Social Indicators Research (2007) 82: 111–145
DOI 10.1007/s11205-006-9029-7

LAURENS CHERCHYE, WILLEM MOESEN, NICKY ROGGE and TOM VAN PUYENBROECK

AN INTRODUCTION TO 'BENEFIT OF THE DOUBT' COMPOSITE INDICATORS

(Accepted 23 May 2006)

ABSTRACT. Despite their increasing use, composite indicators remain controversial. The undesirable dependence of countries' rankings on the preliminary normalization stage, and the disagreement among experts/stakeholders on the specific weighting scheme used to aggregate sub-indicators, are often invoked to undermine the credibility of composite indicators. Data envelopment analysis may be instrumental in overcoming these limitations. One part of its appeal in the composite indicator context stems from its invariance to measurement units, which entails that a normalization stage can be skipped. Secondly, it fills the informational gap in the 'right' set of weights by generating flexible 'benefit of the doubt'-weights for each evaluated country. The ease of interpretation is a third advantage of the specific model that is the main focus of this paper. In sum, the method may help to neutralize some recurring sources of criticism on composite indicators, allowing one to shift the focus to other, and perhaps more essential stages of their construction.

KEY WORDS: composite indicators, data envelopment analysis, performance benchmarking, technology

1. INTRODUCTION

The mere variety of composite indicators reflects their recognition as tools for policy evaluation and communication. Yet despite their increasing prevalence, composite indicators remain the subject of controversy. The lack of a standard construction methodology, and particularly the inescapable subjectivity involved in their construction, are invoked by opponents to undermine their credibility. Subjective choices are indeed pervasive when answering the many questions bound up with a composite indicator (see Booysen, 2002): what is the overall phenomenon one purports to summarize; which sub-indicators should be included; how should they be

An abridged version of this paper was presented at the Workshop on European Indicators and Scoreboards, organised by DG Education and the Joint Research Centre within the auspices of CRELL, in Brussels, October 24–25, 2005.

aggregated; how to deal with missing or low quality data; to what extent can one assess how country rankings are influenced by all the foregoing questions, etc.?

In this paper, we will not delve into all of these matters. Some of them are fundamental, as they relate to the substantive content of any composite indicator: is it just a contrivance to summarize several data dimensions, or does one really aspire to summarize a complex, multi-faceted phenomenon such as human development, social inclusion, the knowledge-based economy, competitiveness,...? We will take it here that *summarizing* is one of its two essential purposes, the other one being the idea of *comparing* several countries (or the evolution of a country over time, and the like). We will also take it that composite indicators bear, although limitedly, on public debate. Because they are so easy to use as communication tools, they inevitably do show up in media headlines and in press releases of well-respected international organizations, so at least increasing awareness of specific issues in society. In such cases, they often have an hit-parade appearance. And most probably, this feature only aggravates uneasy feelings about composite indicators in scholarly circles.

Evidently, there are reasons to suspect composite indicators. Summarizing inevitably entails reducing available information, and therefore, possibly, obscuring *essential* information. Other reasons, including the ones we will be concerned with here, are more methodological. To introduce them on a very general level, consider the following table, which provides the raw data to create a very simple 'technology creation' composite index by averaging over two sub-indicators. The sub-indicators we use in this simple example are (i) patents granted to residents (1998, per million people), and (ii) the receipts of royalties and license fees (in 1999 US$ per 1000 people). The figures, for Sweden and the US, are taken from Table I in Desai et al. (2002).

A first thing to note is that we only consider two countries, but already face the problem that one cannot rank them unless one aggregates the sub-indicators: the US is better than Sweden on the patent dimension, while the

TABLE I

Two Indicators of technology creation

	Patents	Royalties
US	289	130.0
Sweden	271	156.6

reverse holds for royalties. As is well-known, this problem is endemic when, as usual, more sub-indicators are taken into account. Micklewright (2001) warns us of the danger that, lacking a good composite index, excessive public attention may eventually be again focused on just one or a few dimensions, thus abolishing the original desideratum of portraying a multidimensional phenomenon. Indeed, this could undermine the credibility of performance evaluation. We thus move on to consider some possible composite indicators, viz. of the following familiar form:

$$(1) \qquad CI_c = \sum_{i=1}^{m} w_{c,i} \cdot y_{c,i}^n$$

with CI_c the composite index for country j, $y_{c,i}^n$ the (possibly normalized) value for country j on indicator i (i = 1,... m) and w_i the weight assigned to indicator i. In general, weights are bounded in that $0 \leq w_{c,i} \leq 1$ and $\sum_{i=1}^{m} w_{c,i} = 1$. (For most of this text, we will adhere to a summation of sub-indicators as the aggregation rule. We discuss an alternative in the concluding section). Specifically, we consider:

(a) The arithmetic average of the above numbers
(b) The arithmetic average of the above numbers, yet with the royalties data expressed in another currency (say in XXX, with 1US$ = 0.5 XXX)
(c) The arithmetic average of *normalized* numbers, using the 're-scaling' formula

$$X_{j,i}^n = \frac{X_{j,i} - X_i^{\min}}{X_i^{\max} - X_i^{\min}}$$

(d) The arithmetic average of *normalized* numbers, using the 'distance to group leader' formula

$$X_{j,i}^n = \frac{X_{j,i}}{X_i^{\max}}$$

(e) Same as (a), but with unequal weighting (75%-weight for patents, 25% for royalties)
(f) Same as (d), but with unequal weighting (75%-weight for patents, 25% for royalties)

Table II shows the results. Using the raw data to construct a CI seems a questionable undertaking: the eventual country ranking is contingent upon the units of measurement (compare a and b), even if the switch to another currency is in itself a perfectly plausible transformation.

TABLE II
Confusing composites

	CI-a	CI-b	CI-c	CI-d	CI-e	CI-f
US	209.5	177.0	0.5	0.915	249.3	0.957
Sweden	213.8	174.7	0.5	0.969	242.4	0.953

In fact, getting rid of measurement units – notably when these differ across dimensions – is one reason why CI practitioners employ normalization methods. However, this doesn't really solve the problem. Cases c and d show two conventional methods ($X_i^{\max}$ resp. $X_i^{\min}$ in the formulas refer to the highest resp. lowest value for the sub-indicator i over all countries in the data set). A first general remark is that normalization obscures the original purpose of the indicator: one is no longer summarizing the original data, but rescaled scores, or distances to goalposts, or z-scores, and the like. Evidently, this also bears on the inter-country score comparisons. Even if scenario d would be the only one considered, and even if one agrees that these two dimensions suffice to capture technology creation performance, it is clearly difficult to attribute a clear meaning to the statement that Sweden's global performance in technology creation is "5% better" than that of the US.

There is, however, an observation that is still more worrying. A comparison of CI-c and CI-d with starting point CI-a reveals the well-documented criticism that, keeping the weighting system fixed, the eventual rankings still depend on the particular (and so-called 'preliminary') normalization option taken. Ebert and Welsch (2004) criticize the dependency of eventual ranks on the normalization/aggregation procedure from a measurement-theoretic point of view. In a well-defined mathematical sense, a composite indicator is *not meaningful* when the resulting country ordering changes if the original data are transformed in such a way that there informational content is not fundamentally altered. In practice, however, most composite indicators are prone to precisely this deficiency. It is obvious that countries with lower rankings due to a specific normalization procedure may invoke this dependency to question the credibility and the use of composite indicators. Removing the requirement to normalize the data would eliminate this dependency and, thus, an important criticism.

