of the remaining second-order bias, we have performed Monte Carlo simulations for three distributions. First, we considered the uniform distribution with support on the unit interval as it is an interesting reference case in the context of the first-order correction term. Second, we used the log-normal distribution (with log values distributed normally with mean 0 and standard deviation σ_y). We varied σ_y from 1.5 to 0.25 to infer how it affects the magnitude of the bias from grouping (and the performance of the first-order correction term). Third, we used the beta distribution with values of its parameters equaling 0.5, 1, 3, 5, 10, and 25 (with 36 combinations in total). In contrast to the log-normal distribution, its variance and kurtosis can vary independently from the skewness (Deltas, 2003), and therefore it allows disentangling the separate impact of these three moments on the magnitude of the bias from grouping.[8] In addition, the beta is a flexible distribution allowing various shapes of the density function, including bimodality and left and right skewness.

For each distribution, 20,000 independent samples of size $n = 10{,}000$ have been drawn.[9] For each of these samples, the Gini was computed after grouping the data into K groups of equal size for $K = 2,3,\ldots,10,20,30,40,50$, and next compared to the Gini obtained without grouping.[10] The average values of the Gini (and its standard deviation in the Monte Carlo simulation), the first-order correction for grouping, and the covariance for the uniform and log normal distributions are shown in table 1.

Table 1 shows that grouping leads to a downward bias of the Gini and that the bias is decreasing in the number of groups. Its magnitude is large compared to the standard deviation of the Gini and differs across the different distributions. For the log-normal distributions, it seems that the

[8] Strictly speaking, these are normalized central moments, but for brevity, we loosely refer to "moments."

[9] Monte Carlo simulations for the uniform and log-normal distributions with smaller sample sizes ($n = 100$ and 1,000) confirmed our findings based on $n = 10{,}000$, and thus suggest that the asymptotic formula in equation (11) might be reasonable in practice.

[10] These groupings are of "equal size" for $K = 2,4,5,8,10,20,40,50$. For other values of K, these groupings are approximately of "equal size."

TABLE 1.—THE BIAS OF THE GINI DUE TO GROUPING: A SIMULATION EXERCISE

	Uniform				Log Normal (s.d. = 1.5)				Log Normal (s.d. = 1)				Log Normal (s.d. = 0.75)				Log Normal (s.d. = 0.5)				Log Normal (s.d. = 0.25)			
Groups	Mean	s.d.	First Order	Cov	Mean	s.d.	First Order	Cov	Mean	s.d.	First Order	Cov	Mean	s.d.	First Order	Cov	Mean	s.d.	First Order	Cov	Mean	s.d.	First Order	Cov
Full	0.333	0.002			0.711	0.007			0.520	0.004			0.404	0.003			0.276	0.002			0.140	0.001		
50	0.333	0.002	0.333	0.000	0.709	0.007	0.709	−0.002	0.519	0.004	0.520	−0.001	0.404	0.003	0.404	−0.000	0.276	0.002	0.276	−0.000	0.140	0.001	0.140	−0.000
40	0.333	0.002	0.333	0.000	0.708	0.007	0.708	−0.003	0.519	0.004	0.519	−0.001	0.403	0.003	0.404	−0.000	0.276	0.002	0.276	−0.000	0.140	0.001	0.140	−0.000
30	0.333	0.002	0.333	0.000	0.706	0.006	0.707	−0.004	0.518	0.004	0.519	−0.002	0.403	0.003	0.403	−0.001	0.276	0.002	0.276	−0.000	0.140	0.001	0.140	−0.000
20	0.332	0.002	0.333	0.000	0.701	0.006	0.703	−0.008	0.516	0.004	0.517	−0.003	0.401	0.003	0.402	−0.002	0.275	0.002	0.276	−0.001	0.140	0.001	0.140	−0.000
10	0.330	0.002	0.333	0.000	0.684	0.006	0.691	−0.020	0.507	0.004	0.512	−0.008	0.396	0.003	0.400	−0.004	0.271	0.002	0.274	−0.002	0.138	0.001	0.139	−0.001
9	0.329	0.002	0.333	0.000	0.679	0.005	0.688	−0.023	0.505	0.004	0.511	−0.009	0.394	0.003	0.399	−0.005	0.270	0.002	0.274	−0.003	0.138	0.001	0.139	−0.001
8	0.328	0.002	0.333	0.000	0.673	0.005	0.684	−0.027	0.501	0.004	0.509	−0.011	0.392	0.003	0.398	−0.006	0.269	0.002	0.273	−0.003	0.137	0.001	0.139	−0.001
7	0.327	0.002	0.333	0.000	0.665	0.005	0.679	−0.031	0.496	0.004	0.507	−0.013	0.388	0.003	0.396	−0.008	0.267	0.002	0.272	−0.004	0.136	0.001	0.139	−0.001
6	0.324	0.002	0.333	0.000	0.653	0.005	0.672	−0.038	0.490	0.004	0.504	−0.016	0.384	0.003	0.394	−0.009	0.264	0.002	0.272	−0.005	0.135	0.001	0.138	−0.002
5	0.320	0.002	0.333	0.000	0.635	0.004	0.662	−0.047	0.479	0.003	0.499	−0.021	0.376	0.003	0.392	−0.012	0.259	0.002	0.270	−0.006	0.132	0.001	0.138	−0.002
4	0.312	0.002	0.333	0.000	0.607	0.004	0.647	−0.060	0.461	0.003	0.492	−0.027	0.363	0.003	0.387	−0.016	0.251	0.002	0.268	−0.008	0.128	0.001	0.137	−0.003
3	0.296	0.002	0.333	0.000	0.554	0.003	0.623	−0.078	0.426	0.003	0.479	−0.037	0.338	0.002	0.380	−0.022	0.234	0.002	0.264	−0.011	0.120	0.001	0.135	−0.005
2	0.250	0.002	0.333	0.000	0.433	0.002	0.577	−0.100	0.341	0.002	0.455	−0.049	0.273	0.002	0.364	−0.030	0.191	0.001	0.255	−0.016	0.099	0.001	0.132	−0.007

Full: the Gini calculated without grouping. First-order: Gini after the first-order correction. Cov: covariance term from equation 10.
Each simulation exercise is based on 20,000 independent samples of size $n = 10{,}000$.

bias is increasing with the value of the standard deviation σ_y, but we postpone a more comprehensive discussion of this issue to the beta distributions. With respect to the performance of the first-order correction, we confirm that it removes all bias for the uniform distribution and removes a large share of the bias for the log-normal distributions. While the covariance terms are always negative for the log-normal distributions—implying that the first-order correction "undercorrects" (that is, the first-order corrected Gini is lower than the one calculated from individual data)—the first-order correction performs better for log-normal distributions with lower σ_y.

