the review of income and wealth

Review of Income and Wealth
Series 57, Number 2, June 2011

ASSESSING THE ROBUSTNESS OF COMPOSITE INDICES RANKINGS

by Iñaki Permanyer*

Fulbright Visiting Scholar at Cornell University, Department of City and Regional Planning

Ranking objects in terms of different attributes is a crucial practice that is typically sensitive to the choice of attributes' weights. In this paper we present rigorous methods to assess the extent to which the weight-based rankings are robust to the choice of alternative weights. Empirical illustrations are provided, showing the robustness of country rankings arising from the values of the UNDP Human Development Index, the Gender-related Development Index, or the Human Poverty Index among others. The ideas and techniques presented in this paper can be used to assess the reliability of multiattribute rankings.

JEL Codes: C02, I31, O12, O15

Keywords: composite index, weighting scheme, sensitivity analysis, robustness, Human Development Index, UNDP

1. Introduction

The evaluation and posterior ranking of objects by means of multicriteria comparisons is a pervasive activity in many spheres of human life. To name a few examples: a list of projects have to be evaluated by different referees to decide which is best suited to receive economic support; a list of students have to be ranked to decide which ones deserve a fellowship or to join a certain institution; the social services of a given community need to evaluate the needs of the poorest individuals to know who deserves financial aid; a list of countries has to be ranked in terms of aggregate well-being to decide how to allocate scarce resources; and so on. In all these examples, the different objects are evaluated and ranked using multiple criteria. These criteria might include the opinion of different experts or referees, valuable attributes or characteristics of the objects to be ranked, or both things at the same time. Finally, these multiple criteria or attributes are usually aggregated to obtain a summary numerical measure that will be used to rank the different objects.

In practice, it is very common to summarize this multidimensional information using the so-called weighted means; that is, when aggregating the values of the different attributes, a specific weight is given to each of them according to their relative importance. This is a simple and intuitive way of summarizing information that allows for the possibility of giving more emphasis to those components of the

Note: I would like to thank Kaushik Basu and Erik Thorbecke for their valuable comments on this manuscript. I would specially like to thank two anonymous referees whose comments have greatly improved the contents and presentation of this paper. Support from the Generalitat de Catalunya and the Fulbright Commission is gratefully acknowledged.

*Correspondence to: Iñaki Permanyer, Fulbright Visiting Scholar at Cornell University, Department of City and Regional Planning, Cornell University, 106 W. Sibley Hall, Ithaca, NY 14853, USA (ip52@cornell.edu).

evaluating process that are considered to be more relevant. For example: one might want to give more weight to the opinion of a renowned and experienced referee than to a novel one; or one might consider that health is more important than the number of TVs per inhabitant when evaluating the aggregate well-being of a list of countries; and so on.

However, the choice of specific weights for alternative attributes or dimensions is an extremely delicate issue. Depending on the context we are dealing with, there might be important ethical and normative implications in the choice of one weight or another. As Anand and Sen argue: "Since any choice of weights should be open to questioning and debating in public discussions, it is crucial that the judgements that are implicit in such weighting be made as clear and comprehensible as possible and thus be open to public scrutiny" (Anand and Sen, 1997, p. 6). Due to the importance of this choice, in the literature there is a plethora of techniques to provide more or less reasonable criteria to obtain the "ideal" weighting scheme. To name some of most well-known: one has data driven techniques (including, but not limited to, data envelopment analysis (DEA; see Despotis, 2005); frequency based methods (see Desai and Shah, 1998), principal components analysis, factor analysis, cluster analysis (see Hirschberg *et al*., 1991), regression based weights (see Schokkaert, 2007), or normative weighting techniques (public opinion, expert opinion, or equal weighting; see Chowdhury and Squire, 2006; Stapleton and Garrod, 2007)). The interested reader can find a survey of weighting techniques in the OECD's *Handbook on Constructing Composite Indicators* (Nardo *et al*., 2005).

The most disturbing problem about choosing appropriate weights is that the ranking of objects can eventually be altered when choosing different weighting schemes. If this happens, the reliability of the rankings might be put into question and raise objections or concerns. This is particularly true in the case of international country rankings, that have typically attracted the attention of the media, the academic community, and the different national governments. Confronted with such a daunting task, if a decision maker is uncertain about the merit or appropriateness of a specific weighting scheme, she might prefer to allow for a certain degree of underspecification in the process of choosing weights. In this paper, we will mainly focus on the country rankings arising from indices like the Human Development Index (HDI) and other United Nations Development Program (UNDP) composite indices, but our results are applicable to many other contexts as well.

Facing the problem of finding weighting schemes that would meet conflicting ethical/normative imperatives, Sen (1992) proposed an alternative approach consisting of allowing for a certain degree of underspecification for those weights in which full agreement had not been reached. Foster and Sen (1997, p. 206) state that while "the possibility of arriving at a unique set of weights is rather unlikely, that uniqueness is not really necessary to make agreed judgements in many situations." For instance: if three decision makers claim that the weight that should be attributed to the education component of a well-being composite index is 0.2, 0.3, and 0.4, then they all agree that the weight should not fall below 0.2 nor exceed 0.4. This approach can be very useful when decision-makers are uncertain about the appropriateness of a given weighting scheme. However, this comes at the cost of

obtaining partial orders in which certain couples of objects cannot be ranked vis-à-vis each other. An important limitation of this approach is that nothing is known about the degree of incompleteness of the partial ranking that arises when the weights are allowed to move within a given range. Moreover, it is clear that the larger the decision-maker's uncertainty and the corresponding degree of weights underspecification, the smaller the number of couples of objects that can be robustly ranked vis-à-vis each other. In this context, we contend that it would be very useful for a decision maker to have tools to measure precisely these trade-offs. The main purpose of this paper is to provide very precise answers to these relevant issues by exploring the extent to which the ranking that arises from the choice of a specific weighting scheme is sensitive (i.e. robust) to small/local variations of its values. We contend that the tools presented in this paper can be very useful for policy makers and empirical practitioners to assess the degree of reliability of the rankings they are dealing with.

In order to explore the extent to which the ranking of objects is robust to the choice of alternative weighting schemes, many empirical studies have typically opted for taking a list of different "reasonable" weighting schemes and comparing the corresponding results. This approach is known in the literature as "sensitivity analysis" and has been used extensively in empirical studies. However, we contend that the choice of any of those lists in sensitivity analysis is arbitrary in itself, since one might argue that there can always be other more or less reasonable criteria that would suggest inclusion of yet another significant weighting scheme to the list. Even worse, one might wonder whether the inclusion of other weighting schemes to the list might dramatically alter the corresponding ranking or not. Hence, "traditional" sensitivity analysis is an inherently incomplete technique that only scratches the surface of the problem. In this paper we propose a more holistic approach in which all possible weighting schemes are taken into account to truly assess the robustness of a given ranking.

In recent years there have been different contributions in the literature that have dealt with the issue of sensitivity analysis for composite indices. Saisana *et al.* (2005), for instance, have used Monte Carlo simulation techniques to derive probability distributions of the values of composite indices when some factors that are necessary for their derivation (weighting schemes, aggregation function, normalization method, selection of subindicators, and so on) are allowed to change. On the other hand, Cherchye *et al.* (2008) have studied the robustness of HDI-like rankings when the weights are allowed to vary in different regions *and* the aggregation function is allowed to be any *S*-concave index. Finally, Foster *et al.* (2009) follow a similar approach to the one presented in this paper by exploring the extent to which a ranking is incomplete when the weights are allowed to move in certain sets of "admissible weights." However, as we will later see, these sets of admissible weights and the corresponding robustness measure they derive are somewhat arbitrarily chosen, and their results are only valid for linear aggregation functions. In this paper we will discuss in detail the implications of choosing different sets of admissible weights and derive a robustness function that is valid for *any* of them. In particular, this allows us to expand the robustness analysis to the case where we use any member of the generalized weighted means to aggregate between dimensions.

