United Nations University
WIDER
World Institute for Development Economics Research

Research Paper No. 2004/24

Income-based Measures of Average Well-being

Steve Dowrick*

March 2004

Abstract

Internation comparisons of average national incomes omit important information about leisure, home production, health, etc. They are also bedevilled by index number problems. This paper suggests ways of combining working hours and life-expectancy with income comparisons, and shows that the fixed-price indexes of real income, such as those in the Penn World Table, substantially understate the income gaps between the poorest and richest countries.

Keywords: income comparisons, well-being, life expectancy, exchange rate bias, Geary-Khamis bias, the Afriat index

JEL classification: D50, I31, O47

* Australian National University

This study has been prepared within the UNU-WIDER project on Social Development Indicators (Measuring Human Well-being) directed by Professor Mark McGillivray.

UNU-WIDER acknowledges the financial contributions to the 2002-2003 research programme by the governments of Denmark (Royal Ministry of Foreign Affairs), Finland (Ministry for Foreign Affairs), Norway (Royal Ministry of Foreign Affairs), Sweden (Swedish International Development Cooperation Agency—Sida) and the United Kingdom (Department for International Development).

ISSN 1810-2611 ISBN 92-9190-607-7 (internet version)

The World Institute for Development Economics Research (WIDER) was established by the United Nations University (UNU) as its first research and training centre and started work in Helsinki, Finland in 1985. The Institute undertakes applied research and policy analysis on structural changes affecting the developing and transitional economies, provides a forum for the advocacy of policies leading to robust, equitable and environmentally sustainable growth, and promotes capacity strengthening and training in the field of economic and social policy making. Work is carried out by staff researchers and visiting scholars in Helsinki and through networks of collaborating scholars and institutions around the world.

www.wider.unu.edu *publications@wider.unu.edu*

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Camera-ready typescript prepared by Adam Swallow at UNU-WIDER
Printed at UNU-WIDER, Helsinki

1 Limitations of national income accounting data

International comparisons of living standards or development are most commonly made in terms of gross domestic product per person—whether in newspaper articles examining the latest country rankings or in economics journals analyzing the relative performance of countries. Such comparisons are open to criticism on the grounds that GDP is more properly regarded as a partial measure of aggregate output than as an indicator of either current or future well-being. International GDP comparisons make no allowance for environmental differences, for resource depletion, for leisure, for household production of goods and services, for black-market activities or for external costs and benefits associated with production and consumption.

For example, World Bank measures adjusted for differences in the purchasing power of currencies show GDP per capita in Australia as close to that in Finland. Some part of the difference is due to higher expenditures by Finns on domestic fuel and power (2.6 per cent of GDP) in comparison with Australians whose warmer climate requires lower expenditure for domestic comfort (1.4 per cent of GDP). Also problematic for welfare interpretations of national income is the observation that if polluting industries cause illnesses requiring expensive medical treatment, both the output of the polluting industry and the expenditure on medical services will be counted as positive contributions to GDP. Thus, aggregate output and income, as measured in the national accounts, may be relatively high whilst actual well-being is low.

Comparisons of GDP per capita take no account of differences in hours of work or hours of leisure. Nor do they take account of the value of production for own use. Peasant farming activities are particularly problematic for national accountants, since much of the production may be directly consumed by the farming household and will therefore escape the measurement of market transactions. Furthermore, a large amount of household activity in both rural and urban societies is concerned with the unmeasured production of goods and services such as child-care, education, food preparation, cleaning, etc. for own-household consumption.

The failure of measured GDP to capture well-being accurately is not surprising given that the definitions and measurement practices of national accounts have been designed with a quite different purpose in mind, namely to provide the accounting framework for the operation of monetary and fiscal policies. Dowrick and Quiggin (1998) argues that the System of National Accounts was designed with Keynesian short-run demand management in mind, hence its focus on the gross investment flows and government and market output which constitute the domestic side of the circular flow of funds in the familiar Macroeconomics I diagram. From the point of view of a central bank assessing the supply of money in relation to unemployment and inflation, GDP is a useful measure of the level of market activity without any inherent welfare implications.

These problems have long been recognized by economists. Eisner (1988) provides a survey of the problems and of various attempts to overcome them through amended and extended systems of national accounts. More recently, addressing the problem of valuing non-marketed productive activities, Folbre (2002) estimates that the average non-cash cost of bringing up a child in the United States in 2000 was around US$20,000—valuing the supervisory time plus foregone wages of the parents. Folbre and Nelson (2000) discuss the welfare implications of transferring activities from families to the market.

Considerable research effort is required to properly calculate the adjustments required to convert national accounts data into a measure that has a clear relationship with well-being. Such research is a luxury good that can be afforded only by the richer countries. So it may be useful to examine *ad hoc* adjustments to national accounts data that allow us to make more meaningful cross-country comparisons with readily available data. In the next section I look at two potential adjustments—one for hours of work and one for life expectancy. Then in Section 3 I go on to look at biases in the methods commonly used for calculating international income comparisons.

2 Income measures and alternative indicators of well-being

National governments devote considerable resources to the measurement of GDP in line with the internationally agreed standards of the System of National Accounts. Furthermore, an international programme has been in progress for over thirty years to enable cross-country comparisons of real income and expenditure. At five-year intervals, under the auspices of the International Comparison Project (ICP), detailed price surveys have been conducted in varying groups of countries. The results are published as tables of prices and real quantities for around 150 categories of goods and services that are purchased in each surveyed economy. A wealth of information has been generated on the price and quantity structures of the participating countries, enabling international comparisons of real GDP and its components at purchasing power parity. This information has been extrapolated across non-survey countries and across time to form the Penn World Table—see Summers and Heston (1991)—and in different forms has been analyzed and published by international organizations such as Eurostat, the OECD, and the World Bank.

Given the ready availability of economic statistics based on the national accounting definitions of GDP and income,[1] researchers are naturally tempted to use such data for international comparison of both economic performance and welfare. Even if GDP is not designed for the latter purpose, it is not unreasonable to enquire whether it might be a useful short-cut proxy for a measure of well-being, or whether it can be converted into a useful measure with readily available data.

2.1 Adjusting for hours of work

Recorded hours of employment in the total population vary according to rates of participation in the labour force and according to average hours of work. These rates are influenced by national differences in income levels, by differences in the age-structure of the population and by legal or cultural factors that influence participation, including gender roles and gender discrimination. Other things being equal, we expect a country with high participation rates and high hours of work to record a higher level of output, but this will not necessarily reflect higher well-being if the additional market income is offset by the sacrifice of leisure and home production.

---

1 For the purposes of this paper I use the ICP definition of GDP per capita as my primary measure. The national accounting identity defines gross domestic income to be the same as gross output, so I use the terms income and output interchangeably. But note that this definition is different from that of net national income, which subtracts capital depreciation and adjusts for international income transfers.

The data on recorded hours of work typically suffer from the same drawbacks as the data on GDP, failing to record time spent on productive activities that fall outside the market sector, such as household production and black-market activities. But it is precisely the common nature of these drawbacks that make GDP per hour worked a better indicator of well-being than the ratio of GDP to population. This will be the case if the average value of non-market production is the same as the average value of recorded labour market activity.

Table 1 lists in descending order the average weekly hours of work per head of population for 24 OECD countries in 1990. The variation in recorded hours is remarkably large given that the OECD is a relatively homogeneous group of countries in terms of income levels. The average person in Japan is recorded as working almost double the hours of the average person in Spain. The Table also records indexes of GDP per person and of GDP per hour worked, with each index normalized to have an average value of 100.