The results for the beta distributions have been summarized using the response surface methodology (Hendry, 1984). This method summarizes the 36 Monte Carlo simulations (each consisting of 20,000 independent samples of size $n = 10{,}000$) by treating each of the 36 sets of simulations as a single observation in an OLS model; this is done separately for each value of $K = 2,3,\ldots,10,20,30,40,50$. More exactly, for each value of K, we first calculate the average bias (before and after applying the first-order correction term) for each set of 20,000 simulations and next use these 36 averages as the dependent variable in an OLS model. We explain these biases as a function of the normalized variance, normalized skewness, and normalized kurtosis of the beta distribution.[11]

Table 2 gives the resulting OLS estimates for different values of K, which are in line with our findings for the uniform and log-normal distributions in table 1. The R^2's indicate that we explain a major share of the biases. We find that the bias of the Gini due to grouping is an increasing function of the variance and that it is hardly affected by the skewness or kurtosis. The relative importance of the latter moments increases slightly for the second-order bias, but the variance remains the most important factor. The much lower coefficient estimates in the right panel reflect that the first-order correction removes a major part of the bias, and the reduction in the size of all coefficients estimates when K increases reflects that the bias due to grouping and the second-order bias are decreasing functions of K.

While the response surface methodology is useful for summarizing the Monte Carlo simulations, two interesting features are not revealed in table 2.[12] First, the first-order correction term removes a major share of the bias due to grouping in all 36 simulations, including distributions with very different shapes from the typically right-skewed income distributions. This is evident from comparing the columns Gini and FOC for the beta distributions in table 3 (for completeness, we also present the corresponding summary percentages for the Monte Carlo simulations based on

[11] We use the normalized versions of these moments to ensure that results are scale free (Deltas, 2003). The variance is divided by the square of the mean, and the skewness and kurtosis, respectively, by the cube and the fourth power.

[12] As we mentioned earlier, we also found that the first-order correction gets it exactly right for a beta distribution with parameters 0.5 and 1.

TABLE 2.—BIAS OF THE GINI DUE TO GROUPING: RESPONSE SURFACE ESTIMATES USING THE BETA DISTRIBUTION

	Dependent Variable: Bias of Gini Due to Grouping					Dependent Variable: Second-Order Bias				
Groups	Variance	Skewness	Kurtosis	Constant	R^2	Variance	Skewness	Kurtosis	Constant	R^2
50	0.0003***	0.0000**	−0.0000**	0.0001***	0.9555	0.0001***	0.0000***	−0.0000***	0.0000**	0.8851
40	0.0004***	0.0000**	−0.0000**	0.0001***	0.9555	0.0002***	0.0000***	−0.0000***	0.0000**	0.8828
30	0.0007***	0.0000**	−0.0000**	0.0002***	0.9553	0.0003***	0.0000***	−0.0000***	0.0000**	0.8793
20	0.0016***	0.0000*	−0.0000*	0.0004***	0.9549	0.0006***	0.0000***	−0.0000***	0.0001**	0.8742
10	0.0059***	0.0000	−0.0000	0.0013***	0.9531	0.0020***	0.0000***	−0.0000**	0.0003**	0.8624
9	0.0072***	0.0000	−0.0000	0.0017***	0.9526	0.0023***	0.0000***	−0.0000**	0.0003*	0.8600
8	0.0090***	0.0000	−0.0000	0.0021***	0.9521	0.0028***	0.0000**	−0.0000**	0.0004*	0.8574
7	0.0116***	0.0000	−0.0000	0.0027***	0.9513	0.0035***	0.0000**	−0.0000**	0.0005*	0.8539
6	0.0154***	0.0000	−0.0000	0.0036***	0.9502	0.0045***	0.0000**	−0.0000**	0.0006*	0.8496
5	0.0217***	0.0000	−0.0000	0.0051***	0.9486	0.0060***	0.0000**	−0.0000**	0.0008*	0.8440
4	0.0327***	0.0000	−0.0000	0.0078***	0.9461	0.0084***	0.0000**	−0.0000**	0.0011*	0.8354
3	0.0554***	0.0000	−0.0000	0.0136***	0.9415	0.0127***	0.0000**	−0.0000**	0.0016*	0.8207
2	0.1157***	−0.0000	0.0000	0.0294***	0.9314	0.0219***	0.0000*	−0.0000*	0.0028	0.7889

The response surface estimates are based on 36 observations, obtained from the underlying Monte Carlo simulations using the beta distribution with all combinations of parameters equaling 0.5, 1, 3, 5, 10, and 25. The dependent variable in the left panel is the bias of the Gini due to grouping: $G_n - G_n^K$. The right panel uses the second-order bias: $G_n - G_n^K \{[K^2(n^2 - 1)]/[n^2(K^2 - 1)]\}$. Variance: normalized variance (divided by square of mean); skewness: normalized skewness (divided by cube of mean); kurtosis: normalized kurtosis (divided by fourth power of mean). Significance levels are ***1%; **5%; *10%.

TABLE 3.—BIAS OF THE GINI DUE TO GROUPING: SUMMARY PERCENTAGES OF THE SIMULATION EXERCISES

	Log Normal						Beta Distribution, except Beta(0,5;0,5)							
	Gini			FOC			Gini			FOC			Beta(0.5;0.5)	
Groups	min	mean	max	min	mean	max	min	mean	max	min	mean	max	Gini	FOC
50	99.68	99.84	99.92	99.72	99.88	99.96	99.89	99.93	99.96	99.93	99.97	100.00	99.97	100.01
40	99.54	99.77	99.88	99.60	99.83	99.94	99.84	99.89	99.94	99.90	99.96	100.00	99.95	100.01
30	99.29	99.62	99.80	99.40	99.74	99.91	99.73	99.81	99.89	99.84	99.92	100.00	99.91	100.02
20	98.67	99.26	99.56	98.91	99.51	99.81	99.43	99.60	99.75	99.68	99.85	100.00	99.79	100.04
10	96.19	97.63	98.41	97.16	98.62	99.41	98.02	98.51	99.00	99.01	99.50	100.00	99.18	100.18
9	95.54	97.18	98.07	96.74	98.39	99.30	97.61	98.18	98.77	98.83	99.41	100.00	98.98	100.22
8	94.69	96.57	97.60	96.19	98.10	99.15	97.06	97.74	98.44	98.60	99.29	100.00	98.71	100.28
7	93.53	95.72	96.93	95.48	97.71	98.95	96.27	97.09	97.96	98.28	99.12	100.00	98.32	100.36
6	91.88	94.47	95.93	94.51	97.17	98.67	95.11	96.13	97.22	97.83	98.87	100.00	97.70	100.50
5	89.40	92.53	94.32	93.13	96.39	98.25	93.27	94.57	96.00	97.16	98.51	100.00	96.69	100.72
4	85.35	89.22	91.47	91.05	95.17	97.57	90.08	91.79	93.75	96.08	97.91	100.00	94.80	101.13
3	77.90	82.76	85.65	87.63	93.10	96.36	83.72	86.08	88.89	94.19	96.84	100.00	90.69	102.03
2	60.93	66.76	70.35	81.24	89.01	93.79	67.70	70.92	75.00	90.27	94.56	100.00	78.54	104.72

All results are presented as a proportion of the Gini calculated without grouping. Gini: Gini index; FOC: Gini after the first-order correction; mean (min/max): the mean, minimum, and maximum value across the simulation exercises, that is, 5 log-normal and 35 beta distributions. The final column shows the results for the beta distribution with parameters 0.5 and 0.5.