In Section 2, we will start exploring the consequences of taking the whole set of weighting schemes into account when ranking objects using the ordinary weighted arithmetic means. In particular, this information will be used to construct the so-called *Robustness Function* and the *Confidence Value* associated to a given weighting scheme, that measure the extent to which the ranking arising from the choice of that specific weight is sensitive/robust to changes in its original values. We provide empirical illustrations using the values of the Human Development Index, the Gender-related Development Index (GDI), and the Gender Empowerment Measure (GEM). In Section 3 we extend the ideas presented in the previous section to the context where the composite index that is used to rank objects is a generalized weighted mean. Among many other things, this allows one to explore the level of robustness associated to a given weighting scheme for the country ranking that arises from the use of other UNDP composite indices, like the Human Poverty Index. Section 4 is devoted to a critical discussion of the notions and results presented in the previous sections, together with some concluding remarks.

2. Robust Comparisons

First, let us introduce some general notations that will be used throughout the paper. We assume that we are ranking $n \in N$ objects $\{C_1, \ldots, C_n\}$ (typically countries) and that for each of these objects we have $k \in N$ different attributes (N is the set of natural numbers). These attributes are measured by quantitative individual indicators that are later used to rank the objects. For instance, in the case of the Human Development Index, one has $k = 3$ attributes, namely: health, education, and income. We will denote by I_{ij} the value of attribute "j" for object "i". The vector $I_{i*} := (I_{i1}, \ldots, I_{ik})$ contains the information of the k attributes for object i, and is called the *achievement vector*. Thus, the information used to rank the n objects is an $n \times k$ matrix. If we assume that each attribute contributes in the same direction to the evaluation of the different objects for the purpose at hand,[1] each weight should be non-negative, so the whole set of normalized weighting schemes for k attributes is

$$\Delta_k := \left\{ (w_1, \ldots, w_k) \in R^k \mid \sum_{i=1}^{k} w_i = 1 \quad \textit{and} \quad w_i \geq 0 \ \forall i \in \{1, \ldots, k\} \right\} \tag{1}$$

which is the standard $(k-1)$-dimensional simplex of the Euclidean space in R^k. The vertices of the simplex will be denoted by $\boldsymbol{e}_1 = (1,0, \ldots, 0), \ldots, \boldsymbol{e}_k = (0, \ldots, 0, 1)$. This will be our weights domain. Moreover, for each $i, j \in \{1, \ldots, n\}$, we will define the following sets

$$\Delta_k^{ij} := \{ (w_1, \ldots, w_k) \in \Delta_k \mid C_i \textit{ is not ranked below } C_j \} \tag{2}$$

$$\Delta_k^{ji} := \{ (w_1, \ldots, w_k) \in \Delta_k \mid C_j \textit{ is not ranked below } C_i \} \tag{3}$$

[1]For example: if we are assessing an individual's well-being, the different attributes are assumed to contribute positively to their well-being. If this is not the case, the indicators measuring the corresponding attribute can be appropriately rescaled.

which will be used to assess the robustness of the ranking between C_i and C_j. In words: Δ_k^{ij} (Δ_k^{ji}) is the set of weighting schemes for which object C_i (C_j) is not ranked below object C_j (C_i). Now, recall that the definition of these sets will basically depend on the functional form of the composite index we are using to collapse the k individual indicators into a single value. In this section, we start assuming that the function used to aggregate the different attributes/indicators and rank the corresponding objects is the weighted arithmetic mean $\mu_{\boldsymbol{w}}$, with $\boldsymbol{w} = (w_1, \ldots, w_k)$, where

$$\mu_w(I_{i1}, \ldots, I_{ik}) := \sum_{j=1}^{k} w_j I_{ij}. \tag{4}$$

This is a particularly simple and intuitive way of comparing objects that is used in many circumstances. Consider, among others, the UNDP Human Development Index, the Gender-related Development Index, the Gender Empowerment Measure, or a myriad of other quality-of-life related indices. However, in other circumstances, it might make more sense to use the generalized weighted means to compare different objects (take, for instance, the multidimensional well-being, inequality, or poverty indices that want to be sensitive to the tails of the distributions; see Foster *et al.*, 2005). In Section 3, we will extend our results to the generalized weighted means.

Now, using the weighted arithmetic mean $\mu_{\boldsymbol{w}}$, it is clear that

$$\Delta_k^{ij} := \{(w_1, \ldots, w_k) \in \Delta_k \mid \mu_w(I_{i*}) \geq \mu_w(I_{j*})\} \tag{5}$$

$$\Delta_k^{ji} := \{(w_1, \ldots, w_k) \in \Delta_k \mid \mu_w(I_{i*}) \leq \mu_w(I_{j*})\}. \tag{6}$$

Moreover, one has that $\Delta_k = \Delta_k^{ij} \cup \Delta_k^{ji}$. In order to determine these two sets precisely, we need to find $\Delta_k^{ij} \cap \Delta_k^{ji}$. Now, if $\boldsymbol{w} \in \Delta_k^{ij} \cap \Delta_k^{ji}$, one must have that $\mu_w(I_{i*}) = \mu_w(I_{j*})$ (that is, the objects C_i and C_j have the same score), so it is straightforward to check that

$$\Delta_k^{ij} \cap \Delta_k^{ji} := \left\{(w_1, \ldots, w_k) \in \Delta_k \mid \sum_{l=1}^{k} w_l (I_{il} - I_{jl}) = 0\right\}. \tag{7}$$

Hence, $\Delta_k^{ij} \cap \Delta_k^{ji}$ is a linear hyperplane embedded in Δ_k, so the sets Δ_k^{ij}, Δ_k^{ji} are convex polyhedra of R^k that can be described by giving the list of corresponding vertices. It is important to point out that the set $\Delta_k^{ij} \cap \Delta_k^{ji}$ is empty when either I_{i*} or I_{j*} strictly vector-dominates the other[2] (in that case, there is no weighting scheme that could eventually reverse the ranking, so Δ_k^{ij}, Δ_k^{ji} must be either the empty set or the whole simplex Δ_k). In order to simplify the notation, from now on we will simply write $H(i,j)$ instead of $\Delta_k^{ij} \cap \Delta_k^{ji}$.

In some cases, one might be interested in having a graphical representation of the sets Δ_k^{ij}, Δ_k^{ji}. Consider the following illustrative example that uses data from the

[2]We say that $(x_1, \ldots, x_k)$ vector-dominates $(y_1, \ldots, y_k)$ when $x_i \geq y_i$ for all $1 \leq i \leq k$. When $x_i > y_i$ for all $1 \leq i \leq k$, we say that $(x_1, \ldots, x_k)$ strictly vector-dominates $(y_1, \ldots, y_k)$.

