Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford

OPHI WORKING PAPER NO. 106

Composite Indices, Alternative Weights, and Comparison Robustness

Suman Seth* and Mark McGillivray**

October 2016

Abstract
Composite indices are widely used and can be highly influential. Yet most remain controversial owing to inter alia the arbitrary selection of component weights. Several studies have proposed testing the robustness of rankings generated by composite indices with respect to alternative weights but have not provided sufficient guidance on the choice of these alternatives. This paper proposes a holistic yet theoretically novel approach for selecting sets of alternatives weights and assessing comparison robustness that is applicable to linear composite indices with any finite number of dimensions. This approach is applied to robustness testing of inter-temporal country improvements generated by arguably the world's most influential composite development index, the UNDP Human Development Index (HDI). More than two-thirds of HDI country improvements between 1980 and 2013 were found to be not robust to the selected set of alternative weights.

Keywords: Composite indices, weights, rank robustness tests, inter-temporal comparisons, Human Development Index

JEL classification: D63, C6, I31, O10

Acknowledgements
This paper has benefitted from early discussions with James Foster. All errors in this paper remain our own.

Citation: Seth, S. and McGillivray, M. (2016). "Composite indices, alternative weights, and comparison robustness." *OPHI Working Paper 106*, University of Oxford.

* Economics Division, Leeds University Business School, Maurice Keyworth Building, Leeds, LS2 9JT, UK, and Oxford Poverty and Human Development Initiative (OPHI), University of Oxford, Oxford OX1 3TB, UK, S.Seth@leeds.ac.uk
** Alfred Deakin Institute, Deakin University, Locked Bag 20001, Geelong, Victoria 3220, Australia, and Oxford Poverty and Human Development Initiative (OPHI), University of Oxford, Oxford OX1 3TB, UK, Mark.McGillivray@deakin.edu.au

This study has been prepared within the OPHI theme on multidimensional measurement.

ISSN 2040-8188 ISBN 978-19-0719-491-9

1. Introduction

Composite indices might not rule our world, but they are highly influential in it. This is particularly the case with composite international indices that seek to assess the achievements of countries by various criteria. As Høyland, Moene, and Willumsen (2012:1) observe, "one can hardly open a newspaper without finding a reference to an international index;" they also refer to the "tyranny" of the rankings produced by these indices. The vast majority of composite indices fall into a class of what Ravallion (2011) describes as 'mashups'. Ravallion (2011:1) defines a mashup as a "composite index for which existing theory and practice provides little or no guidance for its design … (with) an unusually large number of moving parts, which the producer is essentially free to set." Ravallion (2011:1) points to a number of pitfalls in the use of these indices, noting that "clearer warning signs are needed for users" of them.

A key moving part of most composite indices, which producers are free to set, is their component weights. Most are set arbitrarily, with the most common practice being to set equal weights for each

---

The Oxford Poverty and Human Development Initiative (OPHI) is a research centre within the Oxford Department of International Development, Queen Elizabeth House, at the University of Oxford. Led by Sabina Alkire, OPHI aspires to build and advance a more systematic methodological and economic framework for reducing multidimensional poverty, grounded in people's experiences and values.

Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford
3 Mansfield Road, Oxford OX1 3TB, UK
Tel. +44 (0)1865 271915 Fax +44 (0)1865 281801
ophi@qeh.ox.ac.uk http://www.ophi.org.uk

component. This choice of weights is due to an uncertainty about the correct weights arising from a lack of theoretical or other guidance.[1] Uncertainty about the setting of weights has clear implications for the interpretation of a composite index. Perhaps the most obvious implication is for index rankings.[2] Index rankings are a function *inter alia* of the weights, and, if there is uncertainty about the correct weights, it must follow that there is uncertainty over the veracity of these rankings.

Uncertainty over composite indices' weights has been acknowledged in a number of previous studies. These studies have sought to analyze the robustness of ranks provided by equally weighted composite indices to alternative weights (Cahill 2005; Cherchye, Ooghe, and Puyenbroeck 2008; Foster, McGillivray, and Seth 2009, 2013; Permanyer 2011; and Zheng and Zheng 2015). Each study looked at the well-known UNDP Human Development Index (HDI), although their analyses are applicable to most composite indices. They did not propose replacing equal weights with alternative weighting schemes, instead advocating tests for the robustness of composite indices rankings to their assigned equal weights so as to aid interpretation, in broadly the same way that significance tests are used in statistical analysis.

This paper further explores the robustness of linear composite index comparisons for the types of indices that fall into the Ravallion mashup class. Its fundamental premise is that in the absence of rigorous scientific guidance on the setting of weights for the components of composite indices, assigning equal weight to their components is broadly defensible provided that, in Ravallion's words, 'warning signs' are provided as to the implications of this for the rankings they yield. Two main objectives are pursued in this paper.

The first is to address a difficulty encountered by previous studies: the selection of a set of alternative weighting schemes for assessing rank robustness. This selection is a requirement of the tests proposed by these studies, yet none provide sufficient guidance for such selection. This paper proposes a general yet theoretically novel approach that both selects alternative weights and assesses comparison robustness. Our approach is founded on the normative assumption that a consensus has been reached

---

1 The setting of equal weights is the norm with composite international indices, including the Ease of Doing Business Index, Country Policy and Institutional Assessment (CPIA) Index, the Environmental Performance Index, Child Well-being Index, Commitment to Development Index, Human Development Index, Economic Resilience Index, Economic Vulnerability Index, Environmental Sustainability Index, Genuine Progress Measure, Global Peace Index, Index of Economic Freedom, Global Peace Index, and the Physical Quality of Life Index. The proponents of the Environmental Sustainability Index, for example, argued for equal weights on the grounds that "that no objective mechanism exists to determine the relative importance of the different aspects of environmental sustainability" (Esty et al. 2005: 66). A comprehensive list of composite indices can be found in Bandura (2008).

2 This was one of five aspects of mashup indices that, according to Ravallion (2011), are in need of more attention. The other four are their conceptual foundations, the tradeoffs they embody (a sound discussion on this issue may be found in Decancq and Lugo 2013), the contextual factors relevant to country performance, and the sensitivity of the implied rankings to changes in the data.

on the minimum and the maximum allowable weights that should be assigned to each component. This consensus then yields a particular set of alternative weights with respect to which the robustness of pairwise comparisons should be tested. These comparisons can be of cross-sectional rankings or inter-temporal changes in index scores for the unit of analysis under consideration. In order to make our approach amenable to practical applications, we provide tractable conditions that should be tested in the assessment of robustness. The theoretical approach that we develop in this paper is applicable to linear composite indices or their monotonic transformations with any finite number of components.

Having developed a normative theoretical framework for selecting alternative weighting schemes and testing robustness of pairwise comparisons, the paper turns to its second objective, which is to evaluate the prevalence of robust country-specific inter-temporal HDI comparisons for six successful subperiods during the years 1980 to 2013. First introduced in the UNDP *Human Development Report 1990* (UNDP 1990), the HDI is a composite index that combines country achievements in health, education, and income. The annual publication of these reports is eagerly awaited, in large part due to release of the latest HDI data. HDI scores have since been published annually by the UNDP, in the annual publications of the *Report*. Country HDI rankings receive enormous attention in the media, policy circles, and elsewhere. Changes over time in the HDI scores of individual countries also receive attention. Many other composite indices are also used to assess changes over time for the units of analysis in question.

To the best of our knowledge, testing the robustness of inter-temporal comparisons of the HDI or other composite indices of its general type has not previously been attempted. This is an important exercise as it must be remembered that the HDI is not only used to assess human development achievements among countries, it is also used to assess country-specific changes in these achievements over time.[3] These changes are arguably just as important for countries as their achievements relative to others at a particular point in time. The above-mentioned uncertainty is equally applicable to these changes, and, as such, warning signs are also required for their use and interpretation.

