CENTRE FOR HEALTH PROGRAM EVALUATION

WORKING PAPER 125

A Critique of the World Health Organisation's Evaluation of Health System Performance

Professor Jeff Richardson
Director, Health Economics Unit, Monash University

Dr Iain Robertson
Senior Research Fellow, Health Economics Unit, Monash University

Dr John Wildman
Research Fellow, Health Economics Unit, Monash University

October, 2001
ISSN 1325 0663
ISBN 1 876662 45 X

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding of these disciplines within the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au

Or by downloading from our website
Web Address http://chpe.buseco.monash.edu.au

ACKNOWLEDGMENTS

The Health Economics Unit of the CHPE is supported by Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR ACKNOWLEDGMENTS

The authors would like to acknowledge the assistance of the World Health Organisation for the provision of the data required for a reclamation of their study.

Table of Contents

List of Tables

List of Figures

.

A Critique of the World Health Organisation's Evaluation of Health System Performance

Jeff Richardson[1]
Iain K Robertson[2]
John Wildman[3]

Abstract

The World Health Organisation (WHO) approach to the measurement of health system efficiency is briefly described. The article then focuses upon four issues. First, it is argued that the choice of objectives gives undue prominence to the equity of financing and, more generally, inappropriately imposes a particular set of values upon all countries. Secondly, the importance weights attached to the objectives are somewhat misleading as an indication of the importance of the dimensions in the country rankings. Depending upon WHO objectives they are an also an incorrect measure of system performance. Thirdly, the model for combining the different objectives into a single index of system performance is problematical and alternative models are shown to alter system rankings. Fourthly, the econometric modelling of system performance does not provide a reliable basis for the evaluation of efficiency and the ranking of system performance.

It is concluded that, despite these problems, the study is a landmark in the evolution of system evaluation, but one which requires significant revision.

1 Professor Jeff Richardson, Director, Health Economics Unit, Centre for Health Program Evaluation, Monash University

2 Dr Iain Robertson, Research Fellow, Health Economics Unit, Centre for Health Program Evaluation, Monash University

3 Dr John Wildman, Research Fellow, Health Economics Unit, Centre for Health Program Evaluation, Monash University

A Critique of the World Health Organisation's Evaluation of Health System Performance

1 Introduction

Background

In the World Health Report 2000, the World Health Organisation (WHO) describes a new approach to the pursuit of its charter to improve the health of the people of the world. The proximate objective of the report is to

> *'present new concepts and measures which lay the empirical basis for assessing health system performance',* (WHO 2000 p 144)

The longer term strategy appears to be: (i) to establish a definition of health systems and the purpose of those systems; (ii) to develop a methodology for quantifying the efficiency of health systems that is applicable to all its member countries; (iii) to rank member countries using this metric; and (iv) to use the ranking as a stimulus to country specific research and reform.

The Report describes three phase of health system development: (i) the creation of large-scale health systems; (ii) the establishment of 'Health for All, 2000' priorities and the promotion of primary care; and (iii) a focus upon the efficiency of existing health systems. The Report intends to encourage this third phase, and to provide assistance with the benchmarking of performance.

The second of these stages commenced over two decades ago, when the WHO initiated the Health for All 2000 strategy. This emphasised the inter-relationships between different sectors of the countries of the world (food production, education, employment and wealth creation, environmental control, security and others) in addition to health services in the generation of health and the prevention and treatment of illness. Countries were grouped on the basis of the stage of economic development and the pattern of disease. These country groups established sets of health priorities, and targets for health improvement. For example, European countries had targets for a 25 percent reduction in the rate of coronary heart disease mortality and morbidity, whilst developing countries sought reductions in the rate of childhood malnutrition and mortality from gastroenteritis. The idea that specific health services should be provided closest to users and by the simplest method of service delivery led to the promotion of primary care as the basis of health services.

It proved to be difficult to achieve a world-wide or inter-sectoral consensus on specific goals and policies to achieve them[4]. Political and economic instability, and the lack of global consensus on the direction of world-wide development has meant that progress in health improvement has been patchy, and in some parts of the world, notably in Africa and some former Eastern block countries, health indices have deteriorated. Since grand-scale strategies have not proved to be universally effective, the third phase described by the WHO might be seen as a more modest but targeted strategy to focus attention upon a limited number of systemic objectives, and to provide system managers with a statistical measure of efficiency with respect to these objectives which will help managers drive system reform.

The Measurement of Efficiency

The conceptual steps in the construction of the WHO index are shown in Figure 1. Five objectives are nominated as evaluative criteria. These are measured and form the basis for two parallel streams of analyses. In the first, a summary measure of a nation's health is examined, namely, the Disability Adjusted Life Year Expectancy or DALE. For the second stream, an overall index of systemic performance is constructed from the five objectives (including health). In both streams the key variable (DALE or the performance index) is explained econometrically by the level of a country's health expenditures and education. These are assumed to fully represent the inputs into the 'production' of national health and system achievement. Efficiency is defined as the difference between a country's actual performance and the performance that is possible with the country's health expenditure and education. What is in principle, technically possible is determined empirically by the performance of the most efficient country. Details of the technical analysis are discussed in Section 5 below.

Goals and their Measurement

The health system is defined in the WHO analysis as:

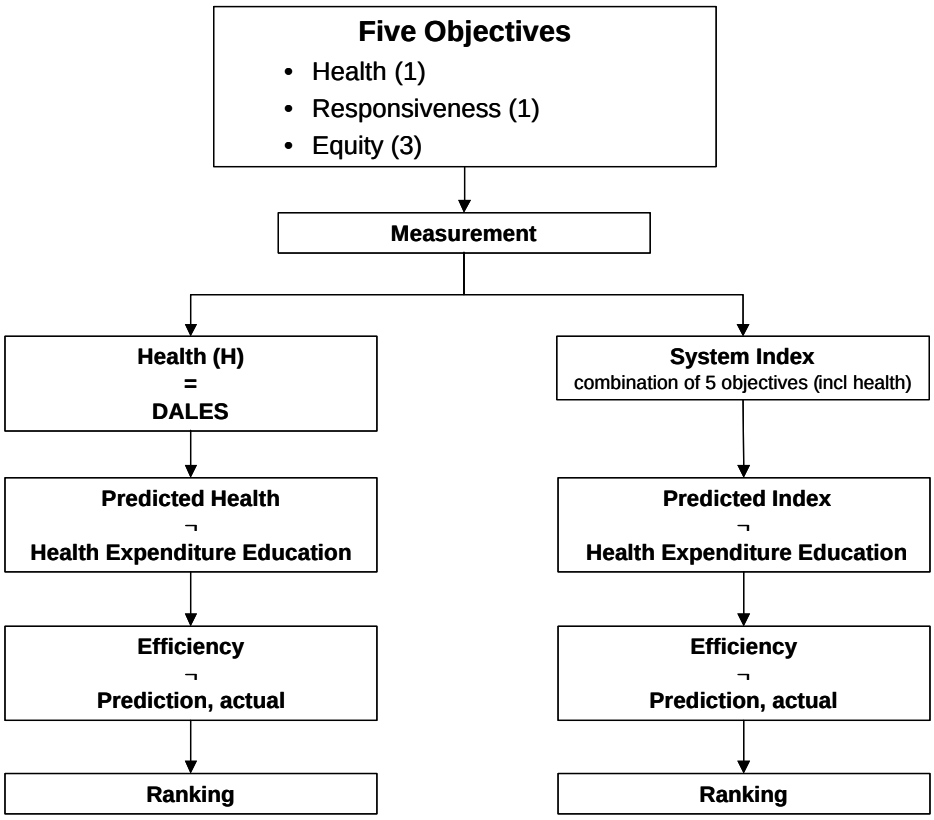
> *'… all the people and actions whose primary purpose is to improve health.'*

The five systemic objectives are:

- Maximising population health;
- Reducing inequalities in population health;
- Maximising health system responsiveness;
- Reducing inequalities in responsiveness;
- Financing health care equitably.

[4] Experience in different countries demonstrated that improvements in health were not dependent on economic development and wealth creation on the classic Western pattern. It had been believed that reduced population growth and infant mortality required Western-style economic growth, but the experience in Bangladesh in the '70s and '80s demonstrated that reduced fertility rates and infant mortality rates could be achieved despite no major improvement in GDP. The critical factor appeared to be increased education of females: schooling of females gives them the ability to take over control of their reproductive lives, to understand the public health messages available, and negotiate the flow of resources within their communities that allows improvement in their own health and that of their children, even though the overall community resources are not greatly increased.

