



# WORLD HEALTH ORGANIZATION

---

## *The Methods and Data used in the World Health Report 2000: A response to the commentary made by the Brazilian Delegation to the Executive Board, 17<sup>th</sup> and 19<sup>th</sup> January 2001*

### **1) Introduction.**

We would like to thank the Brazilian delegation for the time and effort devoted to reviewing the World Health Report 2000 (WHR)(World Health Organization 2000). We strongly believe that constructive input will make it possible to continue the development of the methods and data sources for measuring health systems performance. The ultimate objective, of course, is to understand better how to improve the performance of systems and increase the well-being of populations, but to do this it is necessary to define what are the social goals to which health systems contribute, how these goals can be measured and the current and potential contribution of health systems to these social goals.

The Brazilian delegation made verbal comments at the technical presentation to the Executive Board (EB) on 17 January 2001, and used overhead slides in their presentation of 19 January. Copies of the overhead slides were circulated to delegates to the EB. In addition, the delegation provided WHO with the report of a workshop convened by the Oswaldo Cruz Foundation and the Ministry of Health in Brazil on 14-15 December 2000 which discussed the WHR (Oswaldo Cruz Foundation & Ministry of Health 2000). This document will address questions raised in the presentations by the Brazilian delegation to the EB as well as questions raised in the report of the workshop in Brazil. We hope that this response is simply the beginning of ongoing dialogue on ways to measure health system attainment and performance.

### **2) General Comments: Measuring the social goals to which health systems contribute versus assessing the contribution of the health system to these goals.**

Underlying many of the comments from the Brazilian delegation lie two alternative approaches to thinking about the performance of health systems: narrow or broad accountability. The framework for health systems performance assessment presented in the WHR 2000 is based on a broad view of accountability. It has the following logical steps:

- a) **What are the main social goals to which health systems make a major contribution?**  
Clearly, the dominant social goal that health systems contribute to is the health of the population. Many factors contribute to levels of health including some that health systems may be able to only marginally influence. Nevertheless, the defining goal for a health system is unequivocally to contribute to improved health for all. Health systems may also contribute to economic growth and thus consumption levels of the population. Health systems may also contribute to

education and other social goals. Through a long process of consultation and discussion, two other social goals have been included in the framework: the responsiveness of the health system to the expectations of the people and who bears the burden of paying for the health system.

- b) **To which social goals can health systems make a large enough contribution to warrant routine monitoring?** We believe that health systems by improving health can also contribute to improved economic growth. The Commission on Macroeconomics and Health has been formed to review the evidence on this important linkage. Notwithstanding this, it may not be feasible or necessary to include in periodic performance assessment the quantification of the contribution of the health system to economic growth. Through a process of consultation, WHO included in the framework for health systems performance assessment three social goals for routine monitoring: health, responsiveness and fairness in financial contribution.
- c) **Develop measures of these social goals that adequately capture what society holds to be important.** Rather than starting with existing indicators or available data, the framework has been developed by asking what are the dimensions of these social goals that should be measured. For example, to measure the level of population health it is important to capture mortality, morbidity and impairments. This aspect of the logical development is important because it sets a clear vision of what are the best measures that can be developed.
- d) **Develop indicators, instruments, data systems and analytical methods to come as close as possible to the best measures of health, responsiveness and fairness in financial contribution.** To have valid, reliable and cross-population comparable measures of health, responsiveness and fairness in financial contribution available on a regular basis for all Member States will require much collective effort to refine and improve proxy measures, survey instruments, data collection systems and analytical methods. This effort will benefit from ongoing and continuous debate, experimentation and policy use. In some cases, there is rapid progress such as the development of a common survey instrument for measuring the state of health of an individual and responsiveness.
- e) **Assess the current and potential contribution of a health system to the attainment of the key social goals.** Having measured the attainment of three social goals, the next step is to evaluate the contribution of the health system as compared to other systems to their levels of attainment. Efficiency of the health system is the extent to which the health system has made the maximum achievable contribution to these social goals given available resources. This definition of efficiency entails the broad concept of accountability. For example, to the extent that the stewards of the health system can reduce tobacco consumption by arguing effectively for taxes on tobacco then the maximum achievable levels of health should reflect this. Measuring efficiency or how close systems are coming to their potential contribution depends on being able to assess the maximum possible contribution of the health system given available resources and non-health system factors. This is an area for vigorous debate and development of alternative methods. Perhaps the most promising strategy is through the detailed analysis of the cost-effectiveness of major intervention options. An easier but less precise method is to use frontier production analysis as was used in the WHR 2000.

These five logical steps are important for three reasons. First, by first defining the outcomes that society cares about and then assessing how well health systems are doing in contributing to them, the focus of attention remains on the outcomes. The assessment of performance encourages decision-makers and

society more generally to debate, innovate and implement strategies to make the greatest impact on the outcomes themselves. Second, separating the measurement of outcomes from the assessment of the contribution of the system is just good scientific practice. Hypotheses about determinants of outcomes should not as a rule be used to alter how the outcomes are measured. For example, the Brazilian delegation has argued that we should not measure health but sub-components of health that are more likely to be casually related to specific activities of health systems. Following this logic, one might argue that we should not measure child mortality rates but only the child death rate from immunizable diseases. This we believe is not a good idea. The extent to which systems can contribute to reducing child mortality should be evaluated scientifically after measuring child mortality and changes in it. It should not be done by redefining the indicator of interest to be only child mortality due to immunizable diseases. Third, if a subset of health outcomes that are hypothesized to be more closely related to the activities of the Ministry of Health are used rather than measuring the total experience of health in a population, we will reinforce the tendency of the stewards of health systems to think narrowly about activities that are 100% in their control. Efforts to improve outcomes may require broader thinking and effective advocacy for changes in diet, tobacco consumption or road safety regulations.

The general logic of the framework helps explain some of the questions the Brazilian delegation raised about healthy life expectancy and health inequality. For example, if we accept the Brazilian suggestion to measure health inequalities only between socio-economic classes, we are assuming that the only health inequalities that are important and that can be avoided are those between social classes. By measuring the full distribution of health across individuals, it is possible to subsequently test any hypothesis about the causes of the observed health inequalities and not just restrict the analysis to the possible relationship between social class and health. We could well find that income differences are not the only important determinant of health inequalities, but that factors such as education, geography, migrant status etc. contribute. It is not appropriate to assume that we know the answers already by only analysing health inequalities according to socio-economic status.

The approach used in the WHR is the approach typically undertaken by economists in the analysis of inequalities in income. There are many different indicators of income inequality, including the Gini coefficient, based on interpersonal comparisons. By measuring interpersonal differences in income without prejudging the determinants, policy makers can then ask who the poorest groups are and what types of policies would reduce poverty and inequalities in income. We believe this is the correct approach to reducing health inequalities – to first measure how much inequality is present, and then ask what are the determinants.

Similar comments apply to the use of healthy life expectancy (healthy life expectancy) as the indicator of the level of population health. It will, of course, be influenced by many factors and not just the operation of the health system. But the important thing is to measure the goals which people believe are important for health systems. This requires knowing how to measure health levels of populations, how this changes over time and how it can be influenced by health system structures and policies. If factors outside the health system also improve health, so much the better. But it is still important to know whether health levels in a population improve or fall over time, and how the health system can ensure that they improve to the greatest extent possible.

