Calculating the concentration index when income is grouped[◊]

Philip Clarke[a,*] and Tom Van Ourti[b,c]

[a] *School of Public Health, The University of Sydney, New South Wales 2006, Australia*
[b] *Erasmus School of Economics, Erasmus University Rotterdam, PB 1738, 3000 DR Rotterdam, The Netherlands*
[c] *Tinbergen Institute, Burgemeester Oudlaan 50, 3062 PA Rotterdam, The Netherlands*

November 2009

Abstract

The problem introduced by grouping income data when measuring socioeconomic inequalities in health (and health care) has been highlighted in a recent study in this journal. We re-examine this issue and show there is a tendency to underestimate the concentration index at an increasing rate when lowering the number of income categories. This tendency arises due to a form of measurement error and we propose two correction methods. Firstly, the use of instrumental variables (IV) can reduce the error within income categories. Secondly, through a simple formula for correction that is based only on the number of groups. We find that the simple correction formula reduces the impact of grouping and always outperforms the IV approach. Use of this correction can substantially improve comparisons of the concentration index both across countries and across time.

JEL Classification: C2, D31, I19
KEY WORDS: concentration index, errors-in-variables, instrumental variables, categorical data, first-order correction

[◊] We are grateful for comments received from Teresa Bago d'Uva, Hans van Kippersluis, an anonymous referee, and participants at seminars given at Australian National University, Tilburg University and Erasmus University Rotterdam. We also acknowledge funding from the NETSPAR project "Income, health and work across the life cycle" and thank EUROSTAT for access to the ECHP. Part of this research was undertaken while Tom Van Ourti was a Postdoctoral Fellow of the Netherlands Organisation for Scientific Research – Innovational Research Incentives Scheme – Veni. Philip Clarke is supported by a Sydney University Fellowship. Part of this work was undertaken during a stay at the Melbourne Institute of Applied Economic and Social Research, and Economics RSSS at the Australian National University, the hospitality of which is gratefully acknowledged. The usual caveats apply and all remaining errors are our responsibility.
[*] Corresponding author. Tel: +61-2-93515424; fax: Fax: +61 2 9351 7420.
E-mail addresses: philipc@health.usyd.edu.au (Philip Clarke), vanourti@ese.eur.nl (Tom Van Ourti).

1 Introduction

The concentration index has become the standard measure to quantify income-related inequalities in health economics (Wagstaff and van Doorslaer, 2000). It can be estimated using grouped/aggregated data or micro data sets that contain information on an individual's income and his/her health (care) status. Micro datasets are generally preferred to grouped datasets as the former result in consistent estimation of the concentration index, since point estimates from grouped datasets ignore information on within group association between income (rank) and health (care) status (Kakwani *et al.*, 1997). But also when income is reported in categories, estimating the concentration index using individual level data neglects this within group association.

There are many such examples of inequality studies that have involved grouped data or surveys where the income variable used to rank individuals is reported in categories. These include – among others – Gerdtham *et al.* (1999) who use Swedish health survey data with an income measure with six categories; van Doorslaer *et al.* (2000) use Finnish and Danish data with categorical income data; Wagstaff (2002) and Meheus and van Doorslaer (2008) use aggregate data in wealth quintiles; Humphries and van Doorslaer (2000) and Wagstaff and van Doorslaer (2004) use Canadian data with income deciles; and van Doorslaer *et al.* (2006) use Canadian and Australian data with a limited number of income categories. While this paper focuses on the calculation of the concentration index when income is grouped, exactly the same issue arises when this type of inequality measures is calculated with any categorical indicator of socioeconomic status, such as education and occupation. As we show later, it follows that the health inequality measure will be influenced by the number of groups. Further it is possible to apply our proposed correction methods when categories can be further subdivided into additional groups based on levels of socioeconomic status.

The issue of dealing with income grouping when measuring the concentration index has been highlighted in a recent study in this journal by Chen and Roy (2009). However, the focus of this study is confined to calculating potential bounds on the concentration index and the implications of existing estimators for efficiency of statistical inference. To date the broader question of the consequences (and solutions)

for grouping of the income variable in categories (or using grouped data) for the estimated health inequality measures have not been addressed.

The remainder of the paper is organised as follows. In the next section, we show that categorising or grouping the data creates a form of the classical errors-in-variables problem in which an individual's ranking is measured with error within, but not between groups. While the impact of grouping on the Gini coefficient has been extensively explored in the context of income inequality measurement (e.g. see Gastwirth, 1972; Rasche, 1980; Lerman and Yitzhaki, 1989), findings from this literature can only be extrapolated to the concentration index if concentration curves are globally convex or concave, which is unlikely to hold. We propose two procedures: the first we term the *IV approach* which involves finding an instrumental variable to reduce the error in ordering individuals within each of the income categories, and the second, which we refer to as the *overall correction approach* was recently put forward by Van Ourti and Clarke (2009). The third section presents an empirical examination of this issue using data from the European Community Household Panel (ECHP) and the Medical Expenditure Panel Survey (MEPS). Using these datasets, we then illustrate the impact of income grouping upon the point estimate of the concentration index and explore approaches to reducing the influence of grouping. The final section concludes and discusses the wider relevance and applicability of our correction methods.

2 Analytical framework

2.1 Background

The concentration index is defined as twice the area between the concentration curve and the diagonal. The bounds of this measure are –1 and +1 with a negative (positive) value representing pro-poor (pro-rich) inequality.[1] Kakwani *et al.* (1997) have shown that the concentration index can be calculated using a so-called convenient OLS-regression.

