

## **Cross-national stability of disability weights: the European Disability Weights Project**

Michaël Schwarzinger <sup>1</sup>, Marlies EA Stouthard <sup>2</sup>, Kristina Burström <sup>3</sup>,  
Erik Nord <sup>4</sup> and The European Disability Weights Group <sup>5</sup>

<sup>1</sup>Department of Public Health, Hôpital Henri Mondor, AP -HP, Créteil, France ([schwarzi@univ-paris12.fr](mailto:schwarzi@univ-paris12.fr)),

<sup>2</sup>Institute of Social Medicine, AMC, Amsterdam, The Netherlands ([M.E.Stouthard@amc.uva.nl](mailto:M.E.Stouthard@amc.uva.nl)),

<sup>3</sup>Department of Public Health Sciences, Division of Social Medicine, Karolinska Institute, Stockholm, Sweden ([kristina.burstrom@smd.sll.se](mailto:kristina.burstrom@smd.sll.se)),

<sup>4</sup>National Institute of Public Health, Oslo, Norway ([erik.nord@folkehelsa.no](mailto:erik.nord@folkehelsa.no)),

<sup>5</sup>members listed at the end.

The scientific work underlying this paper was done at the centres that have participated in the European Disability Weights Project, in alphabetical order: National Institute of Public Health (formerly DICE), Copenhagen, Denmark; Health Services Management Centre and the Department of Public Health and Epidemiology, University of Birmingham, England; Department of Public Health, Hôpital Henri Mondor AP -HP, Créteil, France; Department of Public Health, Erasmus University Rotterdam, the Netherlands (project coordination); Institute of Social Medicine, Academic Medical Center, University of Amsterdam, the Netherlands; National Institute of Public Health, Oslo, Norway; National School of Public Health, Institute of Health 'Carlos III', Madrid, Spain; Department of Public Health Sciences, Karolinska Institute, Stockholm, Sweden

Address for reprints and correspondence:

Dr Michaël Schwarzinger, Department of Public Health, Hôpital Henri Mondor, 51 avenue du Maréchal de Lattre de Tassigny, 94010 Créteil, France; e-mail: [schwarzi@univ-paris12.fr](mailto:schwarzi@univ-paris12.fr)

This study was supported by a grant from the BIOMEDI II Programme of the European Union (project number BMH4 -98-3253)

## **Abstract**

**Background** Disability weights represent the relative severity of disease stages to be incorporated in summary measures of population health. The universality of disability weights in Western European countries was investigated with different valuation methods.

**Methods** Disability weights for fifteen disease stages were empirically elicited in panels of health care professionals or non-health care professionals with an academic background following a strictly standardised procedure. Three valuation methods were used: a visual analogue scale (VAS); the time trade-off technique (TTO); and the person trade-off technique (PTO). Agreement among England, France, the Netherlands, Spain, and Sweden on the three disability weights sets was analysed by means of an intraclass correlation coefficient (ICC) in the Generalisability Theory framework. Agreement among the two types of panels was similarly assessed.

**Results** A total of 232 participants were included. Similar rankings of disease stages across countries were found with all valuation methods. The ICC of country agreement on disability weights ranged from 0.56 [0.52-0.62] with PTO to 0.72 [0.70-0.74] with VAS and 0.72 [0.69-0.75] with TTO. The ICC of agreement between health care professionals and non-health care professionals ranged from 0.64 [0.58-0.68] with PTO to 0.73 [0.71-0.75] with VAS and 0.74 [0.72-0.77] with TTO.

**Conclusions** Overall the study supports universality of disability weights in Western European countries with VAS and TTO methods

which focus on individual preferences, but to a lesser extent with PTO method which focuses more on societal values in resource allocation.

**Key words** : cross -national comparison, outcome measures, valuation methods, Disability-Adjusted Life Years, Quality -Adjusted Life Years.

## Background

Summary measures of population health combine information on mortality and non-fatal health outcomes in order to represent the health of a particular population as a single measure. [1] They are used traditionally for comparative judgements of average levels of population health between populations and over time. Summary measures of population health were recently used with an explicit link to health resources allocation, e.g. disability-adjusted life expectancies (DALE) computed among other measures for the evaluation of the performance of health systems in the World Health Report 2000, [2] or disability-adjusted life years (DALY) for burden of disease estimates and cost-effectiveness analyses. [3-5]

All summary measures of population health are built on three critical inputs: mortality by age, sex and condition; epidemiological data on non-fatal health outcome by age, sex and condition; and valuations of health states (disability weights) that assess the relative severity of a year lived in a particular condition. Whereas mortality and epidemiological data may be seen as objective measures, even if scarcity and heterogeneity of data may question their accuracy, valuations of health states are undoubtedly subjective measures.