### **3) Health Levels: Healthy Life Expectancy**

a) **Healthy Life Expectancy versus Life Expectancy.** *The Brazilian delegation criticized the use of healthy life expectancy as an outcome indicator on the grounds that it was not sensitive to changes in the health system, particularly in the short run. At the same time, the Brazilian delegation observed that the correlation between healthy life expectancy and life expectancy was very high, so suggested that there was no advantage in using healthy life expectancy over life expectancy.* First, societies care about more than simply avoiding premature mortality. Measures of population health must reflect the non-fatal aspects of health such as morbidity and impairments. In fact, increasingly the focus of health systems is providing care to improve the health state of individuals including alleviation of pain, decreasing anxiety, improving affect etc. Healthy life expectancy is a better measure of population health simply because it captures the full health experience of the population and not just mortality. Second, we are not convinced that healthy life expectancy cannot change relatively quickly. Consider the case of Oman where health system changes contributed to the massive decline in child mortality (from 330 per 1000 to 30 per 1000 in 3 decades) and hence to a very substantial change in healthy life expectancy.

b) **Complexity of healthy life expectancy:** *Related to this were complaints about the complexity of calculating healthy life expectancy, the process of choosing health state weights, and the difficulty of using the "person trade-off" technique.* We agree that the analytical and conceptual elements underlying summary measures of population health such as healthy life expectancy are more substantial than for measures of mortality such as life expectancy, but complexity is not a reason to ignore progress in measuring health in a more appropriate way. Some conceptually simple measures such as income per capita which are widely used are in fact extremely complicated to calculate. The intricacies of national income accounts are known only to a few yet the measure of overall economic performance is widely used.

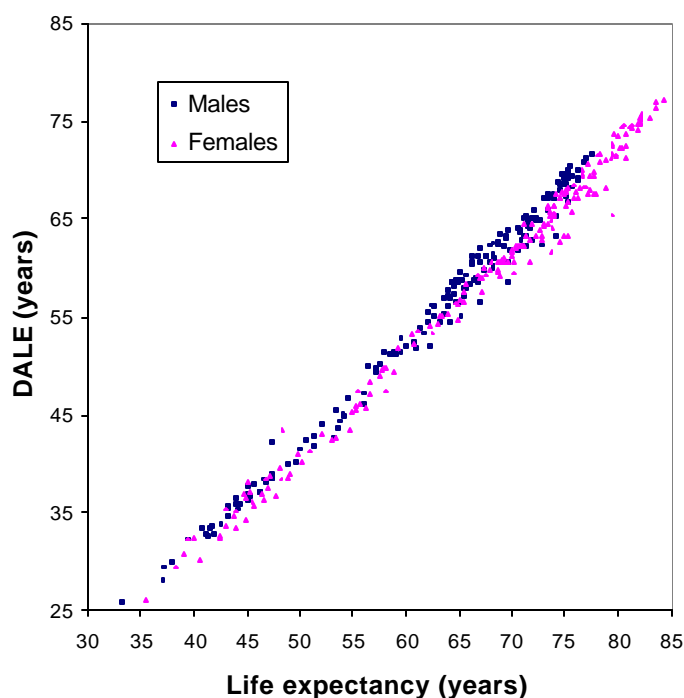
Given that it is important to measure non-fatal health levels, a clear, unambiguous framework is required. Building on more than a decade of intensive work on the development and standardization of summary measures of population health, through a WHO committee of experts WHO has assembled and synthesized various views on how summary measures of should be constructed and how they might be used to inform policy. A WHO book providing guidance on recommended standards for summary measures of population health will be published in 2001 (Murray CJL et al. 2000).

c) **Disability vs. Illness.** *The Brazilian delegation claimed that someone who is disabled is not necessarily sick.* This is an important semantic distinction related to the use of the word health and the word disability. Within the context of the new International Classification of Functioning, Disability and Health, we believe that to avoid semantic confusion DALE should be referred to only as healthy life expectancy. It has been designed to be a summary measure of population health capturing all aspects of fatal and non-fatal population health.

d) **Measuring the Non-Fatal Component of Healthy Life Expectancy:** *The delegation asserted that the non-fatal component of healthy life expectancy was calculated as a constant function of life expectancy.* This is incorrect. For each country, the best available evidence on age-specific mortality was reviewed to calculate a national life table. The age-specific severity-weighted prevalence of ill-health was assessed by analysing the wealth of data

available on major diseases and injuries held by WHO technical programmes and through collaboration with scientists outside WHO. Good examples are the extensive data sets on TB, maternal conditions, injuries, diabetes, cancer, STIs, etc. A detailed description of the methods used to estimate the severity-weighted prevalence of disability by cause, age and sex for each country is given in GPE Discussion Paper 16 (Mathers et al. 2000).

Figure 1 shows the correlation between healthy life expectancy and life expectancy by sex for each Member State.



**Figure 1. Healthy life expectancy (DALE) by total life expectancy at birth, by sex, 191 countries, 1999**

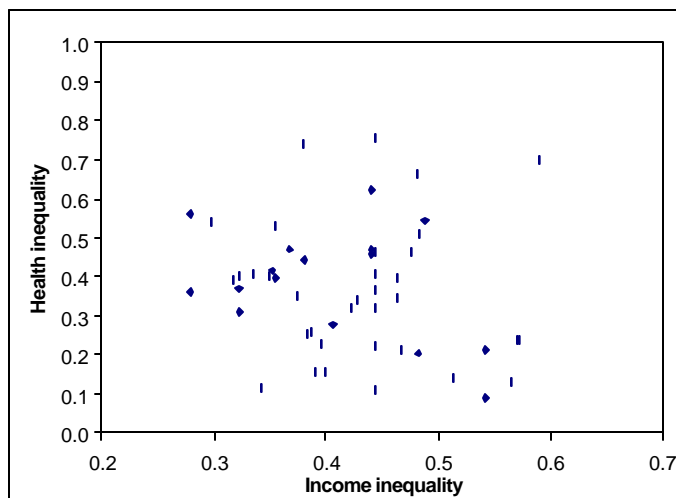
The equivalence line is shown to illustrate the unfortunate fact that those countries with the lowest levels of life expectancy also experience the highest prevalences of ill-health. This overall result is consistent with the hypothesis of compression of morbidity. In other words, as mortality declines, people spend less time ill as well. This finding is extremely important because one explanation for the compression of morbidity may be that health system interventions in the low mortality populations are lowering the prevalence of ill-health. It is also notable that while women live longer they spend on average a larger fraction of their life span in states of poor-health. There are large variations in healthy life expectancy at moderate levels of life expectancy. For example, for countries with the same life expectancy of 70, healthy life expectancy varies from 57 to 63, a non-trivial variation in health levels.

- e) **Capacity Strengthening.** *In the workshop report, it was argued that healthy life expectancy, require a commitment and investment in the development of data and methodology. We agree. We also believe that this should be a concern of the health system*

regardless of the chosen indicator or indicators.

#### 4) Health Inequalities

- a) **Health inequalities vs. inequalities in socio-economic status.** *The Brazilian delegation claims that the WHO indicator is strongly affected by the extent of inequalities in SES within the population.* The correlation between health inequality and income inequality on our cross-country sample is  $-0.16$ , which is very low (see Figure 2 below); therefore, the claim is at least incorrect across countries. However, even if there were a correlation, it would not be a bias but would simply indicate that inequalities in SES might be a determinant of inequalities in health as discussed above. Present evidence does *not* suggest that inequalities in health are related to inequalities in SES in the cross country comparisons. Independent of the cross-country relationship, it might still be true that the more disadvantaged groups in terms of SES also have the worst health in certain settings. But as argued above, it is still important to measure the outcome of interest – health inequalities – before thinking about the determinants and what policies would reduce inequalities.