[1] Erreygers (2009) has shown that for any variable of interest with a finite upper value, the bounds of the concentration index need not be -1 and +1. All the results in this paper also apply to his corrected concentration index, to Wagstaff's (2005) normalized concentration index, and to the generalized concentration index (Wagstaff *et al.*, 1991) since mean health and its lower and upper bound are not affected by income grouping.

Let us start with the situation where we have micro data at hand and every individual $i = 1,\ldots,n$ reports a variable of interest m_i such as health (care) and his/her actual income level y_i with $y_c \leq y_d$ for $c < d$. It is easy to show that the concentration index of m_i – $C(m_i)$ – equals $\hat{\alpha}_1$ in the underneath 'convenient' OLS-regression:

$$2\sigma_R^2 \frac{m_i}{\bar{m}} = \alpha_0 + \alpha_1 R_i^y + \varepsilon_i \tag{1}$$

where $\bar{m}$ is the average of m_i, R_i^y is the fractional rank of y_i, $\sigma_R^2 = n^{-1}\sum_{i=1}^{n}\left(R_i^y - 0.5\right)^2 = \left(12n^2\right)^{-1}\left(n^2 - 1\right)$ is the variance of R_i^y, ε_i is an error term with mean zero, and α_0, α_1 are parameters to be estimated.[2] It is common practice to use the fractional rank as proposed by Lerman and Yitzhaki (1989), i.e. $R_i^y = n^{-1}\left(i - 0.5\right)$, but in order to allow for individuals having the same actual income level, it is better to use:

$$R_i^y = \frac{p(y_i) + 0.5\left[q(y_i) - p(y_i)\right]}{n} \tag{2}$$

where $q(y_i) = \sum_{k=1}^{n} 1(y_k \leq y_i)$ and $p(y_i) = \sum_{k=1}^{n} 1(y_k < y_i)$ equal the number of individuals having at least income y_i, including and excluding y_i respectively (Van Ourti, 2004; Chen and Roy, 2009).

When micro data is available but income is only recorded in a limited number of categories, equation (1) still applies, but the calculation of the fractional income rank differs: there are K different income categories r_i such that $r_i = j$ if $\psi_{j-1} < y_i \leq \psi_j$ with $j = 1,\ldots,K$ and ψ_{j-1} and ψ_j are the bounds of each income category j. The fractional income rank then becomes:

$$R_i^y = R_j^r = \frac{q(\psi_{j-1}) + 0.5\left[q(\psi_j) - q(\psi_{j-1})\right]}{n} = \frac{\sum_{k=1}^{j-1} n_k + 0.5 n_j}{n} \tag{3}$$

where n_j is the number of individuals in income category j.

Finally, in case of grouped data, we can still apply equation (3), but must replace equation (1) by equation (4). The grouped data estimator for $C(m_j; K)$ now equals $\hat{\beta}_1$:

[2] We note that one should use the population formula of the variance of the fractional rank (and not a small-sample adjustment) since OLS is used as an arithmetic (and not as a statistical) device.

$$2\sigma_{R^K}^2 \frac{m_j}{\overline{m}} \sqrt{n_j} = \beta_0 \sqrt{n_j} + \beta_1 R_j^r \sqrt{n_j} + \zeta_j \qquad (4)$$

where m_j is the average variable of interest within income category j, $\sigma_{R^K}^2 = n^{-1} \sum_{j=1}^{K} n_j \left(R_j^r - 0.5 \right)^2$ is the variance of R_j^r, ζ_j is an error term with mean zero, and β_0, β_1 are parameters to be estimated.[3]

The first purpose of this paper is to show how income grouping could impact on the point estimate of the concentration index. Figure 1 provides some intuition on the effect of grouping. The solid lines show a hypothetical Lorenz (panel a) and concentration curve (panel b). Consider the effect of grouping the population into tertiles:[4] both the Lorenz and concentration curve are now composed of straight dotted lines as the within group variation in income or health (care) no longer contributes to the respective curve (Lambert, 2001). In case of the Lorenz curve, grouping of the income variable always leads to an underestimation as the straight dotted lines that approximate the Lorenz curve are bound to lie inside the original curve (see panel a in Figure 1 in which the difference between the grouped and original curves is shaded grey). However, with the concentration curve this need not result in an underestimation, see for example panel b in figure 1 in which the inflection in the concentration curve means that the straight line lies both outside and inside the original concentration curve and so the concentration index based on income grouping will be greater (or lesser) depending on the degree to which it compresses inequalities above or below the dotted line (error also indicated in grey). While the concentration curve illustrated in panel b may be an unusual one it illustrates that there is no a priori guidance on the sign or magnitude of income grouping upon the point estimate of the concentration index.

[Figure 1 about here]

[3] We note that the point estimate of the grouped data estimator equals that of the individual level estimator with a limited number of income categories, but the variances will differ since the grouped data estimator neglects the variability of the variable of interest within the income categories.

[4] For simplicity, we assume that everyone within the first and third tertile has the same value for the variable of interest, i.e. income for the Gini and health (care) for the concentration index. There is only variation within the second tertile.

2.2 Correcting the impact of income grouping

Chen and Roy (2009) have recently examined the impact of grouping on the concentration index. Following Gastwirth (1972), they suggested methods for estimating non-parametric bounds for this measure. However, this approach which is based on the lowest and highest potential within-income-group-correlations between health (care) and income mean that the bounds can be very wide and so it is difficult to know how they will be employed to inform empirical research.

