Unitat d'Història Econòmica
UHE Working Paper 2012_09

Beyond GDP: Methodological and measurement issues in redefining "wealth"

Giuseppe Munda

Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, Edifici B, 08193, Bellaterra (Cerdanyola), Spain

E-mail: Giuseppe.Munda@uab.cat

23/11/2012

Universitat Autònoma de Barcelona

Giuseppe Munda, 2012
Beyond GDP: Methodological and measurement issues in redefining "wealth"
UHE Working Paper 2012_09
http://www.h-economica.uab.es/wps/2012_09.pdf

Unitat d'Història Econòmica
Departament d'Economia i Història Econòmica
Edifici B, Campus UAB
08193 Cerdanyola del Vallès, Spain
Tel: (+34) 935811203
http://www.h-economica.uab.es

UAB
Universitat Autònoma
de Barcelona

Beyond GDP: Methodological and measurement issues in redefining "wealth"

Giuseppe Munda

Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona, Edifici B, 08193, Bellaterra (Cerdanyola), Spain

E-mail: Giuseppe.Munda@uab.cat

Abstract: In the last decades; a growing stock of literature has been devoted to the criticism of GDP as an indicator of societal wealth. A relevant question is: what are the perspectives to build, on the existing knowledge and consensus, alternative measures of prosperity? A starting point may be to connect well-being research agenda with the sustainability one. However, there is no doubt that there is a lot of complexity and fuzziness inherent in multidimensional concepts such as sustainability and well-being. This article analyses the theoretical foundations and the empirical validity of some multidimensional technical tools that can be used for well-being evaluation and assessment. Of course one should not forget that policy conclusions derived through any mathematical model depend also on the conceptual framework used, i.e. which representation of reality (and thus which societal values and interests) has been considered.

Key words: Well-Being, Sustainability, Multi-Criteria Evaluation, Composite Indicators, Complexity

Códigos JEL: C43, C44, D63, E01, Q01

1. Introduction

The debate on the misuse of Gross Domestic Product (GDP) as an indicator of societal wealth is almost as old as GDP itself (for a recent overview, see e.g. van den Bergh, 2009). In recent years, a growing stock of literature has been written about concepts such as *quality of life* and *well-being* above all, after the influential Stiglitz, Sen and Fitoussi (2009) report which proposed the use of the concept of well-being as a proxy for measuring societal prosperity. This debate has also invested the OECD and the European Commission, which devoted a number of recent conferences to the issue of well-being or happiness in the framework of the "Measuring Progress" framework.

A relevant question is hence the following: what are the perspectives to build, on the existing knowledge and consensus, alternative measures of prosperity? A starting point may be to connect well-being research agenda with the sustainability one. This allows us to draw upon results already established in the literature and widely accepted by the

political and scientific communities; as stated by Arrow *et al.,* (2012), *"… Much of the literature on sustainable development has taken human wellbeing to be the object to be sustained."* One shared non-controversial result of the sustainability literature is that sustainability is a multidimensional concept, which should at least include economic, social, environmental and institutional dimensions. The next point to deal with is whether there is a multidimensional measurement framework able to cope with all these issues simultaneously.

The objective of "green accounting" is to furnish information on the sustainability of the economy, but there is no established doctrine on how the different, and at times even contradictory, variables and indicators are to be combined so that they are made immediately useful for policy making in the same way that GDP or other macroeconomic statistics are (see e.g. Barbier and Markandya, 1990; Chichilnisky, 1996; Faucheux and O'Connor, 1998; Funtowicz *et al*, 1999; Horwarth and Norgaard, 1990, 1992; Musu and Siniscalco, 1996; Pearce *et al*., 1996). It is precisely the existence of multiple dimensions, along with that of their multiple interrelationships, that explains the difficult task of analysing sustainability. Taken as a whole, there is no generally accepted way of framing the analysis within which a study of sustainability should be performed.

A point of scientific controversy present in the debate on sustainability measure is the use of monetary or physical indexes. Examples of monetary indexes are Daly and Cobb (1989) ISEW (Index of Sustainable Economic Welfare), the so-called El Serafy approach (Yusuf *et al*., 1989), Pearce and Atkinson (1993) Weak Sustainability Index and Genuine Savings (Atkinson and Hamilton, 2007). Examples of physical indexes are HANPP (Human Appropriation of Net Primary Production) (Vitousek *et al,* 1986), the Ecological Footprint (Wackernagel and Rees, 1995) and MIPS (Material Input Per unit of Service) (Schmidt-Bleek, 1994). Although these approaches may look different, they all have one common characteristic:

1. These indexes are based on the hypothesis that a common measurement rod needs to be established for aggregation purposes (e.g. variables expressing money, energy, space, and so on). This creates the need to make very strong assumptions on conversion coefficients to be used and on the acceptable degree of compensability (e.g. until which point better economic performance may be justified at the expense of environmental destruction or social exclusion?). The mathematical aggregation convention behind an index thus needs an explicit and well-thought formulation.

2. These indexes are somewhat confusing if one wishes to derive policy suggestions. For example, by looking at ISEW, we could know that a country has a worse sustainability performance than the one pictured by the standard GDP, but so what? Since ISEW is so aggregated, it does not provide us with any clear information on the cause of this bad performance, and it is thus useless for policy-making (while conventional GDP is at least giving some information on the economic performance). The same applies to the Ecological Footprint, which sometimes can even give misleading policy suggestions; for example, given that diet is used, it would imply that a more energy intensive agriculture might reduce the Ecological Footprint of e.g. a city, but in reality – if CO2 or energy consumption are factored in - its environmental performance would be much worse.

3. All these approaches belong to the more general family of composite indicators and as a consequence, some assumptions used for their construction are common to them all. Notwithstanding the limits just mentioned, a conclusion that we can borrow from the sustainability literature is that composite indicators could be an adequate approach to measure overall performance regarding multidimensional concepts such as sustainability or well-being *provided the temptation of a single metric is resisted.*

In the next Section, I will defend the idea that incommensurability, as a theoretical foundation, and multi-criteria analysis, as a possible practical framework, are the basic measurement principles for assessing multidimensional concepts such as well-being. In Section 3, sensitivity and uncertainty analysis is presented as an essential tool to increase the transparency of results and to help the framing of the debate around the use of a conceptual framework. Section 4 deals with the issue of how to implement well-being evidence based policy. Finally, in Section 5, some conclusions are drawn.

2. Measuring Multidimensional Concepts

2.1 Why? Complexity and the Incommensurability Principle

The world is characterised by deep *complexity*. This obvious observation has important implications on the manner in which policy problems are represented and policy-making is framed. To take any particular dimension as the true, real or total picture amounts to *reductionism*, whether physical or sociological (Giampietro *et al*, 2006). As a consequence, any attempt to fit the real world in a closed model leads to a simplification, which is violence to the description of reality. One should note that the construction of a descriptive model of a real-world system depends on very strong

assumptions about (1) the *purpose* of this construction, e.g. to evaluate well-being or sustainability (2) the *scale* of analysis, e.g. a city, a region or a country and (3) the set of *dimensions, objectives* and *indicators* used for the evaluation process. A reductionist approach for building a descriptive model can be defined as the use of just one measurable indicator (e.g. GDP per capita), one dimension (e.g. economic), one scale of analysis (e.g. region), one objective (e.g. the maximisation of economic efficiency) and one time horizon.

The previous discussion can be summarized by using the philosophical concept of *incommensurability* (Chang, 1997; Rabinowicz, 2012). From a philosophical perspective, it is possible to distinguish between the concepts of *strong comparability* (there exists a single comparative term by which all different actions can be ranked) implying *strong commensurability* (a common measure of the various consequences of an action based on a cardinal scale of measurement) or *weak commensurability* (a common measure based on an ordinal scale of measurement), and *weak comparability* (irreducible value conflict is unavoidable but compatible with rational choice employing, for example, multi-criteria evaluation) (Martinez-Alier *et al.,* 1998; Munda, 2004; O'Neill, 1993).

In terms of formal logic, the difference between strong and weak comparability, and one defence of weak comparability, can be expressed in terms of Geach's distinction between *attributive and predicative adjectives* (Geach, 1956). In Geach's own words: *"There are familiar examples of what I call attributive adjectives. Big and small are attributive;* x *is a big flea does not split up into* x *is a flea and* x *is big, nor* x *is a small elephant into* x *is an elephant and* x *is small; for if these analyses were legitimate, a simple argument would show that a big flea is a big animal and a small elephant is a small animal. Again, the sort of adjective that the mediaevals called alienans is attributive;* x *is a forged banknote does not split up into* x *is a banknote and* x *is forged, nor* x *is the putative father of* y *into* x *is the father of* y *and* x *is putative. On the other hand, in the phrase a red book red is a predicative adjective in my sense, although not grammatically so, for is a red book logically splits up into is a book and is red.*
I can now state my first thesis about good and evil: good and bad are always attributive, not predicative, adjectives" (Geach, 1956, p. 32).

Although Geach's arguments were developed in the context of moral philosophy, they have an extraordinary explicative power for evaluation and assessment too. In fact, evaluation is all about an object ***a*** being declared better, worse or equal than another

object ***b***. Now by developing further Geagh's logic, it is possible to prove easily that strong comparability (and then commensurability) is a very weak theoretical foundation for evaluation tools, when multidimensionality is considered. In fact, now the question is: when commensurability is logically possible and correct? The distinction between attributive and predicative adjectives gives us a clear answer.