The remaining sections of this paper are as follows. Section Two presents its framework for assessing the robustness of composite index comparisons, including its approach to the selection of alternative weights. Section Three outlines the formulation of the HDI and how it has changed over time, as well as the debates over its component weights. Section Four presents the findings of our empirical analysis. Section Five provides concluding remarks.

---

[3] The UNDP, for example, claimed that "Advances in the HDI have occurred across all regions and almost all countries .... all but 3 of the 135 countries have a higher level of human development today than in 1970 ..." (UNDP 2010: 27).

2. Selecting a Set of Reasonable Alternative Weights for Testing Robustness

The need for testing the veracity of composite index comparisons is well recognized in the academic literature, with a number of corresponding tests having been proposed and applied. Cherchye et al. (2008), for instance, proposed a test for the robustness of comparison of any given composite index score to simultaneous changes in the index's weights, component variable normalization, and aggregation methods, obtaining conditions that kept the comparison under the original weighting scheme preserved. Foster et al. (2009, 2013) proposed an epsilon-contamination model to devise an approach for choosing a set of alternative weighting schemes to be used in assessing composite index comparison robustness. In this approach, one was assumed to have only partial confidence that the initial weighting scheme was correct and any other weighting schemes could be feasible alternatives. The level of confidence one placed on the initial weight determined the size and the shape of the set of alternative weighting schemes. Permanyer (2011) also envisaged the need for robustness testing by considering a set of alternative weighting schemes around the initial weighting scheme and applied the Foster et al. (2009, 2013) approach as an example. The Foster et al. (2009, 2013) approach determined only a particular shape of alternative weighting schemes, which were homothetic contractions of the entire set of weighting schemes. There is also the difficulty in determining the level of confidence one places on the initial weighting scheme, which in turn determines the shape and the size of the alternative weighting scheme.

A normative framework for determining an alternative set of weighting schemes requires a process of strong justifications. Zheng and Zheng (2015) sought to avoid this requirement. Using a fuzzy set theoretical framework, they avoided starting with any initial weighting scheme by considering all possible weighting schemes to be potential alternatives while proposing a robustness measure for gauging the strength of pairwise comparisons. It should however be noted that all possible weighting schemes include those that assign the entire weight to one dimension and zero weight to the remaining dimensions. In these cases, the entire ranking is determined by any one dimension. Allowing such possibilities (entire weight to one dimension and zero weight to the remaining) however goes against the spirit of multidimensionality, which should reflect strictly positive contributions of multiple dimensions to the final index score. There is strong justification, therefore, for not considering these extreme weighting schemes as meaningful alternatives. It can be similarly argued that no alternative weighting schemes should assign excessively high or low weight to any of the multiple dimensions.

In what follows in this section we present a normative theoretical framework that addresses two questions: (i) how one may determine a set of alternative weighting schemes for checking the robustness of pairwise comparisons when there is no a priori reason for treating different dimensions differently and (ii) what relevant and tractable criteria must be satisfied to assess the robustness of these pairwise comparisons with respect to the chosen set of alternative weighting schemes.

We develop in what follows a flexible and holistic approach for determining a set of feasible alternative component weighting schemes. Components will hereinafter be referred to as 'dimensions', given our intended application of this approach to the HDI and that this terminology is typically used in its description. We emphasize, however, that our approach is applicable to any composite index of the general form of the HDI with any finite number of components. We begin with the assumption that there is no a priori reason or justification to assign a higher weight to one dimension than the others or to allow weights on different dimension to vary to different extents. We thus simply assume here that if we allow weights to vary then the weight on every dimension should be allowed to vary to the same extent. The second assumption we make is that the process of choosing a set of reasonable alternative weighting schemes is subject to a general consensus that the weight on any dimension should not be allowed to be higher than a particular value (*maximum weight*) and should not be allowed to be lower than a particular value (*minimum weight*). When no consensus is reached, weights may be allowed to vary to the largest extent possible, although this is undesirable because in this case any dimension can be assigned the entire weight or no weight at all.[4]

Agreement on a maximum allowable weight and a minimum allowable weight yields a continuum of feasible alternative weighting schemes with respect to which the robustness of pairwise comparisons should be evaluated. If a pairwise comparison evaluated with the initial equal weighting scheme is not altered for any of these feasible alternative weighting schemes then the initial weighting scheme is stated to be robust. Checking robustness using the continuum of alternative weighing schemes appears to be a daunting task. We show however that the restrictions on maximum and minimum allowable weights form a bounded set of alternative weighting schemes and the robustness of pairwise comparisons can be conducted under certain tractable conditions. Our proposed approach includes the set of alternative weighting schemes proposed by Foster et al. (2009) as a particular subcase when the initial weighting scheme weighs every dimension equally.

Let us now formally present the approach. We assume that there are a fixed number of $D \in \mathbb{N}\setminus\{1\}$ dimensions, where $\mathbb{N}$ is the set of positive integers. Let $\mathcal{X} \subseteq \mathbb{R}^D$ denote the nonempty set of performance vectors to be ranked, where each performance vector is represented as x in $\mathcal{X}$. A performance vector reflects the performance of a country or any other unit in D dimensions. We denote any dimension by subscript d and the weight assigned to the dimension by w_d. The weight assigned to

4 According to Foster et al. (2009), this is a situation where one has no confidence in the initially chosen equal weighting scheme.

one dimension in comparison to any other dimension represents the relative importance of the former compared to the latter. For example, if all dimensions are equally weighted (i.e., $w_d = 1/D$ for all d), then they are considered equally important relative to each other. Similarly, in the case of three dimensions (i.e., $D = 3$), if the assigned weights are $w_1 = 0.6$, $w_2 = 0.3$, and $w_3 = 0.1$, then the first dimension is considered twice as important as the second dimension and considered six times more important than the third dimension.

Weights assigned to D dimensions are summarized by vector $w = (w_1, \ldots, w_D)$. We refer to a vector of weights as a weighting scheme. Two natural assumptions that we make about the weights are (i) they are non-negative (i.e., $w_d \geq 0$ for all d) and (ii) they sum to one (i.e., $\sum_{d=1}^{D} w_d = 1$). We denote all possible D-dimensonal weighting schemes by $\mathcal{W}$, such that $\mathcal{W} = \{(w_1, \ldots, w_D) \mid w_d \geq 0 \; \forall d, \sum_{d=1}^{D} w_d = 1\}$. Using a performance vector x in X and a weighting scheme w in $\mathcal{W}$, a composite index is defined as $C(x; w) = \sum_{d=1}^{D} w_d x_d$. We denote the D-dimensional initial equal weighting scheme by w^0 and $C(x; w^0)$ denotes the corresponding composite index, evaluated at the initial weighting scheme. Now for any two performance vectors x and y in $\mathcal{X}$, y has an equal or higher composite index value than x at w^0 whenever $C(y; w^0) \geq C(x; w^0)$.

The initial comparison $C(y; w^0) \geq C(x; w^0)$ is stated to be robust with respect to a non-empty set of alternative weighting schemes Δ, if $C(y; w) \geq C(x; w)$ for all w in Δ;[5] whereas the comparison is stated to be not robust with respect to Δ, if $C(y; w) < C(x; w)$ for some w in Δ. What is Δ and how is it defined? Suppose there is a consensus that the weight on any dimension should be not be lower than $\alpha \in [0, 1/d)$ and the weight on any dimension should not be higher than $\beta \in (1/d, 1]$. Then, $\Delta = \{w_1, \ldots, w_d \mid \alpha \leq w_d \leq \beta \; \forall d, \sum_{d=1}^{D} w_d = 1\}$ and indeed $\Delta \in \mathcal{W}$ is bounded. In fact, Δ is a convex hull of a finite number of weighting schemes. What are these finite number of weighting schemes? In order to answer this question, we resort to the majorization theory of measurement.

Definition 1 Any weighting scheme $w' \in \mathcal{W}$ is *not more unequal* than any other weighting scheme $w \in \mathcal{W}$ if and only if $w' = wB$, where B is a D-dimensional doubly stochastic matrix (Marshall and Olkin 1979: 22).