**Figure 2. Income Inequality vs. Health Inequality index, Member States**

- b) **Value of the inequality index.** *Brazil states that the value of the inequality index depends on the unit of measurement.* It is true that the absolute value of the measure depends on the unit of measurement (years, months, days, minutes). However, the relative position of countries does not change when the unit of measurement changes. Scale invariance is a property of purely relative measures of inequality such as the Gini coefficient. The WHO index is not a purely relative measure. To draw an analogy, assume that we were interested in measuring the age of an individual. We could say that someone was 30 years old, or 360 months old, or 10950 days old. Even though the absolute value of the answer varies, it still accurately reflects

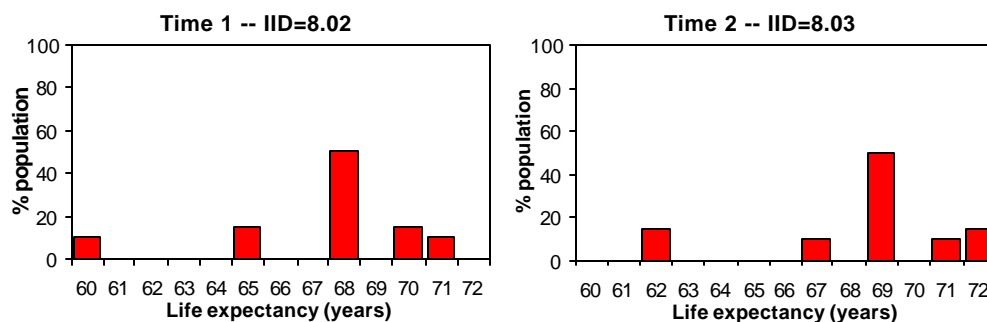
the underlying concept of interest – the age of the individual. Therefore, the claim that the value of the measure changes as the unit of measurement changes is not of any substantive interest.

- c) **Choice of parameters for the inequality index** *The Brazilian delegation states that parameters for the inequality index were chosen in an unusual way.* The choices of the values for the parameters are *normative*, in other words a question of values. For example, choosing between two distributions of income for a population is fundamentally a value choice. Such a choice is *not* a statistical or mathematical issue, but simply a matter of preferences. The century long literature on alternative measures of income distribution has extensively addressed the normative nature of judgements about income distribution. To develop a summary measure of the distribution of health, rather than arbitrary selecting the values of the parameters, we conducted a survey of over 1000 WHO staff and other health professionals from over 120 countries. The choices of the parameters are based on the expressed preferences of these individuals. A detailed analysis of the results revealed no significant trends in the responses by socio-economic characteristics of the respondents. Sensitivity analysis has also shown that the relative position of countries does not change significantly when the values of the parameters change.
- d) **Sampling frame of the surveys used** *In their written critique, Brazil claims that we did not take into account the sampling design of the DHS surveys used.* This is incorrect, as we used the sampling weights provided by the DHS, which make the DHS sample a nationally representative sample. Therefore, there is no bias in the calculation of the inequality index. It was based on nationally representative samples of the population.
- e) **Calculation of uncertainty intervals.** *The Brazilian delegation claims that the way in which uncertainty intervals were calculated is confusing.* The techniques used for the WHR are standard statistical methods. Statistical simulation was based on the variance-covariance matrix of the parameter estimates from the extended beta-binomial model. For more detail on the methods used to compute confidence intervals, see: King G, Tomz M, Wittenberg J. Making the most of statistical analyses: improving interpretation and presentation (King et al. 2000).
- f) **Assumptions behind the inequality index** *Brazil states that the assumption behind the inequality index is that inequalities in health correlated with socio-economic status are a redistribution problem.* This assumption is not made in the analysis, nor are any claims about the nature of socio-economic inequalities in health. The measurement approach used for the WHR focuses on variations in health across individuals pure and simple. The next step would be to determine the causes of inequalities and what can be done about them. This might well involve focusing on particular disadvantaged groups in society, but that is a subsequent step related to the difference between measurement and attribution discussed above.
- g) **Cross-population comparability.** *The Brazilian delegation is concerned that the WHO index is not suited for assessment of inequalities over time or across countries.* On the contrary, one of the strongest attributes of the index is that it can easily be used for cross-country and over-time comparisons because it is independent of any characteristics of a country or time-period that might hamper comparisons. In the design of the index, one of the major concerns was the feasibility of cross-population comparability. The analogy to measuring

income inequalities can illustrate the advantages of the WHO approach. Rather than measuring the Gini-coefficient, one could assess income inequalities by comparing the average incomes of different social classes. Such an approach to measuring income inequality has not been used because it excludes income inequality within social classes and because variation in the social classes across populations and changes in the composition of social classes would hamper meaningful comparisons. By measuring the entire distribution of health across individuals, comparability is ensured and all potential sources of inequality included.

- h) **Simulation examples.** *The Brazilian critique also provides two simulation examples that according to the authors demonstrate their points.* First, we believe that their calculation of the inequality index is *incorrect* for all examples stated. According to their simulated populations in Example 1, the index value for time t1 should be 7.04 and for time t2 6.54. In Example 2, the index for population 1 should be 3.48 and for Population 2 = 3.34. These four values are all different from those calculated by the authors of the critique.

Apart from the incorrect calculation of the index, the conclusions drawn from their examples are also incorrect. Figure 3 depicts their simulated Example 1.



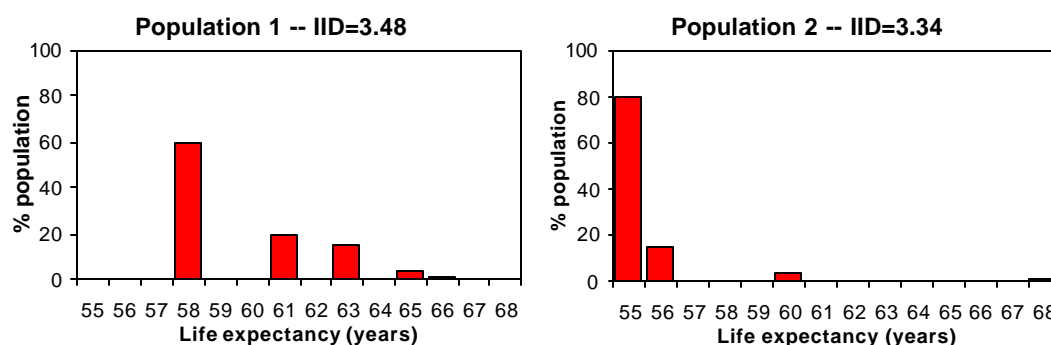
**Figure 3. changes in distributions of life expectancy**

The two parts depict a population in time 1 and time 2. The inequality index increases slightly from time 1 to time 2 to reflect the fact that in time 2 a larger fraction of the population is at the extreme values of 62 years and 72 years, than in time 1. Therefore, the index responds as it is expected to: as inequality in health increases, so does the index.

This example, as presented by the Brazilian delegation, has been constructed so that socio-economic inequalities also increase from time 1 to time 2. They then claim that the health inequality index is strongly related to inequalities in SES. However, there is simply a special case - one can easily think of a scenario where from time 1 to time 2 inequalities in SES have decreased or not increased using the same distribution of health outcomes. This is independent of the fact that in their example the distribution of health is more unequal in time 2 than in time 1, measured both by the standard deviation (an absolute measure) and the WHO index of inequality. The example below demonstrates why their claim is wrong.

**Simulation example 2**– *In this example, Brazil is claiming that the health system of Population 1 has "better performance" than Population 2. They further imply that the fact that the WHO inequality index is larger for Population 1 (i.e. the WHO index shows*

that health inequalities are greater in Population 1) means that the WHO index is erroneous. Life expectancy is higher in Population 1 by 4.2 years; hence, in terms of average level of health, Population 1 achieves more than Population 2 (though its performance is not necessarily higher because performance depends on attainment in relation to resource availability). However, there is much more dispersion in the distribution of life expectancy in population 1 than population 2 (shown in Figure 4 below) and inequality measured both by the standard deviation and by the WHO inequality index (II) is higher in population 1. In terms of inequality in health, the health system in population 2 achieves more. Ideally, we hope that health systems would be able to increase both the average level of health and reduce inequalities at the same time, but the move from population 1 to 2 represents a reduction in average levels, combined with a reduction in inequalities. On balance what is preferred will depend on the relative weights for health levels versus health inequalities.