Table 1 Recorded hours of work and GDP in the OECD, 1990

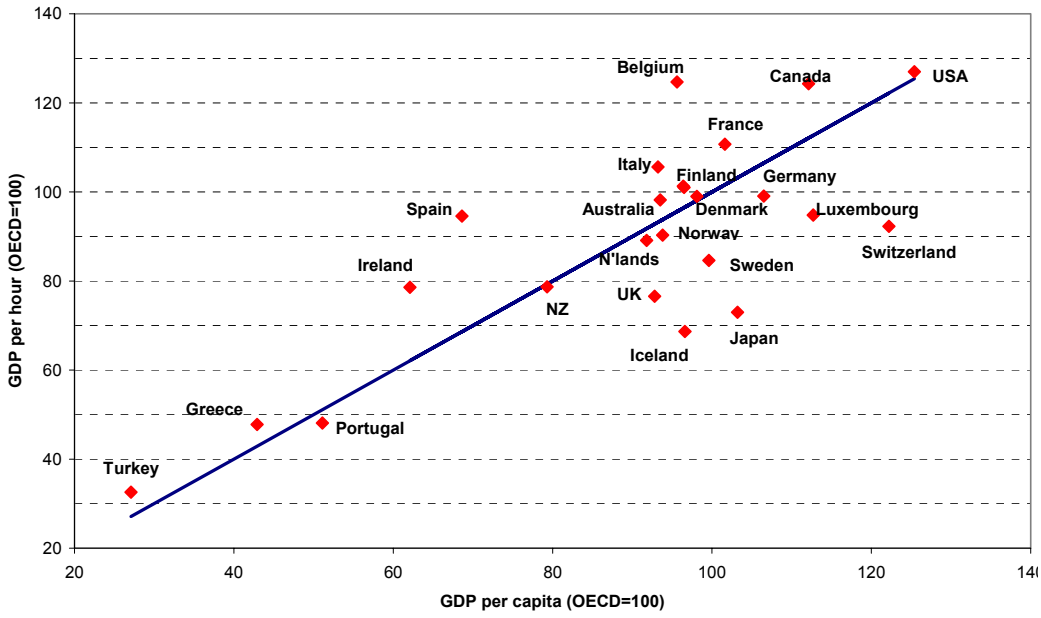
	Hours per person	**GDP per person**	**GDP per hour**		**Hours per person**	**GDP per person**	**GDP per hour**
		OECD = 100	OECD = 100			OECD = 100	OECD = 100
Japan	23.8	103.2	73	USA	16.6	125.4	127
Iceland	23.6	96.6	68.7	Austria	16.1	96.5	101
Switzerland	22.2	122.2	92.3	Australia	16.0	93.5	98.2
UK	20.4	92.8	76.6	Finland	16.0	96.4	101.3
Luxembourg	20.0	112.7	94.8	France	15.4	101.6	110.7
Sweden	19.8	99.6	84.6	Canada	15.2	112.1	124.3
Germany	18.1	106.5	99.1	Greece	15.1	42.9	47.8
Portugal	17.8	51.1	48.1	Italy	14.8	93.2	105.6
Norway	17.5	93.8	90.3	Turkey	14.0	27.1	32.6
Netherlands	17.3	91.8	89.1	Ireland	13.3	62.1	78.6
New Zealand	16.9	79.3	78.7	Belgium	12.9	95.6	124.7
Denmark	16.6	98.1	99	Spain	12.2	68.6	94.6

Source: OECD National Accounts and OECD Labour Force Statistics. Indices calculated by author.

GDP per person is 50 per cent higher in Japan than in Spain. But this does not mean that the average Japanese is that much better off than the average Spaniard who has considerably more time available for both leisure and for home production. Indeed, anecdotal evidence suggests that many Japanese people consider that they are substantially overworked and fail to enjoy the living standards of other high income countries, a judgement backed by the study of Castles (1992) who analyzes time-use data and other indicators to compare living standards between Tokyo and Sydney. These judgements are backed by the index of GDP per hour which shows that labour productivity is nearly one third higher in Spain than in Japan.

The Japan-Spain example is extreme, but other inter-country income comparisons also vary considerably depending on whether income is measured per person or per recorded working hour. These comparisons are illustrated in Figure 1 which is a scatter plot of GDP per hour versus GDP per person. Countries lying above the 45 degree line are those with below average hours of work.

Figure 1 Scatter plot of GDP per hour versus GDP per person



We can see that although Canada and Belgium are well below the US in terms of GDP per capita, they are almost equal in labour productivity. A number of other countries appear to be substantially better off when we take account of hours of work, particularly Ireland and Spain, whilst those which appear substantially worse off include Iceland, Japan, the UK, Luxembourg, Sweden and Switzerland.

Output per hour, as a measure of labour productivity, may not be closely correlated with average well-being if there are large differences across countries in dependency rates. Equally, GDP per hour may be a poor measure of relative welfare if low hours of work do not reflect a voluntary choice of leisure/home-production but are imposed by high unemployment or by social norms that restrict participation by groups such as women. So whilst we may not want to replace GDP per capita with GDP per hour as our preferred income measure, it may be instructive to compare the two measures as in Figure 1. Such comparisons are, however, more difficult to compute with accuracy for non-OECD countries for which the hours of work data compiled by the ILO are often incomplete.

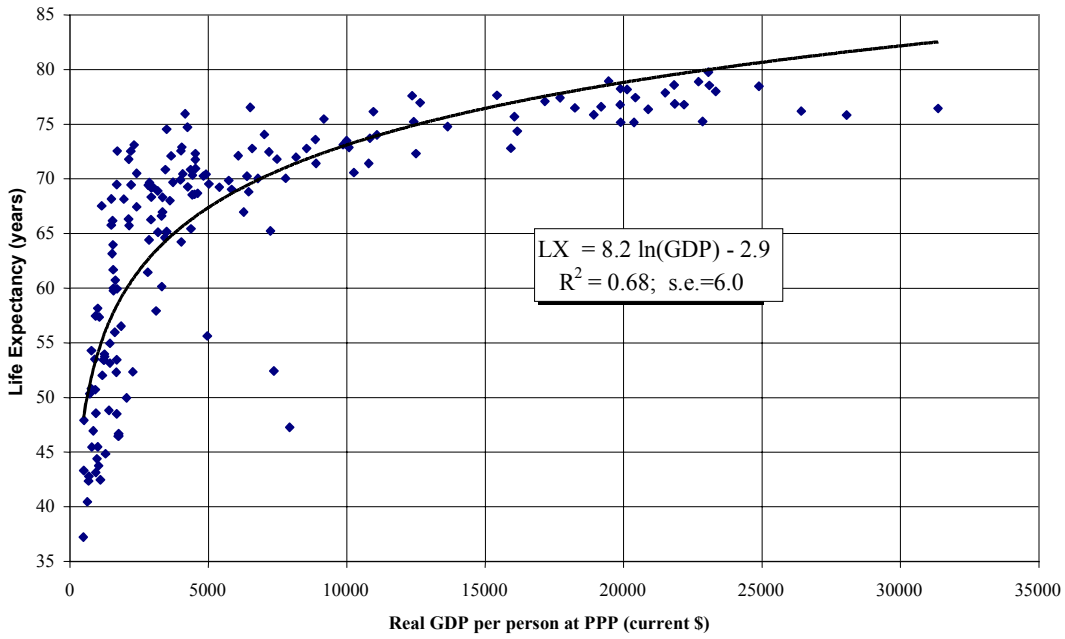
2.2 Adjusting for life expectancy

An approach that has been particularly popular in the development literature is to report multiple indicators of social development including GDP or GNP per capita, life expectancy, educational attainment, literacy rates, etc. The best-known composite index is the Human Development Index (HDI), combining measures of income, life expectancy and literacy. It is regularly updated and amended in the UNDP's annual Human Development Report.

These approaches are intuitively appealing as a solution to the limitations of purely income-based measures, though the composite indicator approach inevitably raises serious questions concerning the rather arbitrary choices of scaling and weighting methods. These issues are addressed in depth in the literature on composite well-being indices (see for example UNDP 1990-2003). For the purposes of this paper, however, it is of interest to examine how closely the much-criticized income measures are correlated with commonly used alternative indicators. If the correlation is high, then income-based measures may serve as a reasonably good proxy for a variety of measures of development.

Life expectancy is probably the most important and widely used indicator of development. Figure 2 is a scatter-plot displaying the relationship between GDP per capita and the life expectancy of a newborn across the 171 countries for which World Bank data was available, averaged over the years 1994 to 1998. We can observe immediately that a child's expectancy of life varies hugely according to their country of birth and that poorer countries tend to have much lower life expectancy.

Figure 2 Real GDP per person and life expectancy across 171 countries, 1994-98



The relationship between income and life expectancy is strongly positive but clearly non-linear. Figure 1 displays the OLS regression line and parameter estimates for the semi-log model:

$$LX_i = \beta \ln(GDP_i) + constant + \varepsilon_i \qquad (1)$$

where *LX* represents life expectancy and GDP is real GDP per capita. This very simple model suggests that variations in income 'explain' over two-thirds of the cross-country variation in life expectancy. If we interpret the relationship as causal, the estimated value for the slope coefficient, β = 8.2, implies that a twelve per cent increase in real income would increase life expectancy by one year.