The lack of a gold standard for health state valuation has led to the development of various valuation methods. [6] The 1996 Global

Burden of Disease study (Global Burden of Disease) represented a milestone in the development of summary measures of population health as it established a single set of several hundred disability weights relating to 107 conditions using the same evaluation method. [7,8] The choice of the specific values of an international panel of about ten health experts was supported by high correlations of their disability weights for 22 hypothetical indicator health states with those of eight panels from National Burden of Disease teams or a World Health Organization (WHO) conference on Burden of Disease methods. [9] Since then the assumption of cross-national stability of disability weights has been further supported by studies using similar valuation protocols, [10,11] whereas agreement between possible informants in health showed contradictory results. [12-15]

One of the primary objectives of the European Disability Weights (EDW) project was to assess the cross-national stability of valuations of health states when elicited using different methods. [16] In the EDW study, a visual analogue scale (VAS) measured the severity of health states relative to the anchoring endpoints of the scale (worst and best imaginable health states). The time trade-off technique (TTO) measured the extent to which respondents would be willing to give up an amount of lifetime to avoid a hypothetical condition and be in full health, and the person trade-off technique (PTO) elicited directly the health decision maker's trade-off between severity of illness, the size of the health gain and the number of people helped. [6] Hypothetical

health states were valued in panels of two possible informants in health, i.e. health care professionals and the general public with an academic background. We report here on the agreement of disability weights from five Western European countries: England, France, the Netherlands, Spain, and Sweden, using VAS, TTO and PTO.

## Methods

The valuation of health states in the participating Western European countries followed a standardised protocol with back and forth translation from English for all valuation materials. [16] Key points of the valuation procedure were fixed to limit construct -irrelevant variance:

1. The scenarios to be valued were presented consistently in the form of a disease label, a brief clinical description of the disease stage, and a generic health state profile (EQ -5D extended with a cognitive dimension); [17-19]
2. three valuation methods were used: visual analogue scale (VAS), time trade-off (TTO), and person trade -off (PTO);
3. a structured protocol which allowed for discussion and deliberation was followed in all panel sessions;
4. panel sessions in each country were led by a trained facilitator from that country.

Two sources of variance in the valuation of health states were retained in our interrater reliability study of each valuation method: 1) the country; 2) the type of panel according to medical background of participants.

## PANEL PARTICIPANTS

At least two panels of health care professionals (almost all medical doctors) and two panels of 'non -health care professionals' each consisting of around ten participants, were planned for each country.

Incentives to participate were given to health care professionals (medical doctors were paid in England and Spain, and received Medical Continuing Formation credits in the Netherlands) whereas the 'non-health care professionals' were recruited generally on local academic webs (they were also paid a small amount in England). Panels took place in five European countries: England, France, the Netherlands, Spain, and Sweden, between March and September 2000.

#### DISEASE STAGES SELECTION AND DESCRIPTION

A list of diseases accounting for almost 80% of years of life lost due to premature mortality and 80% of years lived with disability in the Established Market Economies Region (including all Western European countries) was extracted from the Global Burden of Disease study. [9] These diseases were then selected to cover:

1. The main chapters from the ninth version of the WHO International Classification of Diseases,
2. different dimensions of disability,
3. very mild to very severe health states.

External health care professionals and public health experts participated in both the subdivision of selected diseases into homogenous disease stages with respect to functional status, treatment and prognosis, and the elaboration of a brief clinical description for each disease stage. [16]

Fifteen disease stages were selected for the panel valuation procedure: the stages selected covered the full range of disease severity, from the common cold to a final year of an unspecified fatal disease. All selected disease stages were described on a separate sheet with the name of the disease, the position of the selected disease stage among the other stages, a brief clinical description and a health state profile according to a health-related quality of life instrument EQ-5D extended with a cognitive dimension, i.e. EQ-5D+C.[17-19] EQ-5D+C has six dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression, cognition) each of three levels of severity (no problem, some problems, extreme problems). Consistency of profiles was checked across disease stages within diseases and across diseases. Figure 1 shows an example of a separate sheet displaying a disease stage description.

## VALUATION METHODS

Pilot studies conducted in participating countries tested innovative societal valuation methods [20] after the GBD societal valuation protocol had been criticized at an early stage of the project on ethical grounds.[21,22] Agreement on the valuation protocol was reached by consensus, and the three valuation methods are described below in the order of their use in panels. In VAS, all fifteen disease stages were valued, in PTO and TTO then nine chronic disease stages were valued.