Let us consider the basic example of apples and oranges, we all learn at primary school. Normally we are thought that we cannot sum up them unless we find a common unit of measurement, i.e. their price or the fact that they both belong to the set of fruit. In summary the search for commensurability always imply to look for a more general category (set) that can contain all the characteristics of the objects we wish to compare.
Let us consider the following statements:
a) *"X is an old car, all cars are means of transport, and therefore X is an old mean of transport";*
b) *"X is a good car, all cars are means of transport, and therefore X is a good mean of transport ";*

I believe that everybody would agree on the validity of statement a), but very a few would accept statement b) as a correct way of reasoning. Being good or bad depends on the notion of quality used, which depends on the use connected to the object to be evaluated. For example, it is hard to defend that a car is a good mean of transport to travel inside a city's historical centre. This discussion can be generalised as follows: commensurability is correct only if the adjectives used are predicative ones. An adjective *A* is predicative if it passes the two following logical tests (Martinez-Alier *et al.,* 1998):

(1) *if x is AY, then x is A and x is Y;*
(2) *if x is AY and all Y's are Z's, then x is AZ.*

Adjectives that fail such tests are attributive. Geach claims that *"good"* is an attributive adjective. In many of its uses it clearly fails (2): *"X is a good car, all cars are means of transport, and therefore X is a good mean of transport"* is an invalid argument.

At this point a question arises: is then the search for the *"common rod of measurement"* (such as money, energy or space) a non-sense? The answer is simple: one measurement rod makes sense if it is connected with one objective only; if a multiplicity of objectives has to be considered, to compress all dimensions into only one is fully

wrong. For example, money values are worth to be used if they are connected to one objective and one institution only, i.e. economic efficiency and markets. They fail to incorporate other objectives and values. The same argument applies to e.g. ecological footprint measures that fail to consider economic scarcity and human preferences obviously. The economic value is different from e.g. ethical, environmental or artistic-cultural values.

A gastronomic example may clarify this issue. In choosing my diet I can decide that my objective is to minimise the content of calories and of course I can use Kcal as a common measurement rod correctly. If other objectives are considered too, e.g. to maximise taste or to minimise cost, the reductionism of using Kcal only is not consistent with the existence of two or three different objectives. In conclusion, weak comparability implies *incommensurability* i.e. there is an irreducible value conflict when deciding what common comparative term should be used to rank alternative actions. It is in terms of such descriptions that well-being evaluation and assessment takes place. According to Stiglitz, Sen and Fitoussi (2009, pp. 14, 15) *"To define what well-being means a multidimensional definition has to be used. Based on academic research and a number of concrete initiatives developed around the world, the Commission has identified the following key dimensions that should be taken into account. At least in principle, these dimensions should be considered simultaneously:*
i. Material living standards (income, consumption and wealth);
ii. Health;
iii. Education;
iv. Personal activities including work
v. Political voice and governance;
vi. Social connections and relationships;
vii. Environment (present and future conditions);
viii. Insecurity, of an economic as well as a physical nature."

In a multidimensional framework, a country is not evaluated as good or bad as such, but rather, in relation to different descriptions. A country could have, at one and the same time, a *"good income"* and a *"bad environment",* a *"high level of health"* and a *"bad governance".* The use of these value terms in such contexts is attributive, not predicative.

The basic idea of multi-criteria evaluation is that in evaluation problems, we have first to establish objectives, i.e. the direction of the desired changes of the world (e.g.

maximise economic performance, minimise environmental impact, minimise social exclusion, etc.) and then find useful practical indicators (called criteria) which measure if the options considered are consistent with the objectives chosen (Figueira *et al.*, 2005; Munda, 2008; Nijkamp *et al,* 1990; Roy, 1996). Since in general, objectives are in conflict, multi-criteria mathematical aggregation rules look for so-called *compromise solutions.* In summary, incommensurability does not imply incomparability; on the contrary incommensurability is the only rational way to compare various objects under different methodological assumptions than maximisation or optimisation (Arrow, 1997; Sen, 1997; 2000; Sen and Williams, 1982). Weak comparability can be implemented by using multi-criteria evaluation.

The *discrete multi-criterion problem* can be described in the following way: *A* is a finite set of *N* feasible actions (or alternatives); *M* is the number of different points of view or evaluation criteria g_m *m=1, 2, ... , M* considered relevant in a policy problem, where the action ***a*** is evaluated to be better than action ***b*** (both belonging to the set *A*) according to the *m-th* point of view if $g_m(a)>g_m(b)$, *W* is a set of criterion weights

$W=\{w_m\}$,m=1,2,..., M, with $\sum_{m=1}^{M} w_m = 1$, which can be importance coefficients or trade-offs. *It is evident that the discrete multicriterion problem and the aggregation of individual indicators to build a composite are completely equivalent problems.* In synthesis, the information contained in the impact matrix is:

- *Intensity of preference* (when quantitative criterion/indicator scores are present).
- *Number* of criteria/individual indicators in favour of a given object (country, region, city, etc.) to be ranked.
- *Weight* attached to each single criterion/individual indicator.
- *Relationship* (i.e. relative ordering) of each single object with all the other objects to be ranked.

Combinations of this information generate different aggregation conventions, i.e. mathematical manipulation rules of the available information to arrive at a preference structure generating a ranking. The aggregation of several criteria/individual indicators implies taking a position on the fundamental issue of compensability. *Compensability* refers to the existence of trade-offs, i.e. the possibility of offsetting a disadvantage on some indicators by a sufficiently large advantage on another indicator, whereas smaller advantages would not do the same. Thus a preference relation is non-compensatory if no trade-off occurs and is compensatory otherwise. The use of weights with intensity of

preference originates compensatory aggregation methods and gives the meaning of trade-offs to the weights. On the contrary, the use of weights with ordinal indicator scores originates non-compensatory aggregation procedures and gives the weights the meaning of importance coefficients (Bouyssou and Vansnick, 1986; Keeney and Raiffa, 1976; Podinovskii, 1994; Roberts, 1979).

2.2 How? Mathematical Aggregation Rules

The proliferation of composite indicators in recent decades is a clear symptom of the increasing quantification of policy-making (see e.g. Saltelli 2007) (the so-called *evidence based policy*) and their operational relevance in economic, social and environmental statistics in general (see Banerjee 2005, Cherchye et al. 2007, Cox and others 1992, Cribari-Neto et al 1999, Griliches 1990, Lovell et al. 1995, McGuckin et al. 2007, Srinivasan 2004, Williams and Siddique 2008 and Wilson and Jones 2002, among others). All the major international organizations, such as the OECD, the EU, the World Economic Forum, and the International Monetary Fund, are producing composite indicators in a wide variety of fields (Nardo *et al.,* 2008).

From a formal point of view, a composite indicator is an aggregate of all dimensions, objectives, individual indicators and variables used for its construction (Munda and Nardo, 2009). This implies that what defines a composite indicator is the set of properties underlying its mathematical aggregation convention. Although various functional forms for the underlying aggregation rules of a composite indicator have been developed in the literature, in the standard practice, a composite indicator is very often constructed by using a weighted linear aggregation rule applied to a set of variables. Let us then check which axiomatic conditions govern the applicability of a linear aggregation rule; an essential condition is mutual preferential independence. On this respect, the following theorem holds:

Theorem 1: Given the set of individual indicators G, a subset of indicators Y is *preferentially independent* of $Y^C=Q$ (the complement of Y) only if any conditional preference among elements of Y, holding all elements of Q fixed, remain the same, regardless of the levels at which Q are held. The indicators $g_1, g_2,..., g_m$ are *mutually preferentially independent* if every subset Y of these indicators is preferentially independent of its complementary set of indicators. (Ting, 1971). This means that an additive aggregation function permits the assessment of the marginal contribution of each indicator separately (as a consequence of the preferential independence

condition). The marginal contribution of each indicator can then be added together to yield a total score.

One should note that in operational terms preference independence implies that in constructing a well-being composite indicator, we have to assume that individual indicators such as GDP and urban waste or unemployment rate have no relationship or if environmental dimensions are involved, the use of a linear aggregation function implies that among the different aspects of an ecosystem there are not phenomena of synergy or conflict. Summarising, we may conclude that the assumption of preferential independence is essential for the application of a linear aggregation rule. Unfortunately, it is rarely tested whether preferential independence applies to a given set of indicators prior to aggregating the indicators into a composite indicator, although this assumption has very strong consequences on the results and their interpretation.

Let us now look at another important implication of the use of linear aggregation rule, i.e. the meaning of weights. The common practice in constructing composite indicators is well synthesised in an OECD report, where it is clearly stated: *"Greater weight should be given to components which are considered to be more significant in the context of the particular composite indicator"* (OECD, 2003, p. 10). This concept of weights, from a theoretical point of view, could be related to the symmetrical importance, that is "*… if we have two non-equal numbers to construct a vector in* R^2*, then it is preferable to place the greatest number in the position corresponding to the most important criterion*." (Podinovskii, 1994, p. 241).