Interestingly, this relationship is very close to that estimated on 1980 data by Dowrick, Dunlop, and Quiggin (2003) who find a β coefficient of 9.5 for the 60 countries surveyed in that year by the ICP. Although the currency units in which GDP is measured differ between the two periods, the logarithmic formulation implies that this affects only the constant term, not the slope coefficient.

Of course, these correlations do not establish the direction of causation. But causation is not at issue here. Inasmuch as life expectancy is a crucial element in the measurement of well-being, GDP seems to act as a reasonable proxy and it is worthwhile examining the statistical relationship more closely.

Inspection of Figure 2 shows that the regression systematically over-predicts life expectancy for the very poorest and the very richest countries whilst it under-predicts for countries with average incomes between US$2,000 and US$10,000. Formal diagnostic tests also suggest that this simple semi-logarithmic relationship could be improved: the residuals exhibit heteroscedasticity, the functional form fails the Reset(2) test and, when the data are ordered by real GDP, the sequential application of the Chow test for parameter stability reveals significant structural breaks at low and middle income levels.

Accordingly, the model is re-estimated as:

$$LX_i = \beta_1 \ln(GDP_i) + \beta_2[DUM_i \ln(GDP_i)] + \beta_3 DUM_i + constant + \varepsilon_i \qquad (2)$$

where *DUM* is a dummy variable with a value of unity for countries with income levels below US$6,000 per year and a value of zero for richer countries. Descriptive statistics are given in Table 2 and the regression results are summarized in Table 3.

Allowing for a structural break in the relationship reduces the standard error of estimated life expectancy from 6.0 to 5.5 years, but the diagnostic statistics show that heteroscedasticity is still present. Re-estimation of the relationship on the separate sub-samples reveals a higher standard error for the poorer countries (6.1 years) compared with the richer countries (4.4 years). But the slope coefficients for the independently estimated samples are almost identical to those implied by the pooled regression reported in Table 2.

The coefficient β_2 has a point estimate of 5.8 and is significant at the 1 per cent level. The implication is that the coefficient on GDP is nearly twice as high for the poorer countries in the sample, at 12.6, as it is for the richer countries, at 6.8. The causal interpretation of these estimates is that life expectancy will increase by one year when GDP increases by 15 per cent in a rich country or by just eight years in a poor country.

Table 2 Life expectancy and GDP, descriptive statistics for 171 countries

		Mean	Standard Deviation	Minimum	Maximum
Life expectancy 1994-98	*years*	65.6	10.7	37.2	79.8
GDP per capita 1994-98	*current PPP US$*	6,932.00	7,261	486.00	31,350.00
Log(GDP)		8.30	1.07	6.18	10.35

Source: Global Development Finance and World Development Indicators.
Accessed through http://www.worldbank.org/research/growth/GDNdata.htm
Note: Current PPP dollars are normalized to have the same value as the US$ in the USA.
Variables are averaged.

Table 3 Regression analysis of life expectancy on real GDP per capita

		coefficient	s.e.	t-ratio
Log (GDP per capita)		6.78	1.31	5.17
Log (GDP) x dummy		5.81	1.50	3.86
Constant		9.56	12.9	0.74
Dummy		-44.8	4.16	-3.16
observations	171 countries			
R^2	0.74			
s.e. of estimate	5.5 years			
Heteroscedasticity test (e^2 on predicted value): $\chi^2(1) = 5.85$				

Estimation: OLS using heteroscedasticity-consistent covariance matrix.

Package: Shazam (White 1987).

Data: See Table 1. Dummy variable =1 if GDP < US$6000.

Note: With observations ranked in order of real GDP, a preliminary regression was run without the dummy variable and a Chow Test for parameter stability was applied with sequential breaks. Parameter stability was rejected between the sub-sample of 109 countries with GDP below US$6,000 and the sub-sample of 62 countries with GDP above US$6,000: $F_{2,167} = 22.5$.

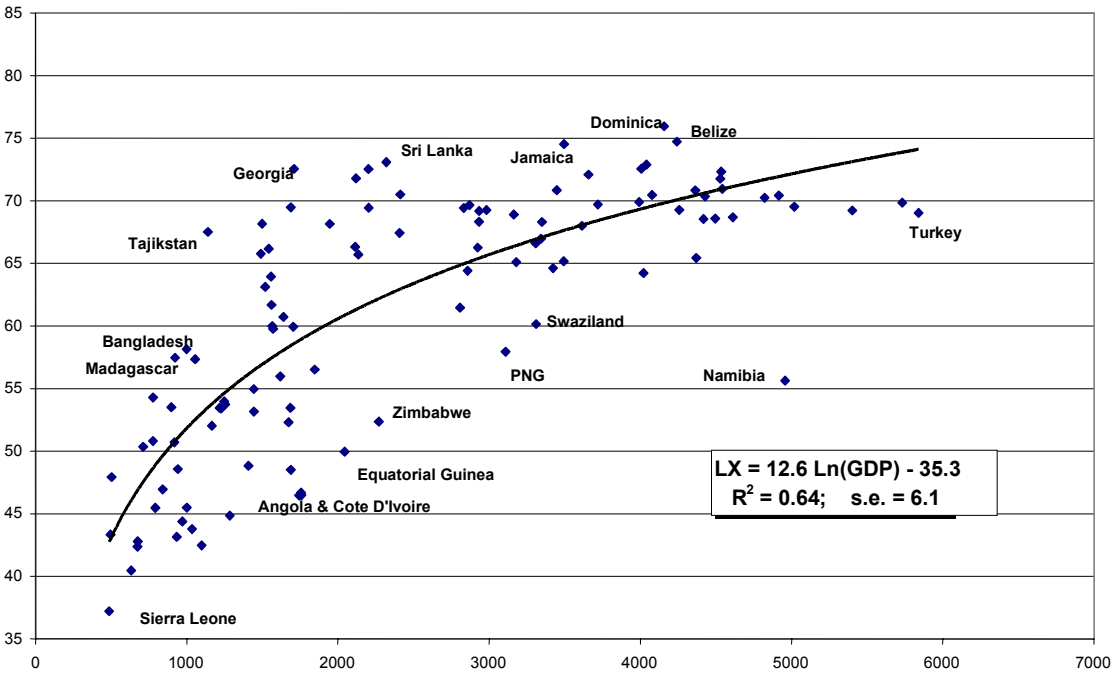
To the extent that real income is a good predictor of life expectancy, the combination of the two variables into a composite indicator seems to be a redundant exercise. But a standard error of over five years in predicting life expectancy is far from trivial, and inspection of Figure 2 indicates that for a substantial number of countries the prediction error is even greater. Indeed, the gap between income and life expectancy may indicate

some of the most interesting important features of the relationship between economic and human development. Figures 3 and 4 display the data and the regression lines separately for the poor and rich sub-samples of the 1994-98 data set.

I have labelled some of the outlying countries on both figures. Amongst the poor countries, the most prominent under-perfomers (in terms of lower than predicted life expectancy, given the level of income) are in sub-Saharan Africa. The over-performers include some Caribbean countries and some former Soviet Union countries. Amongst the rich economies, the under-performers include the three richest countries: Luxembourg, the US and Singapore whilst noticeable amongst the over-performers are Costa Rica, Greece, Sweden and Japan.