Figure 1 Steps in the construction of the WHO country scores and ranking



While conceptually appealing, the definition of the health system cannot be easily measured as many of the people and actions that are dedicated to health are outside the cash economy. Consequently, in the statistical analysis the health system is defined conventionally by paid health services and associated administrative activities. This allows the costs of those services, as reported in national accounts and collected by the WHO, to be related to the measures of health and systemic performance[5] proposed in the report.

The health of the population is measured by the Disability Adjusted Life Expectancy (DALE) from birth as generated by the Global Burden of Disease Study undertaken by the WHO. This takes the published annual population and mortality experience in each country for each age cohort, and converts these into DALEs for each country using disability norms for different diseases. Thus, for any particular country, the amount of disability due to a particular disease is not measured, but assumed to equal the mortality and average disability rates in the sample countries used in the production of the norms.

The distribution of DALEs is not used in the analysis (although presumably this deficiency will be addressed as the performance measure is refined). Instead, small area variation in childhood mortality is used as a proxy for distribution of population health.

[5] It is not clear whether services, such as these, associated the provision of water and food free of contamination are included in the definition of health systems, even though their primary purpose is clearly directed at improving health. We suspect they are not included because of the difficulties in the identification of such expenditures in different countries.

The responsiveness of health systems to their clients is measured using a questionnaire that inquires about the degree of respect for persons (dignity, autonomy, confidentiality) and client orientation (prompt attention, quality of basic amenities, access to social support networks, choice of provider). The initial country estimates, and the distribution of the responsiveness within countries, were derived from convenience samples and best guesses by respondents. It is likely that national household health surveys will be used in future studies. It should be noted that one of the primary functions of the health system is to respond to health problems as identified by the users of the system. It may be helpful in the future development of the health performance index to investigate ways of measuring how closely problem assessment by the health care professionals coincide with that of their clients.

Fair financing is defined as the sum of all costs paid directly or indirectly by households as a percentage of their disposable income. It is discussed further below.

The combined health system performance index is constructed as a simple weighted average of the scores on the five attainment scales. Weights obtained from survey respondents are assigned to each of these as follows:

		%
health (DALEs)	. . .	25.0
health distribution	. . .	25.0
responsiveness	. . .	12.5
distribution of responsiveness	. . .	12.5
fairness of system financing	. . .	25.0

Results

A selection of the final results from the study are reported in Table 1. The ranking of Australia is of particular interest (and especially for Australians!). These indicate that in terms of DALEs Australia ranked second. However, the methodology reduces its ranking with respect to health and systemic performance to the 39th and 32nd positions respectively. These are puzzling results as there is no particular aspect of the Australian health system to explain such a downward revision and it suggests the possibility of one or more defects in the methodology used to generate the results.

Critique

In the March 2001 edition of '*Health Economics*' Alan Williams (2001a) lambastes the authors of the report primarily for their use of unacceptably poor data.[6]

> *'(The index) is based on very little actual data, which is often heavily manipulated to make it usable, and then subjected to a great deal of rather adventurous modelling…'* p 97.

> *'Apart from the DALE calculations, which are all partly factual and partly speculative, the only indicator for the USA which is not imputed is child mortality, and for Denmark they are all imputed. So it was not Real Denmark that was rated below Real USA, but Fictional Denmark that was rated below Fictional USA.'* p 99.

[6] See Murray et al 2000 and Williams 2001a for the authors rejoinder and Williams' further comment.

Williams does concedes that 'At a technical level there is much ingenious analytical work behind this report…' p98.

In the present article we amplify and extend some of Williams' comments on objectives and importance weights in the context of a cross national comparison of health systems but also we consider some of the more technical issues of system modelling more specifically we comment on four aspects of the WHO study. These are:

(i) The objectives;
(ii) The importance weights assigned to objectives;
(iii) The model used to construct the composite performance indices;
(iv) The econometric modelling of predicted health and system performance.

The conclusion we reach is that to maximise its effectiveness the WHO analysis should take into account some serious methodological problems which, we believe, threaten the validity of the results.

Table 1 Health system attainment and performance in selected countries ranked by four measures, (estimates for 1997)

	Attainment of Goals		Performance	
Country	**Health (DALES)**	**Overall goal attainment**	**On level of health**	**Overall health system perform.**
Australia	2	12	39	32
Austria	17	10	15	9
Belgium	16	3	28	21
Canada	12	7	35	30
Denmark	28	20	65	34
Finland	20	22	44	31
France	3	6	4	1
Germany	22	14	41	25
Japan	1	1	9	10
Luxembourg	18	5	31	16
Malta	21	31	2	5
Netherlands	13	8	19	17
New Zealand	31	26	80	41
Oman	72	59	1	8
Portugal	29	32	13	12
Russian Fed	91	100	127	130
Spain	5	19	6	7
Sweden	4	4	21	23
Switzerland	8	2	26	20
Turkey	73	96	33	70
UK	14	9	24	18
USA	24	15	72	37

Source: WHO, World Health Report 2000

2 Objectives

One of the great contributions of the WHO study is that it makes explicit the existence of multiple objectives rather than the two—health outcome and 'equity'—acknowledged in most economic analyses. It further challenges orthodoxy by suggesting that, quantitatively, fairness may be of greater importance than health outcome per se. The selection and treatment of objectives, however, are contestable for a number of reasons. We raise one specific and one general concern. The former is with respect to the measurement and inclusion of the fairness of financing in the composite index and the latter arises from the assumption of universally applicable weights.

First, the WHO methodology imposes the same objectives upon all countries. This is legitimate if the purpose of the WHO is to evaluate health systems *according to the values and ethical judgements of the WHO*. However it would be unfair and misleading to describe a county's health systems as inefficient because it did not perform well when judged by objectives which were not their own. This issue is the essential element in the debate between the proponents of the WHO model (Murray et al 2001) and Blendon et al (2001) who show that for 17 industrialised countries the WHO ranking of performance does not correspond with a ranking based upon consumer satisfaction—a different but defensible systemic objective. The self evident fact that the rank order of a country's performance may change when objectives change is illustrated later in Table 4.

Secondly, it is assumed that the same loss of DALYs occurs for the same disease in every country. As highlighted by Reidpath et al (2001) this implies the same loss of wellbeing for the quadriplegic in Cameroon who must drag herself through the mud to reach a latrine as for the wealthy Australian quadriplegic in her electric wheelchair and in an environment modified to assist the disabled. Third, and the issue discussed in Section 3, the importance weights and model used to combine attainment scores are problematical. Fourth, we argue below that fair financing should not be included in the study.

Fair Financing

Fair financing is measured by a ratio described as follows:

> *'the numerator includes all costs attributable to the household, including those it is not even aware of paying, such as the share of sales or value added taxes it pays on consumption, which governments then devote to health, and the contribution via insurance provided, and partly financed, by employers. The denominator is a measure of the household's capacity to pay… Total non food spending is taken as an approximation of the household's discretionary and relatively permanent income.'*

The approach here is consistent with the conventional assumption that fair financing should be a criterion for evaluating health systems (Wagstaff and Van Doorslaer, 1994). However, the assumption is contestable. The criterion conflates two separate social objectives. The first is the achievement of health and health related fairness. The second concerns the distribution of income. Health per se may be affected by the distribution of income, and access to health services will be affected by the existence of financial barriers. However access is a different issue than fair financing. Private out-of-pocket expenditures are borne exclusively by sick people and this does raise a legitimate ethical issue: should the sick have to pay more; should there be a 'tax

on the sick'? But the majority of funds in most countries are derived from tax revenues and the composition of government taxation is unrelated to any objective relating to health per se or health related fairness.

It might be argued that, while not strictly related to health, equitable financing is a legitimate objective. The argument should, however, be unpersuasive for two reasons. First, the source of taxation does not determine the pattern of government expenditures. (Taxes are seldom tightly hypothecated and insulated from other taxes.) Conversely, particular expenditures cannot be causally linked to particular taxes. Attribution rules are arbitrary (as with the attribution of fixed costs). For example, a common and appealing rule is to apportion health expenditures to taxes in proportion to total tax revenues from each source. This, however, implies that the evaluation of fairness of government tax based health revenues is equivalent to the evaluation of national taxation and this should not be considered as an objective of the health system per se.

Even when compulsory levies are nominally associated with the health sector, the apparent fairness of the contribution may be illusory. This is illustrated in Table 2. A nominal health insurance premium is levied in two countries in a way which implies greater equity in country A. In this country the poor pay nothing. In Country B the percentage contribution by the poor is double the contribution of the rich. *Prima facie*, this appears to be inequitable. However in this second country the rich are independently taxed at a far higher rate than in Country A. The overall impact of these taxes leaves the rich in Country A with 60 percent of their gross earnings which is significantly more than the 45 percent retained by the rich in Country B. The poor in Country A are left with 70 percent of gross income, less than the 90 percent in Country B: the country with the superficially inequitable contribution to health insurance. The general point is that equity cannot sensibly be segmented and especially when contributions are involuntary. The higher insurance premium for the poor in Country B may be an explicit recognition of their low level of general taxation.