**Figure 4. Changes in distribution of life expectancy – 2**

This illustrates that the average level of health and inequality in health should be measured and evaluated separately, as done in the WHR. They should not be mixed together as in the Brazilian example as this complicates evaluation. But purely from the perspective of inequality, the II describes both populations accurately.

## 5) Responsiveness: Level and Distribution

- a) **The number of respondents.** *The Brazilian delegation argued that the validity of the results was low for all countries because the numbers of key respondents used to estimate responsiveness was low, and for some countries the scores were estimated from other variables shown to be correlated with observed responsiveness. For Brazil, there were 33 key informants. We accept that the first attempt at measuring the responsiveness of different health systems was limited by the small data set. This was partly due to the fact that it was not possible to use the results of existing surveys asking similar questions as they had not been standardized across different countries. The results in the WHR should be seen as the genesis of an important idea and the associated measurement methods. This was highlighted in the Report by showing the relatively large uncertainty around the estimated responsiveness scores. For example, the uncertainty interval around the score for Brazil, ranked 130-131, overlapped with countries ranked between 75 and 165 showing clearly that it is not possible to be sure that Brazil ranked exactly 130<sup>th</sup>.*

In recognition that considerably more work needs to be undertaken on responsiveness, there are already face-to-face random sample household surveys underway in 40 countries, self-administered postal surveys in another 31, and key informant interviews prepared for as many of the 191 Member States who wish to be involved. The responsiveness questionnaire items used in these surveys has been carefully developed and field-tested in a series of pilot investigations in 10 different countries. The psychometric properties of the question items have been established and standard protocols were used for translation and back-translation of the questionnaire. The multi-country survey programme will provide a tremendous increase in the empirical knowledge of comparable levels of health system responsiveness. The overlapping design of the multi-country survey programme means that validity of the low-cost rapid appraisal method using key informants can be empirically assessed in comparison to the gold standard random sample household surveys.

- b) **Responsiveness and cultural and political diversity.** *The Brazilians indicated that the responsiveness scores did not control for cultural, religious and other forms of diversity between different countries.* In fact, every effort has been made in the development of the responsiveness survey instrument to ensure that comparisons on specific domains of responsiveness can be made meaningfully. The instrument used for the key informants survey for the WHR 2000 and the new responsiveness survey instrument focus on a generic set of interactions between citizens and health systems, which would be common across cultures and which would illustrate the various elements of responsiveness. The draft instrument was compiled after a review of more than 12 internationally known patient satisfaction and public opinion instruments (see GPE Discussion Paper 32) (De Silva 2000). These questionnaires had been developed for use in many different countries with differing organizational, cultural and political systems, although most of them were used in countries in the developed world. This required us to adapt the approach to also include developing country needs. In addition, certain questions focused on the ability of the health system to accommodate religious diversity.

On the other hand, adjustments were made to the results to accommodate the different proportions of educated people and government employees in the various samples of key informants. This was because responses differed according to education and whether someone was a government employee. Whether there are, indeed, significant differences in attitudes to responsiveness across or within countries is an empirical question – the surveys now under way will allow the hypothesis to be tested.

There is an even more important question than suggested by Brazil. Responsiveness differs from consumer satisfaction in a very important way. Satisfaction reflects people's perceptions of how the system performs relative to their expectations – surveys often show that the rich are less satisfied than the poor not because they are treated worse, but because they expect more. Responsiveness needs to abstract from expectations and measure what actually happens when someone comes into contact with the system. This is a challenge that is being addressed in the current surveys partly by the innovative use of specially designed vignettes.

- c) **Responsiveness and non-users of the system** *The workshop report argued that responsiveness measured only the opinions of users of the system.* This was not relevant to the 1999/2000 key informant survey used in the WHR 2000 – the key informants gave their

opinions as experts with knowledge of the whole society and not only as users of the system. Current household surveys have focused on measuring the responsiveness of the system to people who have used it in the past 12 months. Methodologically, there is a need to limit the time period, both because the overall attainment and health system performance indicators which require responsiveness data are estimated for a particular year, and because of the problem of poor recall if people visited the system more than 12 months ago. The current surveys are designed to elicit utilization rates of different types of services (e.g., hospital inpatient, GP's, etc.) for sub-groups within the population. The survey data on socio-economic characteristics of the respondents will make it possible to weight results to take into account differential utilization for different sub-groups of the population.

#### **6) Index of Fair Financing Contribution (IFFC)**

- a) **Health needs:** *The Brazilian delegation argued that health needs should be included in any concept of fair financing and its measurement.* The IFFC used in the WHR measures the inequality of health financing taking into account all payment mechanisms – taxes direct and indirect, social security, direct payments for services etc. It is independent of the household's health level so it might be that a poor household did not have any out of pocket payments for health because it could not afford them. So it is possible that two health systems have the same IFFC score – in the first everyone can afford health services but in the second part of the population cannot. We believe, however, that this is appropriate. In the second case, the population will show poorer levels of health and greater inequalities in health, ceteris paribus, than in the first – so the problem of poor access will be reflected in poorer health outcomes and in lower overall attainment. To include it in the IFFC would be double counting.
- b) **Vertical equity:** *It was argued by the Brazilian delegation that the IFFC used in the WHR ignored vertical equity – i.e. it did not take as the ideal situation that people with a greater capacity to pay should pay more.* This criticism is incorrect. The IFFC incorporates both vertical and horizontal equity (people with the same capacity to pay should pay equal amounts). A perfectly fair health financing system is defined using the IFFC as: the ratio of total health system contribution of each household to that household's capacity to pay is identical for all households. The capacity to pay is the household's effective income above subsistence spending.

To illustrate the point, we use the same table included in the report of the Brazilian workshop to illustrate how perfectly fair financing is also progressive; thus, it incorporates vertical equity (even though we used total expenditure in the place of gross income in our calculations to get closer to the idea of permanent non-subsistent income).

Income group	Gross income	Food exp.	Food/EXP	Disposable income	Health exp.	Health/disposable income	Health/gross income
1	10000	9500	0.95	500	25	0.05	0.003
2	20000	15000	0.75	5000	250	0.05	0.013
3	90000	30000	0.33	60000	3000	0.05	0.033
4	150000	45000*	0.30	105000	5250	0.05	0.035
5	300000	70000	0.23	230000	11500	0.05	0.038

\* the Brazilian numbers on gross income and food are used except for this one. The change is made because the Brazilian numbers imply that food expenditure as a proportion of gross income is identical for people earning 150,000 and 90,000. We know this is not the case – food expenditure as a % of gross income falls as income increases.

The situation depicted in the table yields a FFC index of 1, the maximum level of fairness by our definition. All households pay a fixed percentage of their non-subsistence income (called disposable income by the Brazilians). The last column shows that this system is also progressive – richer households pay a higher proportion of their gross income than poorer households. So we do not argue that all progressive systems are fair, as the Brazilian submission implies. In a progressive system, some households in income group 1 may be impoverished by catastrophic spending for health even the average spending in the group is a lower share of income than average spending in the higher income groups. We argue that fairness as a construct should reflect three concerns: avoiding catastrophic payments for health, horizontal and vertical equity. A system that approaches equality of shares of capacity to pay is also progressive because subsistence expenditure as a share of Income declines with income.

- c) **Taxes earmarked for health system financing.** *The Brazilian delegation's report noted that taxes earmarked for health were not taken into account in many countries as they are not included in the surveys (Living Standards Measurement Surveys) used to calculate the index of fair financing (IFFC).* This is not correct. In all cases, in addition to survey information, attributable taxes have been identified through other sources and included to trace back the share paid by individual households. To calculate the IFFC, efforts were made to identify and include all the taxes potentially used on health. In most countries, public spending on health comes from general tax and social security contribution. These can be estimated from salary records and knowledge of tax and social security systems. Other sources such as value added or sales tax and most excise duties are captured in household expenditure estimates from household expenditure surveys. Other taxes, such as corporate income tax, import taxes and so on, were incorporated assuming that their distribution among households were the same as income tax and value added taxes.