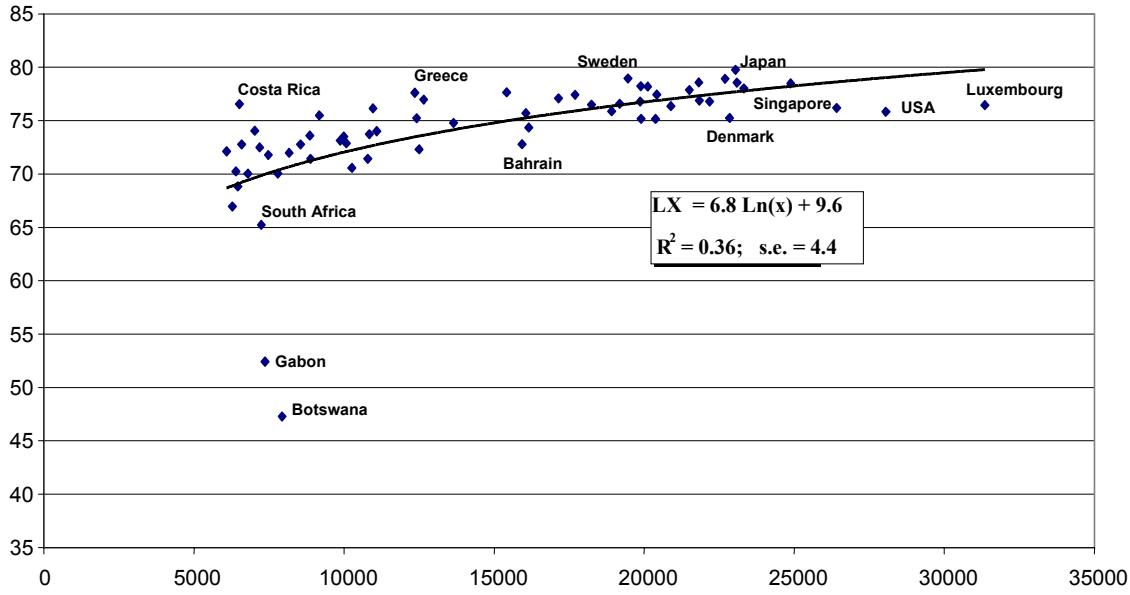
The size of the variations in life expectancy amongst countries at low and middle income levels suggests that per capita income is not always a satisfactory indicator of well-being. Looking at the figure above, it is hard to believe that the average person in Botswana would regard themselves as better off than the average South African because their average income is ten per cent higher, given that their average life expectancy is 18 years less.

Figure 3 Life expectancy and GDP: countries with income under $6000



The deviation of actual from predicted life expectancy can also highlight important questions for investigation. For instance, we might ask why Georgia and Sri Lanka have so much higher life expectancy than other countries of comparable income. Does this reflect a better physical and social infrastructure? Does it reflect a more equitable distribution of income?

Figure 4 Life expectancy and GDP: countries with income over $6000



3 Index number bias in international income comparisons

In the preceding section I have discussed some of the problems arising from the national accounting conventions as to which activities are included in GDP and which activities are excluded. Here I take the composition of the GDP bundle as a given and turn to the theory and practice of comparing expenditures that are expressed in different national currencies for consumers who face very different price structures.

A recent series of papers in the *Journal of Economic Perspectives* has addressed various index number problems in the measurement of the US Consumer Price index—see Schultze (2003), Hausman (2003) and Abraham (2003). These three papers are all addressing the problem of measuring a cost-of-living index, which is dual to an index of real incomes. Because the cost-of-living problem is examined in the context of inter-temporal rather than inter-national comparisons, the authors give more emphasis to problems such as quality adjustment and the introduction of new goods.

Hausman (2003) argues that the bias arising from the incorrect treatment of new goods is a first order problem in contrast to the bias arising out of a failure to account for consumer substitution, which appears as a second order term in his Taylor expansion of the true cost-of-living index. This argument may well be correct in the context of annual changes in the relative prices of goods where a change as high as 10 per cent, such as might be observed between the price of computers and the price of restaurant meals, is exceptional. In the context of international comparisons, however, it is not uncommon for the ICP data to reveal price ratios that differ by 500 per cent or more. In this case we may expect consumer substitution to have first order effects. Moreover, because the ICP redefines for each survey the basket of goods and services for which prices are collected, the problem of new goods is less likely to be significant in the context of cross-country comparisons.

It is well known, due to the work of Balassa and Samuelson, that the conversion of international incomes at currency-market exchange rates induces biased comparisons that tend to understate the relative incomes of poorer countries. It is less well known that the most widely used method of estimating purchasing power parities, the Geary-Khamis method that underpins the Penn World Table, induces the opposite bias—tending to overstate the relative income of poorer countries.

The use of foreign exchange rates (FX) to translate international incomes into a common currency introduces a 'traded sector bias'. Whilst exchange rates tend to equate purchasing power over traded goods and services, much of world production is for domestic consumption only. Wide variations across countries in the prices of non-traded goods and services are not reflected in the market for foreign exchange. So FX-converted incomes do not reflect the purchasing power of consumers in their own countries. Indeed, FX income comparisons tend to exaggerate international income differentials by ignoring the lower cost of living that is typically observed in poorer economies, due to cheaper labour-intensive services in the non-traded sector of low productivity economies.

The most widely used data set on purchasing power parity comparisons of GDP is the Penn World Table, the latest versions of which have been compiled by Summers and Heston (1991). They use the International Comparison Project's price surveys to calculate 'international prices' as the weighted average of the price vectors of all of the countries participating in the survey. The Geary-Khamis (GK) index of real GDP is calculated by valuing each country's per capita GDP bundle at these 'international prices'. The GK purchasing power parities are not calculated directly, rather they are derived from the GK quantity index as the rates of currency exchange which, when applied to nominal GDP, yield the same relative quantities. The GK index is extended across non-ICP countries and over time to produce the full Penn World Table.

The GK approach typically results in substantial revisions to FX valuations of the income of poor countries relative to the rich. For example, the ratio of per capita GDP between the US and Mali, the richest and poorest countries in the 1980 ICP sample, is 58:1 using market exchange rates (see UN and CEC 1987: part I, Table 1). The ICP data reveal, however, that non-traded goods and services are much cheaper, relative to traded goods, in Mali than they are in the USA. The GK measure of the US/Mali real income ratio is almost half that of the FX measure, a ratio of 31:1.

The GK method is, however, just as problematic as the exchange rate approach. PWT analysts have themselves acknowledged that the GK index may impart a bias in the opposite direction.

> The issue arises out of a familiar problem in price and quantity index number construction. …Valuation at other than own prices tends to inflate the aggregate value of the bundle of goods because no allowance is made for the substitutions in quantities toward the goods that are relatively cheap. … The practical importance of this issue … may loom large in comparisons between countries that have widely divergent price and quantity structures. (Kravis, Heston, and Summers 1982: 7)

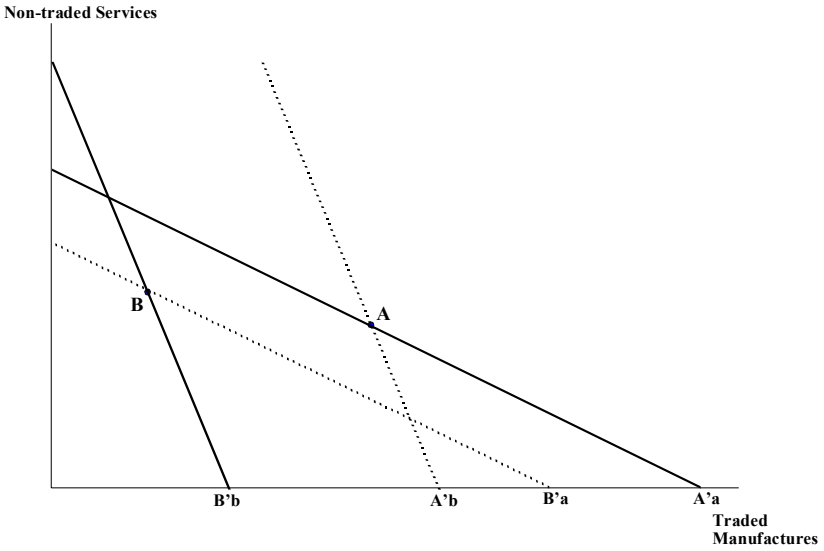
3.1 The economic approach to international income comparisons

In order to better understand these sources of bias in international comparisons, we turn to a model of a two good world where the representative agents of two economies, *A* and *B*, consume both tradable manufactures and non-tradable services, labelled *m* and *s* respectively.[2] The consumption bundle in country *A* is the quantity vector ***A***, where $\boldsymbol{A} \equiv (Q_m^A, Q_s^A)$, and the consumption bundle in *B* is the similarly defined quantity vector ***B***. We normalize the prices of manufactured goods in each country, measured in local currency, to unity. The price vector in country *A* is then defined as ***a***, where $\boldsymbol{a} \equiv (1, P_s^A)$, and ***b*** is the price vector in country *B*.

By assumption the technology of each country is such that labour productivity is higher in the manufacturing sector of country *A*, whilst labour productivity in the labour-intensive service sector is the same across countries. Competition in the tradable sector equalizes the exchange-rate converted prices of manufactures and the productivity-adjusted wage. Both real wages and the price of services are relatively cheap in the low productivity, low wage economy, *B*: $P_s^B < P_s^A$.