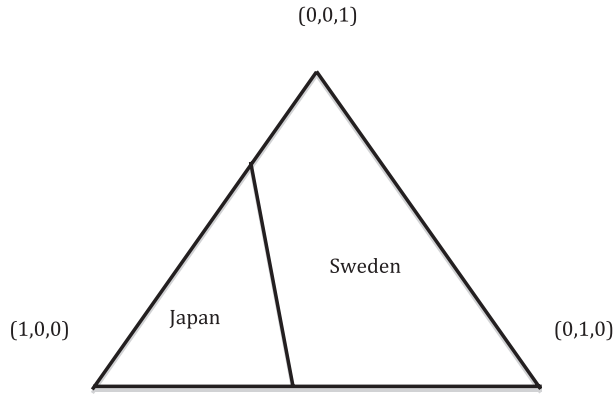

Figure 1. The Simplex Δ_3 Divided in Two Polyhedrical Regions

Human Development Index (year 2006). The three components of the HDI for Sweden are (0.928, 0.974, 0.973), and the corresponding components for Japan are (0.957, 0.949, 0.962). With the classical equal weighting assumption, Sweden gets a score of (1/3)(0.928 + 0.974 + 0.973) = 0.958, so it is ranked above Japan, that gets a score of (1/3)(0.957 + 0.949 + 0.962) = 0.956. However, if one picked $\boldsymbol{w} = (1/2, 1/4, 1/4)$, Sweden would score 0.951 and Japan 0.956, so the ranking would be reversed. More generally, we are interested in knowing which is the set of weights for which Sweden is ranked above Japan and vice versa; they are schematically shown in Figure 1.

The weights for which Sweden (Japan) is ranked above Japan (Sweden) are the ones at the right (left) hand side of the hyperplane. The equation of the hyperplane that divides both sets of weights is equal to $-0.029w_1 + 0.025w_2 + 0.011w_3 = 0$. Similar figures and ideas to the ones discussed in this subsection have already been proposed in Permanyer (2007) and Foster *et al.* (2009).

2.1. *Rank Robustness for a Couple of Objects*

In the previous subsection, we saw that when the achievement vectors I_{i*}, I_{j*} associated to a couple of objects C_i, C_j do not strictly vector-dominate each other, the sets of weights for which one object is ranked above the other Δ_k^{ij}, Δ_k^{ji} are non-empty convex polyhedra separated by a linear hyperplane $H(i,j)$. We introduce now the following intuitive ideas. Suppose that one wants to rank a couple of objects (C_i,C_j) privileging an initial weighting scheme $\boldsymbol{w} = (w_1, \ldots, w_k)$. If this weighting scheme is "very close" to the hyperplane $H(i,j)$, then the ranking between C_i and C_j can be loosely judged as "non-robust," because a slight variation in the values of $\boldsymbol{w}$ could lead to a reversal of the ranking. By the same token, if $\boldsymbol{w}$ happens to be "far away" from $H(i,j)$, the ranking between C_i and C_j can be loosely judged as "robust," because small variations in the values of $\boldsymbol{w}$ would not lead to a ranking reversal. Put in other words: we want to rank couples of objects but allow the weighting schemes to move within a given "reasonable" admissible set of weights neighboring $\boldsymbol{w}$. If the ranking between C_i and C_j is the same for all weights

included in a "large" neighborhood of $\boldsymbol{w}$, then the ranking could be loosely labeled as "robust." Analogously, if the ranking between C_i and C_j is reversed by different weights included in a "small" neighborhood of $\boldsymbol{w}$, then the ranking would be "non-robust."

Given the fact that the choice of weights is such a delicate and controversial issue on which it is so difficult to reach a universal consensus, it seems legitimate to wonder what happens with the corresponding rankings when we take into account not only a specific weighting scheme but also a set of neighboring weights. In other words, we want to measure the extent to which the ranking ensuing from the weighting scheme $\boldsymbol{w}$ is sensitive to small/local variations of its values. This way, we are making room for some underspecification in the aggregation procedure by allowing the weighting schemes to move within a given range that might be deemed "reasonable" by different decision-makers who are uncertain about the appropriateness of the initial weighting scheme (see Foster and Sen, 1997, p. 206).

2.1.1. Neighborhood Systems

In order to operationalize the previous intuitive ideas and make them more precise, we need to introduce some formal definitions.

Definition 1. A *Neighborhood System* for the simplex Δ_k is the following disjoint union:

$$\boldsymbol{N} := \bigcup_{w \in \Delta_k} \boldsymbol{N}_{\boldsymbol{w}}$$

where each $\boldsymbol{N}_{\boldsymbol{w}}$ is a set of closed, nested, dense, and bounded neighborhoods of $\boldsymbol{w}$ (see the Appendix for a more formal definition).

A Neighborhood System is thus a collection of closed neighborhoods for each weighting scheme $\boldsymbol{w}$ in the simplex, with the different properties representing some basic intuitions that seem quite unexceptionable to us for the problem at hand. Basically, each $U \in \boldsymbol{N}_{\boldsymbol{w}}$ represents a set of weighting schemes whose specific definition might depend on what is meant by "a set of weights around or close to $\boldsymbol{w}$." When a decision maker is uncertain about the appropriateness of a given weighting scheme $\boldsymbol{w}$, she might want to consider instead some of its neighborhoods $U \in \boldsymbol{N}_{\boldsymbol{w}}$: the larger the uncertainty about a given $\boldsymbol{w}$, the larger neighborhood she might want to take into account. In this paper, we will mainly use a neighborhood system $\boldsymbol{N}$ in which the sets $\boldsymbol{N}_{\boldsymbol{w}}$ are defined as homothetic contractions of the simplex Δ_k toward a given $\boldsymbol{w}$: it will be referred to as $\boldsymbol{N}^{\mathrm{F}}$ as has been used in Foster *et al.* (2009).

Having defined a neighborhood system to make room for different degrees of weights underspecification, the following natural step is to impose some cardinal structure that allows one to keep track of the pace at which underspecification increases and the different neighborhoods fill the whole simplex. This is a necessary step if a decision maker wants to allow for specific degrees of weight underspecification that can be measured in a ratio scale. For that purpose, we introduce the following definition.

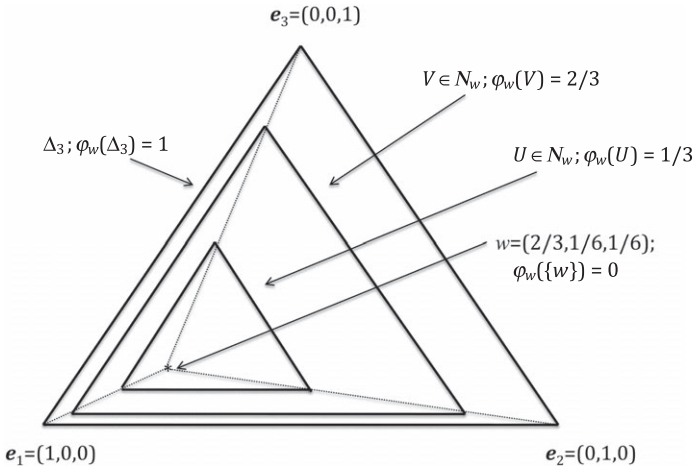

Figure 2. Neighborhood System $\mathcal{N}^F$ for $\boldsymbol{w} = (2/3,1/6,1/6)$

Definition 2. For all $\boldsymbol{w} \in \Delta_k$ and all $U \in \boldsymbol{N}_w$ define $\varphi(U) := m(U)/m(\Delta_k)$, where $m(.)$ is a standard measure of the size of a set[3] in R^k.

By definition, the level of underspecification $\varphi(U)$ associated to any neighborhood $U \in \boldsymbol{N}_w$ will be equal to its relative size with respect to Δ_k. In other words, if a neighborhood $U \in \boldsymbol{N}_w$ contains a proportion "p" of the weights included in the whole simplex, "p" will be its underspecification level. According to this parametrization system, to make an x-fold increase of the underspecification level of a given neighborhood one has to take a new neighborhood that is x times as large as the original one. In Figure 2 we show an illustrative example of $\mathcal{N}^F$ for the case $k = 3$ when $\boldsymbol{w} = (2/3,1/6,1/6)$.

2.1.2. Robustness Levels

With the ideas introduced in the previous section it is now possible to present a formal definition of the robustness of the ranking for a couple of objects C_i, C_j. Assume that the corresponding achievement vectors I_{i^*}, I_{j^*} do *not* strictly vector dominate each other, so that the linear hyperplane $H(i,j)$ is non-empty.