A **doubly stochastic matrix** is a non-negative square matrix with every row and every column summing to one and every doubly stochastic matrix is a convex combination of permutation matrices, where a **permutation matrix** is also a non-negative square matrix with every row and every column having only

---

[5] The argument is in the same spirit of poverty orderings over a range of poverty lines or inequality comparisons using Lorenz orderings. Relevant discussions are provided in Atkinson (1970, 1987) and Foster and Shorrocks (1988a, 1988b).

one element equal to one and the rest of the elements equal to zero. In other words, the set of doubly stochastic matrices is the convex hull of permutation matrices. For D dimensions, the number of maximum possible permutation matrices is $D! = D\times(D-1)\times\cdots\times 1$. Technically, a doubly stochastic matrix can be expressed as $B = \sum_{m=1}^{D!} \omega_m P_m$, where P_m is the m^{th} D-dimensional permutation matrix and ω_m is a weight attached to P_m such that $\omega_m \geq 0$ and $\sum_{m=1}^{D!} \omega_m = 1$. It follows that the relation between w' and w in Definition 1 can be expressed as $w' = wB = w\sum_{m=1}^{D!} \omega_m P_m = \sum_{m=1}^{D!} \omega_m w P_m$, where wP_m is the m^{th} permutation of the weighting scheme w.

Definition 2 Any weighting scheme $w' \in \mathcal{W}$ that is *not more unequal* than any $w \in \mathcal{W}$ by Definition 1 is an element in the convex hull of $D!$ permutations of weighting scheme w.

This concept can be used to identify the finite number of weighting schemes that creates the convex hull of Δ. In order to do so, we must identify the most unequal weighting schemes in Δ. Starting from a particular weighting scheme, the most unequal weighting scheme can be obtained by a finite number of regressive transfers – a concept that is frequently used in the inequality measurement literature.

Definition 3 For any $w, w' \in \mathcal{W}$, weighting scheme w is obtained from a more equal weighting scheme w' by a *regressive transfer* if $w'_l \leq w'_h$, $w_l = w'_l - \epsilon$ and $w_h = w'_h + \epsilon$ for any $\epsilon > 0$ and $w_d = w'_d$ for all $d \neq h, l$.

Let us denote the most unequal weighting scheme in Δ by $\bar{w}$ and so the m^{th} permutation of the weighting scheme is $\bar{w}P_m$. Thus, following Definition 2 and Definition 3, Δ should be the convex hull of $D!$ weighting schemes $\{\bar{w}P_m\}_{m=1}^{D!}$. In practice however the number of distinct permutations of $\bar{w}$ may be *equal or less* than $D!$. For example, when the most unequal weighting scheme has weights of 0.4, 0.3, and 0.2, then there are six (3!) unique permutations: (0.5,0.3,0.2), (0.5,0.2,0.3), (0.3,0.5,0.2), (0.3,0.2,0.5), (0.2,0.5,0.3), and (0.2,0.3,0.5). When the most unequal weighting scheme has weights of 0.5, 0.25, and 0.25, then there are only three unique permutations: (0.5,0.25,0.25), (0.25,0.5,0.25), and (0.25,0.25,0.5). For any arbitrary number of dimensions D, let us denote the number of unique permutations by $\bar{D}$, such that $D \leq \bar{D} \leq D!$. Then Δ is a convex hull of $\bar{D}$ distinct weighting schemes that are denoted by $\{v_m\}_{m=1}^{\bar{D}}$, such that $v_m \neq v_{m'}$ for any $m \neq m'$.

Given the restriction on weights, for any arbitrary number of D dimensions, how many distinct permutations ($\bar{D}$) are there? This depends on the restrictions placed on α and β. Given that weights are bounded between zero and one and they sum to one, the values of α and β are somewhat dependent on each other. For a given value of $\alpha \in [0,1/D)$, β must be bounded between β_l and β_h, such that $\beta_l = (1-\alpha)/(D-1) < \beta < \beta_h = 1-(D-1)\alpha$. Intuitively, once a value of α is chosen, the chosen value of β must lie between β_l and β_h. For example, in case of three dimensions, if $\alpha = 0.2$, then the chosen value of β must be between 0.4 and 0.6. Now once values of α and β are chosen, the following theorem determines the number of distinct permutations $\bar{D}$.

Theorem 1 For any $D \in \mathbb{N}\backslash\{1\}$, for any $\alpha \in [0,1/D)$, and for any $\beta \in (1/D,1]$, $\Delta = \{w_1, \dots, w_d \mid \alpha \leq w_d \leq \beta \ \forall d, \sum_{d=1}^{D} w_d = 1\}$ is the convex hull of $\bar{D}$ distinct permutations of the most unequal weighting scheme $\bar{w} \in \Delta$ such that

$$\bar{D} = \begin{cases} \dfrac{D!}{[D-d]!\times d!} & \text{if } \beta = \dfrac{1-d\alpha}{D-d} \text{ for all } d = 1,2,\dots,(D-1) \\ \dfrac{D!}{[D-(d+1)]!\times d!} & \text{if } \beta \in \left(\dfrac{1-d\alpha}{D-d}, \dfrac{1-(d+1)\alpha}{D-(d+1)}\right) \text{ for all } d = 1,2,\dots,(D-2) \end{cases}.$$

Proof: See the Appendix.

Theorem 1 is quite powerful in the sense that it determines the number of distinct permutations of the most unequal weights in Δ for any arbitrary number of D dimensions once a consensus on the values of α and/or β is reached. The minimum number of unique permutations is obtained when the most unequal weighting scheme is such that all $(D-1)$ dimensions receive the same weight while the remaining one dimension receives a different weight. In this case, the number of unique permutations is $[D!/(D-1)!] = D$. It follows from Theorem 1 that this case occurs when $\beta = (1-\alpha d)/(D-d)$ and either $d = 1$ or $d = (D-1)$ for any $\alpha \in [0,1/D)$. In Table 1, we present three cases where the number of unique permutations of the most unequal weighting scheme is only D. In the first case, weights are allowed to vary to the fullest extent. In this case, $\beta = 1$ and $\alpha = 0$. The most unequal weight $\bar{w} \in \Delta$ is obtained when all $(D-1)$ dimensions are assigned a weight of zero and the remaining dimension is assigned a weight of one. In the second case, suppose one only chooses the value of $\alpha \in (0,1/D)$. The most unequal weighting scheme $\bar{w} \in \Delta$ is obtained when all $(D-1)$ dimensions are assigned weight α and the remaining dimension is assigned weight $\beta = 1-(D-1)\alpha$. In the third case, suppose one chooses only $\beta \in (1/D, 1/(D-1)]$. In this case the most unequal weight $\bar{w} \in \Delta$ is obtained when all $(D-1)$ dimensions are assigned weight β and the remaining dimension is assigned weight $\alpha = [1-\beta(D-1)]$.

Table 1. Cases Where the Number of Unique Permutations of the Most Unequal Weighting Schemes is D

Restrictions on Parameters		Number of Unique Permutations of Most Unequal Weighting Schemes ($\bar{D}$)
1.	$\beta = 1; \alpha = 0$	D
2.	$\alpha \in \left(0, \frac{1}{D}\right); \beta = 1 - (D-1)\alpha$	D
3.	$\beta \in \left(\frac{1}{D}, \frac{1}{D-1}\right]; \alpha = [1 - \beta(D-1)]$	D

Are there tractable conditions for checking whether the pairwise comparison $C(y; w^0) \geq C(x; w^0)$ is robust with respect to Δ? It turns out that one needs to check whether $C(y; v_m) \geq C(x; v_m)$ for all $m = 1, \ldots, \bar{D}$. If $C(y; v_m) < C(x; v_m)$ for at least one m, then the initial comparison $C(y; w^0) \geq C(x; w^0)$ is not robust. The following theorem, which is in the same spirit as in Foster et al. (2009), formally states the condition.