In the self-administered VAS participants were asked to consider the consequences of living with the disease stage for one year. The disease stages were first ranked by decreasing severity, and then scored on a vertical thermometer graded from 0 (the worst imaginable health state) to 100 (the best imaginable health state) considering the consequences of living with the disease stage for one year. The best and the worst disease stages were scored first.

In the PTO panel participants played the role of decision-makers in their country prioritising between two preventive programmes. Several assumptions about the programmes were made explicit in the panel sessions:

- Prevention means the reduction of occurrence in two to four years; programmes are of the same costs and otherwise equal (e.g. age, sex, socio-economic status of groups);
- there are people of various ages in both programmes;
- loss of production for society and burden on family or caretakers were to be disregarded in decisions.

The PTO session began with the following example: "Programme A prevents the occurrence of a rapidly fatal disease in 100 people in your country in 2-4 years time. The identity of these people is unknown. With the programme they will live in normal health for a normal lifetime. Programme B prevents the occurrence of severe vision disorder in a number of people in your country in 2-4 year time. The identity of these people is unknown. With the programme they will

avoid the state and live in normal health for a normal life time.”

Participants determined the number of people in programme B where they were indifferent between the two programmes with the aid of a visual prop that displayed a stepwise procedure increasing the numbers in programme B (100 -200-1,000-10,000 etc). Indifference numbers lower than 100 were also allowed. [21] After the example, participants had to prioritise between the prevention of a rapidly fatal disease and quadriplegia, and then between each of the eight chronic disease stages on the one hand, and quadriplegia on the other. Quadriplegia was thus used as an anchoring state, linking the valuation of chronic states to death. After initial personal valuation, discussion was structured among panel participants by the facilitator who ensured that participants understood and were aware of the implications of their choices. Following discussion panel members had the opportunity to change their responses if they so wished.

In TTO, panel participants had to imagine someone like themselves in full health, and choose between living their remaining 10 years of life in the chronic disease stage or less time in full health. The number of years where the panel participants were indifferent was found using a “ping-pong” procedure, but participants were allowed not to trade off any years of life. [23] The facilitator again ensured that panel participants fully understood the task.

Finally, participants had the opportunity to reconsider their responses after discussion in the panel and were encouraged to compare individual rankings of all disease stages for all three valuation methods and make changes to any responses if they so wished.

## STATISTICAL ANALYSES

While TTO and PTO responses yield disability weights for life years directly, the VAS responses in this study do not, since in the VAS exercise, states of illness were not valued relative to the state of being dead. But as we shall see, VAS raw scores may be transformed into disability weights by means of mathematical relationships to TTO or PTO values. For simplicity of exposition we refer to all valuations as disability weights (DW) in the following. Statistical analyses were performed on the final figures recorded after panel deliberations. Following the GBD study convention, DW were valued to unity for death (or the worst imaginable health state in VAS) and zero for full health (or the best imaginable health state in VAS), and were computed as follows for the three evaluation methods:

VAS:  $DW = 1 - \text{score}/100$ ;

TTO:  $DW = 1 - \text{years}/10$ ;

PTO:  $DW_{\text{quadriplegia}} = 100/\text{number}_{\text{quadriplegia}}$ ,

and  $DW_{\text{disease stage}} = 100/\text{number}_{\text{disease stage}} * DW_{\text{quadriplegia}}$  (quadriplegia was used as an anchoring state instead of death). In case of  $DW > 1$  due to the chained procedure of PTO valuation, i.e. when participants

valued quadriplegia worse than death, DW was truncated to 1 with report of the number of records.

Rankings of the fifteen disease stages based on mean VAS scores were compared across countries with Spearman rank correlation. In a random-effect model, variance components of disability weight were estimated for the random-effects identified in our study:

1. Disease stages (n=15 for VAS and n=9 for TTO and PTO),
2. subjects nested within a type of panel (n=232),
3. types of panel (health care or non-health care professional) nested within country (n=2),
4. countries (n=5),
5. crossed effects of disease stages and other random-effects.