Definition 3. Fix any initial/privileged weighting scheme $\boldsymbol{w} \in \Delta_k$. The *robustness level* of the ranking between C_i, C_j is defined as $\varphi(U)$, where U is the smallest neighborhood belonging to $\boldsymbol{N}_w$ having non-empty intersection with $H(i,j)$. This robustness level will be denoted as ρ.

The intuition behind this measure is the following. Consider the set of enlarging neighborhoods around $\boldsymbol{w}$ ($\boldsymbol{N}_w$). When they become gradually large, one of

[3]Hence, $m(.)$ measures standard lengths, areas, and volumes in one, two, and three dimensional spaces, respectively. A more rigorous definition of the size function is given by $m(U) = \int_{R^k} 1_U \, dx_1 \ldots dx_k$, where $1_U : R^k \to \{0,1\}$ is the indicator function defined for all $(x_1, \ldots, x_k)$ in the k-dimensional Euclidean space R^k that takes a value of 1 if $(x_1, \ldots, x_k)$ belongs to U and 0 otherwise.

them, say U, eventually intersects $H(i,j)$. The robustness level ρ simply measures the relative size of this neighborhood with respect to the whole simplex Δ_k. This way, the number ρ informs the decision-maker about the relative size of the maximal set of weights in the neighborhood system $\boldsymbol{N}_{\boldsymbol{w}}$ that does not reverse the ranking between C_i and C_j. With this definition, we are capturing in a direct way the intuitions about robustness introduced at the beginning of Section 2.1. Another way of motivating the new robustness measure is the following: if each weighting scheme in the simplex is interpreted as a value judgment about the importance of the different k dimensions vis-à-vis each other, then ρ can be interpreted as a *consensus degree* measure indicating the extent to which the ranking between C_i and C_j is supported not only by the original value judgment represented by $\boldsymbol{w}$ but also by its neighboring value judgments represented in $U \in \boldsymbol{N}_{\boldsymbol{w}}$.

In Foster *et al.* (2009), the authors propose another robustness measure[4] denoted by r^*. There are differences between the two approaches; among the most relevant ones are that: (1) r^* has only been defined for a specific neighborhood system $\boldsymbol{N} = \boldsymbol{N}^{\mathrm{F}}$; and (2) r^* is not well defined when the aggregation function is not linear. More details will be given in Section 3.

2.2. *The Robustness Function and the Confidence Value*

In the previous subsections, we focused on the robustness of the ranking between a single couple of objects: C_i and C_j. Now, we turn our attention to the entire collection of comparisons that can be made between the n objects $\{C_1, \ldots, C_n\}$. In this case, we want to assess the extent to which *the whole* ranking associated to $\boldsymbol{w}$ is sensitive (i.e. robust) to small changes in the values of the initial weighting scheme. To start with, we will define the set of all hyperplanes crossing Δ_k.

$$\boldsymbol{H} := \{H(i,j) \quad \forall i, j \in \{1, \ldots, n\}\}. \tag{8}$$

Recall that whenever the achievement vectors I_{i^*}, I_{j^*} strictly vector-dominate each other, the corresponding $H(i,j)$ is empty, so the number of hyperplanes crossing Δ_k is equal to the number of couples of achievement vectors (I_{i^*}, I_{j^*}) for which there is no strict vector dominance. This number will be denoted by $|\boldsymbol{H}|$. If we carry out the analysis of the previous subsection for each couple of comparisons that is not completely robust in our dataset, we obtain a complete list of $|\boldsymbol{H}|$ robustness levels; it will be written as $\{\rho_1, \ldots, \rho_{|H|}\}$.

With all this information at hand, we can now introduce the following definition.

[4]Using the notation in Foster *et al.* (2009), r^* is defined as $r^* := \Delta_0/(\Delta_0 + \Delta_m)$. Using the notation of this paper, r^* can be rewritten as

$$r^* = \frac{\mu_w(I_{i^*}) - \mu_w(I_{j^*})}{\mu_w(I_{i^*}) - \mu_w(I_{j^*}) + \max\limits_{w \in \Delta_k}[\mu_w(I_{j^*}) - \mu_w(I_{i^*}), 0]}.$$

Definition 4. *The Robustness Function associated to a weighting scheme* $\boldsymbol{w} \in \Delta_k$ is a function that for any $r \in [0,1]$ takes the value

$$R_w(r) := \frac{\left|\{\rho_i \in \{\rho_1, \ldots, \rho_{|H|}\} \mid \rho_i \leq r\}\right|}{\left(\frac{n(n-1)}{2}\right)}. \tag{9}$$

For a given value $r \in [0,1]$, the Robustness Function associated to $\boldsymbol{w}$ is counting the share of the $n(n-1)/2$ comparisons whose robustness levels are at most r. Put in other words: the value of $R_w(r)$ is the share of the $n(n-1)/2$ comparisons that can not be ranked unambiguously when the set of admissible weights is the neighborhood $U \in \boldsymbol{N}_w$ with underspecification level $\varphi(U) = r$.

Definition 5. Consider an alternative weighting scheme $\boldsymbol{v} \neq \boldsymbol{w}$ and a number $r^+ \in [0,1]$. We say that *the ranking associated to* $\boldsymbol{w}$ *is more robust up to level* r^+ *than the ranking associated to* $\boldsymbol{v}$ if and only if $R_w(r) \leq R_v(r)$ for all $r \in [0, r^+]$.

In words, for all levels of robustness between 0 and a given r^+, the share of comparisons that exhibit at most a r-level of robustness is always lower for the ranking associated to $\boldsymbol{w}$. The number r^+ can be interpreted as the range of underspecification of the original weighting scheme that one wants to allow for the case at hand. Clearly, the criterion just presented can also be used when we want to compare robustness functions coming from different datasets (even if the initially privileged weights $\boldsymbol{w}$, $\boldsymbol{v}$ are the same). Recall that this criterion to order the rankings associated to different weighting schemes in terms of robustness is incomplete, because the condition $R_w(r) \geq R_v(r)$ or $R_v(r) \geq R_w(r)$ might not hold for all $r \in [0, r^+]$. In those cases, the ranking between $\boldsymbol{w}$ and $\boldsymbol{v}$ in terms of robustness remains ambiguous, so other criteria might be necessary to chose between them. A simple and complete criterion to compare the robustness levels of different rankings is given in the following definition.

Definition 6. The *Confidence Value* of the ranking associated to $\boldsymbol{w}$ is defined as

$$C_w := 1 - \int_0^1 R_w(r)\, dr. \tag{10}$$

This is the area above the robustness function within the unit square. By definition, the confidence value C_w associated to a ranking is a number between 0 and 1 that should be interpreted as a measure of the extent to which the ranking arising from $\boldsymbol{w}$ is reliable or not in terms of its sensitivity to the changes in the values of $\boldsymbol{w}$. Recall that the area above the robustness function, $1 - \int_0^1 R_w(r)\, dr$, can take values between 0 and 1. When the values of C_w are small (near 0), the values of the robustness function $R_w(r)$ are very large, so the ranking arising from $\boldsymbol{w}$ can not be trusted very much because slight changes in the values of $\boldsymbol{w}$ lead to large changes in the corresponding ranking. On the other hand, if the values of C_w are big (near 1), the values of the robustness function $R_w(r)$ are very small, so the

ranking arising from $\boldsymbol{w}$ is highly reliable because the changes in the values of $\boldsymbol{w}$ do not lead to large changes in the corresponding rankings. The confidence value $C_{\boldsymbol{w}}$ can only be equal to the maximal level of 1 when $R_{\boldsymbol{w}}(r) = 0$ for all $r \in [0,1]$—that is, when there is strict vector dominance between all couples of achievement vectors and no weighing scheme can change any ranking. Clearly, if one has that $R_{\boldsymbol{w}}(r) \leq R_{\boldsymbol{v}}(r)$ for all $r \in 0, 1]$, then $C_{\boldsymbol{w}} \geq C_{\boldsymbol{v}}$. However, the opposite is not necessarily true.[5]

At this point, we will briefly comment on some basic properties satisfied by $R_{\boldsymbol{w}}(r)$ and $C_{\boldsymbol{w}}$.