This situation is illustrated in Figure 5 where the solid lines through the consumption bundles ***A*** and ***B*** represent the budget lines for each representative consumer. With consumption of services measured on the vertical axis, *B*'s budget line is steeper, reflecting the relative cheapness of domestic services. With manufactures as the numeraire good, the local currency values of national expenditure per capita are ***A'a*** and ***B'b***, represented by the intercept of each country's budget line with the horizontal axis. With no international capital flows and no depreciation, national expenditure, national income and GDP are all the same.

Figure 5 Revealed preference, Paasche and Laspeyres indexes



Assuming common preferences, full information and rational choice we can use standard economic principles to rank the welfare of *A* and *B*. By construction, the consumption bundle in the low productivity country ***B*** lies inside *A*'s budget set whilst ***A*** lies outside *B*'s budget set. *A* could have chosen the bundle ***B***, but instead chose ***A***. In this situation, the principle of revealed preference tells us that the bundle ***A*** is revealed preferred to ***B***, implying that consumer *A* is better off than consumer *B*. It is in this sense that we can say that country *A* is richer than country *B* or that country *A* has higher GDP per capita than country *B*.

The revealed preference argument can be expressed formally as:

$$L_{AB} \equiv \frac{\boldsymbol{A'b}}{\boldsymbol{B'b}} > 1 \text{ and } P_{AB} \equiv \frac{\boldsymbol{A'a}}{\boldsymbol{B'a}} > 1 \Rightarrow \boldsymbol{A} \text{ is revealed strictly preferred to } \boldsymbol{B} \quad (3)$$

where the first inequality is the condition that ***A*** lies outside *B*'s budget set whilst the second inequality is the condition that ***B*** lies inside A's budget set. In (3) we have also noted that the first ratio, valuing *A*'s bundle at *B*'s prices, is the Laspeyres index, L_{AB}, whilst the second ratio is the Paasche index, P_{AB}. So the revealed preference condition is equivalent to the condition that both the Laspeyres and Paasche indexes exceed unity.

Revealed preference principles enable us to derive a partial ordering. It is an ordering, rather than a cardinal comparison, because whilst it may enable us to say that *A* is better off than *B* (or vice versa) it does not enable us to say that *A* is 10 per cent or 20 per cent better off. The ordering is partial because there may be situations where each consumption bundle lies outside the other's budget set (if $L_{AB} > 1$ and $P_{AB} < 1$) in which case we cannot tell which bundle is preferred. There may also be situations where $L_{AB} < 1$ and $P_{AB} < 1$, i.e. each bundle lies inside the other's budget set, in which case we have to reject the joint hypotheses of common tastes and rational choice.

If we want to make a cardinal comparison, a natural starting point is to consider the Laspeyres and Paasche ratios, alternately comparing the values of the bundles at *B*'s prices or at *A*'s prices. Here we start to confront some of the problems of index number theory. The *L* and *P* ratios will usually be different and there is no obvious reason to choose one over the other. In terms of Figure 5, *B*'s budget line intersects the horizontal axis at ***B***'***b*** and the dotted line parallel to *B*'s budget line through point *A* intersects the horizontal axis at ***A***'***b***. The Laspeyres index is the ratio of the distances of these two points from the origin. A similar construction with a line through point *B* parallel to *A*'s budget line gives the Paasche index.

The diagram has been constructed to illustrate substitution bias: if *A* chooses to consume relatively more of the goods which are cheaper in country *A* than in country *B* then valuing *A*'s bundle at *B*'s prices will tend to exaggerate *A*'s relative welfare and vice versa. Given that substitution bias tends to make the *L* ratio too large and the *P* ratio too small, we might suppose that an unbiased measure would lie between the two ratios. We shall show later that this is indeed the case if the common utility function is

---

2 This is a heuristic version of the modelling of Dowrick and Akmal (2003) where country *B* also produces an intermediate good that is exported. The arguments that follow are applicable in the case of many goods. The two good model is used to enable diagrammatic representation.

homothetic. But at this point it is useful to explain how the foreign exchange and the Geary-Khamis income comparisons are calculated and their biases in relation to the Laspeyres and Paasche ratios.

With manufactures as the numeraire, nominal GDP per capita in each country is given by the intercept of the budget line with the horizontal axis in Figure 5. Abstracting from capital flows and from transport costs, we expect the law of one price to hold for traded goods, i.e. the exchange rate converted price of manufactures is the same in both countries. In this case, the exchange rate converted ratio of GDP per capita is given by:

$$FX_{AB} = \frac{\boldsymbol{A'a}}{\boldsymbol{B'b}} \tag{4}$$

From Figure 5 we can see that this ratio exceeds both the Laspeyres and the Paasche ratios. This illustrates the Balassa–Samuelson result that foreign exchange comparisons tend to understate the relative income of the poorer country because the exchange rate is not influenced by the low price of non-tradables—see Balassa (1964) and Samuelson (1984).

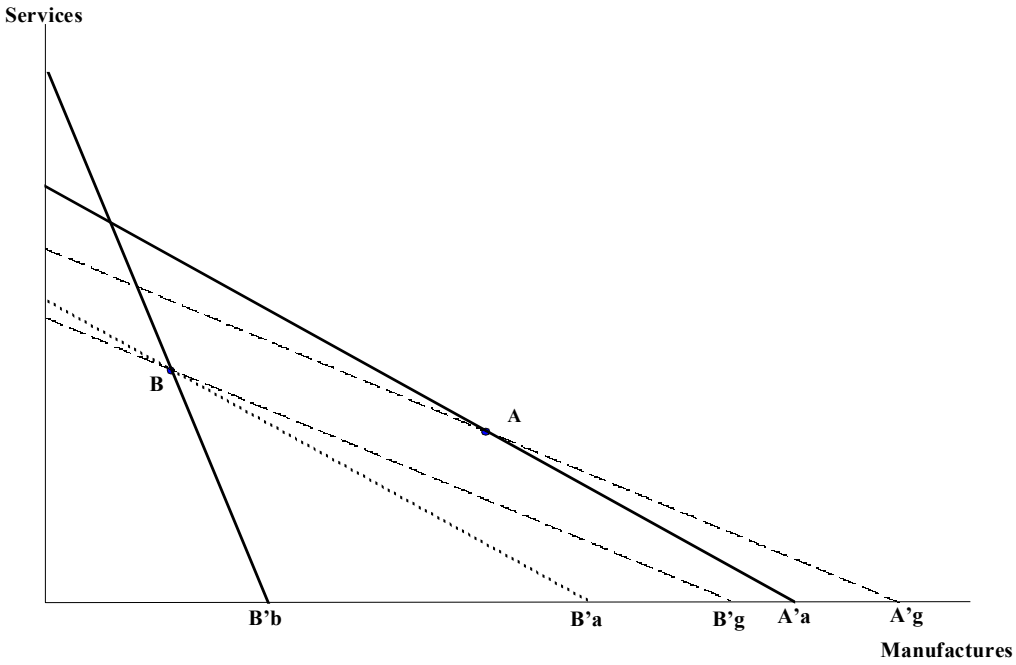
The Geary–Khamis approach attempts to overcome the bias in exchange rate comparisons by valuing the GDP bundle in each country at a fixed price vector, $\boldsymbol{g}$. This 'international price' vector is calculated as the GDP weighted average of price vectors of all the counties in the GK system. The Geary-Khamis quantity index is:

$$GK_{AB} = \frac{\boldsymbol{A'g}}{\boldsymbol{B'g}} \tag{5}$$

The weighting procedure biases the international price vector towards the price structures found in countries with the highest GDP, that is countries with large populations and high per capita incomes. Since the ICP surveys do not include China, their sample is most heavily influenced by the price structures of the rich and populous countries such as the US, Germany and Japan. This means that the international price vector, $\boldsymbol{g}$, which underpins the Penn World Table corresponds to the price structure of a high productivity economy with expensive non-traded services.

Valuing the GDP of poorer countries at rich country prices overstates their relative income levels. In Figure 6 valuation of GDP at the international price vector, $\boldsymbol{g}$, is illustrated by the dashed lines through points $\boldsymbol{A}$ and $\boldsymbol{B}$. It is assumed that the international price vector corresponds to that of a country that is richer than both countries *A* and *B*. We see that in this case the GK ratio, $\boldsymbol{A'g/B'g}$, is even smaller than the Paasche ratio, $\boldsymbol{A'a/B'a}$. Since the latter ratio already overvalues the income level of the poorer country, we can see that the GK method compounds the problem of substitution bias.