the log-normal distributions). Second, $\text{cov}(\delta_i^g, \varepsilon_i)$ was always negative or 0, except for the beta distribution with parameters 0.5 and 0.5 (see the far right column in table 3). While a positive covariance cannot be excluded a priori, our simulations indicate that it will only rarely occur: it was negative or 0 for a wide range of shapes of the density function including left and right skewness but did show up positive in our Monte Carlo simulations for the beta distribution with parameters 0.5 and 0.5. The density function of the latter distribution is symmetric around 0.5 and bimodal with high spikes at the minimum and maximum income levels, and low density at intermediate income levels. For example, the 15% (4%) highest and lowest incomes each account for more than 25% (one-eighth) of total income. Hence, the error of equation 2 and the measurement error defined in equation 5 are on average positive for the lowest and negative for the highest incomes within each income group. This is most easily seen from the most extreme bimodal and symmetric distribution: a density with 50% of mass at the income level 0 and 50% at the income level 1, but will hold true more generally as long as the density function will have sufficient mass at both the minimum and maximum income levels. It should be clear from this discussion that a positive covariance is unlikely to occur in practice, and especially so for income distributions that tend to be right skewed rather than bimodal with spikes of comparable height at the maximum and minimum income levels. But even when a positive covariance might occur, the two far-right columns in table 3 show that the first-order corrected Gini index removes a major share of the dependence on grouping. The fact that it "overcorrects," rather than "undercorrects," in this case, seems unimportant in light of the share of the dependence on grouping it tends to remove.

A final issue concerns whether one can correct for the remaining second-order bias after having used the first-order correction term. While the Monte Carlo simulations gave some idea on the magnitude of the bias, one would in theory need the unobservable underlying individual-level data to get rid of the second-order bias in practical applications. In a related context, Deltas (2003) has noted that "it might sometimes be possible to account, at least partially, for the second-order bias if some information can be obtained about the distribution. In particular, one may be able to estimate the density . . . and then compare . . . with

standard parametric distributions ... to calculate the bias correction term" (p. 231). In the empirical section, we follow this logic but estimate equation (10) from individual-level data rather than relying on standard parametric distributions.

B. Groups of Unequal Size

Until now we have assumed that the K groups are equally sized. Equation (10) is, however, easily generalized to groups of unequal size. Assume that n_u is the number of observations in group $u = 1,\ldots,K$ (with u referring to unequal group size), that $R_u = (n)^{-1}\left(1/2n_u + \sum_{j=1}^{u-1} n_j\right)$ equals the fractional income rank of group u, and that the variance of the latter is defined as $\sigma^2_{R_u^K} = (n)^{-1} \sum_{u=1}^{K} n_u(R_u - 1/2)^2$. We have now sufficient information to derive the equivalent expressions of equations (3) and (4):

$$2\sigma^2_{R_u^K} \frac{y_u}{\bar{y}} \sqrt{n_u} = \alpha^u \sqrt{n_u} + \beta^u R_u \sqrt{n_u} + \varepsilon_u \sqrt{n_u}, \tag{12}$$

$$\begin{aligned} \beta^u = G_n^{K,u} &= \frac{2\sum_{u=1}^{K} n_u y_u R_u}{n\bar{y}} - 1 \\ &= \frac{2\sum_{u=1}^{K} \left[\left(\frac{n_u}{n}\right) y_u R_u\right]}{\bar{y}} - 1 \le G_n = \beta. \end{aligned} \tag{13}$$

Equation (12) is a WLS generalization of equation (3), and equation (13) reduces to equation (4) if all groups have equal size. The relationship between $G_n^{K,u}$ and G_n is established by combining equation (2) with an unequal size generalization of equation (5),

$$R_i^u = R_i + \delta_i^u, \tag{14}$$

where δ_i^u is the measurement error with zero mean and R_i^u is the fractional income rank of group u defined at the individual level. This results in

$$G_n = \frac{\sigma^2_R}{\sigma^2_{R_u^K}} G_n^{K,u} - \frac{\frac{1}{n}\sum_{i=1}^{n} \delta_i^u \varepsilon_i}{\sigma^2_{R_u^K}}. \tag{15}$$

It is straightforward to see that equations (10) and (15) are identical except for the unequal group sizes. It is still the case that the first-order correction term (a) is related to the so-called attenuation bias of the classical measurement error model in that it measures the ratio of the variance of the actual fractional rank and that of the fractional rank of group u, (b) it is easy to calculate, and (c) it depends on only the relative size of the groups. The expression of the second-order bias also still reflects the performance of the first-order correction term and the covariance interpretation remains. The same can be said about the main findings of the Monte Carlo simulations before, although the interplay between the shape of the underlying distribution and the relative size of the groups is now an additional factor.

IV. Empirical illustration

A. Data

In this section, we illustrate the dependence of the Gini index of income on the number of groups and show the performance of the first-order correction term in reducing the bias if applied to income distributions. We analyzed this bias for fifteen European countries and the United States using microdata from the European Community Household Panel (ECHP) and the Medical Expenditure Panel Survey (MEPS).[13] The ECHP was designed and coordinated by EUROSTAT. It contains socioeconomic information for individuals aged 16 or older, uses a standardized questionnaire, and covers fifteen EU member states: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, and the United Kingdom. We use the first wave (the 1994 wave) for all countries except for Austria (joined the survey in 1995), Finland (joined in 1996), and Sweden (joined in 1997). We supplement this with U.S. income microdata from the 2000 wave of the MEPS. We use the first wave of the ECHP because it does not suffer from attrition, and thus has more observations, which is useful for illustrating the first-order correction term and the dependence of the Gini on the number of income groups. Note that all calculations in this section serve only the purpose of illustrating the methods explained in the previous sections, not to deliver any hard evidence on income inequality in the EU and United States.

The key variable for this study is income. The ECHP and MEPS income measures provide annual equivalent disposable (after-tax) household income. Table A1 reports descriptive statistics of equivalent income in each of the countries. As we are analyzing the behavior of estimates of the Gini index for varying grouping sizes, it is reassuring to note that all samples are large (at least 5,500 observations, except for Luxembourg, which has about 2,000 observations).

The analysis takes three steps. First, we calculate the Gini index based on the ECHP and MEPS data sets. Second, we create income categories from the full samples and analyze the effect that follows from these groupings. Third, we illustrate the performance of the first-order correction term in terms of reducing the underestimation. We also present similar evidence on a procedure to address the remaining second-order bias.

B. Gini Index and Number of Income Groupings

We present the estimates of the Gini indices based on the full samples of the ECHP and MEPS in table 4 (see row

[13] We also used Dutch administrative data on more than 5 million individual income tax files for 2004. The findings based on these Dutch administrative data are very much in line with those resulting from the European and U.S. microdata.