In Brazil's case, CLLE (Contribution on Company Net Profit) and COFINS (Contribution for Social security Funding) were included. However, the CPMF (Provisional Contribution on Financial Transactions) was not included as it was only introduced in 1997 and the estimates were based on 1996 expenditure surveys. As the impact of this tax would depend on its distribution among the population in 1997 it is difficult to conclude *a priori* was the possible bias would be.

- d) **Estimating Index of Fairness in Financing.** *The Brazilian delegation pointed out that household survey data were available only for 25 countries so the full micro-method was applied only to them. For the other countries the IFFC was estimated based on a regression with a  $R^2$  of 0.26.* It is true that we were able to locate household health expenditure surveys in only 25 countries at that stage. For those countries a multiple log-linear regression of the calculated IFFC on possible explanatory variables was run and the results used to predict the IFFC for the other Member States. The variables included were based strongly on the appropriate theory, as required for good estimation. A key variable was the fraction of total health spending represented by out-of-pocket payments. Out-of-pocket payments are those most likely to put a household at risk of catastrophic spending and thus those most likely to contribute to an unequal distribution of household financing contribution. Empirical data for out-of-pocket payments was available for nearly all 191 Member States. The overall regression is statistically significant with Root MSE=0.986, Prob(F)=0.000, and the multiple correlation coefficient is 0.51 ( $R^2=0.26$ ).

For those countries where the IFFC was estimated using the relationship between the fraction of total health spending from out-of-pocket payments and other variables, the uncertainty interval is wide. For example, Table 7 shows that Swaziland has an uncertainty interval from 0.790 to 0.962. The width of the uncertainty interval is a highly effective means of communicating to the reader the strength of the estimation method.

Because most countries have regular income and expenditure surveys, it should be possible to calculate the distribution of health financing contribution across households in nearly all Member States. Since the publication of the WHR 2000, the number of countries where such income and expenditure surveys have been made available to WHO has increased dramatically. This strengthened database on fairness in financial contribution will provide a strong database for exploring the determinants of the distribution of household financial contributions to health.

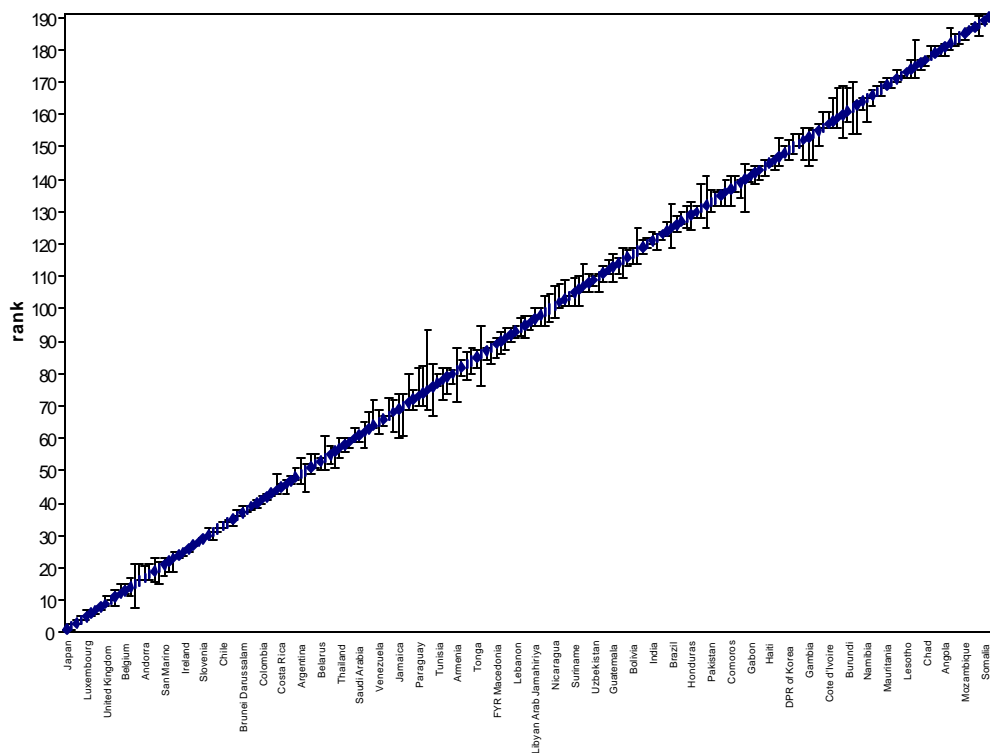
## 7) Overall Attainment

- a) **Sensitivity to changes in weights:** *The Brazilian delegation argued that the rankings of many countries changed with changes in weights and that this undermined the validity of the results.* A very fundamental issue has not been taken into account in much of the Brazilian commentary so we spend some time explaining it here. At no stage did WHO claim that a country's rank was certain. Uncertainty intervals around all the estimated scores were estimated, included in the text, and pointed out to readers - "all the main results are reported with uncertainty intervals in order to communicate to the user the plausible range of estimates for each country on each measure" (WHR, p. 144) (World Health Organization 2000). This is good analytical practice, but we believe it is the first time that an international agency has extended this good analytical practice to its official publications.

The sensitivity analysis undertaken for the World Health Report 2000 was based on a simulation procedure. Overall attainment for each country, and the associated country rank, was estimated by drawing one value randomly from the ranges of possible scores for each component reported in the Annex to the WHR. This was repeated 1000 times, providing 1000 estimates of overall attainment and 1000 possible ranks. From this procedure, the estimated overall attainment for Brazil lay between 67.1 and 70.4 in 80% of the simulations, with ranks of

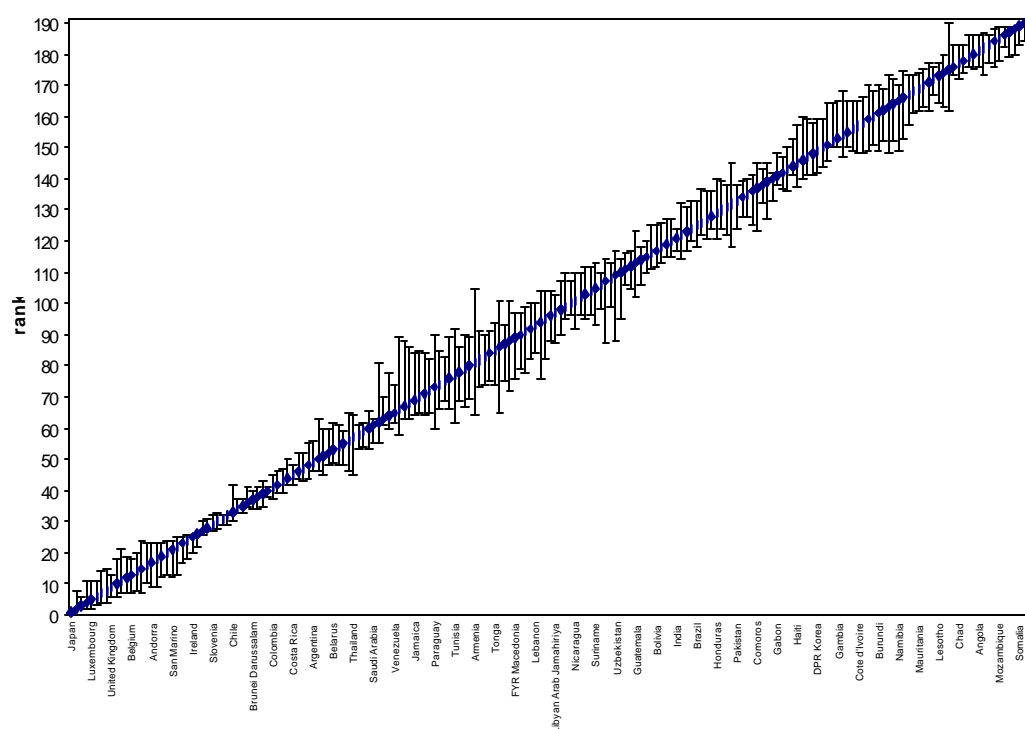
between 118 and 133. The average score was 68.9 and the rank associated with that score was 125. Countries with good data availability had tighter uncertainty intervals than countries where data were less certain.