Figure 6 Substitution bias in the GK index



Dowrick and Akmal (2003) show that this direction of bias is strongly evident in the GK measures which substantially understate the true level of world income inequality.

How can we construct an unbiased or 'true' income comparison? If substitution bias makes the Laspeyres index too high and the Paasche index too low, a natural candidate is the geometric mean of the two: $F_{AB} = \sqrt{L_{AB} P_{AB}}$ which is the Fisher Ideal index. A major drawback for international comparisons, however, is that the Fisher index is not transitive, i.e. $F_{AC} \neq F_{AB} F_{BC}$. So whilst the Fisher index is a natural choice for an unbiased bilateral comparison, it does not provide a consistent multilateral index.

The economic approach to the index number problem is based on the notion that there may be a common utility function generating the different observations. If we can estimate the common utility function, $u(\mathbf{.})$, then the welfare ratio is simply $u(\boldsymbol{A})/u(\boldsymbol{B})$. This procedure does not, however, yield a cardinal index because any particular utility function which fits the data can be subject to a monotonic increasing transformation to yield another function which fits the data equally well, preserving the utility ordering but yielding a different utility ratio.

The non-cardinality of utility functions can be overcome by using the Allen quantity index, $I(\boldsymbol{p})$, which is defined as the ratio of the expenditure functions:

$$I_{AB}(\boldsymbol{p}) \equiv \frac{e[u(\boldsymbol{A}), \boldsymbol{p}]}{e[u(\boldsymbol{B}), \boldsymbol{p}]} \tag{6}$$

where the expenditure function $e[u(\boldsymbol{Q}), \boldsymbol{p}]$ is the minimum expenditure required to attain the utility level $u(\boldsymbol{Q})$ at some reference prices $\boldsymbol{p}$.

Two special cases of the Allen index are worth noting. If country A's prices are chosen as the reference price vector we have the Allen–Paasche index:

$$I_{AB}(\boldsymbol{a}) \equiv \frac{e[u(\boldsymbol{A}), \boldsymbol{a}]}{e[u(\boldsymbol{B}), \boldsymbol{a}]} \geq \frac{\boldsymbol{A}'\boldsymbol{a}}{\boldsymbol{B}'\boldsymbol{a}} \equiv P_{AB} \tag{7}$$

The Allen–Paasche index has a tight lower bound, the Paasche index. The derivation of this result is straightforward. Given utility maximization, the minimum expenditure required to achieve A's utility at A's prices is exactly $\boldsymbol{A'a}$, the value of A's chosen consumption bundle. On the other hand, whilst B's utility could be achieved at A's prices simply by spending $\boldsymbol{B'a}$ to purchase the bundle $\boldsymbol{B}$, there may be some other bundle which is cheaper but generates the same utility level.

Similar reasoning gives the result that the Laspeyres index is the upper bound to the Allen–Laspeyres index evaluated at country B's prices:

$$I_{AB}(\boldsymbol{b}) \equiv \frac{e[u(\boldsymbol{A}), \boldsymbol{b}]}{e[u(\boldsymbol{B}), \boldsymbol{b}]} \leq \frac{\boldsymbol{A}'\boldsymbol{b}}{\boldsymbol{B}'\boldsymbol{b}} \equiv L_{AB} \tag{8}$$

The Allen index is both transitive and cardinal. It is, however, not unique. It depends crucially on the choice of the reference price vector and in the context of cross-country comparisons there is no obvious price vector to choose. The only circumstance under which the Allen index is independent of the choice of reference price vector is where the common utility function is homothetic, $h(.)$. In this case the inequalities in equations (7) and (8) can be combined to yield the result that the Paasche and Laspeyres indexes are the exact upper and lower bounds for the Allen-homothetic index, I^H:

$$P_{AB} \quad \leq \quad I^H_{AB} \equiv \frac{e[h(\boldsymbol{A}), \boldsymbol{p}]}{e[h(\boldsymbol{B}), \boldsymbol{p}]} \quad \leq \quad L_{AB} \tag{9}$$

A special case that satisfies this inequality is the Fisher index which by construction must lie between the P and L indices (and homotheticity ensures that $P < L$). We noted earlier that the Fisher index does not yield a consistent multilateral index because it is not transitive. Afriat (1973) presents a solution to the problem of defining a true multilateral index that can be viewed as a generalization of the Fisher approach. The attractiveness of the Fisher index is that it is a compromise between the Paasche and Laspeyres indices. But it is the specificity of the Fisher compromise—choosing the geometric mid-point—which makes transitivity impossible. The solution proposed by Afriat (1973) comes from asking a more general question: is there any set of real income numbers for our n country problem such that the income ratio for each pair of countries lies somewhere between the corresponding Paasche and Laspeyres ratios? Afriat's requirement that the ratios lie *between* rather than *at the mid-point* of the Paasche and Laspeyres ratios makes it feasible that there may exist such a set of numbers—a 'true index' in Afriat's terminology.

The Afriat index is not just a convenient set of numbers. It is a true welfare measure. Afriat (1981) has a remarkable theorem showing that the existence of such a true index, for a given set of observations on prices and quantities, is equivalent to the existence of a common homothetic preference relationship (or utility function) that rationalizes the data.[3] That is to say, if there exists a set of Afriat index numbers, then there must also exist some common homothetic utility function such that any country's observed consumption bundle maximizes the utility of a representative consumer facing the prices and budget constraint of that country.

If a true multilateral index does exist it will not be unique, but we can establish upper and lower bounds to each of the bilateral ratios. These will be tighter than the Paasche-Laspeyres bounds. We can also establish bounds to the deviation of any observation from the sample average income. Using these true bounds as our benchmark, we can evaluate the degree of bias in both the FX and the GK income measures.

3.2 Applying the economic approach to international income data

Criticism of the economic approach to international income comparisons comes from several angles. There are those who point out that individual preferences can only be aggregated to predict aggregate behaviour if they satisfy quasi-homotheticity, and that there is no evidence to suggest that individual preferences do satisfy that condition. Then there are those who argue that preferences are heterogeneous even within a nation, so the assumption of common preferences across countries is preposterously counter-factual.

My response to these criticisms runs as follows. Yes indeed, individuals do have different preferences which probably do not satisfy the conditions for aggregation. If, however, we find that aggregate patterns of expenditure do satisfy tests for common preferences then we can use the economic approach to value the aggregate bundles. This allows us to assess the relative welfare of a notional representative consumer facing the relative prices and budget set of each country.

When John Quiggin and I first tested the hypothesis of common tastes on international average data, applying revealed preference tests to the 1980 ICP data for sixty countries, we found that the hypothesis of common preferences was not rejected—see Dowrick and Quiggin (1994). All of the variation in the composition of national expenditures could have been due solely to differences in relative prices and incomes. Subsequent research reported in Dowrick and Quiggin (1997) and Dowrick and Bruton (2000) showed that the much stronger hypothesis of common homothetic preferences could be sustained for a substantial majority, though not all, of the countries in the ICP surveys for 1980, 1990, and 1993. These findings, some of which are reproduced later, allow us to quantify the biases in both FX and GK indexes.

Hill (2000) has also addressed the problem of measuring substitution bias, adopting two utility-based approaches to establishing bounds on income comparisons. He estimates the parameters of the linear expenditure system, which is derived from the Stone-Geary utility

---

3 This equivalence is explained further by Varian (1983) who proposes a numerical algorithm to test for the existence of a true index given a set of observations on price and quantity vectors.

function, to derive utility numbers for each country. He notes the sensitivity of the income ratios to the choice of the reference price vector, illustrated by his finding that the US/Turkey ratio could be as high as 7 or as low as 3.5, bounds which encompass the GK ratio of 3.7. His other approach is to assume homothetic preferences, implying that income comparisons based on expenditure function ratios are invariant to the reference price vector. This enables him to tighten the bounds on the US/Turkey ratio to the interval (5.4, 4.0), establishing that the GK measure does indeed overvalue the relative income of the poorer country. This latter approach is similar to that used by Dowrick and Quiggin (1997), but whereas Hill examines only bilateral comparisons, Dowrick and Quiggin develop results on the multilateral properties of true Afriat index numbers.