TABLE 4.—GINI INDEX IN THE ECHP AND MEPS: ADDRESSING THE BIAS DUE TO GROUPING BY FIRST-ORDER CORRECTING, SECOND-ORDER CORRECTING, AND USING POVCAL

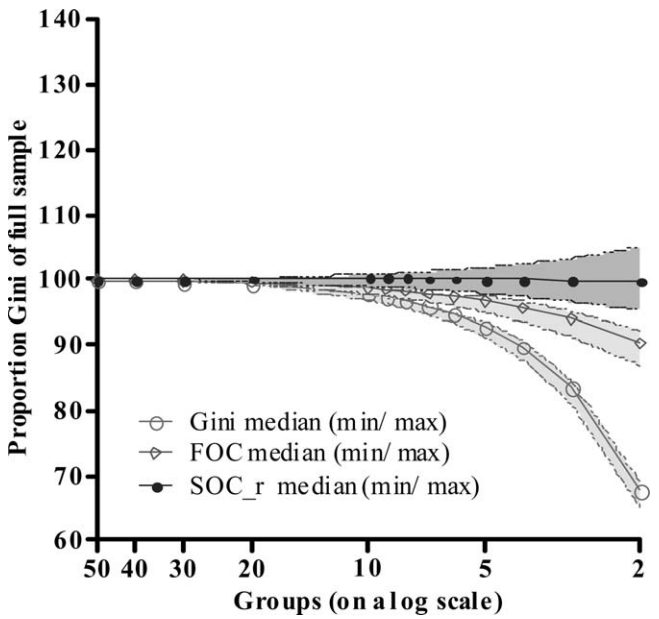
	Sweden				Denmark				Finland				Netherlands				Austria				Belgium			
Groups	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC
Full	0.218				0.233				0.234				0.260				0.280				0.297			
50	0.218	0.218	0.218	0.217	0.233	0.233	0.233	0.233	0.233	0.234	0.234	0.234	0.259	0.260	0.260	0.259	0.279	0.280	0.280	0.280	0.297	0.297	0.297	0.297
40	0.217	0.218	0.218	0.217	0.233	0.233	0.233	0.233	0.233	0.233	0.234	0.234	0.259	0.259	0.260	0.259	0.279	0.279	0.280	0.280	0.296	0.297	0.297	0.297
30	0.217	0.217	0.218	0.217	0.232	0.232	0.233	0.233	0.233	0.233	0.234	0.234	0.259	0.259	0.260	0.259	0.279	0.279	0.280	0.280	0.296	0.296	0.297	0.297
20	0.216	0.217	0.218	0.217	0.231	0.232	0.233	0.233	0.232	0.233	0.234	0.234	0.258	0.259	0.260	0.260	0.278	0.279	0.280	0.280	0.295	0.295	0.297	0.297
10	0.213	0.215	0.219	0.217	0.227	0.230	0.234	0.232	0.228	0.231	0.234	0.234	0.254	0.257	0.261	0.260	0.274	0.277	0.281	0.280	0.290	0.293	0.297	0.296
9	0.212	0.215	0.220	0.217	0.226	0.229	0.234	0.232	0.227	0.230	0.235	0.234	0.253	0.256	0.261	0.260	0.273	0.276	0.281	0.279	0.288	0.292	0.297	0.296
8	0.211	0.215	0.220	0.217	0.225	0.229	0.234	0.232	0.226	0.229	0.235	0.234	0.252	0.256	0.261	0.260	0.271	0.275	0.281	0.279	0.287	0.291	0.296	0.296
7	0.210	0.214	0.220	0.217	0.223	0.228	0.234	0.232	0.224	0.228	0.235	0.234	0.250	0.255	0.261	0.260	0.269	0.274	0.281	0.279	0.284	0.290	0.296	0.296
6	0.207	0.213	0.221	0.217	0.220	0.227	0.234	0.232	0.221	0.227	0.235	0.234	0.247	0.254	0.262	0.260	0.265	0.273	0.281	0.279	0.280	0.288	0.296	0.295
5	0.203	0.212	0.222	0.217	0.216	0.225	0.235	0.231	0.216	0.225	0.235	0.234	0.242	0.252	0.262	0.260	0.260	0.271	0.281	0.279	0.274	0.286	0.296	0.295
4	0.197	0.210	0.223	0.216	0.209	0.222	0.236	0.231	0.209	0.223	0.236	0.234	0.234	0.250	0.263	0.260	0.251	0.268	0.281	0.279	0.265	0.282	0.296	0.294
3	0.183	0.206	0.225	0.325	0.194	0.218	0.237	0.219	0.194	0.218	0.237	NA	0.219	0.246	0.264	NA	0.234	0.263	0.281	NA	0.246	0.277	0.296	NA
2	0.149	0.199	0.229	NA	0.157	0.210	0.239	NA	0.157	0.210	0.239	NA	0.178	0.238	0.267	NA	0.189	0.253	0.282	NA	0.200	0.266	0.295	NA

	Luxembourg				Ireland				Germany				Italy				Spain				France			
	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC
Full	0.304				0.306				0.312				0.330				0.335				0.343			
50	0.304	0.304	0.304	0.304	0.305	0.305	0.305	0.305	0.312	0.312	0.312	0.312	0.329	0.330	0.330	0.329	0.334	0.334	0.335	0.334	0.342	0.342	0.343	0.343
40	0.304	0.304	0.304	0.304	0.305	0.305	0.305	0.305	0.311	0.312	0.312	0.312	0.329	0.329	0.330	0.329	0.334	0.334	0.335	0.334	0.342	0.342	0.343	0.343
30	0.303	0.303	0.304	0.304	0.304	0.304	0.305	0.305	0.311	0.311	0.312	0.312	0.329	0.329	0.330	0.329	0.334	0.334	0.335	0.334	0.341	0.341	0.342	0.343
20	0.302	0.303	0.304	0.304	0.303	0.304	0.305	0.305	0.310	0.310	0.312	0.312	0.328	0.329	0.330	0.329	0.333	0.333	0.335	0.334	0.339	0.340	0.342	0.343
10	0.298	0.301	0.304	0.304	0.298	0.301	0.305	0.305	0.304	0.307	0.311	0.312	0.323	0.326	0.330	0.329	0.328	0.331	0.335	0.334	0.332	0.335	0.339	0.343
9	0.296	0.300	0.305	0.303	0.297	0.300	0.305	0.305	0.303	0.306	0.311	0.312	0.322	0.326	0.330	0.329	0.326	0.330	0.335	0.334	0.330	0.334	0.339	0.343
8	0.295	0.299	0.305	0.303	0.295	0.300	0.305	0.305	0.301	0.305	0.311	0.312	0.320	0.325	0.330	0.329	0.324	0.330	0.335	0.334	0.328	0.333	0.338	0.342
7	0.292	0.298	0.304	0.303	0.293	0.299	0.305	0.304	0.298	0.304	0.311	0.312	0.317	0.324	0.330	0.328	0.322	0.328	0.335	0.334	0.325	0.331	0.338	0.342
6	0.288	0.297	0.305	0.303	0.289	0.297	0.305	0.304	0.294	0.302	0.310	0.312	0.313	0.322	0.330	0.328	0.318	0.327	0.335	0.334	0.320	0.329	0.337	0.342
5	0.283	0.295	0.305	0.303	0.283	0.295	0.305	0.304	0.288	0.300	0.310	0.312	0.307	0.320	0.330	0.327	0.312	0.325	0.335	0.335	0.313	0.326	0.336	0.342
4	0.274	0.292	0.305	0.304	0.274	0.292	0.305	0.300	0.277	0.296	0.309	0.311	0.297	0.317	0.330	0.326	0.301	0.321	0.335	0.335	0.301	0.321	0.334	0.341
3	0.255	0.287	0.305	0.319	0.255	0.287	0.306	0.361	0.257	0.289	0.308	0.334	0.277	0.312	0.330	0.305	0.280	0.315	0.334	NA	0.278	0.313	0.331	0.364
2	0.206	0.275	0.304	NA	0.209	0.279	0.308	NA	0.207	0.277	0.306	NA	0.227	0.302	0.331	NA	0.227	0.303	0.332	NA	0.223	0.298	0.327	NA