Here we take a different approach in that we re-interpret the problem as a classical errors-in-variables problem (for example Wooldridge, 2003, section 4.4.2) since income grouping is equivalent to error in the ranking variable (R_i^y in equation 1) within (but not between) groups. We explore two approaches to deal with this issue. Firstly, an IV-approach to remove the impact of income grouping upon estimates of the concentration index when individual level data are available on potential instruments. Secondly we apply what we term an overall correction approach following a procedure of Van Ourti and Clarke (2009) which was originally derived to deal with the impact of income grouping upon the Gini index. Contrary to the IV approach it can be applied to both grouped data and micro data with a limited number of income categories and requires no additional information except on the number (and relative size) of income groups.

2.2.1 IV approach

In regard to the IV approach the normal conditions for a good instrument must apply: (i) sufficient correlation between the instrument and R_i^y, and (ii) no correlation between the instrument and ε_i in equation 1. When employing the standard IV approach to address errors-in-variables the first condition can be tested, but the second must be maintained. However, despite the technical similarity, our approach differs from the standard IV approach as there is no measurement error at the level of the income

categories, but only within the categories.[5] It follows that neither condition can be observed within income groups, so we suggest the following procedure for deciding on a suitable instrument.

First, the instrument(s) should be defined in such a way that if we were to use individual level income as an instrument (i.e. the unobserved actual income value), our IV estimator would give the same point estimate as the individual level estimator in equation (1)–(2). This can be achieved by making sure that the instrument(s) preserves the ranking across income categories (i.e. all people in a higher income category continue to be assigned a greater rank than individuals in any lower income category), and by ranking the individuals within the income categories by 'another variable' that is correlated with the income rank.[6] Second, the sign of the correlation between the rank of this 'other variable' and the fractional income rank based on the income categories is used to decide whether we rank the individuals by this 'other variable' in an increasing or decreasing manner within the income categories. So for example if the fractional income rank based on income categories is positively correlated with years of education the individuals within each category will be re-ranked in order of increasing years of education.

2.2.2 Overall correction approach

Similarly to the IV approach, the overall correction approach is derived from studying the estimator of the concentration index within a measurement error framework. The main difference with the IV approach is that it is a first-order-correction approach based on the number (and relative size) of income groups only (hence 'first-order'). While this increases the likelihood of some remaining second order bias, it is simple to implement and Van Ourti and Clarke (2009) show that it has good Monte Carlo and empirical performance when applied to the Gini index. It is precisely this reliance upon the number (and relative size) of income groups that makes it a potential candidate for reducing the impact of income grouping upon the concentration index. The intuition of

[5] In addition, the measurement error is not correlated with the rank of the variable measured with error, which holds since the measurement error is uniformly distributed and has zero mean within each income category. See also equation (5) and footnote 7.

[6] Note that if this variable only takes a limited number of values (e.g. years of education), one should apply equation (3) to this 'other variable' to calculate the ranking within the income categories.

their approach consists of comparing the estimator in equation (1) with the one based on a grouping of the data in equation (4), and by next exploiting the properties of the fractional rank and those of OLS as an *arithmetic* tool. For clarity, we first present the procedure – applied to the concentration index – of Van Ourti and Clarke (2009) for income groupings of equal size, i.e. $n_1 = n_2 = \ldots = n_K = n/K$, and next generalize for income groups of unequal size.

Let us start from the observation that $C\left(m_j; K\right)$ differs from $C\left(m_i\right)$ if the fractional income rank R_i^y is associated with m_i within the income groups (see also Figure 1b). The same insight emerges from the difference between the LHS and RHS of equations (1) and (4).[7] The RHS difference is addressed by defining an equation that describes the measurement error

$$R_i^j = R_i^y + \delta_i^j \tag{5}$$

where R_i^j is the fractional rank of group j – $K^{-1}\left(j - 0.5\right)$ – assigned to each individual i and δ_i^j is the measurement error with zero mean[8]. The LHS difference is addressed by multiplying through with the ratio of the variances of the fractional income ranks at the individual and grouped data level, i.e.

$$\frac{\sigma_R^2}{\sigma_{R^K}^2} = \frac{K^2\left(n^2 - 1\right)}{n^2\left(K^2 - 1\right)} \tag{6}$$

After some algebra (consult Van Ourti and Clarke (2009) for more details), one gets an equation that expresses the concentration index estimated from individual level data as a function of the concentration index estimated from equally-sized groupings of these data, i.e.

---

[7] Strictly speaking, one cannot compare equations (1) and (4) as these are based on respectively n and K observations. Nevertheless it is easy to derive an equation that gives the same OLS point estimate for β_1 as the one in equation (4), but that is defined on n observations. With the assumption of income groups of equal size, equation (4) reduces to $\left[K^{-2}\left(K^2 - 1\right)\right]\left[\left(6\bar{m}\right)^{-1} m_j\right] = \beta_0 + \beta_1\left[K^{-1}\left(j - 0.5\right)\right] + \zeta_j$. If we next use individual level data and assign the fractional rank of group j to each individual, i.e. $R_i^j = K^{-1}\left(j - 0.5\right)$, one gets $\left[K^{-2}\left(K^2 - 1\right)\right]\left[\left(6\bar{m}\right)^{-1} m_i\right] = \beta_0 + \beta_1 R_i^j + \psi_i$, which is comparable to equation (1).