Maximum likelihood estimates of the variance components were used to compute the proportion of total variance accounted for by each random-effect. For our interrater reliability study, two intraclass correlation coefficients were computed according to Generalisability Theory, a specific application of analysis of variance. [24] A first intraclass correlation coefficient measured the agreement between countries on disability weights deducted from VAS, TTO, and PTO:

$$\frac{(\sigma^2_{\text{disease}} + \sigma^2_{\text{subject(panel)}} + \sigma^2_{\text{panel(country)}} + \sigma^2_{\text{disease*panel}})}{(\sigma^2_{\text{disease}} + \sigma^2_{\text{subject(panel)}} + \sigma^2_{\text{panel(country)}} + \sigma^2_{\text{disease*panel}} + \sigma^2_{\text{country}} + \sigma^2_{\text{disease*country}} + \sigma^2_{\text{residual}})}$$

The numerator includes the variance components of all random-effects on disability weight other than country-related effects and the residual term, which are added in the denominator. The closer to unity the

intraclass correlation coefficient, the better the agreement of countries on disease stages' valuations. With a comparable design, a second intraclass correlation coefficient was computed to measure the agreement between the two types of panel for all valuation methods. Non parametric bootstrap resampling technique was used to compute 95% confidence intervals, [25] since the complex design of our interrater reliability study did not allow simple computations. [26] One hundred independent random samples were resampled from individual data depending on country and panel type. Significance was attributed at the 5% level, and data were analysed with SAS 8.0 (SAS Institute, Cary NC).

No general statement about the desired level of the reliability coefficient of a test can be made, because the purpose for which the test is used must always be taken into account. [24] When tests intended for important decisions at the individual level, e.g. admission for/discontinuation of a clinical treatment, a reliability coefficient  $\geq 0.90$  may be considered as "good". When tests intended for less important decisions at the individual level, e.g. evaluation of treatment outcome, a reliability coefficient  $\geq 0.80$  may be considered as "good". In our particular case where valuation methods intended for research at the group level, a reliability coefficient  $\geq 0.70$  may be considered as "good",  $\geq 0.60$  and  $< 0.70$  as sufficient, and  $< 0.60$  as insufficient. [27,28]

## Results

A total of 232 participants of five countries: England, France, the Netherlands, Spain, and Sweden, were included in 13 panels of health care professionals and 10 panels of non-health care professionals. Overall, 60% of subjects were females and mean age was 40.4 (std 15.2) years with significant differences in mean age between countries as shown in Table 1. Health care professionals included 84% medical doctors. Health care professionals differed significantly from non-health care professionals in age (48.9 (std 14.1) vs. 32.4 (std 11.3), respectively) and gender (48% female vs. 71%, respectively). Difficulties were reported with at least one disease stage description in panels of either non-health care professionals (8 out of 10) or health care professionals (10 among 13). Similar proportions of panels of non-health care professionals and health care professionals also reported difficulties with prognosis in TTO (74%), the initial example of PTO (35%), and the PTO valuation method overall (17%).

Table 2 shows that disease stages were ranked similarly between countries according to mean VAS scores. The average of the ten Spearman's  $\rho$  between countries two by two was 0.95 with minimum 0.89 (Spain/Sweden). Independence of ranks was rejected at  $p < 0.0001$  in all measures. Similar results were found with mean TTO (average Spearman's  $\rho$  of 0.96) and median PTO (average Spearman's  $\rho$  of 0.92) for the nine chronic disease stages.

Table 3 shows that disease stages accounted for more than 60% of total variance of disability weights from VAS and TTO, whereas this decreased to 36.7% with disability weights from PTO. The contribution of systematic differences between participants in valuation of the nine disease stages doubled from VAS (5.4%) to TTO (9.8%), and PTO (16.4%).

Country-related effects accounted for 1.9% of total variance with VAS, increasing to 3% with TTO and 10.8% with PTO. The agreement between countries fell from 0.72 with VAS and TTO to 0.56 with PTO. Panel type-related effects accounted for 1.3% of total variance with VAS, 1.1% with TTO and 3% with PTO. The agreement between health care professional panels and non-health care professional panels decreased from 0.73 with VAS and 0.74 with TTO to 0.64 with PTO. (Mean VAS disability weights overall and per country are published elsewhere. [16] Mean TTO and PTO disability weights overall and per country may be obtained from the authors).

## Discussion

A total of 232 participants from five western European countries valued disease stages in health care professional and non-health care professional panels. Overall we found a very similar ranking of disease stages across countries whichever valuation method was used. This confirms previous findings based on the valuation of seventeen health conditions, either with VAS through individual interviews of about fifteen key informants in fourteen countries of all World Regions,[29] or with PTO in the GBD study and recent refinements.[9,10] Similar rankings of disability weights are not enough, however, to judge the appropriateness of a universal disability weights set used at a cardinal level in summary measures of population health.