Table 2 Equity Tax and Health Financing

	Country			
	A		**B**	
	rich	**poor**	**rich**	**poor**
Outlay (% of Income)	**%**	**%**	**%**	**%**
Tax	30	30	50	0
Health Ins. Premium	10	0	5	10
+ Out of pocket	0	0	0	0
Pvt Consumption				
+ Saving	60	70	45	90
Total Income (%)	100	100	100	100

Finally, and most fundamentally, the index of fair financing derived here and elsewhere does not provide a correct answer to an interesting question (although it purports to do so). In the present context at least three questions might be asked. First, 'what was the effect upon the distribution of income when a national health system was created?' (The answer might require the cumulation of incremental changes that occurred through time.) The question, however, is only of historical interest, as other elements of taxation might have been adjusted because of the initial effects of health financing legislation.

The second question is 'what is the effect upon the present distribution of income of present health financing arrangements?' As noted above the answer to this question requires the use of attribution rules to link health revenues with particular taxes. However there is generally no causal relationship between particular taxes and health revenues. As the growth of health expenditures rises and falls there is no corresponding increase and decrease in the taxes to which health revenues have been attributed. As also noted, the common pro rata rule reflects, not health related fairness but the fairness of the entire tax system.

Thirdly, it may be asked whether a country's health financing is more or less regressive than health financing in other countries. While the answer to this question again requires the use of an attribution rule, it may be argued that the use of the same rule will lead to a valid comparison. The argument is not, however, correct. The same attribution rule applied to differing taxation systems will produce results reflecting the overall taxation system. In general, the consistent use of an invalid method does not lead to valid conclusions.

The attempt to determine the distributional effect of a health system encounters a conceptual problem that has been neglected in the literature. This is that the impact of one scheme may only be determined by comparison with another scheme. The first rule of economic evaluation is that there must be a comparator. The fair financing literature has not made this explicit and has commonly and implicitly adopted an unrealistic comparator. The assumption that tax revenues can be attributed pro rata to the various components of total taxation is equivalent to assuming that if a national health scheme was to be dismantled then the revenue saving to government would be returned to taxpayers in direct proportion to the size of each tax. Further, it must be assumed that nothing else would happen; people would not find alternative methods for financing health services. If they were to do so then the effect of dismantling the national health scheme would depend upon the financial incidence of the (unknown) new health scheme.

The counterfactual – the demise of the national scheme - could also alter the pattern of spending. The revenue saving could be used to reduce proportional taxes: income taxes on the wealthy could be reduced (as proposed in the USA by President Bush); or social service expenditures could rise. The retention and financing of the NHS would, therefore, stave off a redistribution where the initial effect was neutral, regressive, or progressive.

In sum, economic evaluation, including the evaluation of tax incidence, requires a choice between alternatives and, as sunk costs and sunk benefits are irrelevant, the appropriate choice is between future options. Without a knowledge of these the distributional effect of the status quo cannot be determined.

3 Importance Weights

In principle, the omission of an objective is equivalent to the assignment of a zero importance weight. The criticism that there is no universally accepted set of objectives translates into the criticism that it is nations assign different importance weights to the WHO and other possible objectives.

Prima facie, the weights attached to the different objectives, reported earlier, are implausible. Health achieves an importance weight of only 0.25 and the other four objectives, all concerning fairness, or responsiveness (as distinct from health per se) have a combined importance weight

of 0.75. As Williams (2001b) points out a significant drop in health outcome might be offset by a comparatively modest increase in equity.

The nominal weights are misleading. The effective weight depends not simply upon the nominal weight but also upon the variation in the scores to which they will be applied. For example, if there was no variation in the equity of health outcome then, despite the large importance weight of 0.25, equity would contribute nothing to the variation in the performance indices and, would be of no importance for the ranking of health systems.

In Table 3 an adjustment is carried out to take into account the range of values for each of the objectives. Column 1 reports the nominal importance weights. Column 2 reports the value of the maximum less minimum score for each of the objectives for all countries in the study (Column 2a) and for the 50 countries with the highest performance index (Column 2b). The maximum impact of each objective on the performance index—Column (1) x (2)—is reported in Column 3 and standardised effective weights, (ie weights which sum to 1.00) are reported in Column 4 for the two groups of countries. Results indicate that the effective weight for health is significantly greater than the nominal weight. The weighting for equity in health does not change significantly. Responsiveness more than doubles its importance weight for the top 50 countries but almost halves its importance for all countries. In the standardised responses equity in responsiveness ceases to be of importance and fair financing is of reduced importance and particularly for the top 50 countries.

The effective weights may or may not appear more plausible but the general criticism still applies. It is highly unlikely that there will be agreement over these weights as they reflect social or ethical rather than technical judgements. Countries with a strongly egalitarian or communitarian tradition are likely to place greater emphasis upon equity and less emphasis on responsiveness. Countries which place greater relative importance upon individualism will place greater relative importance upon responsiveness.

Table 3 Importance Weights: Apparent and Effective[(1)]

Objective	Weight	Range (max to min)		Wt.* Range		[(1)] = Effective Weight	
		All	Top 50	All	Top 50	All	Top 50
Column	**1**	**2a**	**2b**	**3a**	**3b**	**4a**	**4b**
Health	0.25	0.912	0.173	0.22	.04	.33	.37
Equity In Health	0.25	0.754	0.107	0.19	0.026	0.29	.24
Responsiveness	0.125	0.414	0.237	.052	0.03	0.08	0.28
Equity in Respons.	0.125	0.586	0.018	0.073	0.002	0.11	.02
Fair Finance	0.25	0.524	.034	.131	0.01	0.20	0.09
Total	**1.00**	**-**	**-**	**.666**	**.108**	**1.00**	**1.00**

Notes: [(1)] The 'effective weight' is the weight which must be assigned to a 100 point scale with endpoints set equal to the maximum and minimum observed score so that the (new) weighted score adds exactly the same to the total score as the old weighted score and range. Its advantage is that the contribution to the final index is immediately apparent from the importance weight without reference back to the range of scores.

The self-evident fact that the rank order of a countries performance index may change with a change in the importance weights is illustrated later in Table 4.

Variability of values does not imply the impossibility of a global analysis of system efficiency. The WHO exercise could be repeated with each country having a unique combination of importance weights. The analytical requirements of such an exercise would be trivial. Importance weights would simply need to be constrained so that the final indices were comparable. For example, with the simple additive model used by the WHO (indices are equal to the weighted average of scores) importance weights should sum to unity and the maximum and minimum country scores be set equal to zero and unity when weights are derived.

Prima facie it may appear that the use of different importance weights in different countries would invalidate comparisons of the indices. This is not true. Indices would now be reinterpreted as the efficiency of each health system relative to its own stated objectives. This would be no more or less confusing, conceptually, than the concept of the gross national product and cross national comparisons would be no more problematical than cross national comparisons of GDP. This aggregate measure does not use either standardised importance weights or a standardised range of products ('objectives'). Rather, importance weights—market prices—reflect the supply and demand conditions which will be unique in each country. Cross national comparisons are possible because the importance weights—prices—reflect the same value, namely the revealed preference for different goods and services and a linkage between the nominal unit of measurement—national currencies—exists because of both trade and the construction of indices of purchasing power parity. Likewise importance weights in the WHO analysis may differ yet permit comparison if they are constructed with the same methodology and endpoints.

4 Modelling the Index of Overall Goal Attainment

Once a set of objectives, scores and importance weights have been determined these must be combined into a single index. This may be done a number of ways and four options are given in equations 1 - 4 (below).