[8] As explained before in footnote 5, the measurement error δ_i^j is not correlated with R_i^j. This can also be seen from noting that R_i^j equals the average R_i^y of group j. In other words, $\sum_{i\in j} \delta_i^j R_i^j = \sum_{i\in j}\left(R_i^j - R_i^y\right) R_i^j = 0$, and hence $\sum_{i=1}^{n} \delta_i^j R_i^j = \sum_{j=1}^{K}\left(\sum_{i\in j} \delta_i^j R_i^j\right) = 0$.

$$C(m_i) = \frac{K^2}{K^2-1}\left[\frac{n^2-1}{n^2}C(m_j;K) - \frac{12}{n}\sum_{i=1}^{n}\delta_i^j\varepsilon_i\right] \quad (7)$$

Assuming $n \to +\infty; K < +\infty$ (i.e. the number and relative size of groups is fixed in the population), equation (7) reduces to $C(m_i) = \left(K^2-1\right)^{-1}K^2\left[C(m_j;K) - 12\operatorname{cov}\left(\delta_i^j,\varepsilon_i\right)\right]$, and thus provides a first-order-correction term $\left(K^2-1\right)^{-1}K^2$ and an expression for the remaining second-order bias $-12\operatorname{cov}\left(\delta_i^j,\varepsilon_i\right)K^2\left(K^2-1\right)^{-1}$.

Equation (6) shows that the first-order-correction does only depend on the number of income groups and can be interpreted as a "grouped data" adjustment of the variance of the fractional rank. The remaining second order bias is a function of the covariance between the measurement error and the error from equation (1); and the smaller its value, the better the overall-correction-approach/first-order-correction performs in empirical applications. Although the exact value and sign of this covariance cannot be known as ε_i is unobservable, we can make some approximate statements. First, this covariance will be smaller the higher the number of groups K. Second, it will be zero if m_i is uniformly distributed over R_i^y within income categories since the variance of ε_i equals zero in this case. Third, Van Ourti and Clarke (2009) also present Monte Carlo evidence that *suggests* that the covariance will be a function of the variance (but not the skewness) of the distribution of m_i over R_i^y. In other words, the covariance is *likely* to be negative for an asymmetric unimodal distribution. For example, an extreme long right or left tail is likely to result in a negative covariance as can be seen from equation (1) and (5).

The generalisation to groups of unequal size is straightforward:

$$C(m_i) = \frac{1}{\sigma^2_{R^K}}\left[\frac{n^2-1}{12n^2}C(m_j;K) - \frac{1}{n}\sum_{i=1}^{n}\delta_i^j\varepsilon_i\right] \quad (8)$$

Equation (8) and (7) are similar (for $n \to +\infty; K < +\infty$ equation (8) reduces to $C(m_i) = \left(12\sigma^2_{R^K}\right)^{-1}\left[C(m_j;K) - 12\operatorname{cov}\left(\delta_i^j,\varepsilon_i\right)\right]$), except that it is impossible in equation (8) to come up with an exact expression for $\sigma^2_{R^K}$. Nevertheless, the first-order-correction remains easy to calculate, and the interpretation of the first- and second-order terms remains unchanged.

3 Empirical application

3.1 Overview of data

To explore how categorical income data impacts on the point estimate of the concentration index, we use one wave of data from 15 countries participating in the European Community Household Panel (ECHP) and the 2000 wave of the *Medical Expenditure Panel Survey* (MEPS) from the United States. Here we only provide a summary of the data as we have reported this in much greater detail including a full list of summary statistics elsewhere (Clarke and Van Ourti 2009).

The ECHP consists of a representative panel of non-institutionalised households in each country. We use the second (1995) wave (which was the first to include all health-related information) for 13 EU member states: Austria, Belgium, Denmark, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain and the United Kingdom. In the case of Finland and Sweden, we use later waves since these countries joined the ECHP in wave 3 (1996) and wave 4 (1997) respectively. The Medical Expenditure Panel Survey (MEPS) provides nationally representative estimates of health status and health care use for the U.S. civilian non-institutionalized population (Agency for Healthcare Research and Quality, 2008).

The key variables for this study are household income, self-reported health and health care use which are defined as follows:

Income: The ECHP and MEPS collect information on disposable (i.e. after-tax) household income, which is all net monetary income received by the household members. The income variable was further divided by the OECD modified equivalence scale in order to account for household size and composition.[9]

Self-reported health (SRH): We follow the methodology adopted by many studies in the income-related health inequality literature by transforming the ordinal SRH responses onto the cardinal HUI scale (Feeny *et al.*, 1995) which takes values

[9] The OECD modified equivalence scale gives a weight of 1.0 to the first adult, 0.5 to the second and each subsequent person aged 15 and over, and 0.3 to each child aged under 15.

between 0 (death) and 1 (perfect health) by using the approach of Jones and van Doorslaer (2003).[10]

Health care use variables: Both the ECHP and the MEPS contain information on health care utilization including: (i) the *number of nights in hospital* during the last 12 months (NIGHTS), (ii) the *number of dental visits during the last 12 months* (DENT) and (iii) the *number of visits to a GP or specialist during the last 12 months* (PHYS).