It is true that this type of uncertainty analysis accounts for uncertainty in measurement rather than sensitivity to the choice of weights. However, sensitivity of the overall attainment score to the choice of weights was explicitly addressed in GPE Discussion Paper 28 (Murray et al. 2000). In that paper, we reported results from an analysis in which we randomly drew 100 different sets of weights and calculated the composite measure using each set. The ranges of values included all the choices of weights proposed by the Brazilian delegation and which they felt invalidated the report.



**Figure 5. Rank intervals resulting from a sensitivity analysis of the weights.**

The rank changes induced by variation of weights within plausible limits are simply much less important than those implied by measurement error in the underlying components. This is clear by comparing the width of the uncertainty estimates on ranks in Figure 1 and 2. There is much less variation in ranks caused by the choice of weights (Figure 5) than due to uncertainty in measurement (Figure 6).



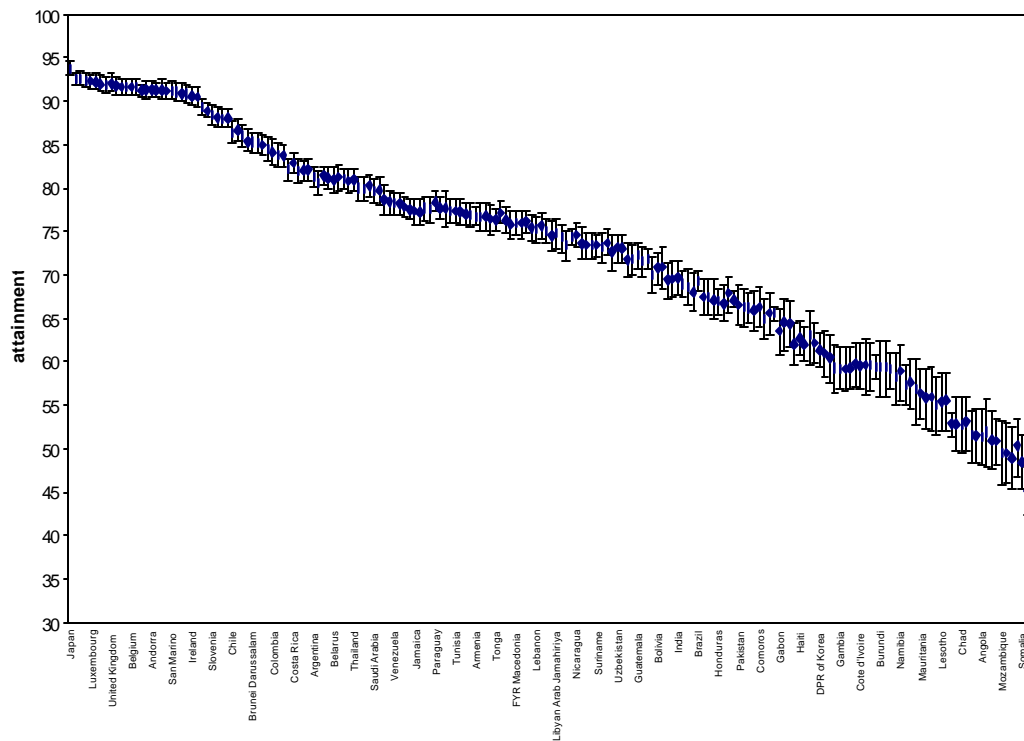
**Figure 6. Rank intervals resulting from measurement uncertainty.**

Now the Brazilian delegation claimed that small changes in weights changed significantly the ranks. But in only 13 cases is the rank interval resulting from changes in weights outside the uncertainty interval reported in the WHR. In six of these cases, the rank interval resulting from the sensitivity analysis is wider by only a single rank; furthermore, in no case is the rank interval resulting from the sensitivity analysis significantly different (in the statistical sense) from the rank interval implied by measurement uncertainty alone.

This implies that jointly measuring sensitivity of rank to variations in the weights *and* measurement uncertainty would add little to considering only measurement uncertainty for all but a tiny handful of countries. Among the latter are Brazil, Nauru and Vietnam, which would have slightly lower possible ranks (and in the case of Nauru, also higher possible ranks) implied by variations in the weights. So the WHR reported uncertainty intervals based on uncertainty around the measurement of variables, and then showed that this was much more important as a cause of uncertainty than the variations in weights which was the focus of the Brazilian critique.

The discussion to this point has focused on ranks. But they are, in fact, much more sensitive to minor changes in weights than the attainment scores themselves. For example, if one country has an attainment of 69.1 and another country 69.2, a small change in either will mean they may change rank (still within the confidence interval). But the scores themselves change very little with changes in weights. To illustrate, Figure 7 shows the uncertainty interval for the composite

measure of attainment. Here it is clear that there is very little sensitivity of the total attainment scores to changes in weights.



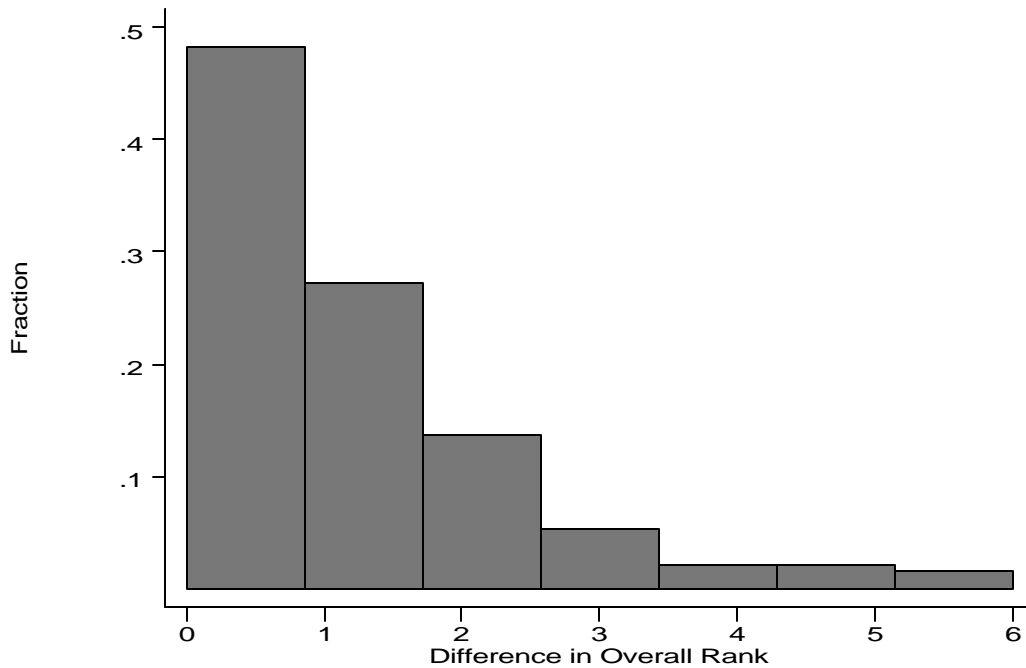
**Figure 7. Attainment intervals resulting from a sensitivity analysis of the weights.**

In addition to this general response, we have replicated the results reported by the Brazilian delegation at the EB, and also contained in the report of the Brazilian workshop, by changing the weights used to compute the composite index as they did:

- 0.24 instead of 0.25 for health level,
- 0.25 for health distribution (unchanged),
- 0.13 instead of 0.125 for responsiveness level,
- 0.16 instead of 0.125 for responsiveness distribution,
- 0.22 instead of 0.25 for fair financing.