Let us further explain how the concept of symmetrical importance is related to the linear aggregation rule. Suppose that country *a* is evaluated according to some indicators values $(x_1(a),...,x_m(a))$. Then the substitution rate at *a*, of the value x_j with respect to the value x_r (taken as a reference) is the amount $S_{jr}(a)$ such that, country *b* whose evaluations are: $x_l(a) = x_l(b), \forall l \neq j,r$; $x_j(b) = x_j(a) - 1$; and $x_r(b) = x_r(a) + S_{jr}(a)$ is indifferent to country *a*. Therefore, $S_{jr}(a)$ is the amount which must be added to the value $x_r(a)$ (reference) in order to compensate the loss of one unit on value $x_j(a)$. Consider now a composite indicator $Y(x_1, x_2,..., x_m)$ and suppose that two countries have equal composite indicator scores. Let

$z(a) = (x_1(a), x_2(a), ..., x_m(a))$ and $z(b) = (x_1(b), x_2(b), ..., x_m(b))$, then as a first approximation one has:

$$0 = Y(z_b) - Y(z_a) = \sum_{i=1}^{m} \left(\frac{\partial Y}{\partial x_i} \right)_{z_a} \left(x_i(b) - x_i(a) \right) = -\left(\frac{\partial Y}{\partial x_j} \right)_{z_a} + S_{jr}(a) \left(\frac{\partial Y}{\partial x_r} \right)_{z_a}$$

which is equivalent to $$S_{jr}(a) = \frac{\left(\frac{\partial Y}{\partial x_j} \right)_{z_a}}{\left(\frac{\partial Y}{\partial x_r} \right)_{z_a}} \quad (1)$$

When the function Y is a weighted average of normalised indicators, i.e.

$$Y(x_1, x_2, ..., x_m) = \sum_{i=1}^{m} w_i x_i \quad (2)$$

then from expression (2) one obtains:

$$S_{jr}(a) = \frac{w_j}{w_r} = const. \quad (3)$$

As a consequence, substitution rates are directly estimated by the weights (Vincke, 1992). This implies a compensatory logic. ‘Compensability’ here refers to the existence of trade-offs, i.e. the possibility of offsetting very low values on several indicators by very high values on just few indicators. Therefore, the use of weights in combination with intensity of preference within a linear aggregation rule originates compensatory aggregation conventions and gives the meaning of trade-offs to the weights. Consequently, there exists an inconsistency between the way weights are used in practice and their theoretical meaning. If weights are to be interpreted as ‘importance coefficients’ (along the lines of ‘symmetrical importance’ of indicators, e.g. place the greatest weight to the most important dimension) then non-compensatory aggregation rules are more appropriate for the construction of composite indicators (Roberts, 1979; Bouyssou, 1986; Bouyssou and Vansnick, 1986; Vansnick, 1986).

We may then conclude that the use of non-linear/non-compensatory aggregation rules to construct composite indicators is compulsory for reasons of theoretical consistency when weights with the meaning of importance coefficients are used or when the

assumption of preferential independence does not hold. Moreover, in standard linear composite indicators, compensability among the different individual indicators is always assumed; this implies complete substitutability among the various components considered. For example, in a hypothetical sustainability index, economic growth can always substitute any environmental destruction or inside e.g., the environmental dimension, clean air can compensate for a loss of potable water. From a normative point of view, such a complete compensability might not be desirable (Markandya and Pedroso-Galinato, 2007, Munda, 1997). A search for alternative mathematical aggregation rules is then needed.

Vansnick (1990) showed that the two main approaches in multi-criteria decision theory i.e., the compensatory and non-compensatory ones can be directly derived from the seminal work of Borda (1784) and Condorcet (1785). The debate on the relative merits of Borda and Condorcet consistent voting rules is a very old one. Indeed according to McLean (1990), these rules were already known in the medieval age, when Ramon Lull (1235-1315) proposed a Condorcet method and Nicolaus Cusanus (1401-1464) proposed a Borda method. In 1986 Kenneth Arrow and Hervé Raynaud published a very influential book titled *"Social choice and multicriterion decision-making",* where the formal analogies between the discrete multi-criterion problem and the social choice one are deeply analyzed. This book is based on the assumption that, in the case where all criteria have ordinal impact scores, if one considers the evaluation criteria as voters, a multi-criteria impact matrix and a voting matrix are identical. As a consequence all results of social choice also apply to multi-criteria decision theory fully (when no intensity of preference is used) and then to the construction of composite indicators too.

A first topic to start with is Arrow's impossibility theorem (Arrow, 1963). A legitimate question arises: does this paradoxical result apply to the general discrete aggregation problem too? Arrow and Raynaud (1986, pp. 17-23) answer this question. Let us assume that a mathematical aggregation convention to arrive at a total ranking of all objects needs at least to satisfy three axioms[1]:

Axiom 1: Unrestricted Domain. The values that can be taken by the indicators are unrestricted and the mathematical aggregation convention must respect unanimity.

[1] The original Arrow's impossibility theorem (Arrow, 1963) is slightly different, above all with respect to the independence of irrelevant alternatives axiom. In the social choice literature formulation, it is called the *axiom of binary independence*, i.e. the social ranking of each pair of alternatives depends only on the preferences of each voter on that specific pair of alternatives. The ranking of any other alternative is irrelevant for this social ranking. Indeed in the version proposed by Arrow and Raynaud (1986) the axiom of independence of irrelevant alternatives is closer to the definition given by Chernoff (1954), which is derived from Nash's bargaining theory. For a deep discussion on the independence of irrelevant alternatives axiom and its various definitions see e.g. Ray (1973) and Bordes and Tideman (1991).

Axiom 2: Independence of irrelevant alternatives. The ranking of the objects (alternatives) in *A* depends only on the objects (alternatives) belonging to *A*. *"This means that it is of no importance for the decision if you have forgotten in the application of the method some (poorly ranked) alternatives: …. The complete set of alternatives is always very large and only a relatively small subset can be identified. It is thus essential that the result of the method on a small set of alternatives not vary if forgotten alternatives are taken into consideration"* (Arrow and Raynaud, 1986, p. 19).

Axiom 3: Positive Responsiveness. The degree of preference between two objects ***a*** and ***b*** is a strictly increasing function of the number of indicators (or weights) that rank ***a*** before ***b***.

The following paradoxical result then applies: the only ranking respecting all these axioms must coincide with the ranking supplied by one of the indicators taken into consideration. A consequence of this theorem is that *no perfect mathematical aggregation convention can exist.* "Reasonable" ranking procedures must then be found. In the framework of composite indicators, this consequence implies two questions: Is it possible to find a ranking algorithm consistent with a set of desirable *properties*[2] ? And on the reverse, is it possible to assure that no essential property is lost? At this point, the question arises: in the framework of composite indicators, can we choose between Borda scoring methods and Condorcet consistent aggregation rules on some theoretical and/or practical grounds?

The following conclusions can be drawn (see Munda, 2012a). Scoring methods present the advantage of always selecting one final solution thus their degree of decisiveness is very high. However, one has to accept that a scoring method always implies to transform (arbitrarily) an original ordinal scale of measurement into a quantitative one, and this implies to always have a compensatory aggregation rule. Compensability, which is based on the concept of intensity of preference, causes a high probability of preference reversal phenomena. Weights should always be in the form of trade-offs. Monotonicity sometimes is lost and neutrality can be relaxed. A strong argument in favour of a Borda scoring rule is that transitivity of the preference relation is never weakened, thus the assumption of individual rationality always applies.

[2] Often this search for clear properties characterizing an algorithm is indicated as the axiomatic approach. However, one should note that properties or assumptions are NOT axioms. As perfectly synthesized by Saari (2006, p. 110) *"Many, if not most, results in this area are merely properties that happen to uniquely identify a particular procedure. But unless these properties can be used to construct, or be identified with all properties of the procedure (such as in the development of utility functions in the individual choice literature), they are not building blocks and they most surely are not axioms: they are properties that just happen to identify but not characterize, a procedure. As an example, the two properties (1) Finnish-American heritage (2) a particular DNA structure, uniquely identify me, but they most surely do not characterize me".*

Condorcet consistent rules are adequate for finding rankings of objects. They present a lower probability of rank reversal than any scoring method. They are not compensatory thus weights can be treated as importance coefficients. A weak point is the high probability of presence of cycles; their solution normally implies *ad hoc* rules of thumb, this problem can be solved by means of the so-called Condorcet (1785), Kemeny (1959), Young and Levenglick (1978) (CKYL) ranking procedure. In the framework of composite indicators, sometimes compensability should be limited and rankings should be supplied; furthermore, transitivity relation can be weakened and neutrality should in principle always be kept. Scoring methods are then, sometimes less adequate than Condorcet based approaches to rank feasible objects.

I offer next a hand-waiving description of a non-compensatory multi-criteria algorithm (for details and formal proofs see Munda and Nardo, 2009 and Munda, 2012a). Given a set of individual indicators $G = \{g_m\}, m = 1,2,...,M$ and a finite set $A = \{a_n\}$, $n = 1,2,...,N$ of countries, let us assume that the performance of a country a_n with respect to an indicator g_m is measured on an interval or ratio scale and that a higher value is preferred to a lower value. Then, a comparison between two countries could either be described by:

$$\begin{cases} a_j P a_k \Leftrightarrow g_m(a_j) > g_m(a_k) \\ a_j I a_k \Leftrightarrow g_m(a_j) = g_m(a_k) \end{cases} \quad (4)$$

where P and I indicate a preference and an indifference relation respectively. Both relations have a transitive property. Let us also assume that weights, w_i, with the meaning of importance coefficients are assigned to the indicators. The question is how to use the available information on indicators and weights to rank in a complete pre-order (i.e. without any incomparability relation) all the countries from the best to the worst one. The mathematical aggregation convention can be divided into two main steps:

1. pair-wise comparison of countries using the entire set of indicators,
2. ranking of countries in a complete pre-order.

In Step 1, an $N \times N$ matrix, E, called 'outranking matrix' can be built (Arrow and Raynaud, 1986; Roy, 1996). An element e_{jk}, $j \neq k$ of the outranking matrix summarises

the result of all pair-wise comparisons between countries *j* and *k* across the M individual indicators, and it is given by:

$$e_{jk} = \sum_{m=1}^{M}\left(w_m(P_{jk}) + \frac{1}{2} w_m(I_{jk}) \right) \quad (5)$$

where P_{jk} and I_{jk} are binary variables representing a preference or indifference relation and are calculated by:

$$P_{jk} = \begin{cases} 1, if\ a_j P a_k \\ 0, otherwise \end{cases}$$

and $I_{jk} = \begin{cases} 1, if\ a_j I a_k \\ 0, otherwise \end{cases}$ (6)

It then holds that

$$e_{jk} + e_{kj} = 1 \quad (7)$$

For Step 2, there are several ranking procedures. The so-called Condorcet-Kemeny-Young-Levenglick (CKYL) ranking procedure can be described as follows. According to CKYL, the ranking of countries with the highest likelihood is the one supported by the maximum number of indicators for each pair-wise comparison, summed over all pairs of countries considered. In practice, call R the set of all $N!$ possible complete rankings of the countries, $R = \{r^s\}, s = 1,2,...,N!$ For each r^s compute the corresponding score φ_S as the summation of e_{jk} over all the $\binom{N}{2}$ pairs j,k of alternatives, i.e.

$$\varphi_s = \sum e_{jk} \quad (8)$$

where $j \neq k, s = 1,2,...,N!$ and $e_{jk} \in r^s$.