Theorem 2 For any $x, y \in \mathcal{X}$, the comparison $C(y; w^0) \geq C(x; w^0)$ is robust with respect to Δ if and only if $C(y; v_m) \geq C(x; v_m)$ for all $m = 1, \ldots, \bar{D}$.

Proof: See the Appendix.

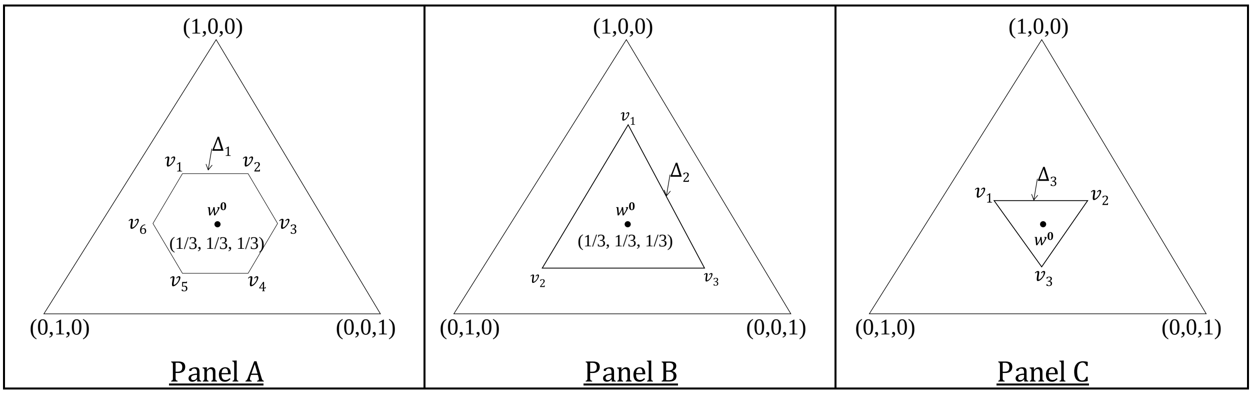
What does Theorem 2 convey? Intuitively, if one agrees on the restrictions on α and/or β, then it boils down to comparing the composite indices at $\bar{D}$ weighting schemes to conclude robustness.

Examples

Let us provide examples involving three dimensions ($D = 3$), where the initial weighting scheme is $w^0 =$ (1/3,1/3,1/3). The set of all weights $\mathcal{W}$ is the convex hull of three extreme weighting schemes (1,0,0), (0,1,0), and (0,0,1).

For the first example, suppose $\alpha = 1/6$ and $\beta = 1/2$. This is a case where $\beta \in ((1 - d\alpha)/(D - d), (1 - (d+1)\alpha)/(D - (d+1)))$ for $d = 1$ or $\beta \in$ (5/12,2/3). The maximum inequality in weights occurs when one dimension receives $1/2$ and the other two dimensions receive weights $1/3$ and $1/6$, respectively. Thus, $\bar{w} =$ (1/2,1/3,1/6). In this case, the number of unique permutations of $\bar{w}$ is $\bar{D} = D!/([D - (d+1)]! \times d!)$ for $d = 1$ or $\bar{D} = 6$ and we denote their convex hull by Δ_1. The six distinct permutations of $\bar{w}$ are: $v_1 =$ (1/2,1/3,1/6), $v_2 =$ (1/2,1/6,1/3), $v_3 =$ (1/3,1/6,1/2), $v_4 =$ (1/6,1/3,1/2), $v_5 =$ (1/6,1/2,1/3), and $v_6 =$ (1/3,1/2,1/6). In order to check the robustness of a pairwise comparison with respect to Δ_1, one needs, therefore, to compare the pair at these six weighting schemes. The shape of Δ_1 is depicted in Panel A of Figure 1.

Figure 1. Examples of Sets of Alternative Weights



For the second example, suppose one only requires that no dimension's weight should fall below $\alpha = 1/6$. Let us denote the set of weighting schemes with this restriction by $\Delta_2 = \{w_1, \ldots, w_D \mid 1/6 \leq w_d \leq 1\ \forall d, \sum_{d=1}^{D} w_d = 1\}$. The most unequal weighting scheme assigns a weight of 1/6 to two dimensions and a weight of 2/3 is assigned to the remaining dimension. This is a situation where there are only $\bar{D} = 3$ distinct permutations of the most unequal weight $\bar{w}$. We present the shape that Δ_2 takes in Panel B of Figure 1, where Δ_2 is a convex hull of $v_1 = (2/3, 1/6, 1/6)$, $v_2 = (1/6, 2/3, 1/6)$, and $v_3 = (1/6, 1/6, 2/3)$. We should point out that setting just a lower bound on weights yields the same set of weights proposed by Foster et al. (2009) for a particular level of confidence with respect to the initially chosen equal weight through the epsilon contamination model.

For the third example, suppose one only requires that the weight on no dimension should surpass $\beta = 0.4$. We denote the set of weighting schemes with this requirement by $\Delta_3 = \{w_1, \ldots, w_D \mid 0 \leq w_d \leq 0.4\ \forall d, \sum_{d=1}^{D} w_d = 1\}$. Note that although no lower bound is set, weights are already bounded from below by zero. The most unequal weighting scheme $\bar{w}$ in this case would be the one that assigns 0.4 to any two dimensions and assigns 0.2 to any one dimension. This is the case where $\beta \in (1/D, 1/(D-1)]$ and $\alpha = [1 - \beta(D-1)$. Therefore, the number of unique permutations is $\bar{D} = 3$ and Δ_3 is a convex hull of $v_1 = (0.4, 0.4, 0.2)$, $v_2 = (0.4, 0.2, 0.4)$, and $v_3 = (0.2, 0.4, 0.4)$. The shape of Δ_3 is presented in Panel C of Figure 1. Hence, to assess the robustness of a pairwise comparison, one needs to compare the composite index at these three vertices.

Using the approach developed above, one may construct a binary test for robustness to consider whether a given comparison is robust or not with respect to a set of alternative weights and report the corresponding prevalence of robust pairwise comparisons, defined as the proportion of all pairwise comparisons able to pass the binary test (as in Permanyer 2011 and Foster et al. 2013). Alternatively or additionally, one may apply a **robustness measure** to assess the level of robustness of each pairwise comparison whenever a pairwise comparison fails the binary test (as in Foster et al. 2009 and Zheng and

Zheng 2015). Section 4 of this paper pursues the first route, conducting a binary test of each HDI inter-temporal comparison and reporting the prevalence of robust comparisons. Prior to that, however, we outline the design of the HDI and debates over its weights in the next section.

The Human Development Index

The HDI has been revised many times since 1990, but two key features have remained unchanged: (i) the aggregation of achievements in health, education, and income and (ii) the weights attached to these achievements, which are each set at one-third. The second of these features has remained controversial from the moment the HDI was first released.

Design

The general formulation of the HDI since 1990 can for any given country be written as follows:

$$HDI = f(x_1, x_2, x_3) \quad (1)$$

where $x_d \in [0,1]$, for each $d = 1$, 2, and 3, is an index of achievement in the d^{th} human development dimension and f is strictly increasing in each x_d. The three dimensions are a long and healthy life (x_1), access to knowledge (x_2), and a decent material standard of living (x_3). In the previous section of this paper we referred to these dimensions simply as health, education, and income, respectively, and shall continue to do so for convenience. In what immediately follows we focus on the HDI used by the UNDP between 1994 and 2009.[6]

Achievement in health is assessed using life expectancy in years, which is formulated as follows:

$$x_1 = \frac{a_1 - a_{1,\min}}{a_{1,\max} - a_{1,\min}} \quad (2)$$

where $a_{1,i}$ is the life expectancy at birth in years, $a_{1,\min}$ is a minimum life expectancy and $a_{1,\max}$ is a maximum life expectancy. Minimum and maximum life expectancy were set at 25 and 85 years, respectively. The objective of this formulation is to normalize life expectancy, with a country's achievement in health being set to zero if $a_1 = a_{1,\min}$ or to unity if $a_1 \geq a_{1,\max}$.