We found that intraclass correlation coefficients measuring agreement between countries were good with VAS and TTO. At first glance, this finding may appear at odds with cross-national comparisons of disability weights focusing on disease conditions separately. Other studies eliciting values for EQ-5D health states with TTO from the general public in United Kingdom and Spain, [30] or in United Kingdom and Japan, [31] showed a high positive correlation of values between countries, but significant differences in values were found for a number of health states. Whereas a great variability in the valuation of health states is observed within countries, [32] the previous approach does not allow one to disentangle systematic differences in

valuation between subjects and between countries. As shown within the Generalisability Theory framework, the subject effect accounted for more variance of disability weights than the country effect for all valuation methods.

In the case of the PTO method, the intraclass correlation coefficient measuring agreement between countries was insufficient. PTO elicited directly health decision-makers' trade-offs between preventive programmes and attempted to get societal preferences between disease stages. Whether respondents actually took a societal view in PTO questions (as opposed to an individual view in VAS and TTO) was not checked, e.g. through follow-up interviews, and is certainly worthy of further research. The PTO method demonstrated a dramatic increase in the systematic effects related to subjects and countries as compared to VAS and TTO. This could be related to different prioritisation behaviours across European people. [33]

We found that the agreement between people of similar academic background but of different medical background was good with VAS and TTO, and sufficient with PTO. This confirms results of an earlier study in the Netherlands. [11] However, agreement between possible informants in health, i.e. individuals in health states, patients' families, healthcare professionals and the general public, showed contradictory results in Western countries using the TTO method, [12-14] or VAS. [15] In the absence of clear agreement between possible

informants on disability weights, the US Panel on Cost-effectiveness in Health and Medicine stated that the general public preferences on health conditions should be used to inform health care resources allocation.[34] Further research should assess differences in valuations between representative samples of the general public and the more educated groups used in our study. This is even more true in the context of developing countries where educated people could share Western values at odds with those of the general public. [35]

The design of four valuation methods may limit comparison with other studies. Framing and anchoring effects were likely to have been present with all three valuation methods. Among other framing effects, VAS scores are prone to sequencing effects (i.e. the worst and the best disease stages were scored first in our study), and the range of health states considered. [23] The anchoring of the TTO in a ten-year timeframe was fixed for all participants to ensure comparability of results. However, TTO disability weights for most disease stages decreased with the age of participants, with older people less willing to give up an amount of life time to avoid a health condition than younger people (data not shown). This may have been of particular relevance in the cross-national comparisons focusing on disease conditions separately since age patterns differed between participating countries. Pilot studies resulted in a “chained PTO” to limit the “rule of rescue” encapsulated by the technique, i.e. valuations take into account the initial disease severity of the programmes’ recipients in

particular in lifesaving programmes. Quadriplegia as the anchoring state had various consequences at the country level. Firstly, 43% of participants thought that the prevention of quadriplegia should receive a higher priority to that of a life-saving program. This finding was not related to age of participant but differed significantly across countries, from 23% in Sweden to 64% in the Netherlands. Secondly, the proportion of participants feeling that the prevention of some disease stages should receive an equal priority to that of quadriplegia varied greatly across countries. For instance more than 40% of participants found the prioritisation of severe depression equal to that of quadriplegia in England and the Netherlands, whereas this proportion fell under 7% for France, Spain and Sweden.

Another limitation of our study is related to the validity of our valuation protocol. Despite great care being taken to ensure the face validity of disease stages, at least one disease stage was questioned in a majority of panels of both health care and non-health care professionals. For instance, discrepancies between the brief clinical description of spinal cord injuries resulting in quadriplegia and its generic health state profile were often noticed. Difficulties were also encountered with TTO and PTO methods in spite of the deliberative panel process led by a facilitator and the high level of education of the participants. If we are to collect values from the general public as recommended then we need to put more effort into ensuring valuation methods are understood as intended by respondents. Alternatively,

data from the relatively simple VAS -method, which does not yield disability weights for life years directly, could be used as a basis for estimating more complicated trade -offs through mathematical relationships established in comparative studies.[36-38]

## Conclusions

Our study supports universality of disability weights in Western European countries with VAS and TTO methods, but to a lesser extent with the PTO method designed here. The assumption of universality of disability weights requires evidence across World regions, as defined by countries' location and possibly by similarities in mortality patterns and cost structures. Our study showed that even within a relatively homogenous and wealthy World region this assumption may not hold true when a societal perspective is taken into account, i.e. when the summary measure of population health is intended to inform health care resource allocation.