We stress at this point that part of the problem relates to the fact that one is using a *fixed* set of weights. Equal weighting, which is just a specific case of fixed weighting, is regularly invoked as the standard in virtue of its simplicity (e.g. by Babbie, 1995). We have just shown that this alleged

simplicity is often thoroughly misleading: fixing weights and modifying (normalized) sub-indicator values is bound to lead to the problems just discussed.[1]

The last two columns illustrate another source of recurring criticisms. To be clear, it is not so much that country scores and rankings also depend on the weights. This is surely true, but the essential (mathematical) reason for this is actually the same as for the normalization issue: now start modifying weights on a fixed set of (normalized) sub-indicator values, and the same problem shows up in a different form. The deeper problem is that subjective judgments about the relative 'worth' of each of the sub-indicators enter through the weights.[2] In fact, in linear composites such as those introduced, moving from CI-a to CI-e entails that a one unit increase in royalties (i.e. 1 US$ per 1000 people; or 1000 US$ per million) can no longer offset a one unit decrease in the patent dimension (i.e. 1 patent less per million people). In order to keep a country's 'technology achievement'-score unaffected in the CI-e case, one instead needs 3000 dollars per million people in order to compensate for the unit decrease in patents.

Two points can be made here. First, it is not at all clear what ('paternalistic') judgments to impute, especially since weighting information stemming from stakeholders is often characterized by strong inter-individual disagreements. Second, one may dislike the idea of compensation itself.[3] In fact, it has been observed (e.g. by Munda and Nardo, 2003), that experts usually don't interpret weights as defining trade-offs between sub-indicators, but rather as 'importance coefficients' (cf. Freudenberg, 2003, p. 10: "Greater weight should be given to components which are considered to be more significant in the context of the particular composite indicator"). Consequently, we will seek to adhere to such an interpretation below.

The rest of this text discusses how data envelopment analysis (DEA) helps to overcome the issues just raised. This approach has already been applied to composite indicators in the context of policy performance assessment. For example, it has been used to gauge countries' performance with regard to aggregate deprivation (Zaim et al., 2001), to provide an alternative weighting system for the Human Development Index (Mahlberg and Obersteiner, 2001; Despotis, 2005), or as a generalized gauge for Sustainable Development (Cherchye and Kuosmanen, 2006). Especially in the European context, where tensions between the centre and member states may also bear on the precise way by which the latters' policies are evaluated, the need for a flexible weighting system may be warranted. Indeed, besides academic contributions (e.g.: European Unemployment policy (Storrie and Bjurek, 2000), Social Inclusion policy (Cherchye et al., 2004), and Internal Market

policy (Cherchye et al., 2005)), the European Commission itself has used the technique to gauge member states' performance with regard to the Lisbon objectives (European Commission, 2004, pp. 376–378).

The data employed for the Technology Achievement Index (TAI), of which the data in Table I are a miniature subset, are used to provide illustrative examples. The main reason for using the TAI is that it figures likewise in the *Handbook on the construction of composite indicators* of Nardo et al. (2005). In that handbook, where the benefit of the doubt approach is also discussed, one can find both the original TAI-data and ample uses of them to illustrate various issues in composite indicator construction. The fact that we dispose of individual expert information about TAI-weights (see Appendix 1) makes this application especially appealing in the current context.

Section 2 describes, for a non-specialist audience, DEA and the related Benefit of the Doubt method in more detail. Its possible elimination of the dependency of the results on preliminary normalization, and its characteristic of offering flexibility under the form of endogenous weighting, may well tone down some of the aforementioned criticisms on composite indicators. We will stress such fundamental intuitions and show some basic formulas, focusing less in this paper on technical/computational aspects of DEA. These are treated at length in various publications (see e.g. Cooper et al., 2004, or Zhu, 2003 for surveys).

In Section 3 we extend the basic model by appending ''sub-indicator share restrictions''. Such restrictions can be interpreted as bounds for the importance of sub-indicators in the composite score. The approach allows for a straightforward pie-chart representation of composite indicators, with the total size of the pie indicating a country's score, and the (bounded) pie shares indicating how each sub-indicator contributes to this overall value. Some different variants of these 'pie share'-restrictions are discussed and illustrated.

Section 4 summarizes and offers some concluding remarks.

2. DEA AND ''BENEFIT OF THE DOUBT''-WEIGHTING

2.1. *Building a benefit of the doubt indicator*

DEA, initially developed by Charnes et al. (1978), is a (linear programming) tool for evaluating the performance of a set of peer entities that use (possibly multiple) inputs to produce (possibly multiple) outputs. The original question in the DEA literature is how one could measure each entity's efficiency,

given observations on input and output quantities in a sample of similar entities and, often, no reliable information on prices, in a setting where one has no knowledge about the 'functional form' of a production or cost function. However broad, one immediately appreciates the conceptual similarity between that problem and the one of constructing CIs, in which quantitative sub-indicators are available but exact knowledge of weights is not. Indeed, and unsurprisingly, the scope of DEA has broadened considerably over the last two decades, including macro-assessments of countries' productivity performance (e.g Kumar and Russell, 2002), and various applications to composite indicator construction (Cherchye et al., 2004, provide a list of such applications). In the latter context, the method has been labeled alternatively as the 'Benefit-of-the-Doubt'-approach (after Melyn and Moesen (1991), who introduced it in the context of macroeconomic performance evaluation).

This label derives from one of DEA's main conceptual starting points: (some) information on the appropriate weighting scheme for country performance benchmarking can in fact be retrieved from the country data themselves. Specifically, the core idea is that a good relative performance of a country in one particular sub-indicator dimension indicates that this country considers the policy dimension concerned as relatively important. Or, conversely, that a country attaches less importance to those dimensions on which it is demonstrably a weak performer relative to the other countries in the set. Such a data-oriented weighting method is justifiable in the typical CI-context of uncertainty about, and lack of consensus on, an appropriate weighting scheme.[4] This perspective clearly marks a deviation from common practices in composite indicator construction, as, for example, captured by the last four variants in Table II. In the words of Lovell et al. (1995, p. 508): ''Equality across components is unnecessarily restrictive, and equality across nations and through time is undesirably restrictive. Both penalize a country for a successful pursuit of an objective, at the acknowledged expense of another conflicting objective. What is needed is a weighting scheme which allows weights to vary across objectives, over countries and through time''.

Admittedly, some may interpret the latter quote as indicating that the cure of flexible weighting is even worse than the disease of fixed (and equal) weighting. A main objective of this and the following section is to show that this is not the case, for at least the following three reasons. First, the benefit-of-the-doubt weighting approach is inherently bound up with the idea that even under such flexible weighting a country can be outperformed by some other country in the sample (see particularly expressions (2)–(4) below). Second, it is precisely due to the flexible nature of weights, i.e. because

weights can adapt to the choice of measurement units, that the normalization problem of composite indicators may be sidestepped (see Section 2.2). And, last but not least, in cases where additional, even rough information on appropriate weights is available, this can often easily be incorporated into the evaluation exercise (see Section 3). In sum, the method may go some length in providing a practical means of implementing the idea expressed by Foster and Sen (1997, p. 206): "while the possibility of arriving at a unique set of weights is rather unlikely, that uniqueness is not really necessary to make agreed judgments in many situations."[5]

We will present the benefit-of-the-doubt formula in a step-wise fashion, in order to convey its underlying intuition clearly.

As stated in the introduction, the eventual purpose of composite indicators is to compare a country relative to the other countries in the set and/or to some external benchmark. The first step highlights this benchmarking objective: a country c's composite index score is not given by a weighted sum of its sub-indicators (as is done in (1)), but rather by the *ratio* of this sum to a (similarly weighted) sum of the benchmark sub-indicators y_i^B. Note that one thus introduces a quite natural "degree" interpretation for the CI-value: a value of 100% implies a global performance which is similar to that of the benchmark values, a value less (more) than 1 refers to worse (better) performance.