Anonymity: Our measures do not depend on the labels of the objects $\{C_1, \ldots, C_n\}$ and each of them has the same importance. This precludes other criteria that might give more preponderance to certain objects (e.g. like prioritizing countries with larger populations).

Common-slope affine transformation invariance: If all scores in all achievement vectors I_{i^*} are scaled up or down by some positive constant or if the same constant is added up to all scores in all achievement vectors I_{i^*}, the values of $R_{\boldsymbol{w}}(r)$ and $C_{\boldsymbol{w}}$ remain unaffected. In this framework, ratio scaling and weighing are formally equivalent and indistinguishable procedures.

Responsiveness: If some $H(i,j) \in \boldsymbol{H}$ is/are relocated further away from $\boldsymbol{w}$ (by modifying the corresponding achievement vectors), then the modified robustness function will take lower values and the modified confidence value will be higher.

Illustrative Examples

In order to illustrate the previous ideas, we show some examples of robustness functions and confidence values using data from the Human Development Index, the Gender-related Development Index, and the Gender Empowerment Measure (all obtained from the UNDP Human Development Report 2006). The HDI uses the same weight for the three components (health, measured by life expectancy; education, measured by literacy and gross enrolment rates; and income, measured by the per capita Gross Domestic Product) to rank 179 countries all over the world. The GDI measures achievement in the same basic capabilities as the HDI does, but takes note of inequality in achievement between women and men. In other words: it is simply the HDI discounted, or adjusted downwards, for gender inequality. The GDI uses the same weighting scheme for the same three HDI components to rank 140 different countries. The GEM is a measure of agency. It evaluates progress in advancing women's standing in political and economic forums. It examines the extent to which women and men are able to actively participate in economic and political life and take part in decision-making. The GEM has three components that are weighted equally: the political participation and decision-making component (measured with the female and male shares of parliamentary seats), the economic participation and decision-making component (measured with female and male shares of positions as legislators, senior officials,

[5]Recall that these criteria of comparing the relative position of the robustness functions curves or the area under these curves is reminiscent of the well-known criteria to rank income distributions in terms of inequality by comparing the relative positions of the corresponding Lorenz curves or the value of the corresponding Gini indices.

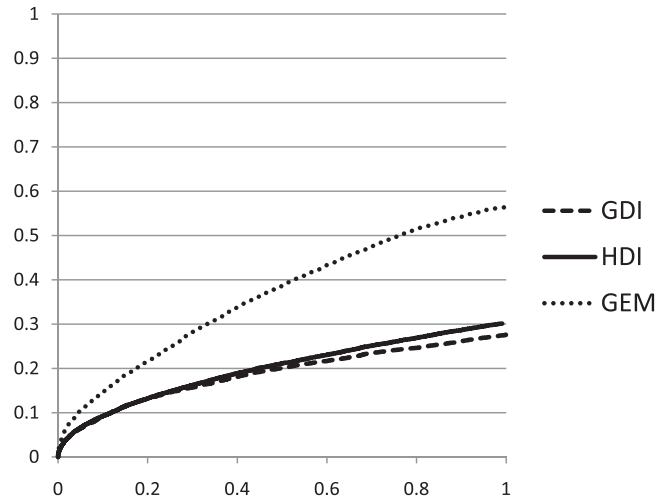

Figure 3. Robustness Functions Using ρ for the HDI, GDI, and GEM Rankings
Source: Author's calculations using UNDP data (2006).

and managers, and shares of professional and technical positions) and the power over economic resources component (measured with the female and male estimated earned income). The necessary data to compute the GEM is available for 107 countries. In Figure 3 we show the graphs of the corresponding robustness functions associated to the weighting scheme $\boldsymbol{w} = (1/3,1/3,1/3)$ using the robustness measure proposed in this paper.

The shapes of the curves are very similar in the cases of the HDI and GDI. This is not surprising for at least two reasons. First, both indices are based on the same three components and measured using exactly the same variables. Second, the difference between the HDI and GDI is only due to the existing gender gaps in the corresponding dimension. However, as shown by Bardhan and Klasen (1999), the penalty imposed for the existing gender gaps using the GDI method is relatively small (the GDI values being on average only 9 percent smaller than their HDI counterparts), so the values of both composite indices are highly correlated. When r approaches its maximum value of 1, the robustness functions for the HDI and GDI approach the values of 0.3 and 0.276: this means that the percentage of comparisons that are not completely robust are 30 and 27.6 percent, respectively. The shape of both curves is very similar: they are concave from $r = 0$ to $r \approx 0.4$ and then they are roughly linear. Moreover, it can be shown that the robustness function associated to the GDI is *not* always smaller than the robustness function associated to the HDI, so the ranking associated to the GDI can not be judged as being more robust than the HDI ranking according to Definition 5. The confidence values for the HDI and GDI rankings are 0.81 and 0.82, respectively (out of a maximum of 1), so the GDI ranking can be considered to be (slightly) more robust according to Definition 6.

On the other hand, the shape of the robustness function associated to the GEM is very different. To start with, the percentage of comparisons that are not completely robust is much larger: 56.3 percent. Moreover, it is clear from the graph

that the robustness function associated to the GEM is always larger than the HDI and GDI counterparts. This means that the ranking associated to the GEM is less robust than the HDI and GDI rankings (according to Definition 5) for all admissible values of r. The confidence value for the GEM ranking is equal to 0.62. The fact that the GEM uses a set of completely different variables that are not so highly correlated between them (as in the cases of the GDI and HDI) might be an explanation for this noticeably different behavior. In Section 4 we will discuss these and other related issues in more detail.

3. Extensions for the Generalized Weighted Means

An important criticism directed against the HDI is the perfect substitutability between alternative dimensions. This way, it is possible to compensate a unit loss of health with a unit gain of education, and so on. More generally, if one uses the arithmetic weighted mean μ_w, the degree of substitution between a couple of dimensions is always the same irrespective of the corresponding levels of achievement. This means that a gain or a loss of a unit in a given dimension will have the same overall effect on the aggregate value of the index, no matter whether the achievement level in that specific dimension is high or low. Clearly, there are many circumstances in which this is a non-realistic and non-desirable property: for instance, a gain of one year of life expectancy is much more important in a country with a life expectancy of 35 than in one with a life expectancy of 80, and so on. One simple way of eliminating this unrealistic assumption is by adopting the *generalized weighted means* μ_w^α. The vast literature on multidimensional well-being, poverty, or inequality measurement provides many examples that illustrate the usefulness of this kind of measures (see, for instance, Foster *et al.*, 2005). They are defined as follows:

$$\mu_w^\alpha(I_{i1},\ldots,I_{ik}) := \begin{cases} \left(w_1 I_{i1}^\alpha + \ldots + w_k I_{ik}^\alpha\right)^{1/\alpha} & for \quad \alpha \neq 0 \\ I_{i1}^{w_1} \ldots I_{ik}^{w_k} & for \quad \alpha = 0 \end{cases}. \tag{11}$$

Parameter α indicates the extent to which μ_w^α emphasizes the upper or the lower ends of the distribution $(I_{i1}, \ldots, I_{ik})$. When $\alpha = 1$, we have the classical weighted arithmetic mean. When $\alpha = 0$ we have the geometric mean, while $\alpha = -1$ yields the harmonic mean. As α approaches (minus) infinity, μ_w^α tends to the maximum (Rawlsian minimum) of the distribution $(I_{i1}, \ldots, I_{ik})$.