Four Weighting Formula (Health)

$$I_j = \Sigma_i w_i S_{ij} \quad \ldots (1)$$

$$I_j = \{1 - \Pi_i (1 + w_i s_{ij})\} .100 \quad \ldots (2)$$

$$I_j = \{1 - \Pi_i (1 - s_{ij})^{wi}\} .100 \quad \ldots (3)$$

$$I_j = H_{ij} (w_h + \Sigma_i w_i s_{ij}) \quad \ldots (4)$$

I_j = Performance index, country j
w_i = importance weight, objective i
H_j = health outcome, country j
S_{ij} = numerical score, objective I, country j
s_{ij} = $S_{ij}/100$

The first option is the one additive model adopted by the WHO. The overall performance index, I, is a weighted average of objective scores where the weights sum to unity as shown in the second column of Table 4. While it is the simplest and most commonly used combination model it has some undesirable properties. For example, a society may consider its health scheme to be very poor if *either* health outcome was very poor and equity very good or if the health outcome was very good but equity very poor. There is no compelling reasons why the rate at which people's assessment of their health system should decline in the way described by the additive model. Analogously there is no reason why the importance of different dimensions of health should have this property. Pain may be so great that at its worst level it may reduce the quality of life (QoL) to zero. Likewise, depression may lead to attempted suicide. With the simple additive formula this cannot occur. The importance weights on psychological and pain dimensions plus other dimensions must all sum to unity and, consequently, with this model each weight must be relatively small. A further property of the additive model is that a reduction in the value of one score does not affect the importance of other scores. With the WHO weights, a country might obtain an index score of 0.75 for responsiveness and the equitable distribution of useless health services. The simple additive model is only a special case of the general combination rule suggested in decision analysis (Von Winterfeldt and Edwards 1993). This is shown as equation 5 below with a special case of the more general formula.

$$I = 1/k \{\prod_i (1 + k\, w_i\, s_i) - 1\} .100 \quad \ldots (5)$$

The additive formula occurs when the overall scaling constant, k, is equal to zero[7]. When the importance weights sum to a value greater than unity the scaling constant, k, is negative and, as the sum of the importance weights increases the scaling constant, k, approaches -1.0. In these cases equation 5 reduces to a multiplicative model, equation (2). With k = -1 equation 5 becomes equation 2. An exponential form of the multiplicative model is defined by equation 3.

Both of these multiplicative models have the disadvantage that, like the additive model, a score of zero for any objective does not reduce the index number to zero. This implies that with either form of the multiplicative model a significant index number could be achieved when health has a score of zero. This problem is overcome in the fourth, 'eclectic' model. In this, the overall index number is directly proportional to health when other weights are fixed. The importance of other objectives depends upon the importance of health. Eclectic, multiplicative, and eclectic expediential rules are also possible but not discussed further here.

The ranking of health systems will depend, upon both the importance weights, and the choice of model. This is illustrated in Table 4 which gives the index number derived from each of the four models for two hypothetical countries, Sweden and the USA. These have (arbitrarily selected) importance weights which indicate that Sweden places a greater emphasis upon equity and the USA upon health per se. The first column of weights represent unadjusted importance scores and in the second column these are standardised. For each of the models defined by equations (1) – (4) both Swedish and US weights are used.

[7]
$$I = 1/k \{1 + k\, w_1\, s_1 + k\, w_2 + s_2 + k^2\, w_1\, w_2\, s_1\, s_2 + \ldots - 1\} .100$$

if k = 0
$$I = w_1\, s_1 + w_2\, s_2 + k\, w_1\, w_2\, s_1\, s_2 \ \ldots .100$$

$$I = w_1\, S_1 + w_2\, S_2$$

Table 4 Four Models of System Performance

	Sweden					USA			
	weight (w_i)		score (s_i)	index (Max 100)		weight (w)		score (s_i)	index (Max 100)
	(1)	(2)				(1)	(2)		
Health (H)	0.7	0.47	75			0.9	0.6	95	
Equity (E)	0.8	0.53	95			0.6	0.4	50	
Index 1									
additive									
w ← Sweden				86	>				71
w_i ← USA				83	<				77
Index 2 =									
multiplicative									
w ← Sweden				89	>				80
w ← USA				86					90
Index 3									
exponential									
w ← Sweden				97	>				93
w ← USA				95	>				96
Index 4									
eclectic									
w ← Sweden			73						70
w ← USA			74						76

Notes: Models 2 and 3 use unadjusted weights;
Models 1 and 3 use standardised weights which sum to unity.

The result of this hypothetical exercise is that the ranking of the two countries with respect to the 'weighted scores for health and equity' depends upon which country's weights are selected and which combination rule is used. With Swedish weights Sweden out-performs the USA with each combination rule. In contrast, with US weights the rank order is reversed in three of the four models. With the simple additive model used by the WHO Sweden unambiguously out-performs the USA. With the remaining three models the outcome depends upon the choice of importance weights. The results in table 4 demonstrate a further general conclusion. With given weights the numerical superiority of one country over the other varies. Thus, with Swedish weights the index number in models 2, 3 and 4 favour Sweden by 9, 4, and 2 percentage points. With US weights its index numbers exceed the Swedish by 4, 1 and 2 percentage points. This suggests that with more than two countries in the ranking the extent to which the choice of weighting system affected rank ordering would also depend upon the choice of combination rule.

5 Validity, Reliability and Goodness of Fit

Efficiency scores were obtained by the WHO for each country with respect to health per se (DALEs) and the achievement of objectives (the composite index). The econometric procedures used in these analyses were replicated as discussed below in order to answer three questions; viz,

1. Does the WHO procedure capture true efficiency (validity)?

2. How well does the procedure perform when there are missing data (reliability)?

3. What is the goodness of fit of the regression equations from which the efficiency indices are calculated? In particular, does the WHO procedure in combination with the full data base of 191 countries out-perform two simple alternative procedures; viz, (i) the assumption that each OECD country has the same performance indicator—the OECD average; and (ii) the derivation of the efficiency index from the error term in a one year OECD specific OLS regression.

WHO Estimation Procedures and Replication of the WHO Model

The efficiency estimates in the World Health Report 2000 are obtained by combining fixed effects panel data methods with a standard corrected ordinary least squares procedure. Efficiency is measured for both outcome variables, DALEs and the composite index. The models used to estimate each output are reduced forms of a trans-log model (a flexible functional form) and are identical for each output.

The WHO criticise data envelopment analysis (DEA) and disposable hull methods for being unable to separate true inefficiency from random variation (Evans et al 2000). The authors also criticise corrected ordinary least squares (COLS) methods for not bounding the data, for being problematic if heteroskedasticity is evident and for suffering from an inability to distinguish between random errors and true inefficiency. Despite this, the COLS method is used to estimate the minimum level of health. Stochastic frontier models are rejected due to the need to impose distributional assumptions on the efficiency term.

Evans et al claim that panel data estimation is more efficient than models using cross-sectional data in extracting information on inefficiencies. They also claim that panel data estimation is preferable to stochastic frontier techniques because there is no need to specify the distribution of the inefficiency terms. Fixed effects estimation is used because the models do not pass a Hausman test, suggesting that the regressors are correlated with the error term, leading to biased estimates using a random effects model.

WHO Model

Evans et al specify the following model for both DALEs and the composite index:

$$Y_{it} = a + X_{it}b + v_{it} - u_{it} \quad \ldots (6)$$

which can be re-written as:

$$Y_{it} = a_i + X_{it}b + v_{it} \quad \ldots (7)$$

The intercept a_i is a country specific effect and can be estimated using the fixed effects method. The frontier intercept is represented by a (the highest level of efficiency) and each country's level of inefficiency is calculated as:

$$\hat{a} = \max(\hat{a}_i) \quad \text{and:}$$

$$\hat{u}_i = \hat{a} - \hat{a}_i$$

where the u_is are the country specific inefficiency terms. Technical efficiency is defined as:

$$TE_i = \frac{E[Y_{it} \mid u_{it}, X_{it}]}{E[Y_{it} \mid u_i = 0, X_{it}]} \quad \ldots (8)$$

Where Y_{it} is the output variable (either DALE or the composite index), u_i is inefficiency and the Xs are the independent variables.

Equation 8 is modified to account for minimum health status (M_{it}) – the level of health that would exist in each country had no health service. The resulting metric is referred to as overall efficiency (E_i):

$$E_i = \frac{E(Y_{it} \mid u_i, X_{it}) - M_{it}}{E(Y_{it} \mid u_i = 0, X_{it}) - M_{it}} \quad \ldots (9)$$

The composite minimum is similarly calculated using the total scores for fair financing and responsiveness distribution plus the weighted normalised DALE.

$$COMPOSITE_{i\min} = 37.5 + 25 * [\frac{(DALE_i - DALE_{\min})}{(DALE_{\max} - DALE_{\min})}] \quad \ldots (10)$$

The $DALE_{min}$ term and the minimum health term for the DALE model are estimated using a COLS procedure. Using data from 25 countries the health measure DALE was regressed on literacy rates in 1908. The model (subtracting the lowest residual from the COLS procedure) was used to predict the minimum level of health for each country for the years 1993-1997 if there were no health service.

To generate confidence intervals around the efficiency measures Evans et al (2000) and Tandon et al (2000) bootstrapped the efficiency measures.