In order to highlight the magnitudes of bias involved it is worth presenting results with respect to an extreme problem: what is the ratio of real GDP per capita in the richest country, the US, relative to that in Mali, one of the very poorest? As discussed above, exchange rate comparisons give a measure of 58, whilst the GK method applied to the ICP data set for 1980 GDP in sixty countries, reduces the ratio to 31.

These ratios for US/Mali GDP per capita are displayed in Figure 7, along with alternative index numbers. The Paasche and Laspeyres indices give the income ratios which are obtained by evaluating the GDP bundles at US prices or Mali prices respectively. Given the very different price structures in the two countries it is not surprising to see that the Laspeyres ratio of 82, is very much higher than the Paasche ratio of 36. The true multilateral index bounds, labelled the Afriat Upper and Afriat Lower, are necessarily tighter. We see that the bounds are tightened very considerably by the assumption of common homothetic preferences, giving a range between 50.2 and 54.6.

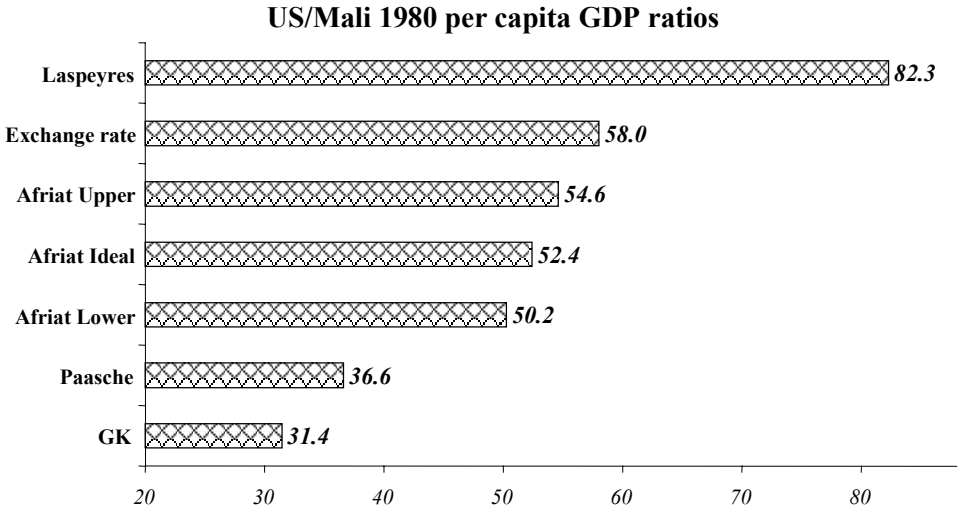


Figure 7 Quantifying the bias in FX and GK indexes

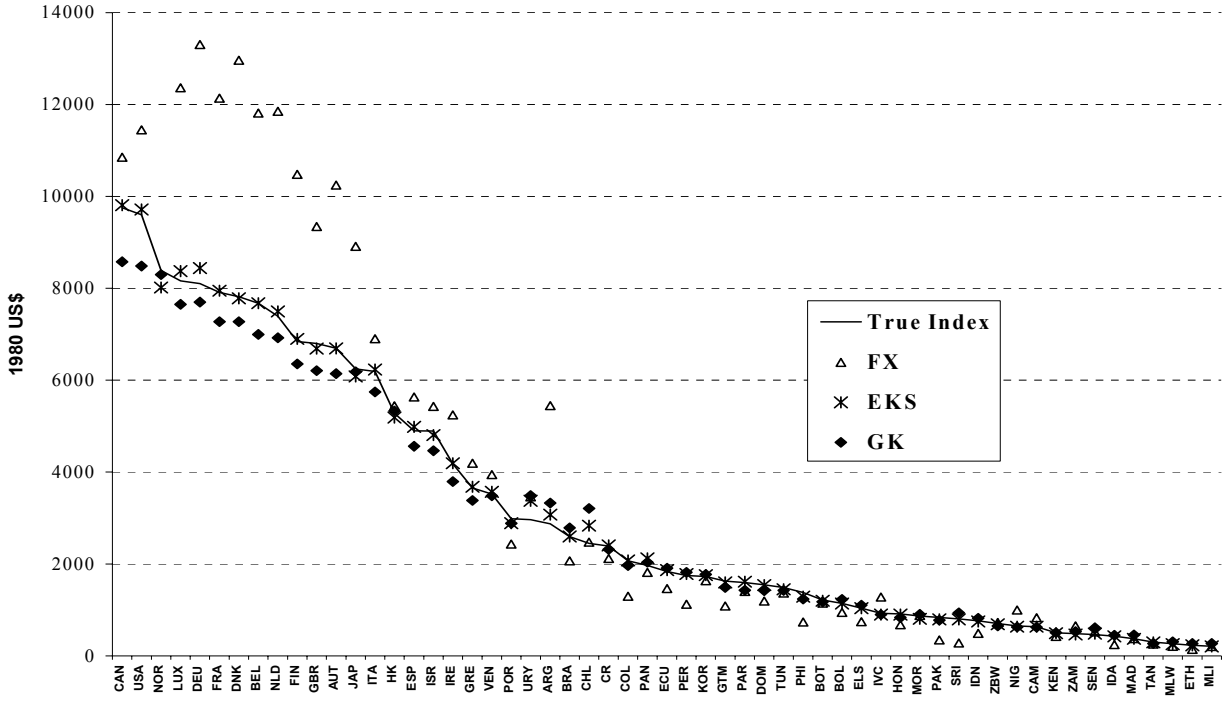


Figure 8 Indexes of GDP per capita

A conservative approach to the evaluation of bias is to measure it relative to the closest true bound. We see that the exchange rate measure lies above the true upper bound, over-valuing US income relative to Mali by more than 6 per cent. We can also see that the GK measure has under-valued US–Mali relative incomes by nearly 40 per cent.

Our preferred choice of true multilateral index numbers is the geometric average of the upper and lower bounds relative to the sample mean. This measure, the Afriat Ideal index, gives a US–Mali income ratio of 52.4.

Comparison of the FX and GK indexes with the Afriat Ideal index for 57 countries is illustrated in Figure 8, where countries are ordered in decreasing true income from left to right. All of the indexes have been normalized to the geometric mean of the FX index, measured in 1980 US$. We can see that the FX index does indeed overstate the income of the richer countries relative to the mean.

Because of the scaling, it is difficult to distinguish the index values for the poorer countries. So the data are re-presented in Figure 9 as log ratios relative to the Afriat Ideal index. It is evident that the FX measure tends to understate the income levels of the poorer countries—sometimes by as much as 60 per cent—with an equal and opposite bias in relation to the rich countries. The bias in the GK index is smaller—rarely exceeding 30 per cent—and the direction of bias is opposite to that of the FX measure. The GK comparisons tend to overstate the incomes of poorer countries and understate the incomes of the richer countries.