Step 1: the benchmarking idea

$$(2) \qquad I_c = \frac{\text{actual overall performance}}{\text{benchmark overall performance}} = \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\sum_{i=1}^{m} w_{c,i} y_i^B}$$

The next question relates to the identification of benchmark performance. For the time being, we concentrate on the case in which benchmarks are to be taken from the observed sample itself. This option gives a clear meaning to the notion of best *practice*: the eventual CI-value will be driven by comparison with other, *existing* observations, rather than with external (and necessarily normative) references. In particular, the benchmark observation specified in the denominator of (3) is itself obtained from an optimization problem, as indicated formally by the appearance of the max operator and its associated argument. It is in fact a country that, employing the weights $w_{c,i}$, obtains the maximal weighted sum. Consequently, this benchmark will be endogenous too: it may well differ from one evaluated country to another.

It should be noted that this selection yields further intuition to the CI-value of 1: if, for some reason or another, a country acts as its own benchmark (that is, if no other outperforming observation is found for this country), then we have in fact retrieved the maximal composite indicator value. Evidently, this upper bound is contingent on the particular choice not to include 'external' benchmark observations. If such external benchmarks exist (e.g. because one uses sub-indicator values of a previous time period), then this upper bound vanishes. Clearly, a CI-value of more than 1 would then bear a natural interpretation as well (see also the concluding section).

Step 2: selecting a country-specific benchmark

$$I_c = \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\max_{y_{j,i} \in \{\text{studied countries}\}} \sum_{i=1}^{m} w_{c,i} y_{j,i}} \tag{3}$$

The following step pertains to the specification of the appropriate weights. Here, the benefit of the doubt-idea enters. The weighting problem is handled for each country separately, and the country-specific weights accorded to each sub-indicator are endogenously determined. The conceptual basis for this option is the data-oriented perspective mentioned above: good relative performance of a country (i.e., relative to other observed countries) on a sub-indicator dimension is considered to be revealed evidence of comparatively higher policy priority, while the reverse position is taken for sub-indicators on which the country performs relatively poorly. Stated otherwise, since one doesn't *know* a country's true (policy) 'weights', one assumes that they can be inferred from looking at relative strengths and weaknesses. Specifically, this perspective entails that the analyst looks for country specific weights which make its composite indicator value as high as possible.[6] In the absence of more verifiable information, this indeed means that each country is granted the benefit-of-the-doubt when it comes to assigning weights. Formally, this point is covered by the new max operator in equation (4). It also follows that this problem must be solved (separately) for each of the countries.

Step 3: selecting country-specific benefit-of-the-doubt weights

$$I_c = \max_{w_{c,i}} \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\max_{y_{j,i} \in \{\text{studied countries}\}} \sum_{i=1}^{m} w_{c,i} y_{j,i}} \tag{4}$$

In the absence of an a priori weighting scheme, the method thus selects the weights which maximize the composite indicator for each country under investigation. To put it differently: any other weighting scheme than the one specified in (4) would worsen the position of the evaluated country vis-à-vis the other countries. This quality explains a major part of the appeal of benefit of the doubt-based composite indicators in real settings. Countries cannot claim that a poor relative performance is due to a harmful or unfair weighting scheme.[7]

Two more features are added. One is a normalization constraint (5a), stating that no other country in the set has a resulting composite indicator greater than one when applying the optimal weights for the evaluated country. Being a scaling constraint, the precise value of this upper bound is, of course, arbitrary. Yet, once again, (5a) highlights the benchmarking idea: the most favorable weights for one country are always applied to all (n) observations. One is in that way effectively looking which of the countries' sub-indicator values are such that they would lead to a worse, similar, or... better composite score, *when applying the most favorable weights for the evaluated country*. If there are indeed countries in the third class, a strong case can be made for the notion of 'being outperformed': despite the fact that one allows for country-specific benefit-of-the-doubt weights, there is then still at least one other country which, using the same weighting scheme, does even better.[8]

Constraint (5b) limits the weights to be non-negative. Hence, the composite indicator is a non-decreasing function of the sub-indicators, and the total composite indicator value is bounded below as well. That is, $0 \leqslant I_c \leqslant 1$ for each country, where higher values represent a better overall relative performance.

$$(4, \textit{repeated}) \qquad I_c = \max_{w_{c,i}} \frac{\sum_{i=1}^{m} w_{c,i} y_{c,i}}{\max_{y_{j,i} \in \{\text{studied countries}\}} \sum_{i=1}^{m} w_{c,i} y_{j,i}}$$

s.t.

$$(5a) \qquad \sum_{i=1}^{m} w_{c,i} y_{j,i} \leqslant 1$$

(n constraints, one for each country j)

$$(5b) \qquad w_{c,i} \geq 0$$

(m constraints, one for each indicator i)

Considering the fact that, by construction, the benchmark observation attains the maximal composite indicator value of 1, the above (fractional) maximization problem can be written in a linear form, which is computationally easier to handle (e.g. by Excel-solvers):

$$(6) \qquad I_c = \max_{w_{c,i}} \sum_{i=1}^{m} w_{c,i} y_{c,i},$$

subject to constraints (5a) and (5b).

As stated above, this method is rooted in DEA. It is indeed easily verified that the model just presented is formally tantamount to the original input oriented DEA model of Charnes et al. (1978), with all sub-indicators considered as outputs and a 'dummy input' equal to one for all the countries. In that reading, the dummy input for each country may be interpreted as a 'helmsman' that pursues several policy objectives, corresponding to the different sub-indicators; see e.g. Lovell et al. (1995). Still, it should be clear from our discussion that an intuitive interpretation may also be obtained simply by regarding the model as a tool for aggregating several sub-indicators of performance, without explicit reference to the inputs that are used for achieving such performance. The problem is then indeed one in a "pure output setting" (a term coined by Cook (2004)), in which the normalization constraint (5a) is interpreted as a scaling or bounding condition (see also Cook and Kress, 1991, 1994). Indeed, the most notable difference between general DEA problems and the problem we just discussed for constructing composite indicators, is that composite indicators typically only look at 'achievements' without taking into account the input-side. A valuable side-remark, which we will not pursue further in this paper, thus emerges: the method just described is fully apt to deal with CI-construction in the prevailing case where input sub-indicators would appear along with achievement sub-indicators. In fact, the DEA-model of Zaim et al. (2001) exploits this characteristic.[9]

2.2. *Unit invariance and sub-indicator shares*

An important feature of the DEA framework and, hence, of the benefit-of-the-doubt model is its unit invariance: the value of the composite indicator is independent of the units of measurement of the sub-indicators. We will not provide a formal proof of this statement here (see e.g. Cooper et al., 2000, p. 39), but the underlying intuition should be clear: the fundamental reason for this unit invariance goes back to the feature that weights are endogenous.

Endogeneity implies flexibility, and this in turn will cause weights to adapt to the units of measurement. By way of an example, recall that the sub-indicator 'number of patents granted' is originally expressed per 1,000,000 people. Assume then that the associated weight derived from problem (4)–(5a)–(5b) is $w_{c,\ \text{patents}}$. If we would instead express the patent indicator per 1000 people (in all countries, of course), the optimal weight will automatically be rescaled (to $w_{c,\ \text{patents}} \times 1000$) to preserve the normalization constraint. Note that one would expect exactly the same thing e.g. in a consumer price index: if one measures apples in kilos, then the proper weight attached is the price per kilo. And if one switches to measuring apples in tons, then the price should change accordingly.