Replication

The procedures described in the World Health Report 2000 were replicated to further examine the results. Table 5 below gives the coefficients reported from Evans et al. (2000), Tandon et al. (2000) and our replication results. Columns 1 and 2 refer to the DALE models and columns 3 and 4 to the modelling of the composite index.

Table 5 Replication of WHO results

Variable	WHO DALE	DALE replication	WHO COMP	COMP replication
Health Expenditure	0.009	0.009*	0.007	0.007*
Schooling	0.063	0.063	0.050	0.049*
Schooling squared	0.022	0.022	0.022	0.023*
Constant	3.813	3.813*	4.112	4.110*
Max (u)	0.213	0.209	0.173	0.170

Note: * significant using a 95% confidence interval

The coefficients in Table 5 show that our results are very similar to those reported by Evans et al (2000), and Tandon et al (2000). Their coefficients are not estimates but the averages of the bootstrapped coefficients. In our analysis we report only the estimated coefficients, since averaging bootstrapped coefficients introduces bias (Stata manuals 2001). Despite this the results from the two procedures are virtually identical. The Pearson correlation coefficients between the two sets of results are 1.00 for DALEs and 0.975 for the composite index. The Spearman rank correlation coefficients for the two variables are 0.9999 and 0.9867.

Re-evaluating estimation procedures

Evans et al (2000) criticise and disregard alternative methods for quantifying efficiency because they cannot differentiate between errors and true efficiency. They recommend the use of panel data because it can overcome this problem. Unfortunately the fixed effects estimator is no better at identifying efficiency estimates than any other method. The a_i term in equation 7 is an estimate of country level fixed effects or observable/unobservable heterogeneity. Any unobserved/observed time invariant factor relating to each country will be included in this estimate. To separate such time invariant effects and efficiency is still very difficult. If the time invariant factor is observable the sample can be stratified by the observable factor and the models re-estimated. Fixed-effects models cannot estimate time invariant variables so the use of dummy variables is unfeasible.

It is likely that OECD countries exhibit different features to non OECD countries and to test this hypothesis we estimated separate models for OECD and non OECD countries.

For the OECD and for the non-OECD countries R^2 are 0.14 and 0.56 respectively; that is, the explanatory power in the OECD model is poor and significantly worse than for non-OECD countries. Results for the non OECD countries are closer to the results for the entire sample and the maximum level of efficiency is also higher for the non OECD countries.

Table 6 Re-estimated models for OECD and non OECD countries

	DALES		Composite Index	
Variable	**OECD**	**Other**	**OECD**	**Other**
Health expenditure	0.016*	0.008*	0.014*	0.006*
Schooling	0.334*	0.075	0.131	0.058*
Schooling squared	-0.026	0.015	0.014	0.018
Constant	3.528*	3.767*	4.039*	4.075*
Max(u)	0.086	0.259	0.063	0.209
R^2				

The results from the full sample the R^2-squared is 0.66. The results from the full sample are clearly being driven by the non-OECD countries and this suggests that countries should be stratified and analysed separately. To reinforce this conclusion an application of the RESET test for misspecification finds that the OECD model is misspecified—the non OECD and full sample models are not. This suggests that the model used to estimate efficiency for the OECD should be different to that used for non OECD countries.

There are two other estimation problems with the WHO approach. First, the analysis has been carried out using fixed effects panel data models because the models estimated did not pass a Hausman test of specification. For estimating marginal effects the fixed effects model return unbiased (although inefficient) estimators as either N or T go to infinity. For the individual specific term the fixed effects estimator is inconsistent, regardless of the size of N (Verbeek, 2000). Such inconsistency casts doubt over results presented using the fixed effects approach.

The second problem is that the treatment of efficiency as time invariant. Khumbrakar and Lovell (1990) state that over long panels, or over any period in which structural change could have occurred, assuming time invariance is a very strong assumption. For many countries, such as those breaking away from the former USSR or those in periods of civil war such issues will be vital. Efficiency needs to be measured using time varying parametric panel data techniques as suggested by Cornwell et al. (1990) or nonparametric based panel data Malmquist indices as reviewed by Hollingsworth et al. (1999). Nonparametric techniques have another advantage in that they can estimate how the frontier technology is changing over time. The parametric methods used by the WHO ensure that one country is always efficient, Malmquist indices allow the potential production technology to vary; it may be that the efficiency frontier is moving further away from the least efficient countries.

To provide estimates of efficiency the WHO needs to consider stratifying the sample by observable characteristics. Unobservable country level characteristics will still corrupt the efficiency measurement. The fixed effects estimator is only useful in long panels—the estimated efficiency term is only consistent as T goes to infinity. The use of long panels require the use of time varying methods of efficiency measurement. The WHO should re-estimate the models using a variety of parametric and nonparametric methods and use the results for model and method validation. Then the results presented will be closer to unbiased and informative measures of efficiency.

1 Validity

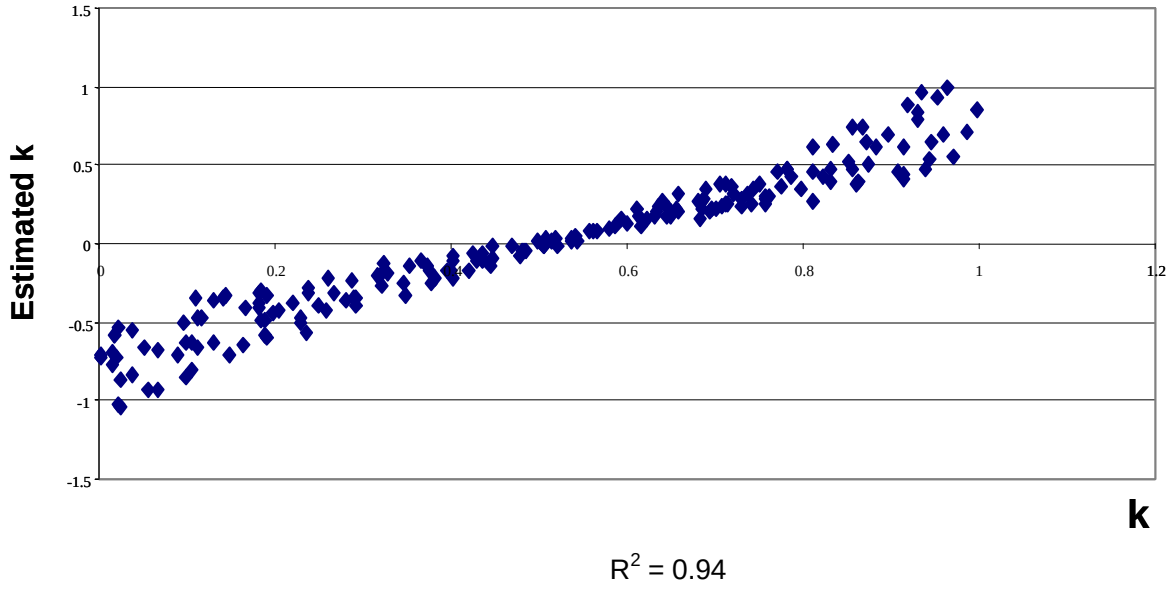
The ability of the WHO procedure to detect and rank efficiency cannot be determined by comparison with a gold standard as no such gold standard exists. Consequently a simulation exercise was carried out using equation 11 and the WHO data base.

$$Y_{it} = a + kb_1 H + b_2 Ed + b_3 Ed^2 \quad \ldots (11)$$

Y = Dales or composite Index
H = Health expenditure
Ed = Education

Actual values of health expenditure and education were used to generate a new set of values for each country's predicted DALEs and composition index. These differed from the actual predicted values as health expenditures were multiplied by a variable, k, which was allowed to vary randomly between zero and unity. The variable therefore represents the efficiency with which health expenditures are translated into DALEs or system improvement. The WHO procedure was then used to estimate the efficiency coefficient, k. The results are given in Table 7 and compared with the true value of k as shown in Figure 2. This reveals a very close correlation between the true and estimated values ($R^2 = 0.94$) and represents a confirmation of the validity of the WHO approach to the modelling of efficiency.

Figure 2 Efficiency coefficient, k versus estimated k



$R^2 = 0.94$

2 Omitted Variables

As noted earlier, the WHO procedure does not, as its authors assert, eliminate bias arising from omitted variables. The effect of this bias was tested in our simulation by eliminating the third variable in equation 11 (the square of education). Results are shown in the final column of Table 7 and compared with true efficiency, k, in Figure 3, this reveals a significant decline in the strength of the association between true and estimated k. In Figure 4 actual values were replaced by rank order for the top 100 countries (for each of five years). This results in the breakdown of the association between true and estimated efficiency. The correlation coefficient is 0.5323.