	United Kingdom				Greece				Portugal				United States			
Groups	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC	Gini	FOC	SOC_r	POVC
Full	0.362				0.367				0.393				0.395			
50	0.361	0.362	0.362	0.362	0.366	0.366	0.366	0.366	0.392	0.392	0.393	0.393	0.394	0.394	0.395	0.394
40	0.361	0.361	0.362	0.362	0.366	0.366	0.366	0.366	0.392	0.392	0.393	0.393	0.394	0.394	0.395	0.394
30	0.360	0.361	0.362	0.362	0.365	0.366	0.366	0.366	0.391	0.392	0.393	0.393	0.393	0.394	0.395	0.394
20	0.359	0.360	0.361	0.361	0.364	0.365	0.366	0.366	0.390	0.391	0.392	0.393	0.392	0.393	0.395	0.394
10	0.352	0.356	0.360	0.361	0.358	0.362	0.366	0.365	0.384	0.388	0.392	0.393	0.387	0.391	0.395	0.394
9	0.351	0.355	0.359	0.361	0.356	0.361	0.365	0.365	0.382	0.387	0.391	0.393	0.386	0.390	0.395	0.394
8	0.348	0.354	0.359	0.360	0.354	0.360	0.365	0.365	0.380	0.386	0.391	0.393	0.383	0.390	0.395	0.393
7	0.345	0.352	0.359	0.360	0.351	0.359	0.365	0.365	0.376	0.384	0.391	0.393	0.380	0.388	0.395	0.393
6	0.340	0.350	0.358	0.359	0.347	0.357	0.365	0.365	0.372	0.382	0.390	0.393	0.376	0.387	0.395	0.393
5	0.333	0.347	0.357	0.359	0.340	0.354	0.364	0.364	0.364	0.379	0.389	0.393	0.369	0.385	0.395	0.393
4	0.321	0.343	0.356	0.358	0.329	0.351	0.364	0.364	0.351	0.375	0.388	0.394	0.357	0.381	0.394	0.392
3	0.298	0.335	0.354	0.336	0.306	0.345	0.363	NA	0.326	0.367	0.385	0.419	0.333	0.375	0.394	NA
2	0.242	0.322	0.351	NA	0.248	0.331	0.360	NA	0.263	0.350	0.380	NA	0.273	0.364	0.393	NA

Gini: Gini index; FOC: Gini after the first-order correction; SOC_r: Gini after the first-order correction and the second-order correction where the latter is derived from the OLS regression in equation (16) on individual-level data; POVC: Gini estimate obtained from the POVCAL computational tool (World Bank, 2008) using the general quadratic Lorenz curve (Villaseñor & Arnold, 1989) or, when b is added, the one proposed by Kakwani (1980b); NA: not available.

FIGURE 1.—GINI, ITS DEPENDENCE ON INCOME GROUPING, AND CORRECTING FOR THIS DEPENDENCE IN THE EU AND UNITED STATES



All results are presented as a proportion of the Gini calculated from the full sample. Gini: the Gini estimated from grouped income data; FOC: the Gini after applying the first-order correction term; SOC_r: Gini after the first-order correction and the second-order correction where the latter is derived from the OLS regression in equation (16) on individual-level data; median (min/max): the median (line), and minimum and maximum value (shaded region) across countries.

"full") and have ranked countries from low to high relative income inequality. These estimates in this study are considered the benchmark estimates against which the effect of grouping the data is evaluated. Similar to the Monte Carlo simulations in section III, we have subdivided the full sample into $K = 2,3,\ldots,10,20,30,40,50$ equally sized (equivalent) income categories and used the average equivalent incomes of each income category to calculate the Gini index using equation 3. The resulting estimates for each value of K are presented in the "Gini" column in table 4 and are expressed as a proportion of the Gini's estimated from the full sample in figure 1 and in the "Gini" column in table 5—$100 \times (G_n^K/G_n)$.

We confirm the findings on the bias due to grouping obtained from the Monte Carlo simulations. First, due to the convexity of Lorenz curves, the Gini index based on grouped data always underestimates the one in the full sample. Second, figure 1 and table 5 reveal that the underestimation, expressed in relative terms, is similar across countries. The range of the underestimation across countries is low, suggesting that the shape of the underlying income distributions is similar across countries but that the spread differs, which is in line with the Monte Carlo evidence that the bias is an increasing function of the variance. Third, the underestimation of the Gini index due to grouping the data increases at an increasing pace when lowering the number of income categories, and matches the findings of Davies and Shorrocks (1989). It seems that most of the action is taking place for twenty or fewer income groups. In the extreme case of two income groups, the Gini index based

TABLE 5.—DEPENDENCE OF THE GINI INDEX ON INCOME GROUPING IN THE EU AND UNITED STATES: SUMMARY PERCENTAGES

	Gini			FOC			SOC_r			POVC		
Groups	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
50	99.72	99.83	99.88	99.76	99.87	99.92	99.86	100.00	100.07	99.80	99.89	100.02
40	99.59	99.76	99.82	99.66	99.82	99.88	99.81	99.99	100.10	99.79	99.89	100.02
30	99.37	99.61	99.70	99.48	99.72	99.81	99.72	99.99	100.15	99.78	99.88	100.02
20	98.84	99.25	99.39	99.09	99.50	99.64	99.52	99.98	100.29	99.75	99.87	100.03
10	96.78	97.65	97.97	97.76	98.64	98.96	98.91	99.96	100.75	99.59	99.80	100.04
9	96.24	97.21	97.54	97.44	98.43	98.76	98.77	99.95	100.86	99.54	99.78	100.05
8	95.52	96.62	96.98	97.04	98.16	98.52	98.61	99.95	100.99	99.47	99.75	100.05
7	94.56	95.79	96.23	96.53	97.79	98.23	98.41	99.94	101.20	99.38	99.70	100.04
6	93.20	94.59	95.09	95.87	97.29	97.81	98.17	99.93	101.44	99.22	99.65	100.08
5	91.12	92.73	93.35	94.92	96.59	97.24	97.83	99.91	101.83	99.00	99.60	100.20
4	87.61	89.54	90.32	93.45	95.51	96.34	97.29	99.90	102.39	98.07	99.43	100.33
3	81.04	83.29	84.15	91.17	93.70	94.67	96.58	99.88	103.20	NA	NA	NA
2	65.09	67.61	68.68	86.79	90.14	91.58	95.30	99.86	104.92	NA	NA	NA

All results are presented as a proportion of the Gini calculated from the full sample. Gini: Gini index; FOC: Gini after the first-order correction; SOC_r: Gini after the first-order correction and the second-order correction where the latter is derived from the OLS regression in equation (16) on individual-level data; POVC: Gini estimate obtained from the POVCAL computational tool (World Bank, 2008); Mean (Min/Max): the mean, and minimum, and maximum value across countries; NA: not available as POVCAL did not provide an estimate for all countries (see also table 4).

on grouped income data is only between 65% and 69% of the one based on the full sample. For five income groups, the underestimation is between 7% and 9%, and for ten income groups, the underestimation still amounts to about 2% to 3%. We can safely conclude that these percentages represent important underestimations since we find that the magnitude of the underestimation is substantial compared to the sampling variability of the Gini index[14] (which confirms the Monte Carlo evidence), and since it is large compared to the evolution of the Gini over time in the full sample.[15]

C. *Reduction of Underestimation after First-Order Correction*

This section discusses the performance of the first-order correction term as applied to income distributions. Tables 4 and 5 and figure 1 also present results for a procedure that corrects for the remaining second-order bias, after having applied the first-order correction term (see also the final paragraph in section IIIA), and tables 4 and 5 also include estimates obtained from the POVCAL software tool (World Bank, 2008).