Achievement in education was formulated as follows:

6 The HDI used between 1990 and 1993 was computed using a deprivation index I_d for each of the three dimensions such that $I_d = 1 - x_d$ for all $d = 1$, 2, 3. Nevertheless, the mathematical formulation was identical to the HDI between 1994 to 2009. The three dimensions in that period were assessed using the same indicators used between 1994 to 2009 – with the exception of education in 1990, which was assessed using the adult literacy rate only. Further details can be found in UNDP (1994).

$$x_2 = \frac{2}{3}\frac{a_{2,1} - a_{2,1,\min}}{a_{2,1,\max} - a_{2,1,\min}} + \frac{1}{3}\frac{a_{2,2} - a_{2,2,\min}}{a_{2,2,\max} - a_{2,2,\min}} \tag{3}$$

where $a_{2,1}$ and $a_{2,2}$ are the adult literacy rate and the gross all-levels school enrolment ratio, respectively.[7] It is clear from equation (3) that these two sub-indicators are normalized in the same way as life expectancy, using maximum and minimum values for both variables. The minimum and maximum values of both education variables were set at zero and 100, respectively.

Achievement in income is formulated as follows:

$$x_3 = \frac{\ln a_3 - \ln a_{3,\min}}{\ln a_{3,\max} - \ln a_{3,\min}} \tag{4}$$

where a_3 is a transformation of a country's income per capita (measured by \$PPP GDP per capita) and $a_{3,\min}$ and $a_{3,\max}$ are the respective minimum and maximum values for income. Various transformations of income per capita were used between 1994 and 2009, each intended to reflect diminishing returns to the conversion of higher income into higher human development. A logarithmic transformation was used since 1999. As with achievements in health and education, achievement in income is set to zero if $a_3 = a_{3,\min}$ or to unity if $a_3 \geq a_{3,\max}$. The minimum and maximum values of income per capita were set at \$PPP 200 and \$PPP 40,000, respectively.

The 1994 to 2009 HDI was formed by the following aggregation method:

$$HDI_A = \frac{1}{3}\sum_{d=1}^{3} x_d. \tag{5}$$

The use of an equally weighted arithmetic mean assumes the perfect substitutability of normalized achievements – that low normalized achievement in any one dimension can be fully compensated by high normalized achievement in another.

The above-mentioned assumption was questioned with the release of the *Human Development Report 2010* (UNDP 2010). The assumption of non-perfect normalized achievement substitution was adopted through formulating the HDI as a geometric mean, as follows:

$$HDI_G = \prod_{d=1}^{3} x_d^{\frac{1}{3}}. \tag{6}$$

The HDI_G formulation in equation (7) may also be presented as

$$\ln HDI_G = \frac{1}{3}\sum_{d=1}^{3} \ln x_d. \tag{7}$$

[7] Mean years of schooling was used instead of the gross school enrolment ratio in the 1994 HDI.

It follows, therefore, that formulation in equation (6) is a monotonic transformation of the linear formulation in equation (5).

It is clear from equations (5) to (7) that the HDI introduced in 2010 retained the same achievements (in health, education, and income) and equal weighting of them as previous HDIs. It did, however, use different indicators for these achievements. Achievement in health was still assessed using life expectancy. Achievement in education was assessed using mean and expected years of schooling instead of the adult literacy rate and the gross all-level school enrolment ratio, and income was assessed using $PPP GNI per capita instead of $PPP GDP per capita. The same general method of transforming these achievements was employed as was used previously.[8] This 2010 HDI formulation has been used in each subsequent *Human Development Report* (UNDP 2011–2015). The minima and maxima values used in the transformations were further revised in 2014 (UNDP 2014). Those for life expectancy were set at 20 years and 85 years, respectively. The minimum values for the transformation of mean and expected years of schooling were set at zero. The maximum values of these variables were set at 15 and 18 years, respectively. The minimum and maximum values of GNI per capita were set at $PPP 100 and $PPP 75,000. The UNDP in 2015 provides a more detailed articulation of the rationale for the setting of the minimum and maximum values for each of these indicators, describing them as 'natural zeros' and 'aspirational goals', respectively (UNDP 2015).[9]

Weights and Robustness of Comparisons

Hopkins (1991) commented that the UNDP essentially invoked Occam's razor in the selection of weights, taking the simplest possible alternative that is likely to attract the least disagreement. Kelley (1991) argued for a higher weight for income on the grounds that it provides a capacity to choose among many other dimensions of human development. Kelley, however, acknowledged that a priori it is difficult to justify any set of weights and for this reason calls for testing the sensitivity of the HDI to alternative weights.

The UNDP responded to these and other concerns over the HDI weighting scheme in *Human Development Report 1993* (UNDP 1993). In that *Report* it acknowledged that in an ideal world the weights would be taken from a meta-production for human development, with the respective contributions of

8 The minimum and maximum values for life expectancy were set at 20 years and 83.2 years. Those for the mean years of schooling were zero years and 13.2 years, for expected years of schooling they were zero years and 20.6 years, and for GNI per capita they were $PPP 163 and $PPP 108,211 (UNDP 2010).

9 For example, the natural zero for expected years of schooling is set at zero on the grounds that "societies can subsist without formal education," while the aspirational value for this indicator of 18 years was chosen on the grounds of it being "the equivalent to achieving a master's degree in most countries" (UNDP 2015:2). The rationale for the aspirational value of GNI per capita is a finding of Kahneman and Deaton (2010), which according to UNDP (2015) is that there is virtually no gain in human development and well-being from annual income beyond PPP$75,000.

health, education, and income to human development being their respective weights (UNDP 1993). Lacking such a function, the UNDP referred to a Principal Components Analysis (PCA) of the HDI undertaken by Tatlidil (1992), which assigned rather similar weightings to the index's achievements. The PCA method sets weights (component loadings) that account for as much variability in the data in question as possible. While PCA might produce interesting results, this criterion for the determination of weights is of course arbitrary, and there is absolutely no guarantee that these weights might resemble those provided by the production function. This did not, however, deter the UNDP from claiming that Tatlidil's analysis "confirms the equal weights" (UNDP 1993: 110).

While concerns over the HDI weights have simmered over time since the early 1990s, the UNDP's adoption of the geometric mean formulation of the HDI in 2010 returned attention more firmly to it. Ravallion (2011: 12–13) commented that:

> Equality of the weights was, of course, an arbitrary judgment, and it might have been hoped that the weights would evolve in the light of the subsequent public debate. But that did not happen. The weights on the three components of the HDI (health, education, and income) have not changed in 20 years, and it is hard to believe that the HDI got it right first go.

Ravallion further added that in the context of equally weighted international indices per se, is it hard to believe that the weights should be the same for all countries and all people within them.

Ravallion's comments attest to the uncertainty over HDI weights given that they have not evolved in light of public debate and that guidance on differential weights for countries and people within them does not exist. As Kelly acknowledged, it remains the case that any set of weights is difficult to justify and therefore this uncertainty could well be a permanent feature of the HDI and the many indices like it. It is this issue that has motivated the HDI rank robustness studies cited above, the results of which we now discuss.

Cahill (2005) used a simple approach to conclude that HDI rankings were robust. Using six alternative weighting schemes, Cahill found the six country rankings to be statistically indistinguishable from the original HDI ranking. Unlike Cahill (2005), Foster et al. (2009, 2013) and Permanyer (2011) proposed relatively sophisticated normative frameworks for determining a set of alterative weighting schemes. But like Cahill, Foster et al. (2009) found similar conclusions on HDI robustness. Nearly 70% of cross-sectional pairwise HDI country rankings were robust between 1998 and 2004 regardless of how the three achievements were weighted, whereas more than 90% of cross-sectional pairwise comparisons during the same period were robust when the weight on each of the three dimensions was allowed to vary between 1/4 and 1/2. Cherchye et al. (2008) found that nearly 75% of pairwise comparisons of

2002 HDI scores were reversible to the simultaneous application of alternative normalizations, aggregation methods, and weights. Zheng and Zheng (2015) found that seven of the 45 pairwise comparisons of the top ten HDI countries in the *2014 Human Development Report* (UNDP 2014) were fully robust or had a 'truth value' of unity.