The final ranking (r^*) is the one which maximises Equation (9), which is:

$$r^* \Leftrightarrow \varphi^* = \max \sum e_{jk} \quad where\ e_{jk} \in R. \quad (9)$$

Let us take into consideration a simple hypothetical example with three countries (A, B, C) to be ranked according to a composite indicator. Let us assume that three dimensions have to be considered, i.e. economic, social and environmental, and that each dimension should have the same weight, i.e. 0.3333.

The following individual indicators are used:

Economic dimension

Indicator: GDP per capita. Weight: 0.167. Objective: maximization of economic growth. Variable: US dollar per year.

Indicator: Unemployment rate. Weight: 0.167. Objective: minimization of unemployed people. Variable: percentage of population.

Environmental dimension

Indicator: Solid waste generated per capita. Weight: 0.333. Objective: minimization of environmental impact. Variable: tons per year.

Social dimension

Indicator: Income disparity. Weight: 0.167. Objective: minimization of distributional inequity. Variable: Q5/Q1.

Indicator: Crime rate. Weight: 0.167. Objective: minimization of criminality. Variable: robberies per 1000 inhabitants.

The impact matrix described in Table 1 can then be constructed.

Indicators	GDP	Unemp. rate	Solid waste	Inc. dispar.	Crime rate
Countries					
A	22,000	0.17	0.4	10.5	40
B	45,000	0.09	0,45	11.0	45
C	20,000	0.08	0.35	5.3	80

Table 1 Impact Matrix of the Illustrative Numerical Example

The pair-wise comparison results can be summarized in the following outranking matrix:

$$E = \begin{bmatrix} & A & B & C \\ A & 0 & 0.666 & 0.333 \\ B & 0.333 & 0 & 0.333 \\ C & 0.666 & 0.666 & 0 \end{bmatrix}$$

By applying the C-K-Y-L rule to the *3!* possible rankings we obtain:

ABC $\varphi_1 = 0.666 + 0.333 + 0.333 = 1.333$

BCA $\varphi_2 = 0.333 + 0.333 + 0.666 = 1.333$

CAB $\varphi_3 = 0.666 + 0.666 + 0.666 = 2$

ACB $\varphi_4 = 0.333 + 0.666 + 0.666 = 1.666$

BAC $\varphi_5 = 0.333 + 0.333 + 0.333 = 1$

CBA $\varphi_6 = 0.666 + 0.666 + 0.333 = 1.666$

The final ranking $r*$ is then CAB.

Note that using one of the standard ways to produce a composite indicator would produce a different result. Let us look at a real-world example: the *"Environmental Sustainability Index"* (ESI). The index for 2005 was produced by a team of environmental researchers from Yale and Columbia Universities, in co-operation with the World Economic Forum and the Joint Research Centre of the European Commission. The aim of the ESI 2005 was to benchmark the ability of 146 nations to protect the environment over the next decades, by integrating 76 data sets into 21 indicators of environmental sustainability (see Esty *at al.,* 2005). The database used to construct the ESI covers a wide range of aspects of environmental sustainability ranging from the physical state and stress of the environmental systems (like natural resource depletion, pollution, ecosystem destruction) to the more general social and institutional capacity to respond to environmental challenges. Poverty, short-term thinking and lack of investment in capacity and infrastructure committed to pollution control and ecosystem protection thus compete to determine the measure of a country's sustainability.

Although the official ESI ranking is based upon the linear aggregation of 21 equally weighted indicators, an attempt has been made, in the methodological appendix, to apply the non-compensatory approach presented here, in order to tackle the issues of weights as "importance measure" and the compensability of different and crucial dimension of environmental sustainability (see the Methodological Appendix in Esty *et al*., 2005). It is important to underline that although both aggregation schemes seem to produce consistent rankings those rankings do not nevertheless coincide. Using the non-compensatory approach, 43 out of 146 countries experience a change in rank

greater than ten positions (none before the 30th ESI rank). When compensability among indicators is not allowed, countries with very poor performance in some indicators, such as Indonesia or Armenia, worsen their rank with respect to the linear yardstick, whereas countries that have less extreme values improve their ranking, such as Azerbaijan or Spain. Table 2 shows the countries with the largest variation in their ranks.

	Aggregation	ESI rank with LIN	rank with NCMC	Change in Rank
Improvement	Azerbaijan	99	61	*38*
	Spain	76	45	*31*
	Nigeria	98	69	*29*
	South Africa	93	68	*25*
	Burundi	130	107	*23*
Deterioration	Indonesia	75	114	*39*
	Armenia	44	79	*35*
	Ecuador	51	78	*27*
	Turkey	91	115	*24*
	Sri Lanka	79	101	*22*
Average change over 146 countries				*8*

Table 2 ESI Rankings Obtained by Linear Aggregation (LIN) and the C-K-Y-L Ranking Procedure: Countries That Greatly Improve or Greatly Worsen Their Rank Position

To give another example, we may consider results obtained by Munda and Saisana (2011), who considered a theoretical framework for measuring regional sustainability based on the three main dimensions - environment, society, economy- and has been based on 29 indicators applied to Spanish regions and selected Greek and Italian regions. Figure 1 plots the non-compensatory/non-linear multi-criteria ranks versus those of the linear aggregation. This graph allows one to see immediately which regions are compensating their deficiencies in some indicators with a relatively good performance in other indicators under a linear/compensatory logic. All those regions are found at the bottom-right part of Figure 1, e.g. Attiki, Kriti, Extremadura and Thessalia. Another apparent feature is that the aggregation method primarily affects the middle rank regions and, to a lesser extent, the most or least sustainable regions. The two aggregation approaches have a Spearman correlation coefficient r = 0.643.

In conclusion, by means of the C-K-Y-L approach non-compensability can be implemented and cycles can be tackled in a general way with no arbitrariness. A criticism often made to this approach is that non-compensability implies the analytical cost of loosing all available information about intensity of preference, i.e. if some

variables are measured on interval or ratio scales, they have to be treated as measured on an ordinal scale. Indeed this criticism is not entirely correct; in fact it is possible to model e.g. degrees of credibility of preference and indifference ralations inside a non-compensatory framework by means of sensibility thresholds (Luce, 1956). To give a simple example, by introducing a positive constant indifference threshold q the resulting preference model is the *threshold model* where a_j and a_k belong to the set A of alternatives and g_m to the set G of evaluation indicators.

$$\left\{\begin{array}{lll} a_j P a_k & \Leftrightarrow & g_m(a_j) > g_m(a_k)+q \\ a_j I a_k & \Leftrightarrow & \left| g_m(a_j) - g_m(a_k) \right| \le q \end{array}\right\} \qquad (10)$$

A survey of mathematical characterisations of preference modelling with thresholds and an advance the state of the art can be found in Munda (2012b).

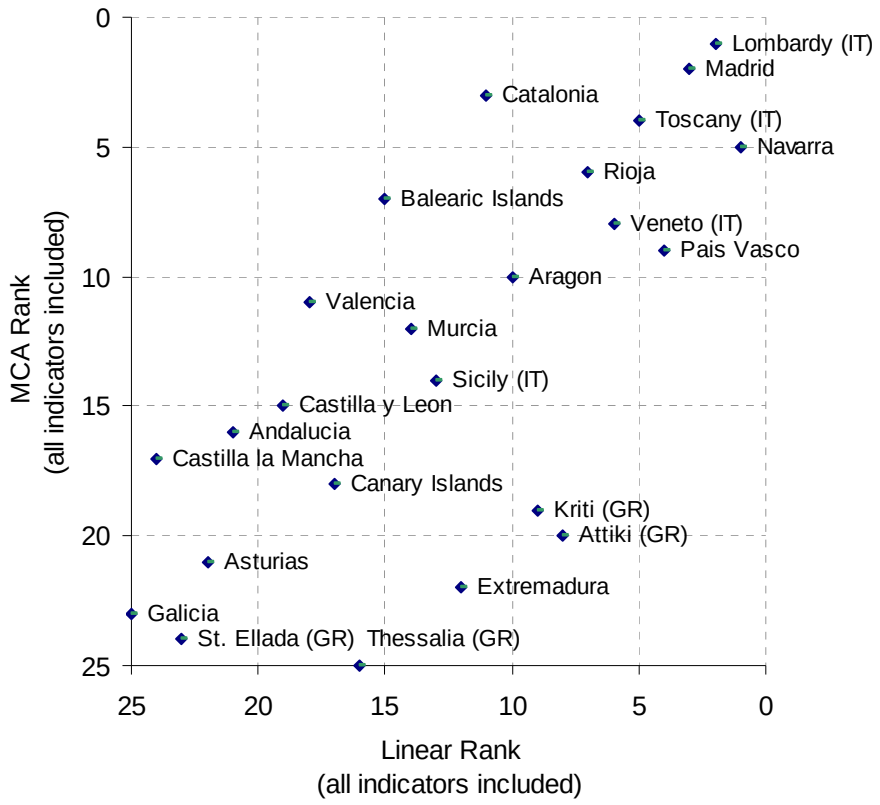

Figure 1 Non-Compensatory Multi-Criteria Aggregation (MCA) of Indicators versus Linear Aggregation of Indicators

However, an important problem to be solved is the computation of the C-K-Y-L ranking scores when many countries are present. One should note that the number of permutations can easily become unmanageable; for example it is 10!=3,628,800. Moulin (1988, p. 312) clearly states that the Kemeny method (that I call the C-K-Y-L approach) is *"the correct method"* for ranking objects, and that the *"only drawback of this aggregation method is the difficulty in computing it when the number of candidates grows*". Indeed this computational drawback is very serious since the Kemeny median order is NP-hard to compute[3]. This NP-hardness has discouraged the development of algorithms searching for exact solutions; thus the majority of algorithms useful in the framework of composite indicators are heuristics based on artificial intelligence, branch and bound approaches and multi-stage techniques (see e.g., Barthelemy et al., 1989; Charon et al., 1997; Cohen et al., 1999).