The results for the first-order correction term (see FOC) are obtained by multiplying the values of the Gini calculated from grouped data (see Gini) by $\sigma_R^2/\sigma_{R^K}^2 = [K^2(n^2-1)]/[n^2(K^2-1)]$. Conditional on observing a grouped data estimate of the Gini, the first-order correction term thus requires information on $n =$ size of the full sample (all observations) and $K =$ number of groups. (Obviously the size of each group is n/K under the assumption that each group is of equal size). The procedure to remove the second-order bias (see SOC_r) uses an empirical estimate of $n^{-1}\sum_{i=1}^{n}\delta_i^g\varepsilon_i$ and next applies equation 10 to the Gini calculated from grouped data.[16] While it is impossible to observe this covariance term without observing the underlying individual-level data, one might obtain an estimate from income distributions with a similar shape that are recorded at the individual level. To this means, we have used the between-country variation in the underlying individual-level data of all sixteen countries to identify one covariance term $\mu_K = n^{-1}\sum_{i=1}^{n}\delta_i^g\varepsilon_i$ for each value of $K = 2,3, \ldots ,49,50$. Next, we have applied this single estimate of the covariance term to correct for the second-order bias in all countries. More exactly, we use the regression model that results from rearranging and dividing equation 10 by G_n,

$$\frac{G_n^K}{G_n} = \frac{n^2(K^2-1)}{K^2(n^2-1)} + \mu_k\frac{12}{G_n}\frac{n^2}{(n^2-1)}, \tag{16}$$

and apply OLS (excluding a constant) on 784 observations: 49 income groupings for sixteen countries. We find that the latter regression fits the data very well (the uncentered and standard R^2 equal 1 and 0.982, respectively), and is therefore preferable over a simple mean or median over the sixteen countries of these covariance terms since it imposes the relationship implied by equation (10) on the covariance terms. All 49 covariance terms are negative, take a low value that increases monotonically with the number of income groupings, and all terms are precisely estimated. We report these 49 terms in table A2.

Finally, we have also calculated results using the POVCAL software tool of the World Bank (2008). Our main goal for reporting these results (see POVC) is to compare our procedure to correct for the second-order bias with an existing method.[17] Note that the first-order correction has far lower information requirements than the parametric functional forms implemented in POVCAL, as the latter also require information on the average incomes per income group. The POVCAL software tool estimates Gini indices from grouped data by fitting a parametric Lorenz curve to the average incomes of each income group. It uses the general quadratic Lorenz curve (Villaseñor & Arnold, 1989) or the functional form proposed by Kakwani (1980b). In order to put our procedure to correct for the second-order bias to a strong comparative test, we present the functional form closest to the benchmark estimate obtained from the full sample data.[18]

Note from tables 4 and 5 and figure 1 that the first-order correction term (FOC) reduces a large share of the underestimation in each of the sixteen countries, but the remaining underestimation is higher at a low number of income groups. Also, application of the first-order correction term never results in an overestimation of the Gini index. Both observations are in line with the Monte Carlo evidence presented before.[19]

[14] We obtained 95% confidence intervals for the Gini index using the bootstrap (see, e.g., Mills & Zandvakili, 1997). For all countries, the Ginis resulting from six or fewer income groupings were not included in these confidence intervals.

[15] For all countries in the ECHP, we have calculated the proportional change in the Gini between the first available and last wave using a balanced panel, and calculated the underestimation that results from grouping the data in the first wave of the balanced panel. We find that in all countries, the proportional change in the Gini over time (eight years for most countries) is smaller than the underestimation, expressed in relative terms, resulting from five income groups.

[16] The extension to groups of unequal size is straightforward and based on equation (15).

[17] We did not calculate nonparametric bounds as the latter provide a range rather than a point estimate of the Gini and are therefore a less interesting point of comparison.

[18] Note that POVCAL reports no results for two income groups since three coefficients need to be estimated for the general quadratic Lorenz curve and the one proposed by Kakwani (1980b). In some cases, POVCAL reports no results for three income groups since the conditions for a valid Lorenz curve were violated, that is, going through (0,0), (1,1), monotonically increasing and convex.

[19] We also confirmed that the first-order correction might be helpful in cross-country comparative research when there are different numbers of income groupings per country, especially when the underlying average incomes per income group are not observed. For example, using the estimates in table 4, we checked how income grouping in one country (and using the full sample indices for the other countries) affects the income inequality ranking of the sixteen countries based on the original full samples and how this effect is neutralized by using the first-order correction. We find that changes in the income inequality ranking occur frequently, especially in case of a low number of income groups, and that the first-order correction often neutralizes the latter effect. We reached similar conclusions when studying the effect of income groupings on longitudinal variation, which, for example, refers to the case where the number of income categories used in a questionnaire changes over time.

Columns SOC_r and POVC in tables 4 and 5 and figure 1 present evidence on the empirical performance of our procedure to address the second-order bias. We find that it reduces the impact of grouping in all countries; the resulting second-order corrected Gini is much closer to the Gini calculated from the full sample. Table 4 also shows that the second-order correction might overcorrect (that is, the second-order corrected Gini is higher than the one calculated from individual data in several countries). Nevertheless, the summary percentages in columns FOC and SOC_r in table 5 show that it always outperforms the first-order correction.

Comparing columns SOC_r and POVC in tables 4 and 5 reveals how our procedure to address the second-order bias compares with the results obtained from the POVCAL software tool. We find that SOC_r always outperforms POVC for two and three income groupings for obvious reasons (see note 18). For four or more income groups, table 5 shows that both methods perform equally well on average (consult columns "Mean"), but POVC has a lower range compared to SOC_r (compare columns "Min" and "Max"). Note also that both methods might give rise to an overcorrection (see the "Max" columns). While the comparison based on the range might seem unfavorable for SOC_r, it should not distract attention from the fact that SOC_r outperforms POVC in four countries for each value of $K = 4,5, \ldots ,49,50$, while this occurs only twice for POVC. Therefore, it thus seems that neither of the underlying assumptions is unequivocally superior in removing the underestimation due to grouping, that is, imposing a specific functional form in case of POVCAL, or imposing an estimate of the covariance term in the other case. The latter is always feasible (for example, by using the values reported in this study), while the former requires information on the average incomes per income group. Nevertheless, there is always an issue of external validity when imposing an estimate of the covariance term. This is likely to be important when the covariance terms are estimated from a single data set, but in this empirical application, we have used income distributions of sixteen countries that differ greatly in degree of income inequality.[20] This does not mean that fewer assumptions are imposed when using POVAL since one has to impose the functional form.[21]

Overall, we conclude that the first-order correction performs well empirically and removes a large share of the underestimation due to grouping (and this is backed by extensive Monte Carlo evidence). When information on the average incomes per income group is available—which is not always the case, such as, for example, for the majority of countries in the UNU-WIDER World Income Inequality Database (UNU-WIDER, 2008)—or when estimates of the covariance term are available (for example, by using the values reported in this study), one might address the remaining second-order bias by fitting parametric functions or imposing a value of the covariance term. Neither method unequivocally outperforms the other, and both methods differ in informational requirements.