4. How Robust are Inter-temporal Changes in the HDI?

How robust are changes in HDI score for individual countries between 1980 and 2013? We address this question using HDI dimension achievements published by the UNDP. These data are available for all indicators for most years from 1990 but are not updated annually, despite HDI values being published for each year since 1990. For this reason, we select data for every five years in the period 1980 to 2015, plus that for 2013, the latest year for which HDI dimensional data are available. This selection provides us with data on all three dimensions for 123 countries.[10]

We commence our investigation by using these data to calculate both the geometric and arithmetic mean HDI formulations, but in both cases employing the dimension achievement indicators and maximum and minimum values used by the UNDP since 2010 and 2014, respectively. So, in essence, the second of these indices is the current HDI but aggregated by taking the arithmetic rather than geometric mean. Below we refer to this formulation as HDI'_A. We consider both formulations primarily to gauge whether the UNDP's move to the geometric mean matters in terms of rank robustness selection.

Table 2. HDI Scores and Dimension Achievements between 1980 and 2013

	HDI Arithmetic (HDI'_A)		**HDI Geometric (HDI_G)**		**Health**		**Education**		**Income**	
(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**
Year	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
1980	0.562	0.160	0.544	0.167	0.650	0.160	0.412	0.176	0.623	0.186
1985	0.581	0.160	0.566	0.166	0.679	0.150	0.442	0.175	0.622	0.189
1990	0.597	0.165	0.584	0.171	0.698	0.155	0.468	0.179	0.626	0.194
2000	0.640	0.174	0.632	0.178	0.724	0.169	0.547	0.187	0.651	0.200
2005	0.666	0.171	0.659	0.175	0.746	0.165	0.586	0.182	0.666	0.198
2010	0.691	0.162	0.685	0.166	0.772	0.151	0.619	0.172	0.683	0.192
2013	0.700	0.159	0.694	0.162	0.785	0.145	0.625	0.170	0.692	0.188

Source: Author calculations using UNDP data. S.D.: Standard Deviation.

Table 2 presents the change in mean HDIs and dimension achievement for the 123 countries between 1980 and 2013. Columns 2 and 4 of the table show that the means of both HDIs have steadily improved between 1980 and 2013, with the corresponding standard deviation remaining between 0.160 and 0.178. Achievements in each dimension are shown in columns 6, 8, and 10 of Table 2. All mean achievements

[10] The data were downloaded from http://hdr.undp.org/en/data in October 2015.

have increased over the years in question, especially in education. Standard deviations range between 0.145 and 0.200.

Improved dimension achievements and overall HDI scores are not, however, observed for all countries. This is shown in Table 3. Columns 2 to 4 of the table report the number of countries in our sample whose achievements in each dimension improved between each period. For example, between 1980 and 1985, achievement in health improved or did not change in 117 countries, while in education and income it improved or did not change in 113 and 69 countries, respectively. The fifth column presents the number of countries in each period for which improved achievement was observed in all three dimensions, while the sixth column presents the number of countries in each period for which achievement increased or did not change. The seventh and the eighth columns show the number of countries for which achievement declined in all three dimensions and either declined or remained the same, respectively. A pleasing finding shown in columns seven and eight is that deterioration in all dimensional achievements is observed in only a handful of countries.

The remaining columns in Table 3 present aggregate results. Columns 9 and 13 present the number of countries in each period for which HDI_A' and HDI_G improved, respectively. Columns 11 and 15 present the number of countries in each period for which HDI_A' and HDI_G deteriorated, respectively. HDIs for all countries improved only between 1980 and 2013, between 2000 and 2010, between 2000 and 2013, and between 2005 and 2013.

If we carefully compare the fifth column with columns 9 and 13, large discrepancies are observed. Between 1985 and 1990, HDI_A' improved for 106 countries and HDI_G improved for 108 countries, but only for 70 countries was it the case that none of the three dimensions deteriorated. For the rest of the 36–38 odd countries, the HDI improvement was accompanied by deterioration in at least one dimension, which means any alternative weighting scheme may reverse the direction of improvement. In fact, it was never the case that all three indicators improved or declined together across years. Which dimension showed the most volatility? We observe from the fourth column that income, unlike health and education, did not increase systematically for many countries.

The question that arises from the preceding observations is how robust were the observed improvements and reductions in the HDIs? The answer certainly depends on how different sub-indices have changed over time as well as on the set of alternative weighting schemes subject to which we should check robustness. The following criteria are used to determine the set of alternative weights. We allow the weights to vary to the extent where any one dimension can be considered to be three times more important than the other two dimensions, but no dimension should be considered as less than

Table 3. The Changes in Component Indices and HDIs across 123 Countries between 1980 and 2013

	Increase in Respective Achievements (Weak)			Increase in All Achievements		Decrease in All Achievements		Change in HDI'_A				Change in HDI_G			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Time Period	Health	Education	Income	Strict	Weak	Strict	Weak	Increase	Robust	Decrease	Robust	Increase	Robust	Decrease	Robust
1980–1985	117	113	69	58	62	0	1	111	81	12	1	116	83	7	1
1980–1990	115	117	76	66	70	0	1	112	85	11	2	115	88	8	2
1980–2000	108	122	82	73	78	1	1	116	95	7	1	118	95	5	1
1980–2005	112	123	94	84	88	0	0	118	99	5	0	119	107	4	0
1980–2010	117	123	100	94	96	0	0	122	111	1	0	122	112	1	0
1980–2013	121	123	105	100	103	0	0	123	111	0	0	123	113	0	0
1985–1990	110	113	83	65	70	2	2	106	81	17	5	108	87	15	5
1985–2000	101	121	89	76	80	1	1	114	96	9	2	116	98	7	2
1985–2005	110	122	99	90	94	1	1	118	106	5	2	118	107	5	2
1985–2010	117	123	107	101	103	0	0	119	113	4	0	119	114	4	0
1985–2013	118	123	106	100	102	0	0	120	114	3	0	120	115	3	0
1990–2000	104	118	98	83	87	1	1	110	101	13	1	112	100	11	1
1990–2005	110	122	107	97	101	0	0	116	105	7	1	116	105	7	1
1990–2010	117	123	109	103	105	0	0	120	112	3	0	119	112	4	0
1990–2013	119	123	110	104	106	0	0	121	114	2	0	121	115	2	0
2000–2005	118	118	106	94	97	0	0	117	109	6	0	116	107	7	0
2000–2010	120	120	109	102	103	0	0	123	114	0	0	123	114	0	0
2000–2013	122	121	108	103	105	0	0	123	116	0	0	123	117	0	0
2005–2010	122	118	101	94	96	0	0	121	110	2	0	122	108	1	0
2005–2013	123	119	101	95	98	0	0	123	109	0	0	123	111	0	0
2010–2013	123	113	102	58	93	0	0	113	99	10	0	113	99	10	0

Source: Author computations using UNDP data.

one-tenth important than the other two dimensions together. We acknowledge that these requirements are set only for exemplary purposes and in practice the requirements should be determined by public debate. Given that weights are bounded, this means that no dimension should be assigned a weight that is more than 0.75 and at the same time no dimension should be assigned a weight that is less than 0.1. The set of alternative weights then is the convex hull of six weights: (0.75,0.15,0.1), (0.75,0.1,0.15), (0.15,0.75,0.1), (0.15,0.1,0.75), (0.1,0.15,0.75), and (0.1,0.75,0.15). We denote the set of alternative weights that is the convex hull of these six points by Δ_1. Note that the shape of the alternative weights would be the same as the shape in Panel A of Figure 1.