3. Uncertainty and Sensitivity Analysis

The current rise in the number and influence of composite indicators in public policy calls for a systematic investigation into their intrinsic accuracy and reliability. In fact composite indicators may also present a number of risks, such as oversimplification, wrong policy conclusions due to model misspecification, and biased results caused by hidden subjective judgments in the design process (Saltelli, 2007). Uncertainty and sensitivity analysis can help to gauge the robustness of the composite indicator, to increase its transparency and to help frame a debate around it (Jamison and Sandbu, 2001; Kratena and Streicher 2012; Maasoumi and Yalonetzky, 2013; Munda and Saisana, 2011; Paruolo et al, forthcoming; Saisana et al., 2005; Saltelli et al., 2000, 2008). *Uncertainty analysis* focuses on how the sources of uncertainty propagate through the structure of the composite indicator and affect its values. *Sensitivity analysis* studies how much each individual source of uncertainty contributes to the composite indicator value/ranking variance. The types of questions for which an answer is sought via the application of these techniques are:

- Does the use of one construction strategy versus another in building the composite indicator actually provide a partial picture of the countries' performance?
- Which countries have large uncertainty bounds in their rank (volatile countries)?
- Which are the factors that affect the countries rankings?

[3] The **complexity class** of **decision problems** that are intrinsically harder than those that can be solved by a **nondeterministic Turing machine** in **polynomial time**. When a decision version of a combinatorial **optimization problem** is proved to belong to the class of **NP-complete** problems, then the optimization version is NP-hard (definition given by the National Institute of Standards and Technology, http://www.nist.gov/dads/HTML/nphard.html).

A plurality of methods (all with their implications) should be initially considered, because no model (composite indicator construction strategy) is a priori better than another, provided that internal coherence is always assured, as each model serves different interests. The composite indicator is no longer a magic number corresponding to crisp data treatment, weighting set or aggregation method, but reflects uncertainty and ambiguity in a more transparent and defensible fashion (Munda, 2008; Nardo *et al*, 2008).

In summary, results obtained by using a composite indicator always depend heavily on the problem's structuring phase. In general main delicate issues are (Munda, 2005):
(1) Mathematical aggregation rule used. This issue has been examined in the previous Section.
(2) Quality of the information available. One should note that even if a data base has been submitted to rigorous quality check, from a pure technical point of view, the following uncertainty sources are always present and have to be taken into account (Nardo *et al*., 2008):

- the consideration of measurement error in the data,
- the imputation of missing data,
- the treatment of outliers and extreme values,
- the transformation of skewed indicators,
- the standardization/normalization of the data (e.g., re-scaling, standardisation).

(3) Indicators chosen i.e. which representation of reality we are using. A set of indicators is not the reality, but it is simply a descriptive model of it. It is important then to check the relevance and the explicative capacity of the theoretical framework used. The way one may choose to deal with this issue is by looking at the sensitivity of results to the exclusion/inclusion of different individual indicators and dimensions. Although, this analysis may look very technical in nature, in reality a social component is also present. In fact to consider or not a given dimension, normally has behind a long story of social, political and scientific controversy (Munda, 2008). To give an example, the environmental dimension nowadays is considered very important in almost any analysis; however this was not true 40 years ago, mainly because the social concerns on the environment in the past were very limited. As a conclusion, we should remember that to include or exclude a given dimension or a set of indicators means to deal or not with peculiar social concerns and social actors.

(4)Weighting of the indicators e.g., equal weighting, factor analysis, expert opinion and so on. This again has a technical and a more socio-political component. The following weighting assumptions may have a general validity:
(1) equal weights to individual indicators (thus dimensions weight is determined by the total number of individual indicators per dimension),
(2) equal weights to the various dimensions (thus weights to individual indicators vary according to their number per dimension) (Munda, 2008),
(3) Factor analysis (The indicators (z-scores) are weighted and aggregated into sub-dimensions using factor analysis (Nardo *et al.,* 2008).
(4) Endogenous weights derived by data envelopment analysis. These weights allow to check how stable is a bottom position of a country – since the best set of weights for that country is used – and then to derive policy priority (Cherchye *et al.,* 2004).

To look at an example, we may continue the ESI example seen in the previous Section and ask: what are the largest factors influencing results of the 2005 ESI? To answer this question, one may focus on the following comparisons (see the Methodological Appendix in Esty *et al*., 2005):

1. Imputation versus no imputation
2. Expert-weighting versus equal weighting of the 21 indicators
3. Aggregation at the dimensions level versus at the indicators level
4. Non-compensatory aggregation scheme versus linear aggregation

Imputation
Imputation should be more influential for countries where missing data are a large problem, although this relationship does not seem to be straightforward. Among the countries that miss almost 33% of their observations, only Guinea-Bissau and Myanmar are strongly affected by the imputations. Without imputation, Syria, Algeria, Belgium and the Dominican Republic improve their ranks between 29 and 37 positions. Conversely, Mali, Guinea-Bissau, Myanmar, and Zambia, decline 27 to 43 positions. Overall, the imputation has an average impact of 10 ranks and a rank-order correlation coefficient of 0.949.

Expert weighting versus equal weighting
The ESI 2005 used equal weights to calculate the country scores from 21 indicators. As alternative weighting schemes a *"budget allocation scheme,"* was tested in which the weights are obtained from experts with a demonstrated understanding of environmental sustainability. Seventeen experts were each given a "budget" of 100

points and asked to allocate them to the 21 indicators according to their personal judgment of the relative importance of the indicators. The average expert weighting is slightly different from the equal weighting used in the ESI, nevertheless, the variance of experts' opinions is rather large, varying from 40-80% of the mean weight. This explains the difference between the ESI ranking and the one obtained when using the average expert weighting set. Overall, the weighting has an average impact of 5 ranks in the simulations and a rank-order correlation coefficient of 0.989.

Aggregation at the Dimension Level v. Aggregation at the Indicators Level

In order to further assess the robustness of the ESI, it was analysed the possibility of equally weighting the five dimensions: Environmental Systems, Reducing Environmental Stresses, Human Vulnerability, Social and Institutional Capacity, and Global Stewardship, instead of the 21 indicators. It was found that by changing the aggregation level, the average shift of the top 40 and the bottom 30 countries of the ESI 2005 is 7 positions and the shift of the remaining countries averages 11 positions. The average impact is 8 ranks and the rank-order correlation coefficient remains very high at 0.964.

Aggregation rule and the compensation issue

The aggregation rule matters mainly for the mid-performing countries. When the assumption of compensability among indicators is removed, countries having very poor performance in some indicators, such as Indonesia or Armenia, decline in rank, whereas countries with fewer extreme values, such as Azerbaijan or Spain, improve their position. Overall, the aggregation rule has an average impact of 8 ranks.

As one can see, sensitivity analysis helps to gauge the robustness of the results obtained, to increase the transparency of the ranking system, to identify how countries or regions that improve or decline under certain assumptions, and to help the framing of the debate around the use of a conceptual framework. Unfortunately, most practitioners compute a composite indicator by a simple weighted summation mathematical model; sometimes it is acknowledged that the ranking obtained is subject to some uncertainty, but this issue is treated as a kind of mathematical appendix for technical readers, and all policy suggestions are derived under the assumption of the linear aggregation model.

Saisana and Munda (2008) believe that to deal with the criticism that rankings are presented as they were under conditions of certainty, while it is well known that this is

not true, it is a key issue for the use of composite indicators in the policy arena. Saisana and Munda (2008) make a simple methodological proposal, i.e. let us consider as the final composite indicator the frequency of all rankings obtained by means of all the simulations carried out by considering the combinations of all the possible methodological assumptions relevant for the construction of a real-world composite indicator. In this way, the ranking presented is the one derived by considering the whole spectrum of uncertainty. The objective is to synthesize and make explicit the uncertainty contained in the rankings. For each country it is indicated the percentage of times it was in a given rank in all the simulations, thus presenting a clear measurement of the degree of uncertainty contained in the ranking obtained. In summary, all possible technical uncertainties are simulated and then aggregated by using a simple social choice aggregation rule, i.e. a Borda scoring method.

To give a simple illustrative example, let us consider a composite for the knowledge economy, called KEI (Saisana and Munda (2008)). In the KEI original conceptual framework of the knowledge economy, a total of 115 individual indicators were selected. The proposed multi-modelling approach was applied to weight and further aggregate the sub-dimensions scores into dimensions and finally into a composite indicator (see Table 3). The computations consisted of about 2,000 simulations (saturated sampling) based on combinations of the:

1. *imputation method* (Missing data were imputed using two different approaches: splines or multiple imputation, 2 datasets were thus used in the analysis described next);
2. *number of sub-dimensions* (all 29 sub-dimensions included or one-at-time excluded);
3. *number of dimensions* (all seven dimensions included or one-at-time excluded);
4. *normalisation* of the 29 sub-dimensions scores (z-scores or min-max);
5. *structure* relating the sub-dimensions to the dimensions (preserved or not);
6. *weighting* method (factor analysis, equal weighting, data envelopment analysis);
7. *aggregation rule* (additive, multiplicative, non-compensatory multi-criteria analysis).