3.2 Empirical difference between the individual level and grouped data estimators of the concentration index

Figure 2 gives an overview of the impact of income grouping upon estimates of concentration indices of self-reported health, hospital nights, dental visits, and physician visits. We summarize the impact of grouping individual level micro data into equally sized groups, i.e. 'income-tiles', by calculating $100\left\{\left[C\left(m_j,K\right)/C\left(m_i\right)\right]-1\right\}$ for 50 to 2 'income-tiles', $K=50,49,\ldots,3,2$.[11] We do this separately for each of the 15 European ECHP countries and the 2000 wave of the MEPS (United States), and arrive at three main findings.

First, we observe both downward (<0) and upward (>0) impacts of income grouping indicating that the underlying concentration curves are neither globally convex nor globally concave. Second and unsurprisingly, we find a much larger and 'random' impact of income grouping across K in those cases where the concentration index based on individual level microdata was insignificant (i.e. $p>0.1$). This is the case for PHYS in the MEPS 2000 data (highlighted in red) and Sweden (highlighted in green) as well as NIGHTS and SRH in France (highlighted in blue) and DENT for Germany.[12] This follows from our 'relative' indicator of the impact of income grouping that gets inflated if the denominator is very small. Third, the pattern of the impact of income grouping across the number of income categories K shows a similar shape for different

[10] It should also be noted that self-reported health is a categorical variable which may impact on the measurement of inequality, but this issue is beyond the scope of this study. See also footnote 1.

[11] The impact of income grouping will *in general* be similar for income groupings of unequal size, but one cannot preclude that it deviates for very peculiar shapes of the distribution of m_i over R_i^y.

[12] We observe the significance level of the individual level concentration indices since we create the problem of income grouping. This is not possible for the applied researcher using grouped data since she/he only observes the significance level of the concentration indices resulting from grouped data.

variables and countries. The impact of income grouping seems only empirically relevant – i.e. dominating the randomness across countries – for 10 or less income groups, and markedly so for 5 or less groups. In the extreme case of 2 income groups, the median degree of downward underestimation across all countries is 26% for SRH, 30% for NIGHTS, 27% for DENT and 20% for PHYS.

[Figure 2 about here]

3.3 Comparative performance of the IV and Overall Correction Approach

While the IV approach has the potential to completely remove the impact of grouping of the income variable, in practice there are typically a limited number of potential candidates for constructing instruments from variables routinely collected in health (care) surveys. In the MEPS and ECHP data sets we identified three variables for constructing instruments: (i) modified OECD equivalence scale (*eqscale*); (ii) the highest educational degree obtained (*educ*). Third, we have the age of each individual (*age*) which should correlate with the fractional income rank given the evidence on the life-cycle behaviour of incomes. We defined instruments from these variables.

[Figure 3 about here]

Figure 3 shows the degree to which each of the three instruments and the overall correction is able to reduce the impact of grouping, defined as $B_j = 100\left\{\left[C\left(m_j, K\right)/CI\left(m_i\right)\right]-1\right\}$. Figure 3a–3c reports these statistics for the United States using the MEPS from the year 2000 for SRH, NIGHTS and DENT[13] and 3d-3g provides a graphical summary of the performance across ECHP countries using the median degree of error across countries. The figures show that the impact of the IV approach varies considerably by both instrument and the variable of interest. For example, IV(age) results in a downward error when applied to NIGHTS in MEPS, but an upward error in the majority of ECHP countries. While in some cases (i.e. Figure 3c) the IV approach appears to remove a considerable proportion of the error, the benefits of the IV correction are not universal – in Figure 3d, IV(eqscale) and IV(age) increase the error relatively the original grouped *C*. In contrast the overall correction approach

[13] Results for PHYS are not shown as *C* was not significant.

appears to always reduce the error relative to the original index when there are low numbers of income groups.

Three conclusions emerge. First, for five or less income groups, the overall correction considerably improves upon the original error and removes a major share of the impact of grouping. In the majority of countries, it also improves matters for higher numbers of income groups. For example, figure 3 shows that it always improves upon the original error in the MEPS. Second, the overall correction approach always outperforms the IV approach in this empirical illustration. Finally, the IV approach improves upon the original error in a few instances, but overall it does not remove the error and often even worsens matters compared to the "original *C*". While it should be acknowledged that only three variables to construct instruments have been tested, most (health) surveys contain relatively few variables that could be used with the IV approach. Moreover, instrument validity across countries is a related concern and is not confirmed in our empirical illustration. More generally, the poorer performance of the IV approach reveals that the impact of grouping can only be reduced through the judicious use of instruments and that using poor instruments can lead to an increase rather a reduction of the impact of grouping.[14] Based on these finds we cannot recommend the IV approach as a practical solution for correcting grouped concentration indexes.

Clarke and Van Ourti (2009) report the performance of the overall correction approach in correcting re-ranking due to grouping of income categories. For comparisons across all the ECHP countries for SRH, NIGHTS, DENT and PHYS between 2-15 categories the correction improved the ranking 27 occasions, lead to no change on 17 occasions and made matters worse 13 times. For the same comparisons across the 10 years of MEPS data the overall correction approach improved ranking on 35 occasions, lead to no change on six occasions and made matters worse only once.

[14] While using several instruments jointly is an advantage of the IV approach, it is not feasible here due to too high collinearity between these instruments. Alternatively, one can replace the single instrument by K variables; each variable equalling the instrument for one income category and taking zero for all other income categories. The intuitive reasoning behind this set of instruments is that it allows using overidentifying restrictions tests, such as the J-statistic of Hansen (1982). Sensitivity analyses for the MEPS show that this approach works in all cases where the validity of the instruments is confirmed by the J-statistic, but that the added flexibility comes at the cost of a smaller reduction of the impact of income grouping.