One clarifying side-remark is in order: not *all* conceivable data transformations preserve the outcome of program (4)–(5a)–(5b). For instance, it is easy to see that automatic rescaling of the weights no longer holds if one normalizes a sub-indicator by substituting original values with the countries' mutual rank on that dimension. Ratio-scale transformations (e.g., scenarios (b) and (d) in our example in the Introduction section; see Table II) will lead to exactly the same outcome as when using the original data. The exact bearing of this remark should perhaps be spelled out: in the above model, normalization is in fact *redundant* for sub-indicators that are measured on a ratio scale (prices, percentages, etc.), even if the units of measurement differ *between* the sub-indicators.[10]

At first sight, the feature that weights adapt to units may just seem to shift the problem of dependency on measurement units to the weights without actually solving it. The above example certainly carries the caveat that, even if the overall score is unaffected, the weights are not: they do depend on the units of measurement. Consequently, one should be cautious when comparing and interpreting benefit of the doubt weights. Also, if one would impose additional restrictions on the weights (i.e., in addition to (5b)), it may well be difficult to give an instantly recognizable meaning to such restrictions.

Two escape routes are, however, feasible. One is to normalize the original data anyway (using an allowable transformation), as is e.g. done in Cherchye et al. (2004). A second one, which we will pursue further on the following pages, is to shift the focus to 'sub-indicator shares', which are completely independent of measurement units.

Sub-indicator shares are in fact the *product* of the original value of the sub-indicator $y_{c,i}$ and the assigned weight $w_{c,i}$.[11] As can be inferred from the preceding discussion, rescaling the sub-indicators will not change the sub-indicator shares value, given that such rescaling is countervailed by

the (inversely) rescaled weights. In point of fact, the product of the two remains unchanged. Furthermore, since measurement units cancel out upon multiplication of a weight with its associated sub-indicator, these sub-indicator shares are pure numbers. Note that this pure number quality was already implicit in (6) and (5a), as the sum of these sub-indicator shares in effect equals a pure number. Again, an analogy can be drawn with a consumer price index, in which one ultimately adds expenditure shares.

Referring back to equation (6), the eventual composite indicator can thus be re-interpreted as a sum of $i = 1 \ldots, m$ sub-indicator shares, one for each achievement dimension. The conceptual interpretation of each of these m variables is very straightforward: each pure number $w_{c,i}y_{c,i}$ indicates by how much dimension i contributes to the overall composite score of country c. Clearly, these m terms may also be interpreted as the 'pie shares' that together constitute I_c: the ith term represents the volume of the pie share of the ith sub-indicator. The total volume of the pie accordingly captures a country's composite indicator score, and the *relative* size of the shares reflects what we have earlier referred to as the relative importance/significance of the sub-indicators.

Figure 1 and Table III show how all this combines into a graphical and tabular representation. The results are shown only for a subset of the countries contained in the TAI dataset. (Additional information can be found in Appendix 2). The figure reveals the benefit-of-the-doubt nature of the exercise: the relative importance of the pie shares/sub-indicators is

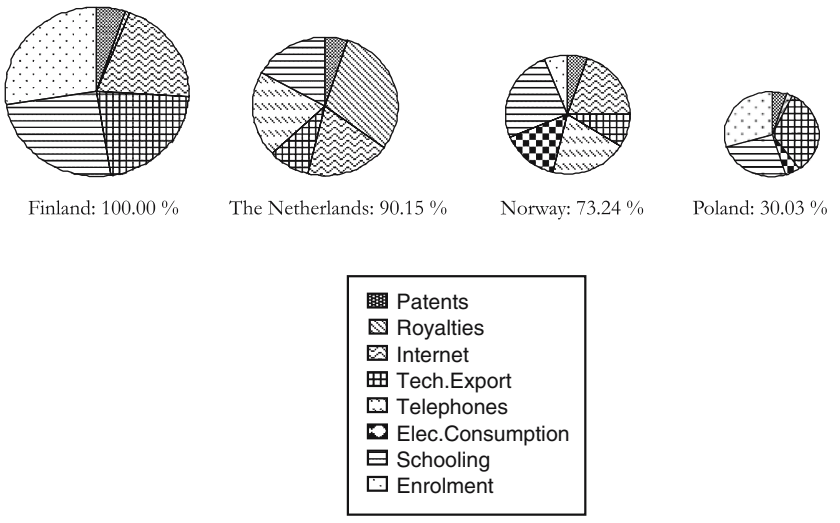

Fig. 1. Pie chart representation of benefit-of-the-doubt (TAI) index for selected countries.

TABLE III

Absolute values and percentage contributions to CI of subindicator shares

Country	Patents	Royalties	Internet	Exports	Telephones	Electricity	Schooling	Enrolment	Score
Sub-indicator shares									
Finland	0.0500	0.0093	0.2000	0.2148	0.0000	0.0000	0.2500	0.2759	100.00%
Netherlands	0.0451	0.2704	0.1712	0.0811	0.1803	0.0000	0.1534	0.0000	90.15%
Norway	0.0366	0.0000	0.1465	0.0659	0.1465	0.1099	0.1831	0.0439	73.25%
Poland	0.0150	0.0000	0.0060	0.0991	0.0000	0.0161	0.0751	0.0890	30.03%
Percentage contribution									
Finland	5.00%	0.93%	20.00%	21.48%	0.00%	0.00%	25.00%	27.59%	
Netherlands	5.00%	30.00%	18.99%	9.00%	20.00%	0.00%	17.01%	0.00%	
Norway	5.00%	0.00%	20.00%	9.00%	20.00%	15.00%	25.00%	6.00%	
Poland	5.00%	0.00%	2.00%	33.00%	0.00%	5.36%	25.00%	29.64%	

different over the four countries considered. And, a fortiori, this holds for their absolute size.

Table III provides more information. The upper part show the respective countries' values of sub-indicator shares, which, as indicated, sum up to their composite score. One infers, e.g., that the absolute values of the pie shares of top-ranked Finland are not always bigger than those corresponding to the other countries that are listed. In fact, one further sees that the listed countries do not even make use of *all* sub-indicators to arrive at their (benefit of the doubt) score: for each country, at least one dimension is left out. The underlying 'revealed evidence'-intuition for these observations is, again, that a country is not likely to put very much weight (and in the limit no weight at all) on dimensions in which it demonstrably has a comparative disadvantage relative to the performance of other countries in the sample.

In fact, if one would run the model (4)–(5a)–(5b) as it stands, one may expect this zero weight phenomenon to occur frequently. The mirror image of this full flexibility implies that many countries get the maximum score of 100% (see Appendix 2). Some observers dislike this particular aspect of (full) benefit of the doubt weighting. The results shown above are however *not* produced by this 'unrestricted' (full) benefit of the doubt model, but rather by a variant in which additional constraints, based on expert views, have been embedded. We will present that particular model in the next section.

The lower part of the table shows the percentage shares. Percentage contributions further reveal how each country is offered (some) leeway in assigning 'importance' to each of the components of the composite index. One notices some similarities (e.g. for patents, or for the schooling indicator), but some huge inter-country differences as well (e.g. for royalties, internet, and others). The next section addresses these findings in more detail.

3. SUB-INDICATOR SHARE RESTRICTIONS

Apart from the non-negativity of the weights (equation (5b)), the formal model hitherto discussed allows weights to be freely estimated in order to maximize the relative efficiency score of the evaluated country. (The weights are only restricted in that they must not make the final score exceed the upper limit of 1). The advantage of such flexibility is that it becomes hard for countries to argue that the weights themselves put them at a disadvantage. However, there are also disadvantages to this full flexibility. In some situations, it can allow a country to appear as a brilliant performer in a way that is difficult to justify. For example, if some zero weights are assigned as

in Table III, and if there is no prior information which backs up this possibility, some of the achievement indicators do not contribute to a country's composite measure. One then faces the risk of basing 'global' performance on a small subset of all (and often meticulously selected) sub-indicators.