However, uncertainty surrounding disability weights may be well considered small when compared to the lack of epidemiological data in many areas of the World, either to compare summary measure of population health across countries in the World Health Report 2000,[39] or to perform cost-effectiveness analyses. [40] In the European Disability Weights study, cross-national comparisons of burden of disease, as measured by disability-adjusted life years (DALYs), showed that differences between European countries per valuation method were negligible in comparison to differences in epidemiological estimates. [16]

## **Competing interests: none declared.**

### **Authors' contribution**

M.L. Essink-Bot, L.J. Gunning-Schepers, P.J. van der Maas, M.E.A. Stouthard and G.J. Bonsel were responsible for the original grant application. P.J. van der Maas, L.J. Gunning-Schepers, J. Pereira, I. Durand-Zaleski, J. Raftery, F. Diderichsen and F. Kamper-Jørgensen were members of the Steering Committee of the European Disability Weights project. P.J. van der Maas acted as the Project Coordinator. The sub-studies on valuation and burden of disease estimation were ultimately designed in the discussions of the European Disability Weights group, to which all members listed below contributed. The respective country teams did empirical data collection in each participating country. Writing Committee, consisting M. Schwarzingler, M.E.A. Stouthard, K. Burström and E. Nord drafted the manuscript of this paper. All members of the European Disability Weights group read and approved the final manuscript.

Members of the European Disability Weights group:

Finn Kamper-Jørgensen, Ulla Christensen, Kim Moesgaard Iburg, (National Institute of Public Health, Copenhagen, Denmark); James Raftery, Claire Packer, Lisa Gold, Suzanne Robinson (from October 1999) (University of Birmingham, England); Isabelle Durand-Zaleski, Michael Schwarzingler (Public Health, Hôpital Henri Mondor, AP-HP, Paris, France); Louise Gunning-Schepers, Gouke Bonsel, Clara Moerman (from 01.01.2000), Marlies Stouthard (Academic Medical Center, University of Amsterdam, The Netherlands); Paul van der Maas (project coordinator), Marie-Louise Essink-Bot (Dept. of Public Health, Erasmus University Rotterdam, The Netherlands); Joaquin Pereira, Ana Baylin (until 01.01.1999), Eduardo

Fernandez Zincke (National School of Public Health, Madrid, Spain); Finn Diderichsen, Kristina Burström, Rickard Ljung (Karolinska Institute, Stockholm, Sweden).

### **Acknowledgements**

We thank Joshua A. Salomon (WHO, Geneva, Switzerland) for general advice, and Bruno Falissard (Department of Public Health, Hôpital Paul Brousse, France) for help and advice with the generalisability study.

## References

1. Field MJ, Gold GM, eds. Summarizing population health: directions for the development and application of population metrics. Washington DC: National Academy Press; 1998.
2. WHO. The World Health Report 2000. Health Systems: Improving Performance. Geneva: World Health Organization; 2000.
3. Politi C, Carrin G, Evans D, Kuzoe FA, Cattand PD. Cost-effectiveness analysis of alternative treatments of African gambiense trypanosomiasis in Uganda. *Health Econ* 1995;4:273-87.
4. Goodman CA, Coleman PG, Mills AJ. Changing the first line drug for malaria treatment -- cost-effectiveness analysis with highly uncertain inter-temporal trade-offs. *Health Econ* 2001;10:731-49.
5. Marseille E, Hofmann PB, Kahn JG. HIV prevention before HAART in sub-Saharan Africa. *Lancet* 2002;359:1851-6.
6. Nord E. Methods for quality adjustment of life years. *Soc Sci Med* 1992;34:559-69.
7. Murray CJ, Lopez AD. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *Lancet* 1997;349:1436-42.
8. Murray CJ, Lopez AD. Regional patterns of disability-free life expectancy and disability-adjusted life expectancy: Global Burden of Disease Study. *Lancet* 1997;349:1347-52.
9. Murray CJ: Rethinking DALYs. In *The Global Burden of Disease. Vol 1: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Edited by Murray CJ, Lopez AD. Cambridge: Harvard University Press; 1996:1-99.
10. Murray CJ, Lopez AD. Progress and directions in refining the global burden of disease approach: a response to Williams. *Health Econ* 2000;9:69-82.
11. Soutthard ME, Essink-Bot ML, Bonsel GJ, on behalf of the Dutch Disability Weights Group. Disability weights for diseases: a modified protocol and results for a Western European region. *Eur J Public Health* 2000;10:24-30.
12. Dolan P. Whose preferences count? *Med Decis Making* 1999;19:482-6.
13. Zethraeus N, Johannesson M. A comparison of patient and social tariff values derived from the time trade-off method. *Health Econ* 1999;8:541-5.
14. Ubel PA, Loewenstein G, Hershey J, Baron J, Mohr T, Asch DA, Jepson C. Do nonpatients underestimate the quality of life associated with chronic health conditions because of a focusing illusion? *Med Decis Making* 2001;21:190-9.