The purpose of this section is to extend the robustness analysis introduced in the previous section to those contexts in which the generalized weighted means are used to average the different achievement vectors I_{i*} and rank the n objects.

As before, we want to assess the robustness of the ranking between objects C_i and C_j (with achievement vectors I_{i*}, I_{j*}) by means of the corresponding Δ_k^{ij}, Δ_k^{ji} and $\Delta_k^{ij} \cap \Delta_k^{ji}$. Now, it is clear that

$$\Delta_k^{ij} := \{(w_1,\ldots,w_k) \in \Delta_k \mid \mu_w^\alpha(I_{i*}) \geq \mu_w^\alpha(I_{j*})\} \tag{12}$$

(13) $$\Delta_k^{ij} \cap \Delta_k^{ji} := \{(w_1, \ldots, w_k) \in \Delta_k \mid \mu_w^\alpha(I_{i*}) = \mu_w^\alpha(I_{j*})\}.$$

Hence, if $\boldsymbol{w} \in \Delta_k^{ij} \cap \Delta_k^{ji}$ and $\alpha \neq 0$, one must have that $\left(\sum_{l=1}^{k} w_l I_{il}^\alpha\right)^{1/\alpha} = \left(\sum_{l=1}^{k} w_l I_{jl}^\alpha\right)^{1/\alpha}$. Then, if $\alpha \neq 0$, it is clear that

(14) $$\Delta_k^{ij} \cap \Delta_k^{ji} := \left\{(w_1, \ldots, w_k) \in \Delta_k \mid \sum_{l=1}^{k} (I_{il}^\alpha - I_{jl}^\alpha) = 0\right\}.$$

On the other hand, if $\boldsymbol{w} \in \Delta_k^{ij} \cap \Delta_k^{ji}$ and $\alpha = 0$, one must have that $I_{i1}^{w_1} \ldots I_{ik}^{w_k} = I_{j1}^{w_1} \ldots I_{jk}^{w_k}$. Taking logarithms to the last expression one has that $\sum_{l=1}^{k} w_l \ln(I_{il}) = \sum_{l=1}^{k} w_l \ln(I_{jl})$. Hence, if $\alpha = 0$,

(15) $$\Delta_k^{ij} \cap \Delta_k^{ji} := \left\{(w_1, \ldots, w_k) \in \Delta_k \mid \sum_{l=1}^{k} w_l (\ln(I_{il}) - \ln(I_{jl})) = 0\right\}.$$

In both cases, as one can see in (14) and (15), $H(i, j) = \Delta_k^{ij} \cap \Delta_k^{ji}$ is a linear hyperplane, so Δ_k^{ij}, Δ_k^{ji} are convex polyhedra as in the previous section. Hence, the concept of *robustness level* for a couple of objects C_i and C_j introduced in Definitions 3 and 4 can also be applied in this context. Analogously, the definitions of Robustness Function $R_w(r)$ associated to a weighting scheme $\boldsymbol{w}$ and of Confidence Value C_w apply as well. The only difference in this context is that the equations of the hyperplanes included in $\boldsymbol{H}$ are now given by (14) and (15).

Illustrative Examples

The previous ideas will now be illustrated with the robustness functions for a couple of empirical examples. The first one uses data from the Human Poverty Index (HPI), which is another well-known index published yearly in the Human Development Reports. In this example, we will focus specifically on the HPI-I (HPI from now on), which is a deprivation index defined for the context of developing countries (one has also the HPI-II, which has been defined for the context of highly developed societies: see the Human Development Reports for more details). The HPI concentrates on the deprivation in the three essential elements of human life already reflected in the HDI: longevity, knowledge, and a decent standard of living. The first component is measured with the probability at birth of not surviving to the age 40 (P_1), the second with the adult illiteracy rate (P_2), and the third is measured as the average of the percentage of the population not using an improved water source and the percentage of children under weight-for-age (P_3). The formula of the HPI is given by

(16) $$HPI := \left(\frac{1}{3}\left(P_1^\alpha + P_2^\alpha + P_3^\alpha\right)\right)^{1/\alpha}.$$

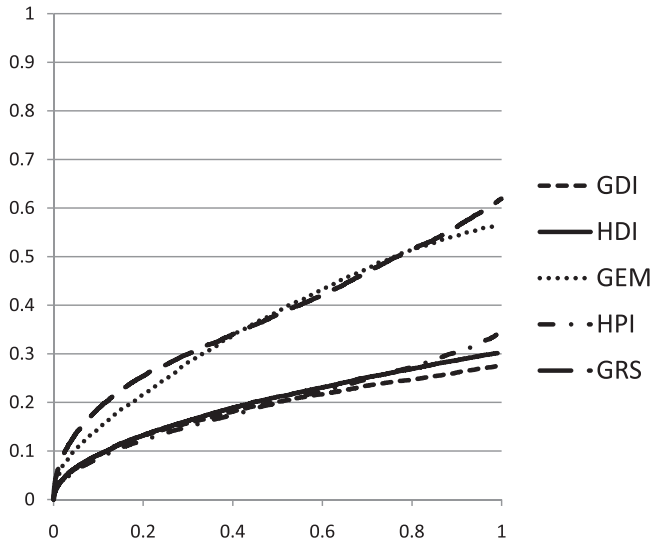

Figure 4. Robustness Functions Using ρ for the GEM, HDI, GDI, HPI, and GRS Rankings
Source: Author's calculations using UNDP data (2006).

For practical purposes, and in order to give more importance to the dimension where the level of deprivation is higher, the value of α chosen by UNDP is 3. Clearly, this is an example of a generalized mean with $\alpha = 3$.

The second index for which we will compute a robustness function is less well-known but is related in spirit to the indices published in the Human Development Reports. It is called Gender Relative Status (GRS) and it measures the average gender gap for the same well-being dimensions included in the definition of the HDI disaggregated by sex. It is defined as

$$GRS := \left(\frac{x_1}{y_1}\right)^{w_1}\left(\frac{x_2}{y_2}\right)^{w_2}\left(\frac{x_3}{y_3}\right)^{w_3} \tag{17}$$

where x_i, y_i for $i = 1,2,3$ are the average achievement levels for women and men in the same three dimensions included in the HDI. This is an example of a generalized mean with $\alpha = 0$. The GRS has been defined as an alternative to the GDI in order to measure the average levels of gender inequality.[6] In Figure 4 we show the robustness functions using ρ for the HPI and GRS indices privileging the initial weighting scheme $\boldsymbol{w} = (1/3,1/3,1/3)$ together with the functions shown in Figure 3 to make comparisons easier.

The robustness function for the HPI is similar to the HDI and GDI robustness functions when the values of r are under 0.6; they are concave. For the values of r between 0.6 and 1, the function is convex and increases more rapidly than its

[6]It should be pointed out that the GDI is *not* measuring gender inequality in itself. It is rather measuring the overall development levels in a given country and correcting them downwards by the existing levels of gender inequality.

counterparts. When r reaches its maximum value of 1, the robustness function approaches the value of 0.35. This means that 35 percent of all possible country comparisons by means of the HPI could be reversed if one were allowed to move the admissible weighting schemes within the whole simplex. This is a similar value to the ones obtained for the HDI and GDI (30 and 27.6 percent, respectively). The confidence value of the HPI ranking is equal to 0.78 (out of a maximum of 1).