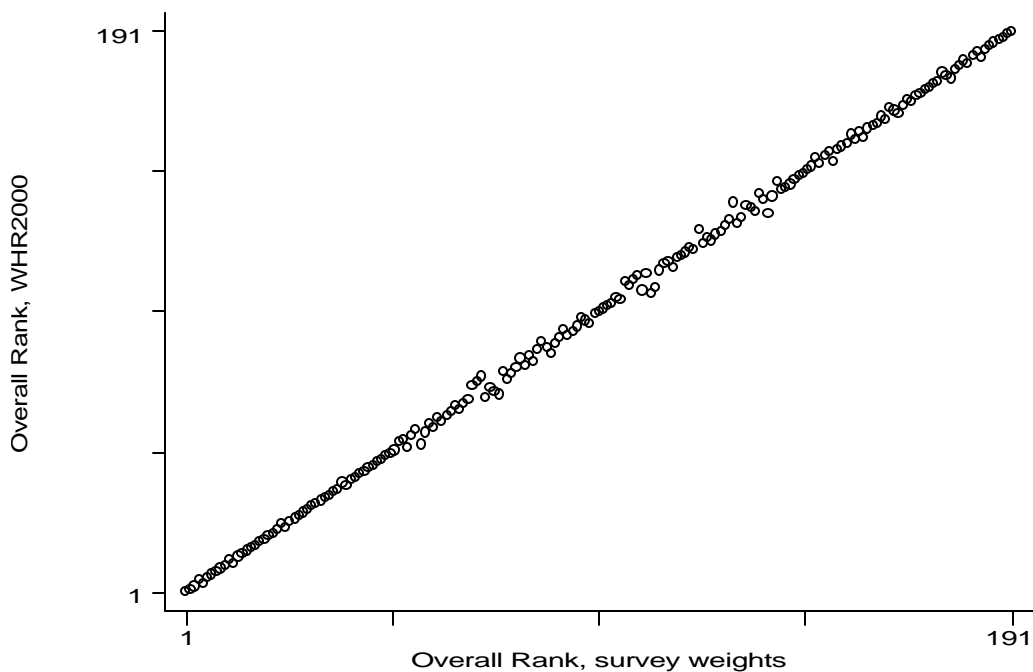
These weights are the unadjusted weights as measured with the survey instrument.

The report of the Brazilian workshop states that making the above changes results in “some countries moving up or down the scale by more than 30 points” (p. 23) (Oswaldo Cruz Foundation (Ministry of Health 2000). In the presentation, on the other hand, the claim was that countries move “up or down” by “up to five positions” and that rank is “indeed very sensitive to minor variations” in the weights.



**Figure 8. Rank differences implied by using the unadjusted survey weights.**

Contrary to those claims, Figure 8 shows that in our replication nearly 90% of countries showed small changes of 0, 1, or 2 ranks, and the largest change observed was 6 (Niue, China, and Bangladesh). This is totally consistent with the uncertainty analysis reported originally in the WHR. The rank correlation coefficient was 0.9996 between the new ranks using the Brazilian suggestions and those in the WHR (Figure 9).

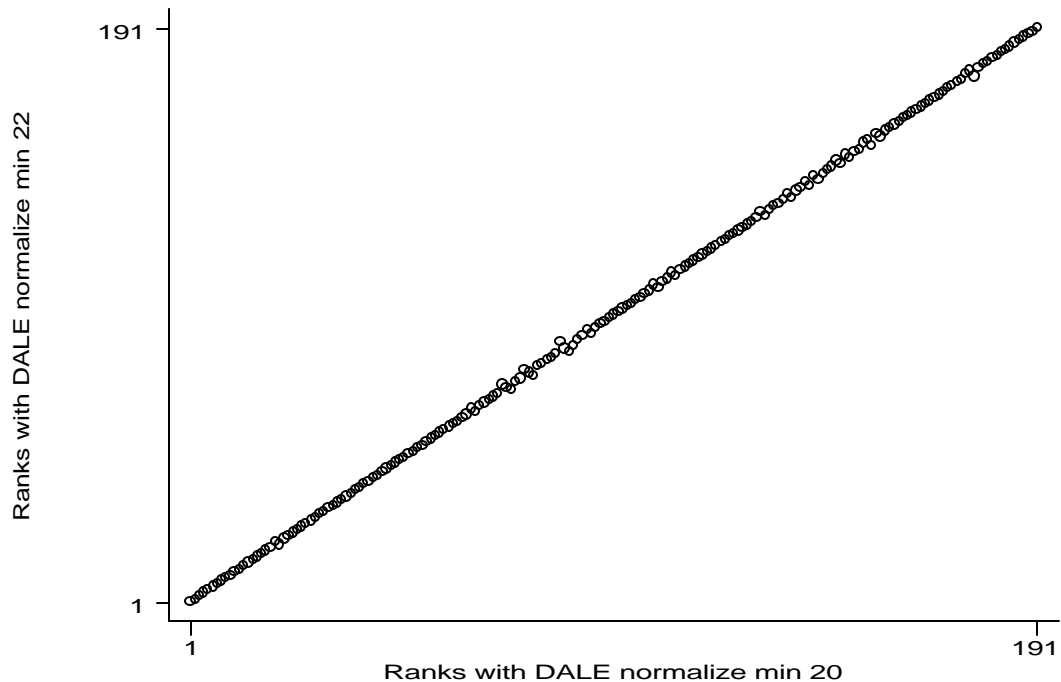


**Figure 9. Rank-rank plot using different weights.**

Accordingly, the gist of our response is that some of the claims made by Brazil were not correct and that the overall thrust of their criticism misunderstood the basic concept of uncertainty intervals.

- b) **False Precision:** *The second point on overall attainment made in the presentation by the Brazilian delegates was that the individual goal attainment scores used to calculate overall attainment were reported with unrealistic levels of precision. It is nothing other than an artefact of the Excel spreadsheet environment that the tables available on the WHO website contain up to 15 digits. All indicators were rounded for calculation of the overall achievement measure and other derivative measures such as overall performance and health performance. Again, the results of the uncertainty analysis reported in the WHR, and the fact that they show clearly that Brazil's overall attainment score was somewhere between 67.1 and 70.4, seem to have been misunderstood in this criticism.*
- c) **Other Scaling Functions.** *The Brazilian delegation criticized the methods used to scale attainment on the individual indicators such as healthy life expectancy in order to create the overall attainment index. Transforming all five indicators to the same scale was necessary for the application of weights. It is incorrect to add healthy life expectancies measured in years and responsiveness scores lying between 0 and 1, for example, in the creation of a single index. The approach adopted in WHR2000 was to transform the scales of each individual indicator into a [0,1]-scale where "0" represented the worst, and "1" the best outcome. The necessary transformations are reported in GPE Discussion Paper 28 (Murray et al. 2000). The Brazilians suggested that other scaling procedures could be used and would give different scores and ranks – in particular, they proposed three options: 1. transform healthy life expectancy by setting the minimum possible to 22 rather than 20 – they claimed this would result in rank changes for some 37 countries; 2. introduce an additional transformation to the WHR technique by means of a scalar multiplication step that ensures all five component distributions have the same mean; 3. use z-scores to transform the indicator scales. Each will be discussed briefly.*

**1) Changing the minimum healthy life expectancy from 20 to 22** does in fact induce about the number of rank changes claimed by the Brazilian delegation (37). Of course this means that 154 countries showed no change in rank. Moreover, most of the 37 changes are of one rank, seven are of two ranks, and one change is of three ranks (Albania). All of these changes are within the rank uncertainty intervals reported in the WHR. Furthermore, since the ranks that emerge from the Brazilian assumptions have a correlation of 0.9999 with the ranks reported in the WHR, the two sets of ranks are statistically indistinguishable. This is illustrated by the fact that a rank-rank plot shows nearly a perfectly straight line (Figure 10)