The frequency matrix of a country's rank in each of the seven dimensions and the overall KEI is calculated across the 2,000 scenarios is presented in Table 3. The objective here is to synthesize and make explicit the uncertainty contained in the country ranking. For each country it is indicated the percentage of times it was in a given rank in all the 2,000 simulations, one can see that e.g. Poland was 100% of times

in the last position, and Sweden 54% of times in the first position and 46% in the second. In the example we are considering, overall we can state that the ranking is very stable; in fact considering the whole 2,000 simulations, all countries are clustered unambiguously. No doubt the top performing countries are Sweden, Denmark Luxembourg, Finland and the USA. Then it follows the group Japan, United Kingdom, Netherlands and Ireland (where Japan and UK are slightly better than the other two). Austria, Belgium, France and Germany form the next group (where Germany is slightly worst than all the other three). All the rest of countries can be considered with a bad performance with respect to knowledge based economy.

Knowledge Economy Index

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10	Rank 11	Rank 12	Rank 13	Rank 14	Rank 15	Rank 16	Rank 17	Rank 18	Rank 19	Rank 20	Rank 21	Rank 22	Rank 23	Rank 24	Rank 25	Rank 26	Rank 27	Rank 28	Rank 29
Sweden	54	46																											
Denmark			55	30	14																								
Luxembourg	36	4	14	25	4	7	7	4																					
Finland		18	23	29	9	11	11																						
USA	11	32	2	4	39	9	4																						
Japan			4	7	18	32	36			4																			
UK			2	5	16	38	39																						
Netherlands								86	4		4			7															
Ireland							4		61	14	4	9	9																
Austria									18	50	18	7		7															
Belgium								11	4	11	57	16	2																
France						4			14	18	11	54																	
EU15											4		57	39															
EU25										4	4	14	32	39	7														
Germany														7	79	4	7	4											
Slovenia															7	41	38	14											
Estonia																4	36	25	21	11	4								
Malta															7	13	9	21	23	27									
Cyprus																36	7	4	23	23	7								
Spain																4	4	32	25	29	7								
Czech. Rep.																			4	7	30	39	5	7	7				
Latvia																					20	36	11	21	7		5		
Italy																					29	18	9	29	9	7			
Greece																				4	4	4	29	18	21	7	14		
Lithuania																						4	41	13	32	11			
Hungary																							2	13	13	57	2	14	
Portugal																			4				4		7	11	61	14	
Slovakia																									4	7	18	71	
Poland																													100

Legend:
Frequency lower 15%
Frequency between 15 and 30%
Frequency between 30 and 50%
Frequency greater than 50%

Table 3 Frequency matrix of the KEI composite country rankings

4. Evidence Based Policy: Evaluation and Benchmarking

At this stage a question may be raised: *is all this effort of any use?* Even if we have a very reliable ranking, which is the utility, for policy-making, of knowing that a country is overall better than another one or vice versa? Let's try to put some light on this issue. First of all, one should note that for the majority of indicators used in assessment

exercises no clear reference point is available, for instance, when GDP is used nobody knows the ideal value of a Country GDP, thus it is quite common to compare with other countries GDP, e.g. the USA one. In order to get a set of reference values, an “ideal point” can be defined by choosing the best values reached in any single indicator. This is a well established technique in multi-criteria evaluation literature (see e.g. Yu, 1985; Zeleny, 1982) and has the advantage of indicating real world ideal values.

Briefly, the philosophy underlying the multi-criteria methods based on ideal point concepts can be summarized as follows: multidimensional evaluation problems are characterized by conflict because of the perceived absence of an obvious “best” option; therefore, the only way to resolve the conflict is to find or invent an ideal point. The only way to decrease the intensity of the conflict is to find or generate alternatives which are as close as possible to the ideal point. The ideal point procedures are characterized by the following axiom of choice: alternatives that are closer to the ideal are preferred to those that are farther away. To be as close as possible to the perceived ideal is the rationale of choice.

One of the traditional ideal point approaches is to compute the mathematical distance of each action from the ideal point and then rank them in terms of their proximity to the ideal. Another possibility is the use of aspiration levels (or goals) which express the desired outcomes of a given policy in terms of a certain level to be aimed at for each objective. The usual way in which aspiration levels are treated is by means of goal programming (Spronk, 1981). An advantage of goal programming is that it always provides a solution, even if none of the goals are realizable, provided that the feasible region is non-empty. This is possible by using deviational variables, which show whether the goals have been attained or not. In the latter case, they measure the distance between the realized and aspired levels. An approach that can be viewed as a generalization of goal programming and ideal point techniques is the "achievement scalarizing functions" method (Wierzbicki, 1982).

A first very simple mathematical procedure can be the application of a normalisation rule known as “distance from the group leader”, which assigns 100 to the leading country and other countries are ranked as percentage points away from the leader (Nardo et al., 2008). This technique can be considered a simple distance function. For example by applying this technique to one of the Spanish regions analysed by Munda and Saisana (2011), the Basque Country, the following results visualised by a radar diagramme are obtained (see Figure 2). As one can see, for each single dimension

considered, the diagrammes present which the policy priority are if a policy action has to be taken to improve the overall performance of the region.

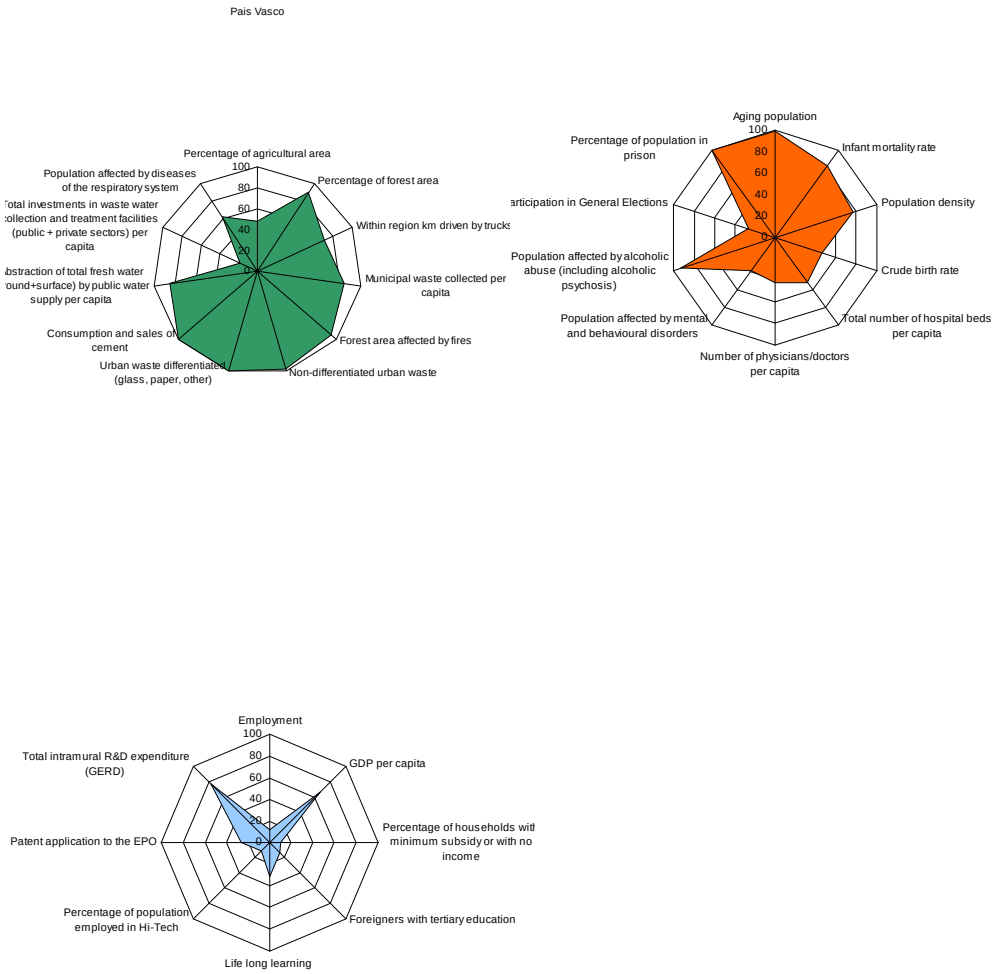

Figure 2 Radar Diagrams for the Basque Country Sustainability Benchmarking

However one should not forget that an important limitation of composite indicators is that they are static in nature. The fact that a country is in a top position in a well-being ranking does not of course imply that this situation will last along time. In fact, when the same process is considered at different temporal scales, the co-existence of opposite causal links may emerge. This is the reason why in a benchmarking exercise, it is essential to look at possible relationships between the composite ranking and some key drivers that might be some components of the composite framework or other complex measures or composites. To give an illustrative example, let us look again at

the KEI composite (Saisana and Munda, 2008). We may ask: is a knowledge based economy a good driver for reducing unemployment? For answering this question the time scale is the key point. In fact job creation could be successfully increased *in the short term,* by a slowdown of the rate of technological progress. As noted by the Kok report[4], this is exactly what has recently happened inside the European Union. But in *a longer time_horizon,* this strategy may easily cause the collapse of the economy given that non-specialized low productivity jobs can easily be substituted by lower wage labour in other parts of the world. Thus, in the short term technological progress and job creation are conflicting objectives but they might be compatible in the long run; this statement can be corroborated by looking at the relationship between long term unemployment rate and the KEI median ranking (see Figure 3). All top countries in the KEI measure are presenting an extremely low long term unemployment rate.