Figure 3 Efficiency coefficient k versus estimated k with an omitted variable

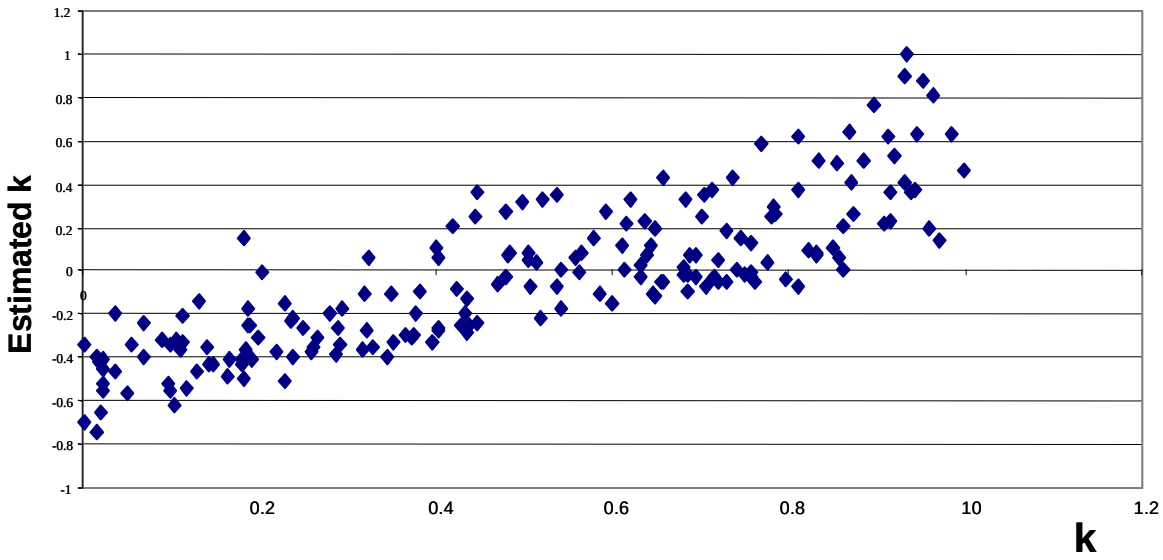


Figure 4 **Rank order of true efficiency (k) versus Rank order of estimated efficiency with omitted variable**

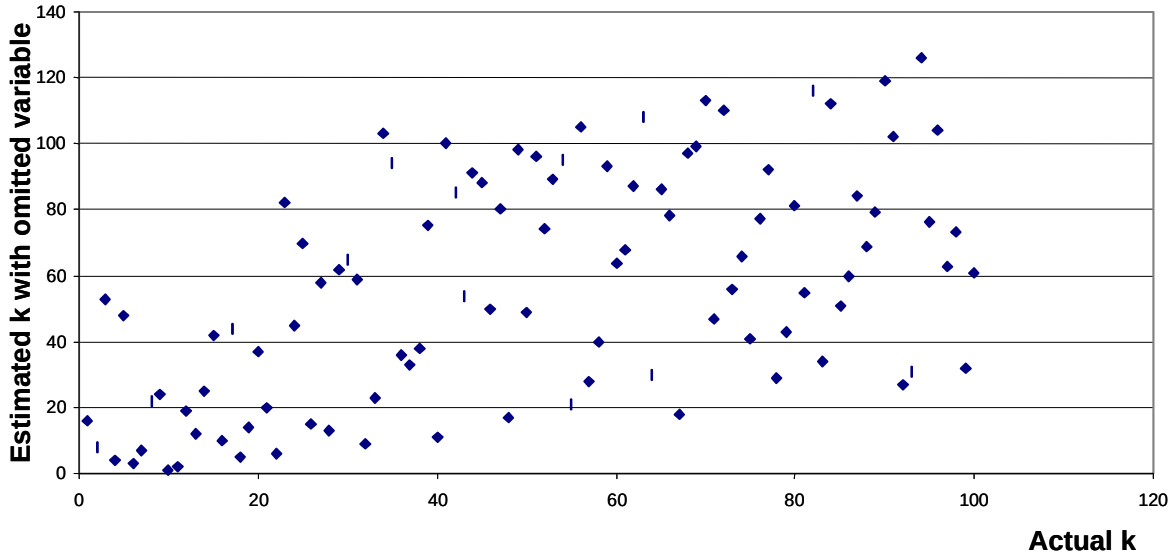


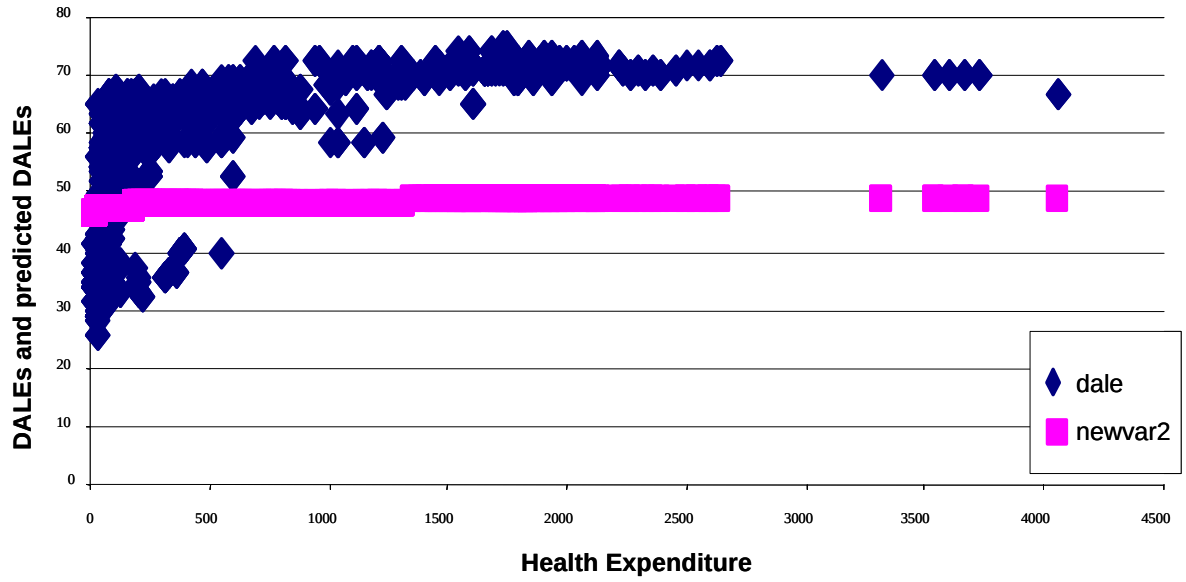
Table 7 Efficiency ranking: True versus two estimates using the WHO methodology

Country	True Efficiency		Estimates					
			Full information			Omitted variables		
	score	rank	score	rank	$R_1 - R_2$	score	rank	$R_1 - R_3$
	True	R1	R2			R3		
Saint Kitts and Nevis	0.999	1	0.856	5	4	0.462	16	-15
Belarus	0.985	2	0.715	10	8	0.627	8	-6
Cote d'Ivoire	0.97	3	0.55	19	16	0.143	53	-50
Switzerland	0.963	4	1	1	3	0.815	4	0
Botswana	0.96	5	0.69	12	7	0.193	48	-43
Canada	0.952	6	0.926	3	3	0.878	3	3
Tonga	0.945	7	0.647	14	7	0.632	7	0
Philippines	0.944	8	0.533	20	12	0.369	22	-14
Tuvalu	0.938	9	0.474	25	16	0.366	24	-15
United States of America	0.935	10	0.966	2	8	1	1	9
New Zealand	0.931	11	0.833	6	5	0.903	2	9
Singapore	0.93	12	0.788	7	12	0.407	19	-7
France	0.919	13	0.886	4	9	0.535	12	1
Turkmenistan	0.915	14	0.442	31	17	0.363	25	-11
Saint Lucia	0.914	15	0.62	16	1	0.231	42	-27
Niger	0.914	16	0.416	34	18	0.616	10	6
Guinea-Bissau	0.908	17	0.463	28	11	0.222	44	-27
Republic of Korea	0.897	18	0.699	11	7	0.765	5	13
Croatia	0.884	19	0.608	18	1	0.506	14	5
Jordan	0.874	20	0.51	22	2	0.257	37	-17
Antigua and Barbuda	0.872	21	0.652	13	8	0.407	20	1
Luxembourg	0.868	22	0.748	8	14	0.639	6	16
Malawi	0.862	23	0.395	36	13	0.003	82	-59
Benin	0.861	24	0.382	38	14	0.209	45	-21
Zimbabwe	0.857	25	0.476	23	2	0.057	70	-45
Austria	0.855	26	0.737	9	15	0.493	15	11
Iran	0.851	27	0.515	21	6	0.104	58	-31
Bahamas	0.834	28	0.637	15	13	0.507	13	15
Honduras	0.832	29	0.468	27	2	0.08	62	-33
Solomon Islands	0.831	30	0.402	35	5	0.074	65	-35
Bhutan	0.823	31	0.42	33	2	0.089	59	-28
Finland	0.812	32	0.613	17	15	0.623	9	23
Marshall Islands	0.812	33	0.455	30	3	0.368	23	10
Dem Republic of the Congo	0.811	34	0.273	54	20	-0.068	103	-69
Haiti	0.797	35	0.342	45	10	-0.035	94	-59
Average Error					8.3			21.0