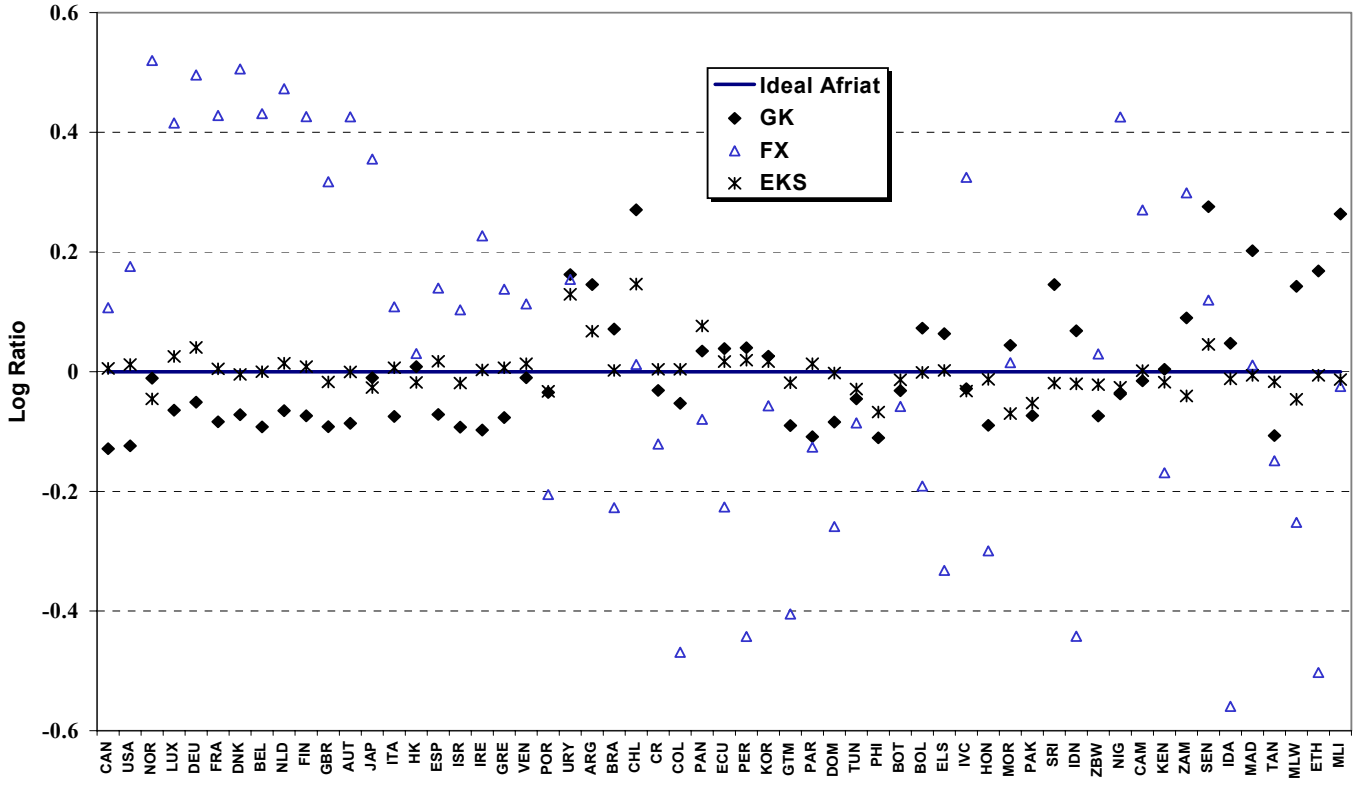


Figure 9 Deviations from Ideal Afriat index

Finally, I have displayed in both Figures 8 and 9 the EKS index, which is constructed for each country as the unweighted geometric average of the bilateral Fisher indexes with respect to each of the other countries. This index turns out to be very close to the Ideal Afriat Index as measured by a standard deviation of the log ratio to the Afriat Ideal index. The same statistic for the GK and FX indexes are much higher, as reported in Table 4. It is also evident from Figure 9 that there is no systematic tendency for the EKS index to overvalue or undervalue incomes by level of development.

Table 4 Standard deviation of log ratio to Afriat Ideal index

GK	FX	EKS
0.104	0.361	0.038

The EKS is of particular interest because it has become the preferred index number method of the OECD in calculating the purchasing power parity incomes of its member countries. Our analysis of the 1980 ICP data suggests that EKS is an unbiased and accurate approximation to the Afriat Ideal index. A disadvantage of the EKS index is that the EKS measures cannot be broken down into components such as private and government consumption and investment, as is done in the Penn World Table using the GK method. On the other hand, the EKS is clearly preferred when it comes to comparing income levels for the purpose of assessing relative well-being.

4 Concluding comments

International comparisons of GDP per capita are fraught with difficulties if we want to use them as indicators of relative well-being. I have discussed the difficulties related to the limitations of national accounts data and the difficulties related to index number problems. Income comparisons can still be valuable if we recognize their limitations. In particular I suggest the following guidelines for using and interpreting international income comparisons.

1. Where data availability allows it, comparisons should be carried out using the EKS index rather than the GK index or exchange rate conversions in order to minimize index number bias. FX measures tend to overstate income differentials, whilst GK measures tend to understate them.
2. Measures of GDP per capita should be contrasted with measures of labour productivity to highlight differences which might reflect variation in levels of unrecorded productive activities. High labour productivity relative to GDP per capita may indicate that a country enjoys more leisure and/or home production—especially if recorded unemployment is low and if access to the labour market is unfettered.
3. Income measures can be used to predict life expectancy and other social indicators. Analysis of deviations from predictions can yield useful insights into important aspects of national well-being.

References

Abraham, K. G. (2003), 'Toward a Cost-of-Living Index: Progress and Prospects', *Journal of Economic Perspectives* 17 (1): 45-58.

Afriat, S. N. (1973), 'On a System of Inequalities in Demand Analysis: An Extension of the Classical Method', *International Economic Review* 14 (2): 460-72.

Afriat, S. N. (1981), 'On the Constructability of Consistent Price Indices between Several Periods Simultaneously', in A. Deaton (ed.) *Essays in the Theory and Measurement of Consumer Behaviour*, Cambridge: Cambridge University Press: 133-161.

Balassa, B. (1964), 'The Purchasing Power Parity Doctrine: A Reappraisal', *Journal of Political Economy* 72 (6): 584-596.

Castles, I. (1992), 'Living Standards in Sydney and Japanese Cities: A Comparison', in K. Sheridan (ed.), *The Australian Economy in the Japanese Mirror*, Brisbane: University of Queensland Press.

Dowrick, S. and M. Akmal (2003), 'Explaining Contradictory Trends in Global Income Inequality: A Tale of Two Biases', *Draft paper for World Institute for Development Economics Research on 'Inequality, Poverty and Human Well-being', May 2003.* http://ecocomm.anu.edu.au/economics/staff/dowrick/dowrick.html.

Dowrick, S. and G. Bruton (2000), 'Quantifying Substitution Bias: True Comparisons across Countries', *4th Biennial Conference of the Pacific Rim Allied Economic Organizations*, Sydney (13 January).

Dowrick, S., Y. Dunlop, and J. Quiggin (2003), 'Social Indicators and Comparisons of Living Standards', *Journal of Development Economics* 70 (2): 501-29.

Dowrick, S. and J. Quiggin (1994), 'International Comparisons of Living Standards and Tastes: A Revealed-Preference Analysis', *American Economic Review* 84 (1): 332-41.

Dowrick, S. and J. Quiggin (1997), 'True Measures of GDP and Convergence', *American Economic Review* 87 (1): 41-64.

Dowrick, S. and J. Quiggin (1998), 'Measures of Economic Activity and Welfare: The Uses and Abuses of GDP', in R. Eckersley (ed.) *Measuring Progress: Is Life Getting Better?*, CSIRO Publishing, Collingwood, Victoria: 93-107.

Eisner, R. (1988), 'Extended Accounts for National Income and Product', *Journal of Economic Literature* 26 (4): 1611-84.

Folbre, N. (2002), 'Valuing Parental Time: New Estimates of Expenditures on Children in the United States', *Meeting of the Allied Social Science Association*, draft paper (28/12/01): 1-23.

Folbre, N. and J. A. Nelson (2000), 'For Love or Money – or Both?' *Journal of Economic Perspectives* 14 (4): 123-40.

Hausman, J. (2003), 'Sources of Bias and Solutions to Bias in the Consumer Price Index', *Journal of Economic Perspectives* 17 (1): 23-44.

Hill, R. J. (2000), 'Measuring Substitution Bias in International Comparisons Based on Additive Purchasing Power Parity Methods', *European Economic Review* 44 (1): 145-62.

Kravis, I. B., A. Heston, and R. Summers (1982), *World Product and Income: International Comparisons of Real Gross Products*, Baltimore: Johns Hopkins University Press.

Samuelson, P. (1984), 'Second Thoughts on Analytical Income Comparisons', *The Economic Journal* 94 (June): 267-78.

Schultze, C. L. (2003), 'The Consumer Price Index: Conceptual Issues and Practical Suggestions', *Journal of Economic Perspectives* 17 (1): 3-22.

Summers, R. and A. Heston (1991), 'The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988', *Quarterly Journal of Economics* 106 (2): 327-68.

United Nations and Commission of the European Communities (UN and CEC) (1987), *World Comparisons of Purchasing Power and Real Product for 1980*, New York: United Nations.

United Nations Development Programme (UNDP) (1990-2003), *Human Development Reports*, New York: Oxford University Press.

Varian, H. R. (1983), 'Non-Parametric Tests of Consumer Behaviour', *Review of Economic Studies* 50 (1): 99-110.

White, K. J. (1987), 'A General Computer Program for Econometric Methods - Shazam', *Econometrica* 46(1): 239-40.