By allowing full freedom, resulting outcomes may in particular contradict prior views on weights (e.g. expert opinions). In practice, it is essential for the credibility and acceptance of composite indicators to incorporate the opinion of experts that have a wide spectrum of knowledge, to ensure that a proper weighting scheme is established. True as this may be, it is at the same time also true that, in the area of composite indicator construction, experts may (strongly) disagree about the precise value of the weights. As is apparent from Table A1 in the Appendix 1, the TAI-case is one example of this recurrent phenomenon.

Fortunately, DEA models are able to incorporate such prior information by adding additional restrictions to the basic problem. This seems especially convenient in the common case where experts disagree on weights. In all probability, this is exactly the setting where the benefit of the doubt approach to CIs seems to be most powerful. When individual expert opinion is available, but when experts disagree about the right set of weights, the method is sufficiently flexible to incorporate 'agreed judgments' by imposing additional (e.g., sub-indicator share) restrictions. And at the point where disagreement remains, i.e. literarily where no further restrictions can be imposed, the informational gap is filled by choosing country-specific benefit-of-the doubt weights.

In our opinion, and with an eye towards practical applications, the latter reasoning may as well be reversed, so as to be more in line with the remark of Foster and Sen cited in Section 2.1. That is: it is easier to let experts agree a priori on *restrictions* than on a unique set of weights. The final result would then reflect what is actually there: *limited* agreement. Evidently, the nature of such restrictions can vary, and we will now briefly survey some alternatives.

As a prior note, we should mention that we will focus on restrictions on the pie shares. To recall, for each country j the i-the pie share equals $w_{j,i}y_{j,i}$, i.e. it is defined as the product of the original value of the sub-indicator $y_{j,i}$ and the corresponding pure weight $w_{j,i}$. In general, one could also consider adding restrictions on the pure weights themselves. All such restrictions are integrated in the original benefit of the doubt framework by adding the additional constraints to the programming problem (see Thanassoulis et al., 2004, for a broader overview of methods for appending 'value judgments'). Yet, the unit invariance of the sub-indicator/pie shares gives an apparent

advantage to restrictions on such shares in the current CI context. For example, this means that we can impose the sub-indicator share restrictions by starting from the original rather than the normalized TAI data. Furthermore, in view of the pie share interpretation introduced above, restrictions on sub-indicator shares allow for an easy and natural representation of prior information about the importance of the CI's components.

3.1. *Absolute restrictions*

Absolute restrictions on sub-indicator shares restrict these shares to vary between a specified range, specified by an absolute upper and lower bound (respectively α_i and β_i).

$$\alpha_i \leqslant w_{j,i} y_{j,i} \leqslant \beta_i \tag{7}$$

Absolute limits on sub-indicator shares could primarily be employed to prevent sub-indicators from being over- or under-emphasized. For example, one could prevent pie shares from being zero. However, one should be cautious when imputing such absolute restrictions. The key difficulty here is that the specification of bounds for one dimension is likely to have 'spill-over' effects to other dimensions. Specifically, given the n equations (5a), setting absolute bounds on a sub-indicator share implicitly affects the values the remaining shares can take. The unfortunate result may be that there is no feasible solution to the programming problem, given the mutual incompatibility of such bounds.[12] Hence, there may be better alternatives around, especially since these may also prevent too much or too little emphasis on particular sub-indicators.

3.2. *Ordinal sub-indicator share restrictions*

Let us now take the opposite, 'minimalist' perspective as regards the informational content of bounds. Assume experts do not agree on numerical bounds, but do agree that sub-indicator X should be "at least as important" as Y, "as important as" Z, "not more important than" W, etc. Golany (1988) proposed imposing ordinal restrictions on the pure weights, but one may apply a similar idea to the sub-indicator shares.

For the TAI, we provide an illustrative example, using the average budget allocation weights (as derived from Table A1) to determine the ordinal sub-indicator shares restrictions in (8). It should perhaps be stressed at this point that the specific form of the bounds in this as well as the following exercises is indeed meant to be illustrative: we aim to show how various

types of 'limited agreement' could be added to the weighting problem. In the ordinal case, one such specific form of limited agreement on the importance of sub-indicator shares could be:

$$(8) \qquad \begin{aligned} & w_{j,6}y_{j,6} \leqslant w_{j,5}y_{j,5} \leqslant w_{j,2}y_{j,2} \leqslant w_{j,3}y_{j,3} \leqslant w_{j,1}y_{j,1} \leqslant w_{j,7}y_{j,7} \\ & \leqslant w_{j,4}y_{j,4} \leqslant w_{j,8}y_{j,8} \end{aligned}$$

With an average weight of 0.063 sub-indicator 6 (''electricity'') is indicated as the least important, while the opposite holds for sub-indicator 8 (''enrolment'') with an average weight of 0.184. Using the average weights from the budget allocation method as a (possible) focal point, we thus partially integrate expert opinion. Of course, since there are no numerical bounds, this type of restriction still allows for quite some leeway in the assignment of 'residual' (benefit of the doubt) weights. For our four countries, the results are shown in Table IV. It is easy to check that (8) holds for each of them.

Introducing these restrictions results in CI values that are equal or lower than those resulting from 'full flexibility' calculations. From a mathematical perspective, this is just what is to be expected: adding restrictions to an optimization problem will never increase the value of its objective function. In non-mathematical terms, we are now allowing for benefit-of-the-doubt only *within* the a priori confines as specified (and assumingly agreed upon) by the panel of experts.

3.3. *Relative restrictions*

Conceivably, experts may agree on more powerful bounds than the ordinal variant just discussed. Relative restrictions on sub-indicator shares impose the *ratio* of sub-indicator share i and share k to vary between an upper and lower bound (see Pedraja-Chapparro et al., 1997). Thus, in the terms used above, such restrictions in fact restrict the *relative size* of two pie shares in the composite indicator. *Relative restrictions may facilitate the translation of expert knowledge when experts express their opinion on the pair-wise relative importance of the sub-indicators*. They would thus capture statements such as ''the pie share of indicator X can at most be double the size of the pie share of indicator Y'', etc.

$$(9) \qquad \alpha_i \leqslant \frac{w_{j,i}y_{j,i}}{w_{j,k}y_{j,k}} \leqslant \beta_i$$

To provide an example, we imposed such relative restrictions on the TAI data set. Our relative pie share constraints are directly inspired by the experts' stated weights of Table A1. The lower bound and upper bound we

TABLE IV

Absolute pie shares of the sub-indicators for the selected countries imposing ordinal share restrictions

Country	Patents	Royalties	Internet	Exports	Telephones	Electricity	Schooling	Enrolment	Score
Sub-indicator shares									
Finland	0.0836	0.0836	0.0836	0.0836	0.0836	0.0836	0.0836	0.4151	100.00%
Netherlands	0.0740	0.0740	0.0740	0.1224	0.0740	0.0275	0.1224	0.1224	69.08%
Norway	0.0230	0.0230	0.0230	0.1054	0.0230	0.0000	0.1054	0.1582	48.43%
Poland	0.0000	0.0000	0.0000	0.1522	0.0000	0.0000	0.1522	0.1522	45.64%

take to illustrate equation (9) are centered around the ratios of the dimension-specific *average* weights as specified by the experts. We then allow for some variation around these 'point values': the lower (upper) bound may be 25% lower (higher) than the mid-point. For m indicators (and as we want to restrict all possible pairs of sub-indicator portions in our example) we end up with $(m(m-1))/2$ restrictions. In the TAI-case, $m = 8$. For example, the relative restriction related to sub-indicators 1 and 2 (i.e., patents and royalties respectively) is:

$$\left(\frac{0.110}{0.107}\right) * 0.75 \leqslant \frac{w_{j,1} y_{j,1}}{w_{j,2} y_2} \leqslant \left(\frac{0.110}{0.107}\right) * 1.25 = 0.7710 \leqslant \frac{w_{j,1} y_{j,1}}{w_{j,2} y_{j,2}} \leqslant 1.2850 \quad (10)$$

Upon adding these restrictions, countries such as Korea and Singapore exhibit a considerable decrease in CI-values and concomitant rank (respectively 17.79% (rank = 20) and 8.65% (rank = 24), compared to 100% in the case of full flexibility; see Appendix 2). As for our earlier selection of four countries, we have the following pies (Figure 2):

Apparently, there are now far less considerable differences in the relative importance of the sub-indicators among the four different countries. Evidently, one merely gets back what one has plugged in by means of the relative restrictions (9). However, the relative pie shares need not be exactly the same (see Table V, e.g. for electricity).