Columns 10, 12, 14, and 16 of Table 3 present the number of countries for which the changes in HDI'_A and HDI_G are robust with respect to Δ_1. Between 1980 and 1985, of the 111 increases in HDI'_A, only 81 were robust with respect to Δ_1, with 30 not being robust. Of the 12 reductions in HDI'_A within the same period, only one was robust with respect to Δ_1. For HDI_G, of the 116 increases 83 were robust. The largest number of robust changes in HDIs were observed between 2000 and 2013, when 116 HDI'_A changes and 117 HDI_G changes were robust. The lowest number of robust changes in HDIs were observed between 1980 and 1985, when 82 HDI'_A changes (81 increases and one decrease) and 84 HDI_G changes (83 increases and one decrease) were robust.

Table 4. The Number of Periods in which Robust Changes in HDI Occurred

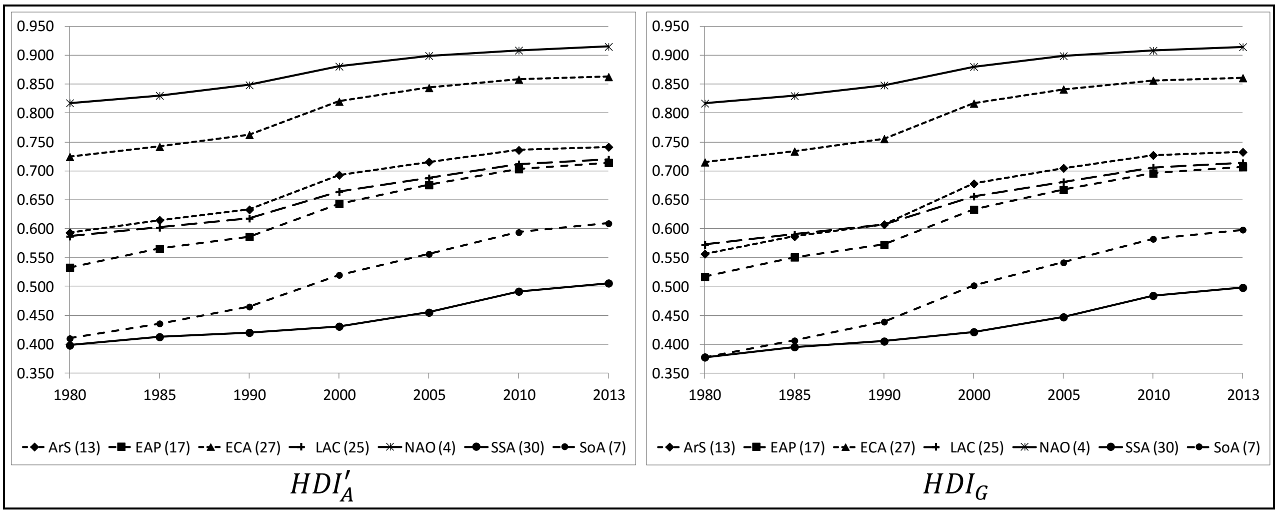
	Change in HDI'_A				Change in HDI_G			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Number of Time Periods	Number of Robust Increases	Share (%)	Number of Robust Decreases	Share (%)	Number of Robust Increases	Share (%)	Number of Robust Decreases	Share (%)
6	36	29.3	0	0.0	38	30.9	0	0.0
5	37	30.1	0	0.0	35	28.5	0	0.0
4	36	29.3	0	0.0	35	28.5	0	0.0
3	9	7.3	0	0.0	12	9.8	0	0.0
2	4	3.3	1	0.8	2	1.6	1	0.8
1	1	0.8	5	4.1	1	0.8	5	4.1
0	0	0.0	117	95.1	0	0.0	117	95.1
Total	**123**	**100.0**	**123**	**100.0**	**123**	**100.0**	**123**	**100.0**

Source: Author computations using UNDP data.

Let us now present a striking result by considering the following six successive periods: 1980–1985, 1985–1990, 1990–2000, 2000–2005, 2005–2010, and 2010–2013. We ask the following question: For how many countries were the changes robust with respect to Δ_1 across all six periods? The answer to this question will tell us how many countries improved gradually but robustly at the same time. Table 4 reports the number of countries whose changes were robust across all six selected periods. The first column reports the number of periods, which range from zero to six. The second and the sixth columns report the number of countries for which the improvements were robust for a number of periods. We surprisingly find that of the 123 countries, only 36 had improvements in HDI'_A that were robust across all six periods. In terms of HDI_G, the number of robust comparisons across all six periods was only slightly higher, 38 of the 123 countries.[11] For a similar number of countries, the HDI'_A and HDI_G improvements were robust across four or five periods. For only one country, the improvements were robust only for one period. We did not find any country among the 123 countries, however, for which all improvements were non-robust with respect to Δ_1.

It is interesting to explore how the number of non-robust comparisons were distributed across different geographic regions. The dataset for 123 countries was divided across seven regions: Arab States (ArS), East Asia and the Pacific (EAP), Europe and Central Asia (ECA), Latin America and the Caribbean (LAC), North America and Oceania (NAO), Sub-Saharan Africa (SSA) and South Asia (SoA). The two graphs in Figure 2 show the change in regional average HDI'_A and HDI_G. Although the pace of improvement varied across different regions between 1980 and 2013, all regional averages showed gradual improvement. Which regions showed the most robust improvements? The question is answered in Table 5.

Figure 2. Change in HDI between 1980 and 2013



Despite the steady improvements, none of the regions had all countries improving robustly in all six periods. The first column of Table 5 reports the seven regions and the second column reports the number of countries in our study falling in that particular group. Of the 36 robust HDI'_A changes and robust HDI_G changes across all six periods, the least number of robust comparisons were observed in the

[11] In fact, in only 24 countries did all three dimension achievements increase across all six periods. This means that if we had allowed the weights to vary to their fullest extents so that any dimension may receive a maximum weight of one and a minimum weight of zero, then only 24 countries would have registered robust improvement across all six periods.

Sub-Saharan African region. Only 6.7% of the HDI_A' changes and 10% of HDI_G changes were robust with respect to Δ_1.

Table 5. Robust Changes in All Periods across Regions

Geographic Region	Number of Countries	Robust Change in All Six Periods			
		HDI_A'	Share (%)	HDI_G	Share (%)
Arab States	13	3	23.1	4	30.8
East Asia and the Pacific	17	9	52.9	8	47.1
Europe and Central Asia	27	9	33.3	9	33.3
Latin America and the Caribbean	25	5	20.0	6	24.0
North America and Oceania	4	3	75.0	3	75.0
Sub-Saharan Africa	30	2	6.7	3	10.0
South Asia	7	5	71.4	5	71.4
Total	**123**	**36**	**29.3%**	**38**	**30.9%**

Source: Author computations using UNDP data.

5. Conclusion

This paper has looked at the robustness of comparisons of composite indices with respect to their chosen weights. These weights are typically chosen arbitrarily and as such there is ambiguity over the comparison of index scores, be they in relation to cross-section rankings or inter-temporal comparisons of index scores for the units of analysis under consideration.

Two objectives were pursued in the paper. The first was to address a difficulty encountered by several previous studies: the selection of alternative weighting schemes for assessing the robustness of comparisons. This selection is a requirement of the tests proposed by these studies, yet none provide sufficient guidance for such selection. This paper proposed a general yet theoretically novel approach for this purpose. This approach is founded on the normative assumption that a consensus has been reached on the minimum and the maximum allowable weights that should be assigned to each component. This consensus then yields a particular set of alternative weights against which the robustness of comparisons is tested.