3 Goodness of Fit

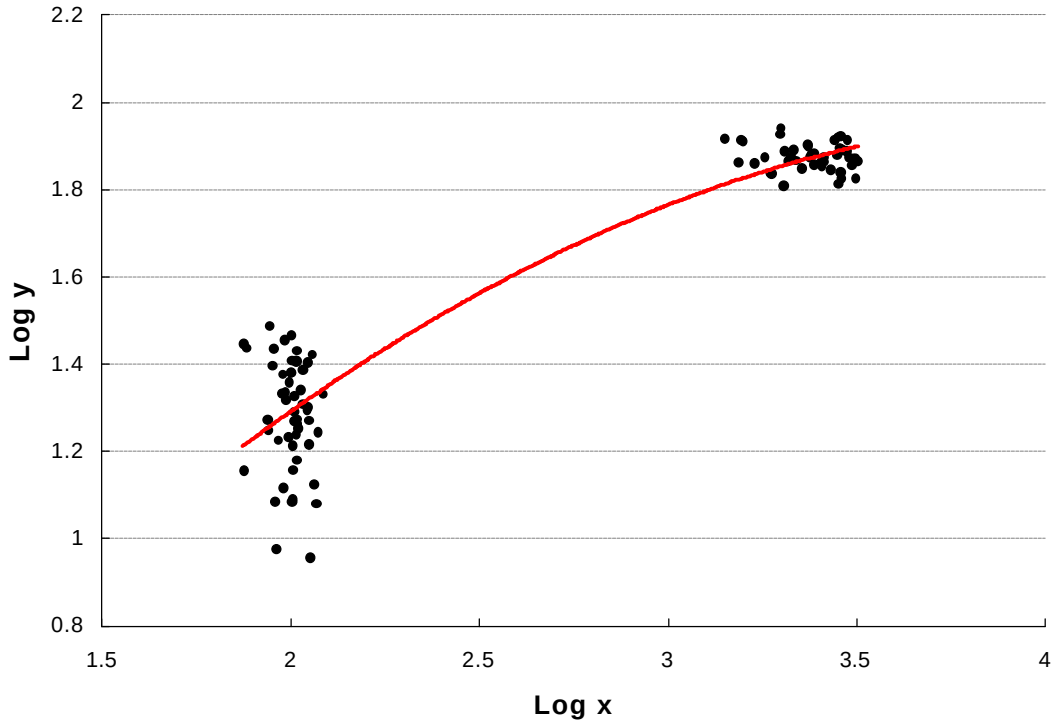
In Figure 5 DALEs are plotted against health expenditures for the full sample of countries used in the WHO study. Our replicated WHO procedure resulted in an apparently good fit between actual and predicted DALEs, with a Spearman correlation coefficient of 0.82. This does not, however, indicate that the goodness of fit is sufficient to obtain a reliable ranking of a country efficiency with respect to the production of DALEs. The reason for this is illustrated in a second simulation summarised in Figure 6. Two sets of data were constructed each with 50 'observations'. The dependent variable on the vertical axis was constrained, in the two sets. The first y value was a randomly simulated normal distribution with a mean of 20 and standard deviation (SD) of 5 and an x value with a mean of 100 and a SD of 10 with no correlation between x and y values. The second set had y values of 75 and SD of 5 and x values of 2450 and SD of 500, again with no constructed association between x and y values. The double log regression run on these data resulted in an adjusted R-squared value of 0.89 which is significantly higher than the corresponding R-squared for the data shown in Figure 6.

Figure 5 DALEs plotted against health expenditure: full sample



The reason for the high R-squared in Figure 6 is that the two sets of data are separated by a large numerical value for both the dependent and independent variables. The double log regression therefore has high explanatory power *between* the two sets of data and this dominates the regression results. By construction, there is no explanatory power within the datasets and, consequently, the regression line of best fit would be an inappropriate basis for the assertion that data had been standardised to take account of variation in the independent variable. To the contrary, the datasets have been constructed without reference to x and 'standardising' with respect to the artefactual relationship between dependent and independent variables may introduce bias into the analysis of country ranking. More specifically, Figure 6 reinforces the conclusion reached earlier that there are two types of behaviour, one for high and one for low values of the independent variable and that this casts doubt upon the validity of a combined analysis of high and low value countries: The artefactual relationship may simply impose a distorted relationship within both datasets.

Figure 6 Regression results from 2 sets of random data



Because of this possible interpretation of the data two hypothesis are tested below. The first is that the predicted value of DALEs and the composite index are poorly correlated with the true values in a sub-set of countries. Secondly the line of best fit derived from the full dataset may have significantly poorer explanatory power than a more restricted analysis.

To test the first hypothesis DALEs and the composite index were predicted for the sub-set of OECD countries with a DALE value exceeding 69. This figure was selected to maximise the homogeneity of the countries analysed. The comparison of the predicted and actual DALEs in Figure 7 reveals a random association. There is a positive relationship between predicted and actual (0.09) values of the composite index in Figure 8. However the R^2 is only 0.22

The remaining two analyses below illustrate the consequences of this poor association in the OECD countries. In Table 8 the predictive power of the WHO procedure in conjunction with the full database is compared with two alternative algorithms. The first is to assume that each countries' DALE is equal to the group average. The second is to predict DALE values using a simple one year double log regression equation. Table 8 reports the error arising from each of these three procedures, that is, the difference between true and predicted DALEs. The WHO procedure results in a significantly greater error than the naï ve assumption that DALE values are equal to the group mean. As shown in column 3 both the WHO procedure and the first alternative algorithm are out-performed by the use of a simple regression model. Identical conclusions may be drawn from Table 9 which gives the results of a similar analysis of the composite index.

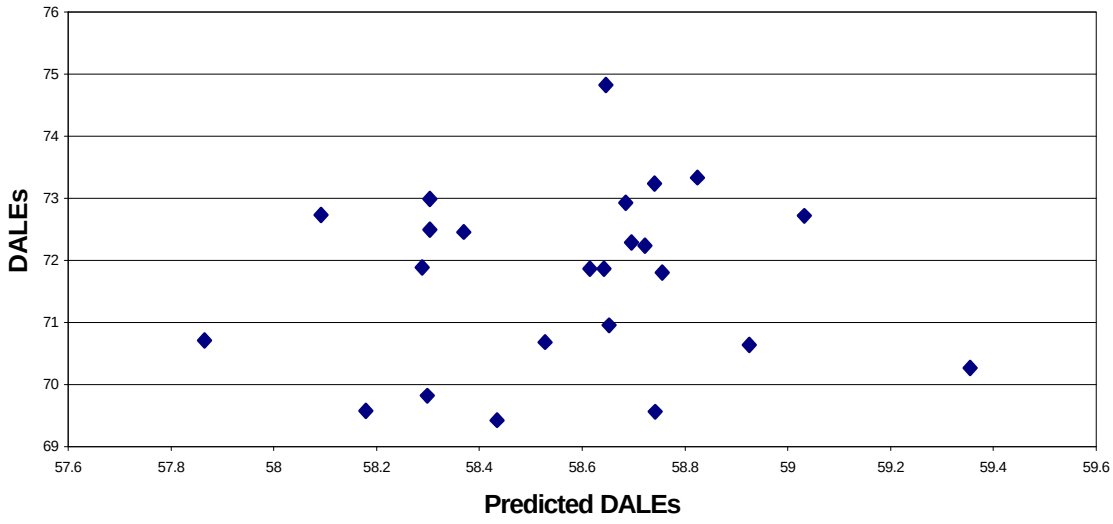
Table 8 Goodness and fit of estimates of DALEs in developed countries