We note without pursuing in greater detail that it may be interesting to ask whether and where the imposed (relative or other) bounds are binding. In the current example, one thus would learn that in the case of Finland, equation (10) attains the lower bound (dividing 0.0769 by 0.0997, or 8.81% by 11.43%, yields 0.771), indicating in fact that royalties is more important for this country. The opposite holds for Poland: 0.0042/0.0033 (or 10.66%/8.30%) implies that Poland could have generated a higher score if the experts had allowed for a relaxed upper bound on the importance of patents, relative to royalties.

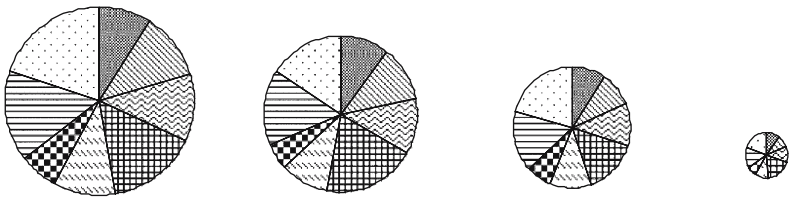

Fig. 2. Pie shares for the selected countries (with relative sub-indicator share restrictions.)

TABLE V

Absolute and percentage contribution of the sub-indicators (relative sub-indicator share restrictions)

Country	Patents	Royalties	Internet	Exports	Telephones	Electricity	Schooling	Enrolment	Score
Sub-indicator shares									
Finland	0.0769	0.0997	0.1015	0.1349	0.0913	0.0587	0.1379	0.1714	87.22%
Netherlands	0.0719	0.0798	0.0813	0.1350	0.0731	0.0376	0.1104	0.1098	69.89%
Norway	0.0406	0.0395	0.0503	0.0668	0.0482	0.0310	0.0729	0.0906	43.99%
Poland	0.0042	0.0033	0.0044	0.0073	0.0040	0.0026	0.0060	0.0075	3.92%
Percentage contribution									
Finland	8.81%	11.43%	11.64%	15.47%	10.47%	6.73%	15.81%	19.65%	
Netherlands	10.28%	11.42%	11.63%	19.32%	10.46%	5.38%	15.80%	15.71%	
Norway	9.23%	8.98%	11.44%	15.19%	10.97%	7.05%	16.56%	20.59%	
Poland	10.66%	8.30%	11.27%	18.72%	10.13%	6.52%	15.31%	19.03%	

3.4. *Proportional sub-indicator share restrictions*

Wong and Beasley (1990), proposed the following type of restrictions, to make it easier for the experts to quantify their opinion in terms of *percentage values*.

$$\alpha_i \leq \frac{w_i y_{j,i}}{\sum_{i=1}^{m} w_i y_{j,i}} \leq \beta_i \tag{11}$$

Importantly, the resulting construction of the composite indicator I_c still remains invariant to the units of measurement. These restrictions may be especially attractive in view of the fact that expert opinion is often collected by a 'budget allocation' approach, in which experts are asked to distribute (100) points over the different dimensions to indicate importance. In point of fact, the data in Appendix A1 are the results of such an approach. The stated 'weights' (which actually are budget shares) are then very easy to incorporate, via the form (11), in the benefit-of-the-doubt model. The only remaining issue is then how to specify bounds, given the observed diversity over individual experts.

To illustrate one particular possibility, we specified the lower and upper bounds by taking respectively the lowest and highest weight assigned over all experts to that sub-indicator. For sub-indicator "patents", this means $\alpha_i = 0.05$ and $\beta_i = 0.20$ (cf. Table A1), thus implying that its pie share should comprise at least 5% and at most 20% of the total pie size. In fact, the data shown earlier in Figure 1/Table III were produced precisely by this approach. The zero-weights assigned to some sub-indicators there can thus be seen to stem from the actual (lower) percentage bounds as forwarded by our expert panel. In fact, when reconsidering the results in Table III (and the information on expert weights in Appendix), one will find that in many cases the upper bounds are effectively binding.

We can check whether or not restrictions are binding for a country by looking at the percentage contribution of each sub-indicator. Just as in the previous case of relative restrictions, binding constraints imply that a country would have done better if these bounds had been relaxed.

We will use the notion of proportional restrictions on the sub-indicator shares (i.e. equation (11)) to illustrate one additional idea.

3.5. *Restrictions pertaining to category shares*

Often, composite indicators are constructed such that their sub-indicators can be classified in p mutually exclusive categories $S_1, \ldots, S_p$. Each

category then represents a certain orientation or focus of the evaluated phenomenon. The TAI itself provides an example of this: the eight sub-indicators are subdivided in four categories: technology creation (with sub-indicators patents and royalties), diffusion of recent innovations (internet and exports), diffusion of old innovations (telephones and electricity) and human skills (schooling and enrolment). Cherchye and Kuosmanen (2006), and Cherchye et al. (2005) show how this can be combined with weight restrictions. Here we apply this idea to restrictions on "category shares". Imposing restrictions on these category shares involves a straightforward extension of earlier restrictions. Formally, one could resort to:

$$\alpha \leqslant \frac{\sum_{i \in Sa} w_{j,i}\, y_{j,i}}{\sum_{i=1}^{m} w_{j,i}\, y_{j,i}} \leqslant \beta,$$

with S_a capturing category a, etc. This type of restrictions imposes bounds on the proportional importance of categories. When no additional restrictions are imposed, the pie shares of each category's constituent sub-indicators can be chosen with (relatively) large leeway.

By way of specific illustration, we determined the category bounds with reference to the average weights for the four categories rather than those for the eight sub-indicators. The dimension "technology creation", for example, gets an average weight of 0.217, etc. Just as before, a certain amount of allowable variation in the importance of each dimension is added, viz. minus 25% (lower bound) and plus 25 % (upper bound) of this average importance. Results are shown in Table VI. Notice that each of our selected countries ignores at least one individual sub-indicator, but without violating the imposed relative *category* share restrictions.

It is easy to see that such categorical restrictions need not be confined to a statement in terms of proportions: a similar extension of equation (9), i.e. to pair-wise bounds, is readily constructed.

Once more, the idea of imposing restrictions on categories arises from the common observation that it is difficult to define weights for individual sub-indicators. Again the gist of our argument holds: agreement on bounds on the level of categories is much simpler to obtain than specific weights for individual sub-indicators. Indeed, in most cases, focusing on the importance of key categories may allow one to obtain stakeholder consensus more swiftly. Imposing restrictions on categories may be taken as a first step in the quest for consensus among experts.