4 Concluding remarks

This paper discusses and illustrates how categorical income data impacts on the point estimate of the concentration index. This issue is conceptually different from the impact of income grouping on the Gini index since the underlying concentration curves need not be globally convex/ concave, and thus the bias can also be upward. We exploit individual level data on health (care) indicators and income to illustrate the impact of grouping by constructing hypothetical income groups. We find an upward impact in some cases, but an overall tendency to underestimate the concentration index at an increasing rate when lowering the number of income categories, which appear empirically relevant when there are less than 10 income groups.

We have proposed two approaches to reduce the impact from income grouping that are derived from a measurement error framework. Firstly an IV approach which involves finding an instrumental variable to reduce the error in ordering individuals within each of the income categories. Secondly we have also put forward the correction factor derived by Van Ourti and Clarke (2009) and termed here the 'overall correction approach' as it only uses information on the number (and relative size) of income groups. In addition, it can be applied to both grouped data and micro data with categorical incomes, while the IV approach can only be used for micro data with income recorded in categories. Our results indicate that the overall correction approach is likely to be the superior method for reducing the impact of grouping in most practical applications where income is recorded in groups and generally improves matters particularly when there are 5 or less income groups.

Although this paper deals with a specific issue in the field of health inequality measurement, we believe the approach has wider applicability. First, concentration indices have a long history outside health economics for analysing distributional issues in taxation (see for example, Lambert, 2001). Second, there are several examples in the health economics literature where concentration indices are calculated with categorical non-income data as the ranking variable such as educational degree or occupational class (see for example, Burström *et al*, 2005; Clarke *et al.*, 2002). Our correction

methods may help reduce the dependence upon the number of categories/groups in both cases.

References

Agency for Healthcare Research and Quality. *Medical Expenditure Panel Survey*. Agency for Healthcare Research and Quality, Rockville, MD, 2008. http://www.ahrq.gov/info/customer.htm

Burström K, Johannesson M, Diderichsen F. Increasing socio-economic inequalities in life expectancy and QALYs in Sweden 1980-1997. *Health Economics* 2005; 14: 831-850.

Chen Z, Roy K. Calculating the concentration index with repetitive values of indicators of economic welfare. *Journal of Health Economics* 2009; 28(1) 169-175.

Clarke PM, Gerdtham UG, Johannesson M, Bingefors K, Smith L. On the measurement of relative and absolute income-related health inequality. *Social Science and Medicine* 2002; 55: 1923–1928.

Clarke P, Van Ourti T. Correcting the bias in the concentration index when income is grouped. *CEPR Discussion paper* 2009-599, Centre for Economic Policy Research, Research School of Social Sciences, Australian National University.

Erreygers G. Correcting the Concentration Index. *Journal of health economics* 2009; 28(2): 504-515.

Gastwirth J. Robust Estimation of the Lorenz Curve and Gini Index. *Review of economics and statistics* 1972; 54: 306-316.

Gerdtham U-G, Johannesson M, Lundberg L, Isacson D. A note on validating Wagstaff and van Doorslaer's health measure in the analysis of inequalities in health. *Journal of Health Economics* 1999; 18: 117-124.

Hansen L. Large sample properties of generalized method of moments estimators. *Econometrica* 1982; 50: 1029-1054.

Humphries KH, van Doorslaer E. Income-related health inequality in Canada. *Social Science and Medicine* 2000; 50: 663-671.

Jones AM, van Doorslaer E. Inequalities in self-reported health: validation of a new approach to measurement. *Journal of health economics* 2003; 22: 61-87.

Kakwani N, Wagstaff A, van Doorslaer E. Socioeconomic inequalities in health: measurement, computation and statistical inference. *Journal of econometrics* 1997; 77: 87-103.

Lambert PJ. *The distribution and redistribution of income: third edition*. Manchester University Press: Manchester, 2001.

Lerman RI, Yitzhaki S. Improving the accuracy of estimates of Gini coefficients. *Journal of econometrics* 1989; 42: 43-47.

Meheus F, van Doorslaer E. Achieving better measles immunization in developing countries: does higher coverage imply lower inequality? *Social Science and Medicine* 2008; 66(8): 1709-1718.

Rasche RH, Gaffney J, Koo AYC, Obst N. Functional forms for estimating the Lorenz curve. *Econometrica* 1980; 48: 1061-1062.

van Doorslaer E, Masseria C, Koolman X. Inequalities in access to medical care by income in developed countries. *Canadian medical association journal* 2006; 174: 177-183.

van Doorslaer E, Wagstaff A, van der Burg H, Christiansen T, De Graeve D, Duchesne I, Gerdtham U-G, Gerfin M, Geurts J, Gross L, Häkkinen U, John J, Klavus J, Leu RE, Nolan B, O'Donnell O, Propper C, Puffer F, Schellhorn M, Sundberg G, Winkelhake O. Equity in the delivery of health care in Europe and the US. *Journal of Health Economics* 2000; 19: 553-583.

Van Ourti T. Measuring horizontal inequity in Belgian health care using a Gaussian random effects two part count data model. *Health Economics* 2004; 13: 705-724.

Van Ourti T, Clarke P. A simple correction to remove the bias of the Gini coefficient due to income grouping. *mimeo*, 2009.

Wagstaff A. Inequality aversion, health inequalities and health achievement. *Journal of health economics* 2002; 21: 627-641.