On the other hand, the robustness function for GRS takes much higher values than the first three functions for all values of r. It is roughly similar to the GEM robustness function, crossing it a couple of times. It is concave for all r below 0.5 and convex from 0.5 onwards. When r reaches its maximum threshold of 1, the value of $R_w(r)$ approaches 0.62, so 62 percent of all possible country comparisons are not completely robust. The confidence value of the GRS ranking is equal to 0.6.

4. Discussion and Concluding Remarks

In this section we will critically discuss different topics related to the meaning and relevance of the results obtained so far. We will start by comparing the shapes of the different robustness functions associated to the different indices. In order to distinguish between them we will write $R_{w,HDI}$, $R_{w,GDI}$, $R_{w,GEM}$, $R_{w,HPI}$, and $R_{w,GRS}$. Looking at Figure 4, we can roughly distinguish two groups of functions on the grounds of their similarity: on the one hand we have $R_{w,HDI}$, $R_{w,GDI}$, $R_{w,HPI}$, and on the other hand we have $R_{w,GEM}$, $R_{w,GRS}$. The first group is characterized by the relatively low values of the functions for all robustness levels (r): roughly speaking, around two thirds of all possible comparisons between couples of countries are fully robust (that is, they do not depend at all on the privileged weighting scheme). This means that these rankings are fairly stable and not very sensitive to the choice of alternative weighting schemes. The confidence values of these rankings are around 0.8 out of a maximum of 1. Within the group, all three curves cross between them, so none of the rankings can be judged as being the most robust according to Definition 5. On the other hand, according to Definition 6, the most robust ranking is the one associated to the GDI, followed by the HDI and then by the HPI. The second group is characterized by relatively high values of the functions for all robustness levels (r): roughly speaking, around two thirds of all possible comparisons between couples of countries are *not* fully robust (that is, they could be reversed by choosing alternative weighting schemes). This means that these ranking are not very stable, so they can be altered substantially by privileging alternative weighting schemes. The confidence values of these rankings are slightly above 0.6. Within the group, both curves cross a couple of times, so none of the rankings can be judged as being the most robust according to Definition 5. On the other hand, according to Definition 6, the least robust ranking is the one associated to GRS.

The fact that the robustness functions for the HDI and GDI are so similar has already been discussed in the empirical example of Section 2.2: both indices roughly measure the same concept using the same variables, the latter being only slightly corrected by the existing gender gaps. Now, one might wonder about the similarity between $R_{w,HDI}$, $R_{w,GDI}$, and $R_{w,HPI}$. One preliminary explanation for this

similarity might be the fact that the strong correlation structure between the variables within the HPI mimic roughly the strong correlation structure for the variables included in the HDI (the correlation structure between the HDI variables has been discussed in McGillivray (1991, p. 1462), and its importance for determining robustness levels has been emphasized in Foster *et al*. (2009, p. 16)). This is due to the fact that the dimensions included in both indices are the same and that, as a matter of fact, the variables included in the HPI can be seen as the mirror images of their counterparts in the HDI. The latter uses the life expectancy at birth while the former uses the probability of not surviving to the age of 40; the latter uses the adult literacy rate while the former uses the adult illiteracy rate; and the latter uses the GDP per capita while the former uses the percentage of the population not using an improved water source and the percentage of children under weight-for-age.

Interestingly, the gender related indices GEM and GRS have very different robustness functions that take much lower values for all possible r. Recall that the fact that the GEM uses a set of completely different variables with respect to the HDI, GDI, and HPI does *not* explain in itself this discrepancy between the respective robustness functions; one might have very different composite indices (regarding the theoretical groundings, the dimensions and variables included, and so on) but with very similar robustness functions. As a matter of fact, there is a great discrepancy between $R_{w,GRS}$ and $R_{w,HDI}$, $R_{w,GDI}$, even if the GRS uses in its definition exactly the same basic variables as the HDI and GDI (life expectancy, literacy rates, and GDP per capita). However, it should be pointed out that while the HDI and GDI are indices that measure overall achievement levels (they can be seen as efficiency indicators), the GRS focuses on the relative attainment between women and men irrespective of the overall achievement levels (so it can be seen as an equality indicator). The large values of $R_{w,GRS}$ mean that there are a lot of countries for which the gender gaps in the well-being dimensions included in the HDI run in opposite directions—that is, they sometimes favor men and sometimes favor women.[7]

Regarding the shape of the different curves, one can see that they can be piece-wise linear, convex, or concave. By construction, there is no *a priori* restriction on the shape of the robustness functions: the specific curvature depends on the pace at which the expanding neighborhoods of $\boldsymbol{N}$ intersect the different hyperplanes in $\boldsymbol{H}$. While this completely explains the robustness of a given ranking, it is a somewhat mechanical explanation that might not be very appealing, so it might be interesting to investigate further about other determinants of robustness.

Determinants and Desirability of Robustness

The issue of finding the determinants of robustness has been explored to a large extent in Foster *et al*. (2009, section V). There, the authors discuss the importance of correlation or positive association between dimensions and some data transformations preserving or altering the levels of robustness. There might

[7]Very often, the gender gap in life expectancy favors women while the gender gap in the earned income component favors men.

be other determinants that would be interesting to explore. Consider, for example, the number of dimensions that are included in the composite index. It seems clear that, other things being equal, the larger the number of dimensions, the more difficult it is for strict vector dominance to occur, so the lower the level of robustness. Related to this point, one might wonder whether the comparison between achievement vectors $(I_{i1}, \ldots, I_{ik})$ and $(I_{j1}, \ldots, I_{jk})$ should always yield the same robustness levels as the comparison between the corresponding d-replicated vectors $(I_{i1}, \ldots, I_{i1}, \ldots, I_{ik}, \ldots, I_{ik})$ and $(I_{j1}, \ldots, I_{j1}, \ldots, I_{jk}, \ldots, I_{jk})$. On this issue, we argue that this is not a clearly unexceptionable property one would definitely like a robustness measure to satisfy. After all, when moving from the k-dimensional space to the replicated kd-dimensional space, the ingredients that are needed to compute the corresponding robustness levels might suffer important modifications that could eventually yield different results (consider, for example, the differences in geometric structure of the simplices Δ_k, Δ_{dk}, or in the neighborhood systems $\boldsymbol{N}$ that are chosen in each space, or the different initial weighting schemes that are privileged and their relative position with respect to the corresponding $H(i,j)$).

Another possible determinant one might be tempted to explore is the number of observations in our dataset (n). *A priori*, there seems to be no clear relationship between n and $R_w(r)$ or C_w. However, the sample of five empirical examples presented in this paper is too small to explore these relationships in a meaningful way. More generally, it would be interesting to perform some Monte Carlo simulation experiments to explore further the possible relationships between certain factors of interest and the levels of robustness. This line of research could be attempted in another paper.