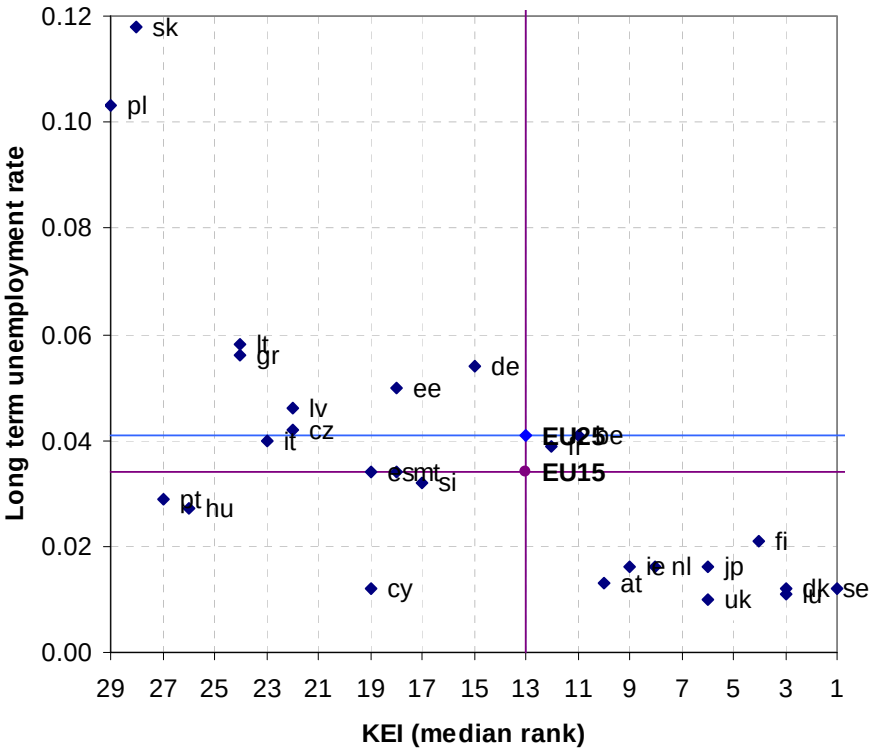

Figure 3 Relationship between the KEI median ranking and long term unemployment rate

[4]Kok W. (2004,) - The High Level Group on Lisbon Strategy (chaired by Wim Kok) (2004) – Facing the Challenge, European Communities, Luxembourg.

5. Conclusions

According to the arguments presented here, the following conclusions can be drawn:

1. Sustainability and well-being research agendas are connected obviously; one should note that the classical definition of sustainability given by the World Commission on Environment and Development (1987) focused on *human needs* as objective to be sustained over time. Of course, human needs are multidimensional in nature and as a consequence well-being is a more adequate comprehensive measure of wealth than traditional GDP.
2. There is no doubt that there is a lot of complexity and fuzziness inherent in multidimensional concepts such as sustainability and well-being. A possible reduction of this complexity, a pre-condition for policy-making, introduces the problem of the descriptors used: indicators and indexes. A well-being policy exercise implies difficult decisions such as the choice of indicators, their policy prioritization and the choice of reference/ideal values; such an exercise is not a technical issue only, it is a socio-political issue too. Behind a list of indicators and a list of targets there would always be a history of scientific research and political controversy. A proper evaluation exercise needs to deal with a plurality of legitimate values and interests found in a society.
3. In a multidimensional framework, multi-criteria evaluation is a very consistent approach for the construction of so-called composite indicators. Often, some indicators improve while others deteriorate. This is the classical conflictual situation dealt with in multi-criteria evaluation; in particular non-compensatory methods are quite relevant, since compensability implies complete substitutability between different types of capital.
4. Composite indicators may also present a number of risks, such as oversimplification, wrong policy conclusions due to model misspecification, and biased results caused by hidden subjective judgments in the design process. Uncertainty and sensitivity analysis can help to gauge the robustness of the results obtained, to increase the transparency of the ranking system, to identify how countries or regions that improve or decline under certain assumptions, and to help the framing of the debate around the conceptual framework used, i.e. which representation of reality (and thus which societal values and interests) has been considered.
5. In the framework of evidence based policy, benchmarking exercises based on real-world reference and ideal points can be very useful. In fact it is possible to evaluate how each single country/region is close or far to each single target and

thus policy priorities can be established. However one should not forget that an important limitation of composite indicators is that they are static in nature. The fact that a country is in a top position in a well-being ranking does not of course imply that this situation will last over time. In fact, when the same process is considered at different temporal scales, the co-existence of opposite causal links may emerge. Surely, more research is needed on this topic. The essence of the time scale problem is perfectly synthesised by Frank Knight (1921, p. 313) *"... We live in a work of contradiction and paradox, a fact of which perhaps the most fundamental illustration is this: that the existence of a problem of knowledge depends on the future being different from the past, while the possibility of the solution of the problem depends on the future being like the past"* .

Acknowledgements I would like to thank Prof. Giovanni Signorello and all participants to the Belpasso International Summer School on Environmental and Resource Economics, *"Sustainable Development Theory and Measurement methods",* for the helpful comments on my lecture which included many ideas presented in this article. This research has been partially developed in the framework of the Spanish Government financially supported project SALMON (HAR2010-20689-02-01).

REFERENCES

Arrow K.J. (1963) - *Social choice and individual values,* 2d edition, Wiley, New York.

Arrow K.J., Raynaud H. (1986) - *Social choice and multicriterion decision making*, M.I.T. Press, Cambridge.

Arrow (1997) - Invaluable Goods, *Journal of Economic Literature*, Vol. 35, No. 2, pp. 757- 763.

Arrow K. J., Dasgupta P., Goulder L. H., Mumford K. J. and Oleson K. (2012) - Sustainability and the measurement of wealth, *Environment and Development Economics* / Volume 17 / Issue 03 / pp. 317 - 353.

Atkinson G., Hamilton K. (2007) - Progress along the path: evolving issues in the measurement of genuine saving, *Environmental and Resource Economics,* Volume 37, Issue 1, pp 43-61.

Banerjee A., Marcellino M., Masten I. (2005) - Leading Indicators for Euro-area Inflation and GDP Growth, *Oxford Bulletin of Economics and Statistics*, Volume 67, Issue Supplement s1, pages 785–813.

Barbier E.B., Markandya A, (1990) -The conditions for achieving environmentally sustainable growth, *European Economic Review,* 34, pp. 659-669.

Barthelemy J.P., Guenoche A., Hudry O. (1989) – Median linear orders: heuristics and a branch and bound algorithm, *European Journal of Operational Research*, 42, pp. 313-325.

Van den Bergh C.J.M. (2009) – The GDP paradox, *Journal of Economic Psychology*, 30, pp. 117-135.

Borda J.C. de (1784) – Mémoire sur les élections au scrutin, in *Histoire de l'Académie Royale des Sciences,* Paris.

Bordes G. and Tideman N. (1991) – Independence of irrelevant alternatives in the theory of voting, *Theory and Decision,* Vol. 30, No. 2, pp. 163-186.

Bouyssou, D. (1986) - 'Some remarks on the notion of compensation in MCDM', *European Journal of Operational Research*, Vol. 26, pp.150-160.

Bouyssou D., Vansnick J.C. (1986) – Noncompensatory and generalized noncompensatory preference structures, *Theory and Decision*, 21, pp. 251-266.

Chang R. (ed.) (1997) - **Incommensurability, Incomparability, and Practical Reason**, Cambridge: Harvard University Press.

Charon I., Guenoche A., Hudry O., Woirgard F. (1997) – New results on the computation of median orders, *Discrete Mathematics*, 165/166, pp. 139-153.

Cherchye L., Moesen W. , Van Puyenbroeck T. (2004) - Legitimately diverse, yet comparable: On synthesising social inclusion performance in the EU. *Journal of Common Market Studies* 42, 919–955.

Cherchye L., Knox L., Moesen, W. anf Van Puyenbroeck, T. (2007) - "One market, one number? A composite indicator assessment of EU internal market dynamics," *European Economic Review,* vol. 51(3), pages 749-779.

Chernoff H. (1954) – Rational selection of decision functions, *Econometrica,* 22 (4), pp. 422-443.

Chichilnisky G. (1996) – An axiomatic approach to sustainable development, *Social Choice and Welfare*, 13(2), pp. 219-248.

Cohen W., Schapire R., Singer Y. (1999) – Learning to order things, *Journal of Artificial Intelligence Research,* 10, pp. 213-270.

Condorcet, Marquis de (1785) – *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la probabilité des voix*, De l' Imprimerie Royale, Paris.

Cox, D., Fitzpatrick, R., Fletcher, A., Gore, S., Spiegelhalter, D. and Jones D. (1992) -'Quality-of-life assessment: can we keep it simple?', *Journal of the Royal Statistical Society,* Vol. 155, pp. 353-393.

Cribari-Neto, F., Jensen, M.J., Novo, A. (1999) - 'Research in econometric theory: quantitative and qualitative productivity rankings', *Econometric Theory*, Vol. 15, pp.719-752.

Daly H.E., Cobb J.J. (1989), *For the common good: redirecting the economy toward community, the environment and a sustainable future*, Beacon Press, Boston.

Esty, D.C., Levy, M., Srebotnjak, T. and de Sherbinin A. (2005) - *2005 Environmental Sustainability Index: Benchmarking National Environmental Stewardship*. New Haven: Yale Center for Environmental Law & Policy.

Figueira, J., Greco, S. and Ehrgott, M. (eds.) (2005) *Multiple-criteria decision analysis. State of the art surveys*. Springer International Series in Operations Research and Management Science, New York.

Funtowicz S., Martinez-Alier J., Munda G. and Ravetz J. (1999) - *Information tools for environmental policy under conditions of complexity*, European Environmental Agency, Experts' Corner, Environmental Issues Series, No. 9.

Geach P. T. (1956) – Good and Evil, *Analysis*, Vol. 17, pp. 32–42 - Reprinted in Foot, Philippa (ed.) (1967). *Theories of Ethics*, Oxford University Press, Oxford, pp. 64–73.

Giampietro, M., Mayumi, K., Munda, G. (2006) - Integrated assessment and energy analysis: quality assurance in multi-criteria analyses of sustainability. *Energy* Volume 31, Issue 1, pp. 59-86.