[20] The fact that we impose these terms to each country (and thus each country being used to estimate these terms) is unimportant since we have sixteen countries.

[21] We also note that we have presented the estimates based on the functional form that is closest to the benchmark estimate obtained from the full sample. In several cases, this choice did not coincide with goodness-of-fit measures reported by POVCAL.

V. Discussion and Conclusion

This paper analyzes the downward bias of the Gini index due to grouped data complicating comparisons of Gini indices calculated from such data. We develop a first-order correction term that results from studying the Gini in a measurement error framework and show that it resembles the so-called attenuation bias in the classical measurement error model and that it is inversely related to the covariance between the fractional rank at the individual and group levels. Besides its simplicity and transparency, the first-order correction allows, in contrast to existing methods, addressing the bias due to grouping in case one has access to estimates of the Gini index based on grouped data without observing the underlying average incomes or income ranges. Instead, it needs only information on the number of individuals in each income group or range. We have also derived an exact and intuitive expression for the remaining and distribution-specific second-order bias, allowing assessing a priori the performance of the first-order correction. We show that the second-order bias is zero for specific distribution functions and that the first-order correction term generally results in a small undercorrection. In addition, Monte Carlo evidence reveals that the first-order correction performs well for a wide range of underlying distribution functions (including bimodality and left and right skewness) and that the second-order bias is increasing in the variance of the underlying distribution.

Using microdata from the ECHP and MEPS on income distributions of fifteen European countries and the United States, we illustrate that the underestimation from income groupings is similar across the sixteen countries. We further illustrate that the underestimation increases at an increasing pace when lowering the number of income categories and that the underestimation is substantial relative to the sampling variability of the Gini index, its evolution over time, and cross-country differences in the value of the Gini. Next, we illustrate the performance of our first-order correction term and show that it reduces the underestimation of the Gini due to income grouping considerably in all countries. We also illustrate that one can address the remaining second-order bias if one is willing to impose additional information. Our results show that this procedure could be used as an alternative to existing correction methods involving fitting conventional parametric forms to the data, but it does not require information on the average incomes per income group.

A final issue concerns the terminology we have used throughout this paper. We have deliberately used "income groupings" to abstract from a situation where the individuals in each income group have the same income. In the latter case, the Gini index estimated from grouped data is not biased, and thus application of our correction term would introduce an upward bias. "Income groupings" instead point to a situation where microdata, official income statistics, and so forth are grouped into a limited number of income groups, and thus neglecting within-income-group income variation leads to an underestimation.

Although the empirical part of this paper deals with the bias due to income groupings of the Gini index, our Monte Carlo simulations suggest that it should be successful in addressing the bias due to grouping in other distributions such as health, education, business concentration, and astronomy. Our simulations encompassed a wide range of distributions, including bimodality, left and right skewness, and the first-order correction improved on the grouped data estimate in all cases. The first-order correction should also be useful for the widely used concentration index that has been applied to taxation (Lambert, 2001) and used to measure inequalities in the health domain (Wagstaff, Paci, & van Doorslaer, 1991; Wagstaff & van Doorslaer, 2000). Its main difference with the Gini is that the fractional rank and the cumulative shares refer to different variables, and thus the bias of the concentration index can be both downward and upward as the underlying concentration curves need not be convex and may have inflection points (Clarke & Van Ourti, 2010).

An important assumption in the theoretical and empirical part of this paper is that we consider measurement error within income groups only. This assumption allows studying the bias due to income groupings of the Gini in isolation but neglects misclassification bias—that an individual might be classified into the wrong income group based on his or her misreported income. It is clear that misclassification and bias due to income groupings might be offsetting each other, and these issues have been analyzed for a Dutch survey for the variance of log incomes, the Theil and Atkinson inequality index by van Praag, Hagenaars, and van Eck (1983). Although we believe future research should analyze the relative importance of both biases in the Gini index, our results show that the bias from income groupings in surveys and administrative data can be considerable.

REFERENCES

Abounoori, Esmaiel, and Patrick McCloughan, "A Simple Way to Calculate the Gini Coefficient for Grouped as Well as Ungrouped Data," *Applied Economics Letters* 10:8 (2003), 505–509.

Abraham, Roberto, Sidney van den Bergh, and Preethi Nair, "A New Approach to Galaxy Morphology. I. Analysis of the Sloan Digital Sky Survey Early Data Release," *Astrophysical Journal* 588:1 (2003), 218–229.

Alker, Hayward, *Mathematics and Politics* (New York: Macmillan, 1965).

Allison, Paul, and John Stewart, "Productivity Differences among Scientists: Evidence for Accumulative Advantage," *American Sociological Review* 39:5 (1974), 596–606.

Basmann, Robert, Kathy Jean Hayes, Daniel Slottje, and John Johnson, "A General Functional Form for Approximating the Lorenz Curve," *Journal of Econometrics* 43:1 (1990), 77–90.

Bigsten, Arne, and Abebe Shimeles, "Can Africa Reduce Poverty by Half by 2015?" *Development Policy Review* 25:2 (2007), 147–166.

Buzzacchi, Luigi, and Tommaso Valletti, "Firm Size Distribution: Testing the 'Independent Submarkets Model' in the Italian Motor Insurance Industry," *International Journal of Industrial Organization* 24:4 (2006), 809–834.

Cameron, Colin, and Pravin Trivedi, *Microeconometrics: Methods and Applications* (Cambridge: Cambridge University Press, 2005).

Clarke, Philip, and Tom Van Ourti, "Calculating the Concentration Index When Income Is Grouped," *Journal of Health Economics* 29:1 (2010), 151–157.

Datt, Gaurav, and Martin Ravallion, "Growth and Redistribution Components of Changes in Poverty Measures: A Decomposition with Applications to Brazil and India in the 1980s," *Journal of Development Economics* 38:2 (1992), 275–295.

Davies, James, Susanna Sandström, Anthony Shorrocks, and Edward Wolff, "The Level and Distribution of Global Household Wealth," *Economic Journal* 121 (2010), 223–254.

Davies, James, and Anthony Shorrocks, "Optimal Grouping of Income and Wealth Data," *Journal of Econometrics* 42:1 (1989), 97–108.

Deltas, George, "The Small-Sample Bias of the Gini Coefficient: Results and Implications for Empirical Research," this REVIEW 85:1 (2003), 226–234.

EUROSTAT, *ECHP UDB Description of Variables: Data Dictionary, Codebook and Differences between Countries and Waves* (Luxembourg: European Commission, 2003).

Fuller, Mike, "The Estimation of Gini Coefficients from Grouped Data: Upper and Lower Bounds," *Economics Letters* 3:2 (1979), 187–192.

Gastwirth, Joseph, "Robust Estimation of the Lorenz Curve and Gini Index," this REVIEW 54:3 (1972), 306–316.