The second objective pursued in the paper was to evaluate the prevalence of robust country-specific inter-temporal comparisons of the influential HDI. Testing the robustness of inter-temporal comparisons of the HDI or other composite indices has not previously been attempted. The results of this evaluation were striking. It found that less than one-third of the inter-temporal HDI comparisons were robust across six subperiods between 1980 and 2013. A geographical breakdown of countries showed that most of the non-robust improvements were observed in Sub-Saharan Africa, where no more than 10% of HDI comparisons were robust across all six periods. This has obvious and serious implications for the use of the HDI in incisively assessing changes in human development achievements over time.

We end this paper by adding voice to previous calls for greater warning signals to be attached to the use of composite indices for which there is insufficient guidance, theoretical or otherwise, in their design. A great risk is that unless greater care and sophistication are used in the reporting of composite indices, their ability to inform could be compromised. It is commonplace in reporting the results of econometric analysis to provide a range of diagnostic and other statistics, including *t*-ratios, so the reader can make judgments about the veracity of these results. No equivalent statistics presently accompany the reporting of composite index scores. It is high time that they did, and the reporting of comparison robustness information would be a useful starting point. We further acknowledge that the proposed robustness approaches should be such that they are intuitive and practically amenable to adaptations and implementations. Our paper has proposed an axiomatic yet intuitive approach that seeks to fulfill this criterion.

References

Atkinson, A. B. (1970). On the Measurement of Inequality. *Journal of Economic Theory, 2*, 244–263.

Atkinson, A. B. (1987). On the Measurement of Poverty. *Econometrica, 55*, 749-764.

Bandura, R. (2008). *A Survey of Composite Indices Measuring Country Performance: 2008 Update*. UNDP/ODS Working Paper. New York: Office of Development Studies, United Nations Development Programme.

Cahill, M. B. (2005). Is the Human Development Index Redundant? *Eastern Economic Journal, 31*, 1-5.

Cherchye, L., Ooghe, E., & Puyenbroeck, T. V. (2008). Robust human development rankings. *Journal of Economic Inequality, 6*, 287-321.

Decancq, K., & Lugo, M. A. (2013). Weights in Multidimensional Indices of Well-Being: An Overview. *Econometric Reviews, 32*, 7-34.

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

Foster, J. E., & Shorrocks, A. B. (1988a). Poverty Orderings. *Econometrica, 56*, 173–177.

Foster, J. E., & Shorrocks, A. F. (1988b). Poverty Orderings and Welfare Dominance. *Social Choice and Welfare, 5*, 179–198.

Foster, J. E., McGillivray, M., & Seth, S. (2009). *Rank robustness of composite indices*. Working Paper 26, Oxford Poverty and Human Development Initiative. Oxford: University of Oxford.

Foster, J. E., McGillivray, M., & Seth, S. (2013). Composite Indices: Rank Robustness, Statistical Association, and Redundancy. *Econometric Reviews, 32*, 35-56.

Høyland, B., Moene, K., & Willumsen, F. (2012). The tyranny of international index rankings. *Journal of Development Economics, 97*, 1-14.

Hopkins, M. (1991). Human development revisited: a new UNDP report. *World Development, 19*, 1469-1473.

Kahneman, D., & Deaton, A. (2010). High Income Improves Evaluation of Life but Not Emotiional Well-being. *Proceedings of the national academy of sciences*, (pp. 16489-16493).

Kelley, A. C. (1991). The Human Development Index: "Handle with Care". *Population and Development Review, 17*, 315-324.

Marshall, A. W., & Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications*. New York: Academic.

Permanyer, I. (2011). Assessing the Robustness of Composite Indices Rankings. *Review of Income and Wealth, 57*, 306-326.

Ravallion, M. (2011). Mashup Indices of Development. *The World Bank Research Observer*, 1-32. doi:10.1093/wbro/lkr009

Tatlidil, H. (1992). *A New Approach for Human Development: Human Development Scores*. Sussex: Institute of Development Studies.

United Nations Development Programme. (1990). *Human Development Report*. New York: Oxford University Press.

United Nations Development Programme. (1993). *Human Development Report*. New York: Oxford University Press.

United Nations Development Programme. (1994). *Human Development Report*. New York: Oxford University Press.

United Nations Development Programme. (2009). *Human Development Report*. Basingstoke: Palgrave Macmillan.

United Nations Development Programme. (2010). *Human Development Report*. Basingstoke: Palgrave Macmillan.

United Nations Development Programme. (2014). *Human Development Report*. New York: United Nations Development Programme.

United Nations Development Programme. (2015). *Human Development Report*. New York: United Nations Development Programme.

Zheng, B., & Zheng, C. (2015). Fuzzy ranking of human development: A proposal. *Mathematical Social Sciences, 78*, 39-47.

Appendix

Proof of Theorem 1

We already know that $\Delta = \{w_1, \dots w_d \mid \alpha \leq w_d \leq \beta \; \forall d, \sum_{d=1}^{D} w_d = 1\}$ In order to prove the first part, suppose $\alpha \in [0,1/D)$ and $\beta = (1 - d\alpha)/(D - d)$ for any $d = 1, \dots, (D-1)$. Starting from the D-dimensional equal weight vector w^0, the most unequal weighting scheme $\bar{w} \in \Delta$ can be obtained by a finite number of regressive transfers following Definition 3. The most unequal weighting scheme $\bar{w}$ is obtained when $(D-d)$ elements are equal to β and d elements are equal to α. It is straightforward to verify that $(D-d)\beta + d\alpha = 1$, which satisfy the restriction on weights $\sum_{d=1}^{D} w_d = 1$. Given that β is repeated $(D-d)$ times and α is repeated d number of times, the number of unique permutation is equal to $D!/[(D-d)! \times d!)]$.

In order to prove the second part, suppose $\alpha \in [0,1/D)$ and $\beta \in ([1-d\alpha]/[D-d], [1-(d+1)\alpha]/[D-(d+1)])$ for any $d = 1, \dots, (D-2)$. Again, starting from the D-dimensional equal weight vector w^0, the most unequal weighting scheme $\bar{w} \in \Delta$ can be obtained by a finite number of regressive transfers following Definition 3. From the above restriction on β however it can be verified that $[D-(d+1)]\beta + (d+1)\alpha < 1$, whereas $(D-d)\beta + d\alpha > 1$. Which means that in order to obtain the most unequal weighting scheme $\bar{w} \in \Delta$, if we assign weight β to $[D-(d+1)]$ dimensions and weight α to $(d+1)$ dimensions, then weights do not sum to one and if we assign weight β to $[D-d]$ dimensions and weight α to d dimensions, then the total weights surpass one. Both cases violate the restriction of weights $\sum_{d=1}^{D} w_d = 1$. Thus the most unequal weighting scheme is obtained when $[D-(d+1)]$ dimensions are assigned weight β, d dimensions are assigned weight α, and the remaining weight γ is assigned to the remaining one dimension, where $\gamma = 1 - [D-(d+1)]\beta - d\alpha$ and it is straightforward to verify that $\gamma \in (\alpha, \beta)$. As β is repeated $[D-(d+1)]$ times and α is repeated d times, the total unique permutation of $\bar{w}$ is $D!/([D-(d+1)]! \times d!)$. ■

Proof of Theorem 2

To prove the sufficiency part, consider any $w' \in \Delta$. Then $w' = \sum_{m=1}^{D} \omega_m v_m$ by definition, where $\omega_m \geq 0$ and $\sum_{m=1}^{D!} \omega_m = 1$. If $C(y; v_m) \geq C(x; v_m)$ for all $m = 1, \dots, \bar{D}$, then it follows that $C(y; w') \geq C(x; w')$. Hence, the comparison $C(y; w^0) \geq C(x; w^0)$ is robust with respect to Δ.

For the necessary part, suppose $C(y; w^0) \geq C(x; w^0)$ is robust with respect to Δ and so that $C(y; w) \geq C(x; w)$ for all $w \in \Delta$. Given that $v_m \in \Delta$ for all $m = 1, \dots, \bar{D}$, it follows that $C(y; v_m) \geq C(x; v_m)$ for all $m = 1, \dots, \bar{D}$. ■