	Absolute value of:		
Country	**Dales, D minus predicted D (WHO)**	**Dales minus Average D**	**Dales minus Predicted D OLS Reg**
Andorra	14.19	0.79	0.55
Austria	13.05	0.11	0.33
San Marino	14.09	0.76	0.39
Belgium	13.23	0.17	0.30
Canada	13.60	0.60	1.08
Switzerland	13.68	1.02	0.78
Germany	11.71	1.06	1.20
Denmark	10.82	2.13	1.60
Spain	14.68	1.29	1.07
Finland	12.15	1.02	0.60
France	14.52	1.64	0.96
United Kingdom	13.60	0.19	0.55
Greece	14.64	1.03	1.28
Ireland	11.52	1.88	1.74
Iceland	12.31	0.74	0.83
Italy	14.25	1.24	0.62
Japan	16.18	3.13	3.28
Malta	12.84	0.99	1.49
Netherlands	13.51	0.53	0.55
Norway	13.25	0.16	0.06
New Zealand	10.99	2.28	1.32
Portugal	11.39	2.13	2.59
Sweden	14.49	1.53	1.74
United States of America	10.91	1.43	1.49
Average	13.15	1.16	0.32

Table 9 Goodness of fit of estimates of the composite index in developed countries

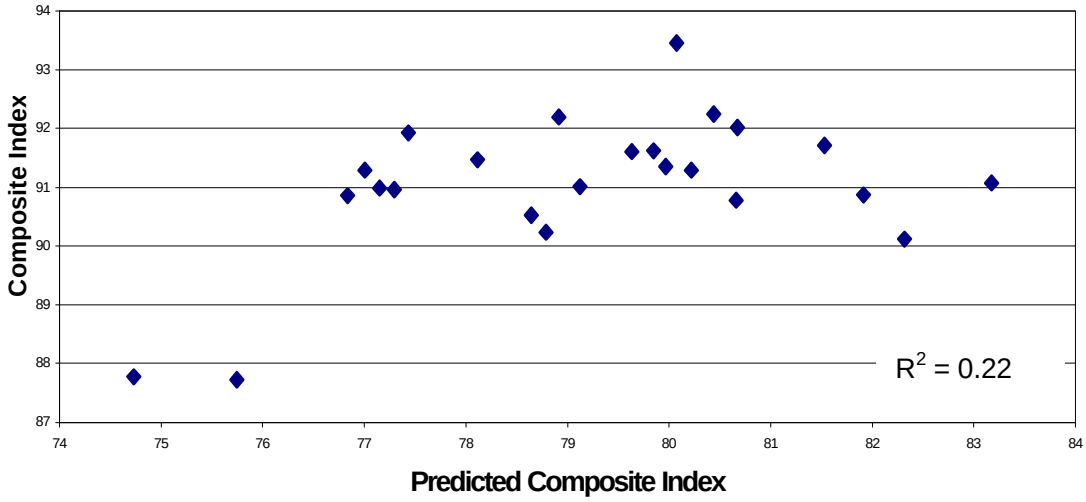
	Absolute value of:		
Country	**Composite, I, minus predicted I (WHO)**	**Composite minus Average I**	**Composite minus Predicted I OLS Reg**
Andorra	13.83	0.06	0.55
Austria	13.36	0.43	0.40
San Marino	14.03	0.18	0.40
Belgium	11.38	0.31	0.19
Canada	10.18	0.67	0.27
Switzerland	11.81	1.21	0.41
Germany	11.06	0.24	0.33
Denmark	8.95	0.18	0.70
Spain	13.66	0.08	0.50
Finland	10.12	0.26	0.34
France	14.49	0.89	0.83
United Kingdom	11.98	0.57	0.87
Greece	11.87	0.52	0.06
Ireland	11.44	0.81	0.40
Iceland	11.89	0.03	0.05
Italy	14.29	0.25	0.45
Japan	13.37	2.41	2.27
Malta	13.04	3.27	2.09
Netherlands	11.78	0.58	0.38
Norway	13.28	1.16	1.21
New Zealand	7.80	0.93	1.14
Portugal	11.98	3.32	2.45
Sweden	11.34	0.97	0.63
United States of America	7.90	0.03	1.92
Average	11.87	0.81	0.34

Figure 7 DALEs versus predicted DALEs (reduced sample[1])



Notes: [1] OECD countries with a DALE value ≥ 69

Figure 8 Composite index versus predicted composite index (reduced sample[1])



Notes: [1] OECD countries with a DALE value ≥ 69

6 Conclusions

Despite our criticisms the WHO study represents a landmark for the evaluation of health systems. It is the most sophisticated cross-national assessment of system performance to date. Further, it has defended a set of importance weights which dramatically states (or, as we ague, overstates) the importance of system objectives other than health outcome.

The present analysis demonstrates that this highly innovative study is seriously flawed in several respects. The choice of objectives is contestable and there is a particularly strong case for omitting the equity of financing from the list and replacing it with an index of access, both financial and geographical. The weights attached to the system objectives have not been validated. It is highly unlikely that a single set of weights or a single set of objectives can be obtained which are a valid reflection of the aspirations of every country in the world. Subsequent analyses may need to explore the option of using country specific values and modelling them in such a way that results may be comparable between nations.

As with the modelling of the quality of life, there has been very little consideration of the validity of the underlying model which combines the scores and weights of the different objectives into a single index of performance. It has been shown here that the ranking of nations may be sensitive to the choice of the combination model. Different models have different properties and the appropriate property for the present exercise needs to be given careful consideration.

Finally, results here suggest that the simultaneous inclusion of every country in the analysis will lead to invalid and unreliable results. While this difficulty is easily overcome—there are sufficient observations to carry out sensible sub-analyses—a more intractable problem may arise because of the paucity of data. Omitted variables may correlate with health expenditure or education to an unknown extent. Consequently model parameters may be distorted and country specific affects change to an unknown extent. It appears likely that this problem may have driven many of the WHO results.

Unlike Williams (2001a) we do not attempt to assess the overall costs and benefits that are likely to follow from the WHO study. Predictions of this type are hazardous and particularly, as Victor Fuchs once noted, when they are about the future.

References

Blendon, R., Kim, M. and Benson, J. 2001, The public versus the World Health Organisation on Health system performance, *Health Affairs*, 20, 3, 10-20.

Cornwell C, Schmidt, P. and Sickles, R.C. 1990, Production frontiers with cross sectional and time-series variation in efficiency levels. *Journal of Econometrics,* 46: 185-200.

Evans, D.B., Tandon, A., Murray, C.J.L. and Lauer, J.A. 2000, The Comparative Efficiency of National Health Systems in Producing Health: An Analysis of 191 Countries. Geneva, Switzerland World Health Organisation, (GPE Discussion Paper Series: No. 29).

Hollingsworth, B., Dawson, P., Maniadakis, N. 1999, Efficiency Measurement of Health Care: A review of non-parametric methods and applications. *Health Care Management Science,* 2(3): 161-172.

Khumbraker, S.C. and Lovell, 1990, C.A.K. Stochastic Frontier Analysis, Cambridge, Cambridge University Press.

Murray, C.J.L., Frenk, J., Evans, D. et al 2001, 'Science or marketing at WHO, A response to Williams, *Health Economics* , vol 10, no 4, pp 277-282.

Murray, C., Kawabata, K. & Valentine, N. 2001, People's experience versus people's expectations, *Health Affairs*, May June, 21-24.

Reidpath, D.D., Allotey, P., Kouame, A., Cummins, R.A. Social cultural and environmental contexts and the measurement of burden of disease: An exploratory study in the developed and developing world. Melbourne: Key Centre for Women's Health in Society, The University of Melbourne, 2001.

StataCorp. 2001, Stata Manuals Release 7. Stata Corporation.

Tandon, A., Murray, C.J.L., Lauer, J.A. and Evans, D.B. 2000, Measuring Overall Health System Performance for 191 Countries. Geneva, Switzerland World Health Organisation, (GPE Discussion Paper Series: No. 30).

Verbeek, M. 2000, *A Guide to Modern Econometrics*, John Wiley & Sons, Chichester, England.

Von Winterfeldt, D. and Edwards, W. 1993, *Decision analysis and Behavioural Research*, Cambridge University Press.

Wagstaff, A., Van Doorslaer, E. 1994, Equity in the finance of health care: Some international comparisons, *Journal of Health Economics*, vol 11, pp 361-388.

Williams, A. 2001a, 'Science or marketing at WHO? A commentary on "World Health Report 2000", *Health Economics*, vol 10 no 2, pp 93-100.

Williams, A. 2001b, 'Science of marketing at WHO? Rejoinder from Alan Wailliams, *Health Economics*, vol 10, no 4, pp 283-286.

World Health Organisation 2000, The World Health Report: *Health Systems: Improving Performance*, World Health Organisation: Geneva, Switzerland.