Guest, Ross, and Robyn Swift, "Fertility, Income Inequality, and Labour Productivity," *Oxford Economic Papers* 60:4 (2008), 597–618.

Hart, Peter, "Entropy and Other Measures of Concentration," *Journal of the Royal Statistical Society, series A (General)* 134:1 (1971), 73–85.

Hendry, David, "Monte Carlo Experimentation in Econometrics" (pp. 937–976), in Zvi Griliches and Michael Intriligator (Eds.), *Handbook of Econometrics* (Amsterdam: Elsevier Science, 1984).

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

——— "On a Class of Poverty Measures," *Econometrica* 48:2 (1980b), 437–466.

——— *Analyzing Redistribution Policies* (Cambridge: Cambridge University Press, 1986).

Kakwani, Nanak, Adam Wagstaff, and Eddy van Doorslaer, "Socioeconomic Inequalities in Health: Measurement, Computation, and Statistical Inference," *Journal of Econometrics* 77:1 (1997), 87–103.

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

Le Grand, Julian, "Inequalities in Health: Some International Comparisons," *European Economic Review* 31:1 (1987), 182–191.

Lerman, Robert, and Shlomo Yitzhaki, "Improving the Accuracy of Estimates of Gini Coefficients," *Journal of Econometrics* 42:1 (1989), 43–47.

Lin, Chun-Hung, "Education Expansion, Educational Inequality, and Income Inequality: Evidence from Taiwan, 1976–2003," *Social Indicators Research* 80:3 (2007), 601–615.

Mehran, Farhad, "Bounds on the Gini Index Based on Observed Points of the Lorenz Curve," *Journal of the American Statistical Association* 70:349 (1975), 64–66.

Milanovic, Branko, "A Simple Way to Calculate the Gini Coefficient, and Some Implications," *Economics Letters* 56:1 (1997), 45–49.

——— "True World Income Distribution, 1988 and 1993: First Calculations Based on Household Surveys Alone," *Economic Journal* 112:476 (2002), 51–92.

——— *Worlds Apart: Measuring International and Global Inequality* (Princeton, NJ: Princeton University Press, 2005).

Mills, Jeffrey, and Sourushe Zandvakili, "Statistical Inference via Bootstrapping for Measures of Inequality," *Journal of Applied Econometrics* 12:2 (1997), 133–150.
Murray, David, "Extreme Values for Gini Coefficients Calculated from Grouped Data," *Economics Letters* 1:4 (1978), 389–393.
OECD, *Main Economic Indicators* (Paris: OECD, 2008).
Ogwang, Tomson, "Bounds of the Gini Index Using Sparse Information on Mean Incomes," *Review of Income and Wealth* 49:3 (2003), 415–423.
——— "An Upper Bound of the Gini Index in the Absence of Mean Income Information," *Review of Income and Wealth* 52:4 (2006), 643–652.
Pradhan, Menno, David Sahn, Stephen Younger, "Decomposing World Health Inequality," *Journal of Health Economics* 22:2 (2003), 271–293.
Pyatt, Graham, Chau-nan Chen, and John Fei, "The Distribution of Income by Factor Components," *Quarterly Journal of Economics* 95:3 (1980), 451–473.
Ryu, Hang, and Daniel Slottje, "Two Flexible Functional Form Approaches for Approximating the Lorenz Curve," *Journal of Econometrics* 72:1 (1996), 251–274.
Sala-i-Martin, Xavier, "The World Distribution of Income: Falling Poverty and . . . Convergence, Period," *Quarterly Journal of Economics* 121:2 (2006), 351–397.
Sheret, Michael, "Evaluation Studies Equality Trends and Comparisons for the Education System of Papua New Guinea," *Studies in Educational Evaluation* 14:1 (1988), 91–112.
UNU-WIDER, *UNU-WIDER World Income Inequality Database, Version 2.0c* (Finland: World Institute for Development Economics Research of the United Nations University, 2008). http://www.wider.unu.edu/research/Database/en_GB/database/ (accessed July 6, 2009).
van Praag, Bernard, Aldi Hagenaars, and Wim van Eck, "The Influence of Classification and Observation Errors on the Measurement of Income Inequality," *Econometrica* 51:4 (1983), 1093–1108.
Villaseñor, José, and Barry Arnold, "Elliptical Lorenz Curves," *Journal of Econometrics* 40:2 (1989), 327–338.
Wagstaff, Adam, Pierella Paci, and Eddy van Doorslaer, "On the Measurement of Inequalities in Health," *Social Science and Medicine* 33:5 (1991), 545–557.
Wagstaff, Adam, and Eddy van Doorslaer, "Equity in Health Care Finance and Delivery" (pp. 1803–1862), in Anthony Culyer and Joseph Newhouse (Eds.), *Handbook of Health Economics* (Amsterdam: Elsevier Science, 2000).
World Bank, *PovcalNet* (Washington, DC: World Bank, 2008). http://iresearch.worldbank.org/PovcalNet/jsp/index.jsp (accessed July 11, 2008).

APPENDIX

TABLE A1.—DESCRIPTIVE STATISTICS OF EQUIVALENT INCOME

	Observations	Mean	s.d.
Sweden	8,889	137,947	63,268
Denmark	5,899	131,497	69,759
Finland	8,171	86,900	50,580
Netherlands	9,351	28,788	15,363
Austria	7,382	214,317	123,594
Belgium	6,664	609,200	507,861
Luxembourg	2,044	866,215	563,721
Ireland	9,890	7,715	7,081
Germany	9,390	31,414	24,164
Italy	17,323	15,943	10,558
Spain	17,757	1,107,543	763,037
France	13,794	94,265	98,806
United Kingdom	10,484	9,431	9,664
Greece	12,423	1,562,758	1,347,131
Portugal	11,445	887,748	750,996
United States	17,399	30,011	23,662

Mean and s.d. are denoted in national currencies.

TABLE A2.—COVARIANCE TERMS TO ADDRESS SECOND-ORDER BIAS

Groups	Covariance	Groups	Covariance	Groups	Covariance	Groups	Covariance	Groups	Covariance
50	−0.0000305	40	−0.0000429	30	−0.0000669	20	−0.0001219	10	−0.0003270
49	−0.0000315	39	−0.0000445	29	−0.0000703	19	−0.0001314	9	−0.0003774
48	−0.0000325	38	−0.0000465	28	−0.0000740	18	−0.0001423	8	−0.0004418
47	−0.0000336	37	−0.0000484	27	−0.0000782	17	−0.0001548	7	−0.0005281
46	−0.0000346	36	−0.0000504	26	−0.0000828	16	−0.0001689	6	−0.0006415
45	−0.0000358	35	−0.0000529	25	−0.0000877	15	−0.0001853	5	−0.0007999
44	−0.0000372	34	−0.0000553	24	−0.0000931	14	−0.0002048	4	−0.0010298
43	−0.0000384	33	−0.0000576	23	−0.0000991	13	−0.0002273	3	−0.0013757
42	−0.0000399	32	−0.0000604	22	−0.0001059	12	−0.0002541	2	−0.0018256
41	−0.0000413	31	−0.0000634	21	−0.0001133	11	−0.0002863		

Covariance equals $\mu_K = n^{-1} \sum_{i=1}^{n} \delta_i^g \varepsilon_i$ and is estimated from equation 16.