15. Suarez -Almazor ME, Conner -Spady B, Kendall CJ, Russell AS, Skeith K. Lack of congruence in the ratings of patients' health status by patients and their physicians. *Med Decis Making* 2001; 21: 113 -21.
16. Essink -Bot ML, Pereira J, Packer C, Schwarzing M, Burstrom K. Cross -national comparability of burden of disease estimates: the European Disability Weights Project. *Bull World Health Organ* 2002; 80: 644 -52.
17. EuroQol -- A new facility for the measurement of health -related quality of life. The EuroQol Group. *Health Policy* 1990; 16: 199 -208.
18. Brooks R. EuroQol: the current state of play. *Health Policy* 1996; 37: 53 -72.
19. Krabbe PF, Stouthard ME, Essink -Bot ML, Bonsel GJ. The effect of adding a cognitive dimension to the EuroQol multiattribute health -status classification system. *J Clin Epidemiol* 1999; 52: 293 -301.
20. Robinson S, Gold L, Moesgaard I, Burg K, and the European Disability Weights group. The development of the PTO for estimating disability weights. International Health Economics Association 2001, York, United Kingdom: Abstract 45A014.
21. Arnesen T, Nord E. The value of DALY life: problems with ethics and validity of disability adjusted life years [Erratum in *BMJ* 2000; 320: 1398]. *BMJ* 1999; 319: 1423 -5.
22. Essink -Bot ML, Stouthard M, Bonsel G, Gunning -Shepers L, van der Maas P. The problems with disability weights. *eBMJ* 1999; 2 December.
23. Drummond MF, O'Brien BJ, Stoddart GL, Torrance GW: *Methods for the Economic Evaluation of Health Care Programmes*. Second ed. Oxford: Oxford University Press; 1997.
24. Streiner DL, Norman GR: *Health measurement scales: a practical guide to their development and use*. Second ed. Oxford: Oxford University Press; 1995.
25. Efron B, Tibshirani RJ: *An introduction to the bootstrap*: Chapman and Hall; 1994.
26. Zou KH, McDermott MP. Higher -moment approaches to approximate interval estimation for a certain intraclass correlation coefficient. *Stat Med* 1999; 18: 2051 -61.
27. Bartram D. The development of international guidelines on test use: the International Test Commission Project. *International Journal of Testing* 2001; 1: 33 -53.
28. Evers A. The revised Dutch ratings system for test quality. *International Journal of Testing* 2001; 1: 155 -82.
29. Ustun TB, Rehm J, Chatterji S, Saxena S, Trotter R, Room R, Bickenbach J. Multiple -informant ranking of the disabling effects of different health conditions in 14 countries. WHO/NIH Joint Project CAR Study Group. *Lancet* 1999; 354: 111 -5.

30. Badia X, Roset M, Herdman M, Kind P. A comparison of United Kingdom and Spanish general population time trade-off values for EQ-5D health states. *Med Decis Making* 2001;21:7-16.
31. Tsuchiya A, Ikeda S, Ikegami N, Nishimura S, Sakai I, Fukuda T, Hamashima C, Hisashige A, Tamura M. Estimating an EQ-5D population value set: the case of Japan. *Health Econ* 2002;11:341-53.
32. Sculpher M, Gafni A. Recognizing diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. *Health Econ* 2001;10:317-24.
33. Mossialos E, King D. Citizens and rationing: analysis of a European survey. *Health Policy* 1999;49:75-135.
34. Gold MR, Siegel JE, Russell LB, Weinstein MC: *Cost-effectiveness in Health and Medicine*. New York: Oxford University Press; 1996.
35. Jelsma J, Chivaura VG, Mhundwa K, De Weerd W, de Cock P. The global burden of disease disability weights. *Lancet* 2000;355:2079-80.
36. Pereira J, Schwarzing M, Moesgaard I, Burg K, Nord E, Stouthard M, and the European Disability Weights group. Constructing a common European disability weight scale for measuring burden of disease: comparison among three methods in the European Disability Weights project. Brussels: European Commission, 2001.
37. Torrance GW. Social preferences for health states: an empirical evaluation of three measurement techniques. *Socio-Econ Plan Sci* 1977;10:129-36.
38. Krabbe PF, Essink-Bot ML, Bonse IGJ. The comparability and reliability of five health-state evaluation methods. *Soc Sci Med* 1997;45:1641-52.
39. Almeida C, Braveman P, Gold MR, Szwarcwald CL, Ribeiro JM, Miglionico A, Millar JS, Porto S, Costa NR, Rubio VO, et al. Methodological concerns and recommendations on policy consequences of the World Health Report 2000. *Lancet* 2001;357:1692-7.
40. Walker D, Fox-Rushby JA. Economic evaluation of communicable disease interventions in developing countries: a critical review of the published literature. *Health Econ* 2000;9:681-98.