TABLE VI

Absolute pie shares of the sub-indicators and the four categories for the selected countries

Country	Patents	Royalties	Internet	Exports	Telephones	Electricity	Schooling	Enrolment	Score
Sub-indicator shares									
Finland	0	0.2708	0.2793	0	0	0.2006	0.2493	0	100.00%
Netherlands	0	0.2446	0.1671	0.0297	0.1812	0	0.2807	0	90.33%
Norway	0.0594	0.0532	0.2516	0	0	0.139	0.1897	0	69.30%
Poland	0.0239	0	0	0.0444	0.0177	0	0.0612	0	14.72%
	Dimension 1		Dimension 2		Dimension 3		Dimension 4		
Finland	27.08%		27.93%		20.06%		24.93%		
Netherlands	27.08%		21.79%		20.06%		31.07%		
Norway	16.25%		36.31%		20.06%		27.38%		
Poland	16.25%		30.16%		12.04%		41.55%		

4. CONCLUDING REMARKS

We recall our starting point for proposing the benefit of the doubt methodology to construct composite indicators: due to insufficiently precise, and probably unverifiable knowledge of the underlying structure of an evaluated composite phenomenon, uncertainty is inherent in the construction of composite indicators. The lack of a standard construction methodology, the disagreement among experts on the importance of the underlying indicators, etc., are just ways in which this uncertainty is manifested. But precisely these methodological aspects have been invoked to undermine the credibility of composite indicators. This defines a clear challenge for those who believe that composite indicators can be a useful tool for communicative purposes, as well as for those who believe that global comparisons of country performance and the closely related idea of benchmarking could eventually promote good policies. Cast against this general background, the preceding pages do certainly not offer a panacea for all problems bound up with composite indicator construction, but some aspects we touched upon may help to prevent getting bogged down in 'merely' methodological discussions.

The model and the pie share extensions discussed in the previous sections certainly do not exhaust the complete range of conceivable uses of the benefit-of-the-doubt approach. Indeed, we have already hinted at the fact that still other types of restrictions (on pure weights as well as on sub-indicator shares) are possible.[13] But the tool may be helpful in more general problem settings as well.

One such more general setting is, for instance, concerned with *dynamic performance evaluation*, i.e. one assesses the performance of a group of countries over time. We mentioned this possibility in Section 2. Clearly, best practices can (and probably do) alter over time. Zaim et al. (2001), and Cherchye et al. (2005) propose a benefit of the doubt aggregate performance index closely related to the Malmquist Productivity Index (MPI, first introduced by Malmquist in 1953) to assess countries' intertemporal performance shifts. Measuring performance change between two periods essentially boils down to comparing the aggregate sub-indicator performance in both periods. The weights for each of these two periods can again be selected endogenously. An appealing feature of the MPI is that it can be decomposed into a ''catching-up''-component and an ''environmental change''-component. The catching-up component indicates the performance change that is effectively due to a country's idiosyncratic improvement: did it get closer to its 'contemporaneous' benchmark or not, and by how much? The environmental change component instead focuses on the conduct of

benchmarks themselves, measuring the favorable or unfavorable change of best practices between the two periods. Note that these components may move in opposite directions: progress may be observed because of a strong country-specific better performance even if there is a less favourable environment than in the original period; or, vice versa, because it can be demonstrated that better practices have become possible, which the country nonetheless only partially exploits.

Other findings that originate from the vast DEA-literature may be readily applicable to the CI-context as well. For instance, and as we already touched upon, value judgements that originate from stakeholders may also be appended via the introduction of exogenous, possibly 'hypothetical' observations, to which all countries could then be compared. Or, to take another example: DEA has been broadened to deal with ordinal variables as well (see Cook, 2004, for a survey), and this clearly is an important extension with an eye towards some existing composite indicators, given that some (partially) build on 'soft' (categorical) survey data (e.g., the World Economic Forum's Competitiveness Index).

It is however also important to stress that, today, some possible extensions are best considered as promising avenues for further research. To give an example of the latter as well, recall the remark in Section 2 that in the general DEA framework sub-indicators are *linearly* aggregated into one single index. An alternative would be the geometric aggregation:

$$CI_c = \prod_{i=1}^{m} (y_{c,i}^n)^{w_{c,i}}$$

which, in fact can be 'linearized' (by taking logarithms) such that one obtains a model that has a formally similar structure as the basic benefit of the doubt model (4)–(5a)–(5b). However, (i) one needs additional (absolute weight) constraints to preserve unit invariance for this model (see e.g. Cooper et al., 2000, pp. 110–111) which may not always lead to feasible solutions, and (ii) the proper interrelationship of (a logged form of) the geometric aggregation form (13) and expert information on the weights has, to the best of our knowledge, not yet been analyzed.

Still, given the current stance of research in this area, the benefit-of-the-doubt approach has some virtues over other, current mainstream approaches to composite indicator construction. As we pointed out, its unit invariance allows us to transcend discussions on the undesirable impact of normalization on eventual country rankings. Its flexible approach to the weighting issue may downplay critical remarks on 'imputed' weighting

systems. Thirdly, and importantly for practitioners, its fundamental interpretation and the concomitant country results are easy to convey (e.g. by using pie charts), a remark which also holds for the kind of information one seeks to distill from the expert community.

Hence, one of its advantages may be that attention can be devoted to stages that are presumably more fundamental in the construction process, to wit, the selection of relevant variables, the search for good data[14], and the quest for (broad) agreement among stakeholders about (bounds on) the relative importance of a composite indicator's constituent components.

APPENDIX 1: THE TECHNOLOGY ACHIEVEMENT INDEX AND EXPERT OPINION ON WEIGHTS

The United Nations' TAI index is developed to capture country performances in creating, adapting and using global technological innovations. Desai et al. (2002) define it as a composite indicator of technological progress that ranks countries on a comparative global scale. The TAI focuses on achievements in four dimensions: creating new technology, diffusing recent innovations, diffusing existing technologies which are still basic inputs to the industrial and the network age and building a human skill base for technology creation and adaptation. Eight sub-indicators capture these dimensions (with two sub-indicators for each dimension): the number of patents granted per 1,000,000 people, the receipt of royalties in US $ per 1000 inhabitants; the number of internet hosts per 1000 people, the exports of high and medium technology products as a share of total goods exports; the number of telephone lines per 1000 people (expressed in logarithms), electricity consumption per capita in kWh (also in logs); the mean years of schooling, and the gross enrolment ratio of tertiary students in science, mathematics and engineering. The eight selected sub-indicators all are 'goods' so that higher values reflect better performance. For extensive explanations on the sub-indicators we refer to Desai et al. (2002). The list immediately shows the different units of measurement across sub-indicators, a recurring issue in the construction of composite indicators. In the calculations of the actual TAI, data are first normalized to overcome this problem. We deviate from this common practice in the main text by aggregating the *original* data. In the original TAI the UN uses equal weights to aggregate the sub-indicators.

Nardo et al. collected opinion from 21 experts about TAI weighting schemes. The weights defined in Table A1 were obtained by using the so-called Budget allocation method.[15] This is a participatory method in which experts have to distribute a budget of 100 points over the sub-indicators allocating more to what they regard to be the more important sub-indicators. It is this information we use to illustrate some possible pie share bounds.