Griliches, Z. (1990) - 'Patent statistics as economic indicators', *Journal of Economic Literature*, Vol. 28, pp.1661-1707.

Horwarth R., **Norgaard R.B.** (1990) – Intergenerational resource rights, efficiency and social optimality, *Land Economics,* 66, pp. 1-11.

Horwarth R., **Norgaard R.B.** (1992) - Environmental valuation under sustainable development, *American Economic Review Papers and Proceedings,* 80, pp. 473-477.

Jamison, D. and Sandbu, M. (2001) - WHO ranking of health system performance. *Science*, 293, 1595-1596.

Keeney R., Raiffa H. (1976) - *Decision with multiple objectives: preferences and value trade-offs*, Wiley, New York.

Kemeny J. (1959) – Mathematics without numbers, *Daedalus*, 88, pp. 571-591.

Knight F. (1921) - *Risk, Uncertainty and Profit*, Houghton Mifflin Company, Cambridge, UK.

Kratena K., Streicher G. (2012) - Spatial Welfare Economics Versus Ecological Footprint: A Sensitivity Analysis Introducing Strong Sustainability, *Environmental and Resource Economics,* Volume 51, Issue 4, pp 617-622.

Lovell, C.A.K., Pastor, J.T., Turner, J.A. (1995) - 'Measuring macroeconomic performance in the OECD: a comparison of European and non-European countries', *European Journal of Operational Research*, Vol. 87, pp.507-518.

Luce R.D. (1956) – Semiorders and a theory of utility discrimination, *Econometrica,* 24, pp.178-191.

Maasoumi E. and Yalonetzky G. (2013) - Introduction to Robustness in Multidimensional Wellbeing Analysis, *Econometric Reviews*, Volume 32, Issue 1, pages 1-6.

Markandya A., Pedroso-Galinato S. (2007) - How substitutable is natural capital?, *Environmental and Resource Economics,* Volume 37, Issue 1, pp 297-312.

Martinez-Alier, J., Munda, G., O'Neill, J. (1998) Weak comparability of values as a foundation for ecological economics. *Ecological Economics*, 26, pp. 277-286.

McGuckin R. H., Ozyildirim A. and Zarnowitz V. (2007) - A More Timely and Useful Index of Leading Indicators, *Journal of Business & Economic Statistics*, vol. 25, pages 110-120.

McLean I. (1990) – The Borda and Condorcet principles: three medieval applications, *Social Choice and Welfare*, Vol. 7, pp. 99-108.

Moulin H. (1988) - *Axioms of co-operative decision making,* Econometric Society Monographs, Cambridge University Press, Cambridge.

Munda G. (1997) – Environmental economics, ecological economics and the concept of sustainable development, *Environmental Values,* vol. 6, No. 2, pp. 213-233.

Munda G. (2004) – "Social multi-criteria evaluation (SMCE)": methodological foundations and operational consequences, *European Journal of Operational Research* Vol. 158, Issue 3: 662- 677.

Munda G. (2005) – "Measuring sustainability": a multi-criterion framework, *Environment, Development and Sustainability ,* Vol 7, No. 1, pp. 117-134.

Munda G. (2008) – *Social multi-criteria evaluation for a sustainable economy*, Springer, Heidelberg, New York.

Munda G. and Nardo M. (2009) - Non-compensatory/non-linear composite indicators for ranking countries: a defensible setting*, Applied Economics* Vol. 41, pp. 1513-1523.

Munda G., Saisana M. (2011) - Methodological Considerations on Regional Sustainability Assessment based on Multicriteria and Sensitivity Analysis, *Regional Studies* Vol. 45.2, pp. 261-276.

Munda G. (2012a) – Choosing aggregation rules for composite indicators, *Social Indicators Research*, Volume 109, Issue 3, pp. 337-354,

Munda G. (2012b) – Intensity of Preference and related Uncertainty in Non-Compensatory Aggregation Rules, *Theory and Decision,* Volume 73, Issue 4, pp. 649-669.

Musu I., Siniscalco D. (1996) – *National accounts and the environment*, Kluwer Academic Publishers, Dordrecht.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. (2008) -*Handbook on Constructing Composite Indicators: Methodology and User Guide.,* Paris: OECD Statistics Working Paper.

Nijkamp, P., Rietveld, P. and Voogd, H. (1990) - *Multicriteria Evaluation in Physical Planning.* Amsterdam: North-Holland.

OECD (2003) - *Composite Indicators of Country Performance: a critical assessment*, Paris: OECD Statistics Working Paper.

O'Neill, J. (1993) - *Ecology, Policy and Politics*. Routledge, London.

Paruolo P., Saltelli A., Saisana M. (forthcoming) - "Ratings and rankings: Voodoo or Science?" *Journal of the Royal Statistical Society: Series A (Statistics in Society);* available at http://arxiv.org/abs/1104.3009

Pearce D.W., Atkinson G.D. (1993) - Capital theory and the measurement of sustainable development: an indicator of "weak" sustainability, *Ecological Economics*, vol. 8, pp. 103-108.

Pearce D, Hamilton G., Atkinson G.D. (1996) – Measuring sustainable development: progress on indicators, *Environment and Development Economics,* 1, pp. 85-101.

Podinovskii V.V. (1994) -Criteria importance theory, *Mathematical Social Sciences,* 27, pp. 237 - 252.

Ray P. (1973) - Independence of irrelevant alternatives, *Econometrica*, Vol. 41, No.5, pp.987-991.

Rabinowicz W. (2012) - Value Relations Revisited, *Economics and Philosophy*, 28, pp. 133-164.

Roberts F. S. (1979) - *Measurement theory with applications to decision making, utility and the social sciences*, Addison-Wesley, London.

Roy, B. (1996) - *Multicriteria Methodology for Decision Analysis*. Dordrecht, The Netherlands: Kluwer.

Saari, D.G. (2006) – Which is better: the Condorcet or Borda winner? *Social Choice and Welfare*, 26, pp. 107-129.

Saisana M., Munda G. (2008) - *Knowledge Economy: measures and drivers*, EUR 23486 EN, European Commission, Joint Research Centre, Ispra, site.

Saisana, M., Tarantola, S., Saltelli, A. (2005) - 'Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators', *Journal of the Royal Statistical Society A*, Vol. 168, pp.307-323.

Saltelli, A., Chan, K. and Scott, M. (2000) - *Sensitivity Analysis*, New York: John Wiley & Sons.

Saltelli, A. (2007) - 'Composite indicators between analysis and advocacy', *Social Indicators Research*, Vol. 81, pp.65-77.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S. (2008) - *Global Sensitivity Analysis. The Primer*, England: John Wiley & Sons.

Schmidt-Bleek, F. (1994) - *Wieviel umwelt braucht der mensch? MIPS Das Mass für ökologisches wirtschaften*, Birkh user, Berlin.

Sen A. (1997) - Maximization and the act of choice, **Econometrica**, 65, pp. 745-779.

Sen A. (2000) - Consequential evaluation and practical reason, **Journal of Philosophy**, 98, pp. 477-502.

Sen, A. and B. Williams, (eds.) (1982) - **Utilitarianism and Beyond**, Cambridge: Cambridge University Press.

Spronk, J. (1981) - *Interactive multiple goal programming for capital budgeting and financial planning*. Martinus Nijhoff, Boston.

Srinivasan S., Stewart G. (2004) - The Quality of Life in England and Wales, *Oxford Bulletin of Economics and Statistics*, Volume 66, Issue 1, pages 1–22.

Stiglitz J., Sen A. and Fitoussi J.P. (2009) – *Report by the Commission on the measurement of economic performance and social progress,* available at http://www.stiglitz-sen-fitoussi.fr/en/index.htm

Ting, H.M. (1971) - *Aggregation of Attributes for Multiattributed Utility Assessment,* Cambridge, MA: M.I.T., Operations Research Center, Technical report N° 66.

Vansnick, J.C. (1986) - 'On the problem of weights in multiple criteria decision making (the non-compensatory approach)', *European Journal of Operational Research*, Vol. 24 pp.288-294.

Vansnick J. C. (1990) - Measurement theory and decision aid- in Bana e Costa C.A. (ed.)- *Readings in multiple criteria decision aid,* Springer-Verlag, Berlin, pp. 81-100.

Vincke, Ph. (1992) - *Multicriteria Decision Aid,* New York: Wiley.

Vitousek P., Ehrlich P., Ehrlich A. and Matson P. (1986) - Human appropriation of the products of photosynthesis, *Bioscience*, 34(6): 368 373.

Wackernagel, M. and W. E. Rees (1995) - *Our ecological footprint: Reducing human impact on the earth,* Gabriola Island, BC and Philadelphia, PA: New Society Publishers.

Wierzbicki, A.P. (1982) - A mathematical basis for satisficing decision making. *Mathematical Modelling,* 3, pp. 391-405.

Williams A. and Siddique A. (2008) - The use (and abuse) of governance indicators in economics: a review, *Economics of Governance*, Vol. 9, No. 2. pp. 131-175.

Wilson, J.W. and Jones, C.P. (2002) - 'An analysis of the S&P 500 index and Cowles's extensions: price indexes and stock returns, 1870-1999', *Journal of Business*, Vol. 75, pp.505-533.

World Commission on Environment and Development (1987) - *Our common future.* Oxford University Press, Oxford.

Young H.P. and Levenglick A. (1978) – A consistent extension of Condorcet's election principle, *SIAM Journal of Applied Mathematics*, 35, pp. 285-300.

Yu, P.L. (1985) - *Multi-criteria decision making: concepts, techniques and extensions.* Plenum Press, New York.

Yusuf, J.A., S. El Serafy and E. Lutz (1989) - *Environmental accounting for sustainable development,* A UNEP World Bank Symposium, Washington D.C.

Zeleny, M. (1982) - *Multiple criteria decision making.* McGraw Hill, New York.