## Figure legends

Figure 1: Example of a disease stage description as presented for valuation.

Legend: Disease stage description included a disease label (dementia); the disease stage to be valued (marked by the arrow), a textual description (in bold) and a generic description of the functional health status (EQ -5D+C; 3 severity levels per attribute; one dot indicates the 'moderate' level).

## Tables

Table 1: Description of panel participants by country

	England	France	the Netherlands	Spain	Sweden	p-value
Number of participants ( $n =$ )	50	46	50	47	39	
Age, mean (std)	41.6(10.6)	44.9(24.9)	40.0(11.1)	32.8(7.3)	43.3(13.5)	0.002
Sex, % females	54	57	60	68	59	0.69
Healthcare professionals, %	58	48	44	51	41	0.52

Table 2: Ranking of fifteen disease stages according to mean visual analog scales scores in 5 Western European countries

Disease states	All (n=232)	England (n=50)	France (n=46)	Netherlands (n=50)	Spain (n=47)	Sweden (n=39)
Common cold	1	1	1	1	1	1
Vision disorder (mild/moderate)	2	2	2	2	2	2
Breast cancer (disease-free stage without sequelae)	3	3	3	3	6	3
Chronic low back pain	4	5	5	4	4	4
Uncomplicated diabetes	5	4	4	5	3	5
Mild dementia	6	6	7	8	7	7
Severe asthma	7	9	6	7	5	9
Colorectal cancer (diagnosis and primary therapy)	8	8	8	9	10	6
AIDS (minor symptoms and HAART)	9	7	10	6	9	10
Severe stable angina (NYHA3)	10	10	9	11	8	11
Acute myocardial infarction	11	11	11	10	11	8
Stroke (moderate permanent impairments)	12	12	12	12	12	12
Severe depression	13	13	13	13	13	13
Quadriplegia	14	14	14	15	14	14
Final year of unspecified fatal disease	15	15	15	14	15	15

Acquired Immune Deficiency Syndrome (AIDS); Highly Active Anti-Retroviral Therapy (HAART); New-York Heart Association (NYHA)

Table 3: Variance components analysis and intraclass correlation coefficient of country and panel agreements with three evaluation methods

Random-effects	Visual analogues scale (15 disease stages)	Visual analogues scale (9 disease stages)	Time trade-off (9 disease stages)	Person trade-off (9 disease stages)
	Proportion of total variance			
Subject nested within panel*	4.3%	5.4%	9.8%	16.4%
Panel nested within country	0.0%	0.0%	0.7%	2.7%
Country	0.7%	0.1%	1.2%	3.9%
Disease	66.5%	65.4%	61.2%	36.7%
Disease effect crossed with:				
-panel nested within country	0.8%	1.3%	0.4%	0.3%
-country	1.9%	1.8%	1.8%	6.9%
Residual	25.8%	26.0%	24.9%	33.2%
	Intraclass correlation coefficient of country agreement**			
	0.72(0.65-0.77)	0.72(0.70-0.74)	0.72(0.69-0.75)	0.56(0.52-0.62)
	Intraclass correlation coefficient of panel agreement**			
	0.73(0.67-0.79)	0.73(0.71-0.75)	0.74(0.72-0.77)	0.64(0.58-0.68)

\*Two types of panel either with healthcare professionals or not

\*\*95% bootstrap confidence interval estimated on 100 independent random samples dependent from country and panel type of the 232 individuals





Figure 1: Example of disease staged description as presented for valuation.

**Dementia**



**Mild dementia**

Moderate dementia

Severe dementia

***Patient with mild loss of recent memory and some problems in planning and organising daily activities, aware of the deterioration in cognitive functioning capable of living independently***

- No problems in walking about
- No problems with washing or dressing self
- Some problems with performing usual activities (e.g. work, study, housework, family or leisure activities)
- No pain or discomfort
- Moderately anxious or depressed
- Some problems in cognitive functioning (e.g. memory, learning ability, concentration, comprehension)