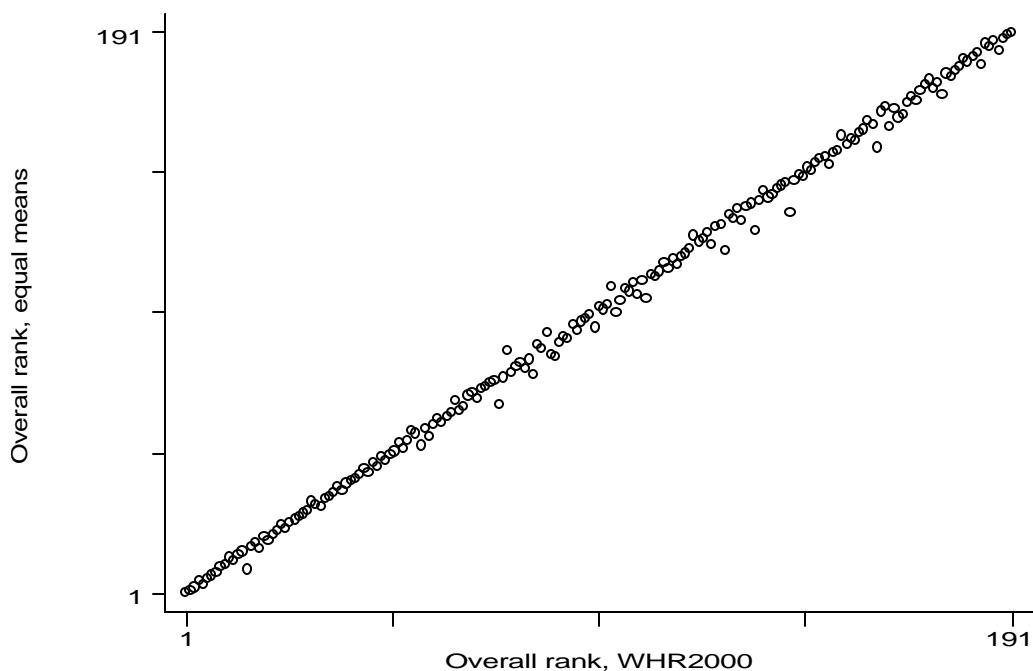


**Figure 10. Rank-rank plot using a different minimum.**

2) *The rationale for the second suggestion, that of transforming the distributions to equal-mean equivalents, is that distributions with higher means will implicitly count for more in the composite measure.* The Brazilian delegation made no arguments as to why this is a problem if in fact the true distributions of each measure have different means. The fact that a given score has a higher mean than another reflects the fact that fewer countries are at the lower end of the [0,1] scale for that index. It implies that the countries in the sample are relatively further from their theoretical minimum value for the index with a higher mean value.

The Brazilian delegation claims that the “real” weights used in the WHR2000 are not those stated in the report but are in fact those that *would have had to be used* with the equal-mean-transformed distributions in order to obtain the composite scores reported in WHR2000. This logic seems quite backwards: on what grounds can one assume the “real” distributions must have equal means?

Even if we were to use the new weights proposed by the Brazilian delegation, they fall well within the limits already analysed and reported in Discussion Paper 28 (Murray et al. 2000). To illustrate this point, we note that the rank order correlation coefficient is 0.9991 (Figure 11).



**Figure 11. Rank-rank plot using equal mean distributions.**

3) *As another suggestion, the Brazilian delegation proposed that the five components of the composite should be combined on the basis of z-scores.* No argument was provided why this would be appropriate. Fundamentally, the combination of the five measures into a composite is a normative judgement not a statistical exercise. Using z-scores is to claim that we should not value the importance of health, responsiveness and fairness in financial contribution for what they are but simply as a function of how variable they are across countries. We can imagine no basis for this assertion. However, our replication of this experiment shows that the rank correlation is 0.9818 – slightly lower than found with other assumptions but still very high.

Another fundamental problem with the use of z-scores to normalize and add the five different indexes is that it is inherently relativistic. In other words, what matters is deviation from the sample mean in a given year. However, if the mean value of the score changes from year to year, then it will be impossible to compare z-score-scaled indexes for a country over time.

- d) **Arithmetic Errors.** *The Brazilian delegation argued that there were arithmetic errors in the computation of the composite index in the WHR2000 Annex Table 9. In particular, they argued that the composite score for any given country does not equal the weighted average of the five individual component scores that were reported.* This is, of course, the case and is clearly described in Discussion Paper 28 (Murray et al. 2000). In fact, the composite score reported in the Annex *cannot* be the exact weighted average of the five individual scores and for the delegation to believe that they should be the same reflects again a misunderstanding of the uncertainty in interval concept.

The reason for this is simple: for all five indexes, the relevant Annex tables report the mean values resulting from 1000 simulations. We did not construct the overall attainment index by taking a weighted sum of these averages. This would be very simple, but it would be incorrect because it would ignore important information about the distribution that is represented by the uncertainty interval.

As reported earlier in this document, we calculated the overall attainment scores for the set of 191 countries, and the associated ranks, 1000 times, drawing randomly from the reported uncertainty interval for each of the component indicators. The mean overall attainment index is the mean of the 1000 separate calculations. It would be equal to the weighted sum of the means reported for each of the components only if all distributions were symmetric around their respective means. Given that some distributions of the component indexes were skewed, the simple method used by the Brazilians to compute the overall attainment index gives incorrect, although relatively close, results.

*Finally, the Brazilian delegation claimed that the simple method produces an overall attainment score for Brazil that lies outside of the uncertainty interval reported in Annex Table 9. We dispute this – our calculations using the simple method show an overall attainment for Brazil of 68.73 which is between the uncertainty intervals (67.1 and 70.4) reported in Annex Table 9.<sup>1</sup>*

## 8) Conclusions.

In this document we have highlighted an apparent difference in approach to the question of accountability – narrow or broad – which seems to have driven some of the Brazilian criticisms. We have also responded to other criticisms which we do not believe are valid, mostly because they reflect a misunderstanding of the methods or the way that they have been applied. At the same time, we recognize that there are weaknesses and we have tried to be open about which areas would benefit from additional work. We very much appreciate the interest of the Brazilian delegation, the experts who advised them, and those who participated in the Brazilian workshop, for the time they spent working through the methods and data. The WHR has raised an enormous amount of interest in the world, and the scientific interest this has generated will result in continual improvement of the methods.

## References

De Silva A (2000) *A framework for measuring responsiveness*. Geneva: World Health Organization (GPE Discussion Paper No. 32).

King G, Tomz M, & Wittenberg J (2000) Making the most of statistical analyses: improving interpretation and presentation. *American Journal of Political Science*, 44(2):341-355.

Mathers C, Sadana R, Salomon J, Murray CJL, & Lopez AD. (2000) *Estimates of DALE for 191 countries: methods and results*. World Health Organization, Geneva, Switzerland (GPE Discussion Paper No. 16).

<sup>1</sup>  $(0.652*0.25+0.762*0.25+0.481*0.125+0.944*0.125+0.623*0.25)*100=68.73$

Murray CJL, Salomon J, Mathers C, Lopez A, & Lozano R. (2000) *Summary measures of population health*. World Health Organization, Geneva (in press).

Murray CJL., Frenk J, Tandon A, & Lauer J (2000) *Overall health system achievement for 191 countries*. World Health Organization, Geneva, Switzerland (GPE Discussion Paper No. 28).

Oswaldo Cruz Foundation (Ministry of Health, B. (2000) Report of the Workshop "Health Systems Performance: The World Health Report 2000" (mimeographed document).

World Health Organization (2000) *World Health Report 2000: Health Systems: Improving Performance*. World Health Organization, Geneva, Switzerland.