Even if it has not been explicitly acknowledged, until now we have implicitly taken for granted that robustness is a normatively desirable property that a composite index should contain. Robust rankings are considered to be reliable and trustworthy, whereas non-robust rankings are considered unstable and unreliable, and might be put into question by rising objections or concerns about the chosen methodology. However, as discussed in McGillivray (1991), McGillivray and White (1993), and Foster *et al*. (2009), a high level of robustness for a composite index is equivalent to a large *redundancy* of its components. These papers rightly argue that in certain circumstances it makes little sense to combine a list of highly correlated indices if any of them basically provides the same ranking as their composite. Hence, we face an apparent paradox in which robustness can be seen simultaneously as a desirable and an undesirable property. Is there a way of getting rid of this uncomfortable contradiction and propose normative criteria that can be used as a guide to identify composite indicators that are robust *and* non-redundant at the same time? Without providing a rigorous or all encompassing answer (that should await further research), we suggest that it might be possible to find composite indices that meet both criteria at the same time. Consider the hypothetical robustness functions depicted in Figure 5: they will be referred to as $R_{w,A}$, $R_{w,B}$, and $R_{w,C}$. In the case of $R_{w,A}$, there is high robustness and high redundancy levels. Recall that, since the highest possible redundancy level takes place when $R_w(1) = 0$, a necessary condition to have non-redundant composite indices is that $R_w(1)$ should not be "very small." In the case of $R_{w,B}$ one has

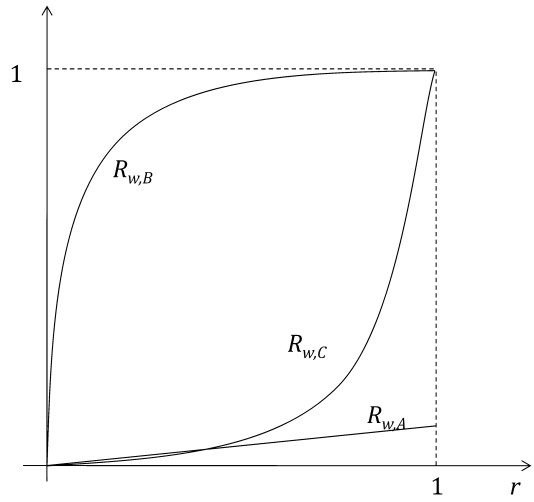

Figure 5. Hypothetical Robustness Functions

low robustness and low redundancy levels. In the third case, $R_{w,C}$ shows high levels of robustness for most values of r, but as r approaches 1, the values of $R_{w,C}(r)$ increase sharply: this way, the necessary condition to have a non-redundant composite index is satisfied. It remains as an open research question to find sufficient conditions to obtain highly robust *and* non-redundant composite indices. However, in order to advance our understanding on these issues, it would be necessary to give a complete and rigorous definition of "redundancy," a topic that is beyond the scope of this paper.

Concluding Remarks

Ranking sets of objects in terms of their attributes is a crucial and extended practice in many areas of human activity, especially those related to decision-making processes in which a scarce resource has to be allocated to the most deserving recipients. However, it turns out that the evaluation techniques that are used to rank the objects are highly sensitive to the weights that are attached to the alternative attributes that are taken into account. This makes the choice of a specific weighting scheme a specially controversial and sensitive issue. In this paper we have presented innovative ways of assessing the extent to which a multiattribute ranking is sensitive to the choice of specific weights by taking into account the whole set of weighting schemes.

The ideas and techniques presented in this paper can be very useful to assist decision-makers when trying to assess the extent to which multiattribute rankings are reliable or not by giving them a complete picture of the ranking problem. However, it should be emphasized that it might not be appropriate to use these techniques as a method of generating new weighting schemes to rank objects. In particular, it might not be a good idea to look for the weighting scheme that provides the highest confidence value or the robustness function with lower

values. Robustness is not a goal in itself: if it were the only criterion to choose among weighting schemes it might lead to non-intuitive results like putting the whole weight to a single attribute. When judging the appropriateness of a composite index, it is also important to take into account its redundancy of composition—that is, the extent to which the information provided by the composite index is the same with respect to its individual components. In this paper we have argued that it might be possible to overcome the "robustness versus redundancy" paradox and find composite indices that are highly robust *and* non-redundant at the same time. However, more research is still needed on this point.

Appendix

Definition 1. A *Neighborhood System* for the simplex Δ_k is the following disjoint union:

$$\boldsymbol{N} := \bigcup_{w \in \Delta_k} \boldsymbol{N}_w$$

where each $\boldsymbol{N}_w$ is a set of neighborhoods of $\boldsymbol{w}$ with the following properties:

(1) For any $U, V \in \boldsymbol{N}_w$ $(U \neq V)$, either $U \subset V$ or $V \subset U$. That is, the neighborhoods of $\boldsymbol{w}$ are nested.
(2) For any $U, V \in \boldsymbol{N}_w$ $(U \neq V)$ with $U \subset V$, there is always some $W \in \boldsymbol{N}_w$ such that $U \subset W \subset V$. That is, the set of neighborhoods of $\boldsymbol{w}$ is dense.
(3) The set $\{\boldsymbol{w}\} \in \boldsymbol{N}_w$. That is, $\boldsymbol{w}$ is considered to be a neighborhood of itself.
(4) The set $\Delta_k \in \boldsymbol{N}_w$. That is, the whole simplex Δ_k is considered to be a neighborhood of $\boldsymbol{w}$.
(5) For any $U \in \boldsymbol{N}_w$, $U \subseteq \Delta_k$ and U is a closed set.

References

Anand, S. and A. Sen, "Concepts of Human Development and Poverty: A Multidimensional Perspective," *Human Development Papers*, Human Development Report Office (UNDP), New York, 1997.

Bardhan, K. and S. Klasen, "UNDP's Gender-Related Indices: A Critical Review," *World Development*, 27, 985–1010, 1999.

Cherchye, L., E. Ooghe, and T. Puyenbroeck, "Robust Human Development Rankings," *Journal of Economic Inequality*, 6, 287–321, 2008.

Chowdhury, S. and L. Squire, "Setting Weights for Aggregate Indices: An Application to the Commitment to Development Index and Human Development Index," *Journal of Development Studies*, 42, 761–71, 2006.

Desai, M. and A. Shah, "An Econometric Approach to the Measurement of Poverty," *Oxford Economic Papers*, 40, 505–22, 1998.

Despotis, D. K., "A Reassessment of the Human Development Index via Data Envelopment Analysis," *Journal of the Operational Research Society*, 56, 969–80, 2005.

Foster, J. and A. Sen, "On Economic Inequality: After a Quarter Century," in A. Sen (ed.), *On Economic Inequality*, Extended Edition, Oxford University Press, 1997.

Foster, J., L. Calva, and M. Székely, "Measuring the Distribution of Human Development: Methodology and an Application to Mexico," *Journal of Human Development*, 6, 5–25, 2005.

Foster, J., M. McGillivray, and S. Seth, "Rank Robustness of Composite Indices," Working Paper 26, OPHI, Oxford University, 2009.

Hirschberg, J., E. Maasoumi, and D. Slottje, "Cluster Analysis for Measuring Welfare and Quality of Life Across Countries," *Journal of Econometrics*, 50, 131–50, 1991.
McGillivray, M., "The Human Development Index: Yet Another Redundant Composite Development Indicator?" *World Development*, 19, 1461–8, 1991.
McGillivray, M. and H. White, "Measuring Development? The UNDP's Human Development Index," *Journal of International Development*, 5, 183–92, 1993.
Nardo, M., M. Saisana, A. Saltelli, S. Tarantola, A. Hoffman, and E. Giovannini, *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Statistics Working Paper STD/DOC(2005)3, OECD, Paris, 2005.
Permanyer, I., "On the Robust Measurement of Well-Being in a Gender Perspective," PhD Dissertation, Mimeo, Universitat Autònoma de Barcelona, 2007.
Saisana, M., A. Saltelli, and S. Tarantola, "Uncertainty and Sensitivity Analysis as Tools for the Quality Assessment of Composite Indicators," *Journal of the Royal Statistical Society*, Series A, 168, 1–17, 2005.
Schokkaert, E., "Capabilities and Satisfaction with Life," *Journal of Human Development*, 8, 415–30, 2007.
Sen, A., *Inequality Reexamined*, Oxford University Press, 1992.
Stapleton, L. and Garrod, G., "Keeping Things Simple: Why the Human Development Index Should Not Diverge from Its Equal Weights Assumption," *Social Indicators Research*, 84, 179–88, 2007.