Wagstaff A. The bounds of the Concentration Index when the variable of interest is binary, with an application to immunization inequality. *Health Economics* 2005; 14(4): 429-432.

Wagstaff A, Paci P, van Doorslaer E. On the measurement of inequalities in health. *Social Science and Medicine* 1991; 33(5): 545–557.

Wagstaff A, van Doorslaer E. Equity in health care finance and delivery. In: Culyer A, Newhouse J (eds), *Handbook of Health Economics*. North Holland: Amsterdam, 2000; 1804–1862.

Wagstaff A, van Doorslaer E. Overall versus socioeconomic health inequality: a measurement framework and two empirical illustrations. *Health Economics* 2004; 13: 297-301.

Wooldridge JM. *Econometric analysis of cross section and panel data*. The MIT Press: London, 2002.

Figure 1: hypothetical example of the impact of categorical income data on the Gini and concentration index

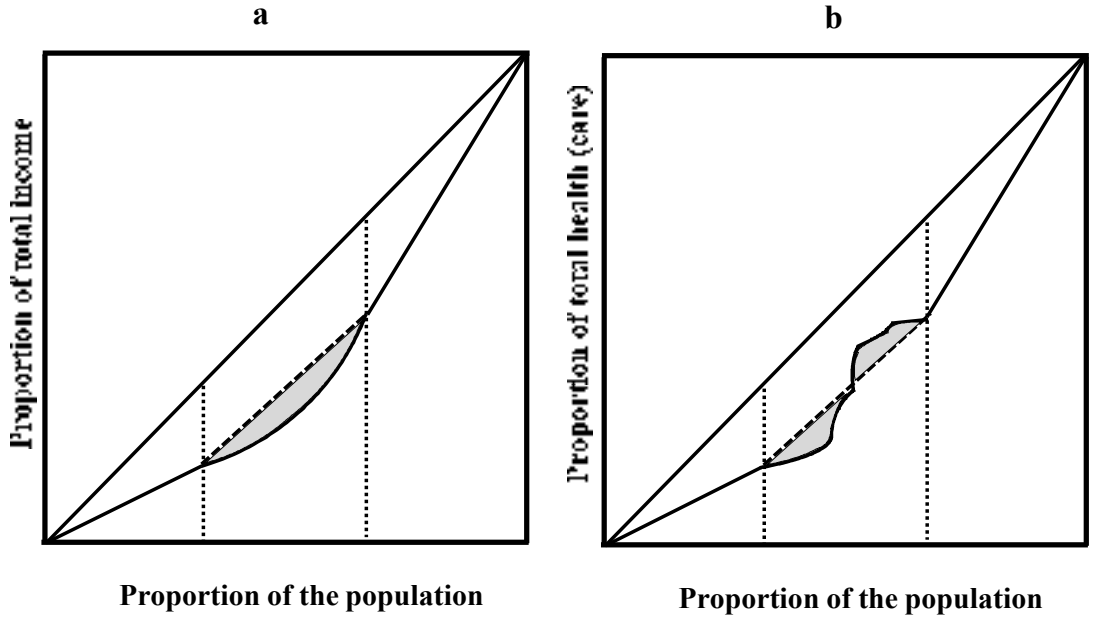
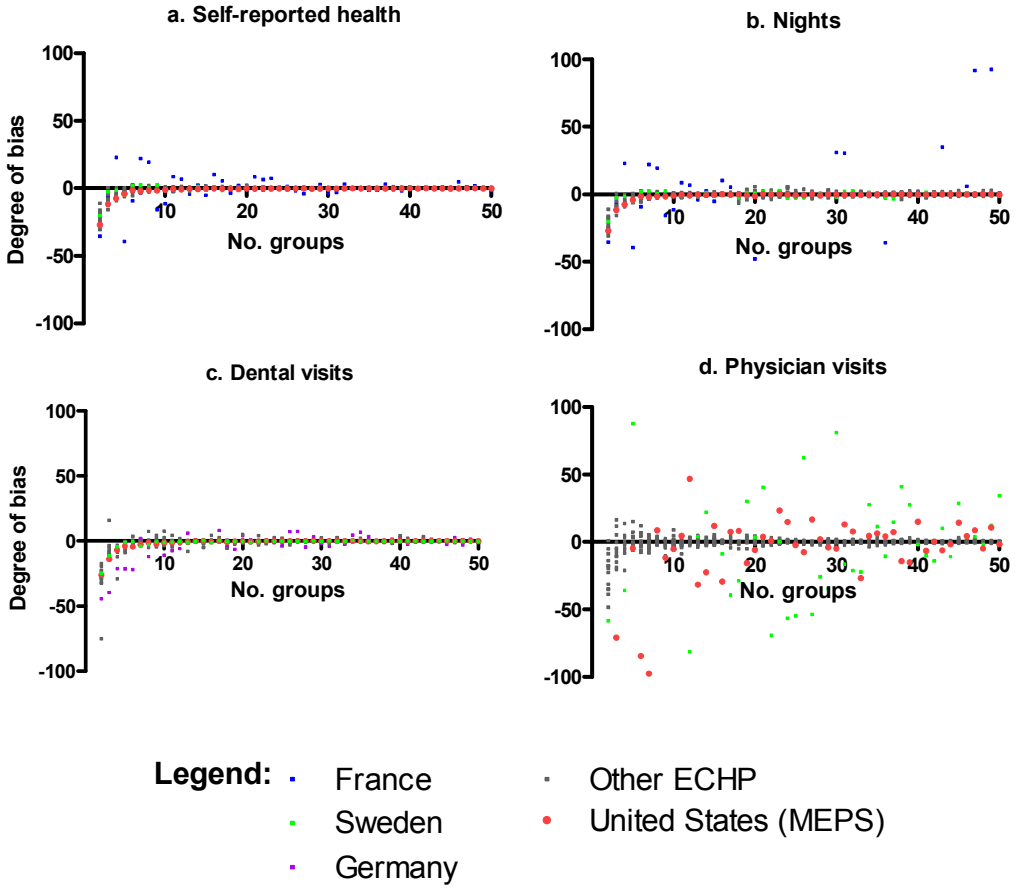
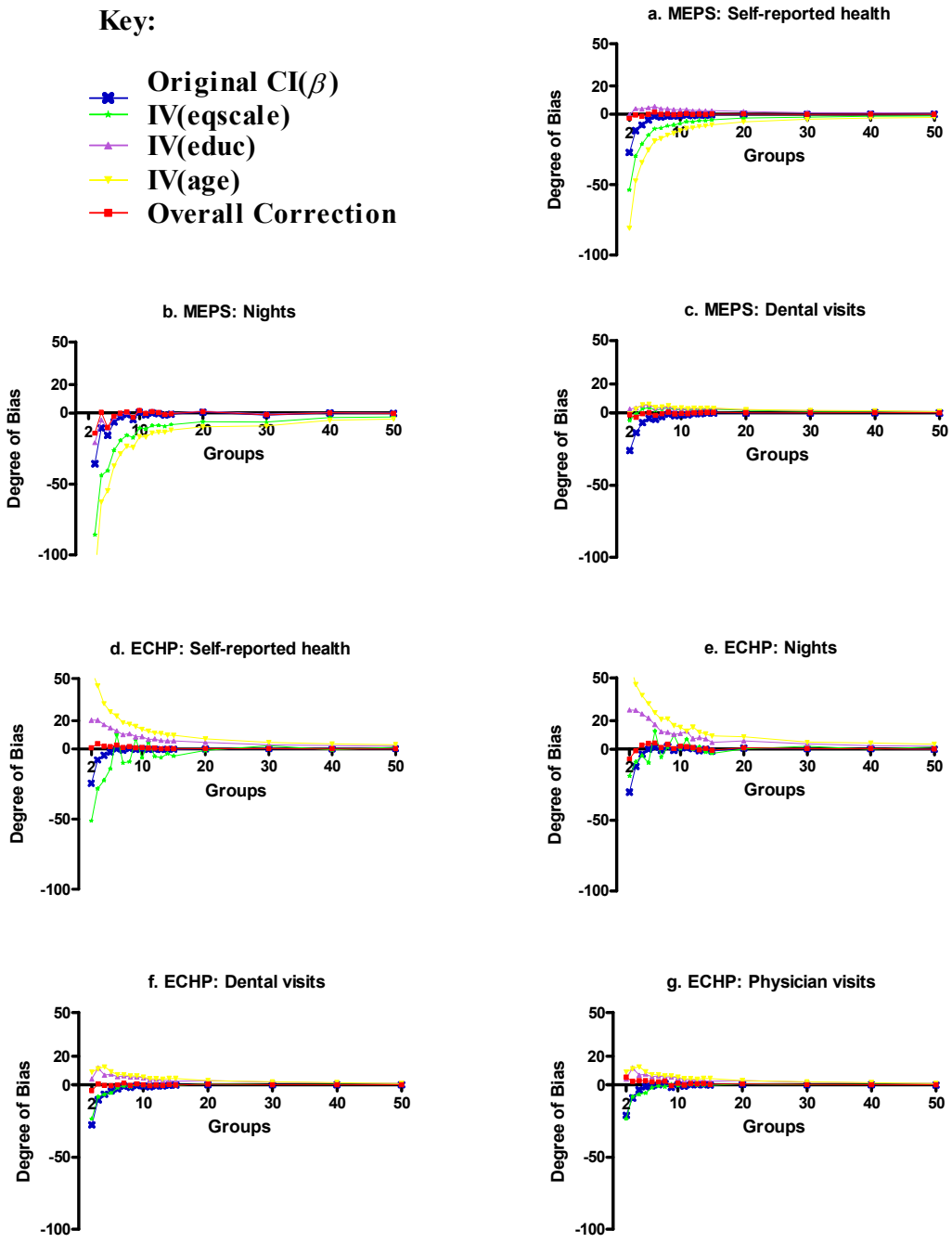


Figure 2: The impact of income grouping on the concentration index based on ECHP and MEPS data



Note: (i) Highlighted dots refer to the impact of income grouping of concentration indices based on individual micro-data that are insignificant at the 10 percent level: these are France for self-reported health and nights, dental visits for Germany and Sweden/US for Physician visits. (ii) The following countries have been excluded due to lack of data: NIGHTS (Germany); DENT (France); PHYS (France).

Figure 3: The impact of income grouping on the concentration index and various correction methods using data from the United States (using MEPS 2000) and ECHP countries



Notes: (i) Median degree of error reported for ECHP countries in figures d. to g. (ii) The following countries have been excluded from comparison due to either lack of data or a non-significant concentration index: SRH (France); NIGHTS (France, Germany); DENT (France, Germany); PHYS (France, Sweden, United States)