TABLE A1

Budget allocation weights for the Technology Achievement Index

	Patents	Royalties	Internet	Exports	Telephones	Electricity	Schooling	Enrolment
Expert1	0.05	0.05	0.05	0.2	0.1	0.05	0.2	0.3
Expert2	0.15	0.15	0.1	0.15	0.1	0.05	0.2	0.1
Expert3	0.2	0	0.1	0.2	0	0	0.2	0.3
Expert4	0.07	0.07	0.1	0.15	0.13	0.06	0.23	0.19
Expert5	0.1	0.15	0.15	0.3	0.05	0.05	0.2	0
Expert6	0.1	0.05	0.2	0.1	0.15	0.05	0.15	0.2
Expert7	0.2	0.2	0.05	0.1	0.1	0.15	0.15	0.05
Expert8	0.1	0.05	0.15	0.2	0.1	0.1	0.15	0.15
Expert9	0.1	0.1	0.15	0.15	0.1	0.05	0.2	0.15
Expert10	0.2	0.05	0.05	0.3	0.05	0	0.05	0.3
Expert11	0.12	0.15	0.12	0.15	0.1	0.12	0.08	0.16
Expert12	0.05	0.05	0.2	0.1	0.05	0.05	0.25	0.25
Expert13	0.1	0.05	0.1	0.25	0.05	0.15	0.1	0.2
Expert14	0.05	0.05	0.05	0.1	0.2	0.05	0.2	0.3
Expert15	0.1	0.3	0.02	0.33	0.03	0.02	0.05	0.15
Expert16	0.1	0.15	0.1	0.2	0.15	0.1	0.1	0.1
Expert17	0.15	0.13	0.11	0.14	0.1	0.1	0.15	0.12
Expert18	0.1	0.15	0.1	0.2	0.1	0	0.15	0.2
Expert19	0.12	0.22	0.09	0.09	0.09	0.09	0.1	0.2
Expert20	0.05	0.02	0.2	0.15	0.15	0.03	0.15	0.25
Expert21	0.1	0.1	0.1	0.25	0.15	0.05	0.05	0.2
Average	0.110	0.107	0.109	0.181	0.098	0.063	0.148	0.184
Max	0.2	0.3	0.2	0.33	0.2	0.15	0.25	0.3
Min	0.05	0	0.02	0.09	0	0	0.05	0

APPENDIX 2: SUMMARY TABLES FOR (BENEFIT OF THE DOUBT) COMPOSITE INDICATOR VALUES

The following table(s) A2 recapture the original TAI-values for 23 countries, and adds to this the index values as provided by the 'full flexibility' benefit of the doubt model (4)–(5a)–(5b); a model with ordinal sub-indicator share restrictions (8) added; one with relative sub-indicator share restrictions (9).

The second part of the table adds the proportional pie share restrictions (11), and its counterpart (12), putting bounds on the four categories. One interesting idea, that we have not pursued in this paper, is checking the robustness of the country rankings (or scores) to this different scenario's by uncertainty/sensitivity analysis (see e.g. Nardo et al. (2005), or Saisana et al. (2005)).

TABLE A2

CI values and country rankings following different scenarios (exact scenarios explained in the main text)

Country	Original TAI	Full flexibility	Ordinal VWRs	Relative VWRs
Finland	74.40% (1)	100.00% (1)	100.00% (3)	87.22% (3)
United States	73.30% (2)	100.00% (1)	92.31% (6)	92.58% (2)
Sweden	70.30% (3)	100.00% (1)	92.96% (4)	92.95% (1)
Japan	69.80% (4)	100.00% (1)	82.06% (7)	63.71% (5)
Korea, Rep. Of	66.60% (5)	100.00% (1)	100.00% (1)	17.79% (20)
Netherlands	63.00% (6)	99.45% (10)	69.08% (16)	69.89% (4)
United Kingdom	60.60% (7)	97.65% (12)	79.29% (9)	52.97% (8)
Canada	58.90% (8)	98.22% (11)	76.25% (10)	27.26% (16)
Australia	58.70% (9)	100.00% (1)	92.34% (5)	36.27% (15)
Singapore	58.50% (10)	100.00% (1)	100.00% (2)	8.65% (24)
Germany	58.30% (11)	92.07% (13)	80.95% (8)	56.26% (6)
Norway	57.90% (12)	100.00% (1)	48.43% (25)	43.99% (12)
Ireland	56.60% (13)	83.10% (16)	70.10% (13)	56.14% (7)
Belgium	55.30% (14)	80.23% (22)	70.28% (12)	46.28% (11)
New Zealand	54.80% (15)	100.00% (1)	48.96% (24)	36.76% (14)
Austria	54.40% (16)	81.95% (19)	69.50% (14)	46.71% (10)
France	53.50% (17)	84.92% (15)	69.22% (15)	50.98% (9)
Israel	51.40% (18)	81.29% (21)	63.61% (18)	43.79% (13)
Spain	48.10% (19)	75.62% (26)	72.06% (11)	25.96% (17)
Italy	47.10% (20)	82.21% (17)	65.32% (17)	12.13% (22)
Czech Republic	46.50% (21)	79.17% (23)	55.82% (21)	17.28% (21)
Hungary	46.40% (22)	85.59% (14)	54.99% (23)	18.17% (19)
Slovenia	45.80% (23)	68.38% (28)	59.32% (20)	18.83% (18)
Slovakia	44.70% (24)	77.50% (25)	59.99% (19)	11.81% (23)
Portugal	41.90% (25)	71.67% (27)	55.73% (22)	5.60% (25)

TABLE A2
Continued

Poland	40.70% (26)	81.67% (20)	45.64% (27)	3.92% (26)
Mexico	38.90% (27)	82.05% (18)	39.30% (29)	0.97% (30)
Argentina	38.10% (28)	65.25% (29)	46.60% (26)	3.02% (27)
Romania	37.10% (29)	79.17% (23)	42.53% (28)	1.37% (29)
Uruguay	34.30% (30)	63.33% (31)	30.36% (31)	0.00% (34)
Thailand	33.70% (31)	63.74% (30)	35.16% (30)	0.90% (31)
Brazil	31.10% (32)	45.14% (33)	25.77% (34)	1.89% (28)
China	29.90% (33)	56.04% (32)	26.26% (32)	0.33% (33)
Colombia	27.40% (34)	44.17% (34)	26.18% (33)	0.83% (32)

Country	Original TAI	Proportional VWRs	Category RVWRs	Category PVWRs
Finland	0.7440 (1)	100.00% (1)	100.00% (1)	100.00% (1)
United States	0.7330 (2)	100.00% (1)	100.00% (1)	100.00% (1)
Sweden	0.7030 (3)	100.00% (1)	98.02% (4)	100.00% (1)
Japan	0.6980 (4)	100.00% (1)	99.75% (3)	100.00% (1)
Korea, Rep Of	0.6660 (5)	62.47% (12)	57.88% (12)	100.00% (1)
Netherlands	0.6300 (6)	90.15% (5)	88.92% (5)	90.33% (7)
United Kingdom	0.6060 (7)	74.95% (7)	69.61% (6)	92.66% (6)
Canada	0.5890 (8)	43.49% (19)	61.68% (10)	66.17% (13)
Australia	0.5870 (9)	61.84% (13)	52.11% (15)	55.41% (17)
Singapore	0.5850 (10)	14.35% (26)	46.29% (18)	59.96% (15)
Germany	0.5830 (11)	81.84% (6)	64.36% (8)	74.44% (9)
Norway	0.5790 (12)	73.24% (10)	64.68% (7)	69.30% (11)
Ireland	0.5660 (13)	73.45% (9)	62.36% (9)	81.50% (8)
Belgium	0.5530 (14)	61.60% (14)	58.83% (11)	71.59% (10)
New Zealand	0.5480 (15)	61.42% (15)	50.80% (16)	54.73% (18)
Austria	0.5440 (16)	72.94% (11)	55.83% (14)	59.72% (16)
France	0.5350 (17)	73.62% (8)	57.85% (13)	67.05% (12)
Israel	0.5140 (18)	56.46% (16)	48.04% (17)	60.95% (14)
Spain	0.4810 (19)	43.56% (18)	25.44% (20)	31.73% (20)
Italy	0.4710 (20)	20.39% (25)	22.10% (21)	27.86% (21)
Czech Republic	0.4650 (21)	33.10% (20)	17.73% (24)	19.58% (24)
Hungary	0.4640 (22)	32.00% (21)	19.35% (23)	23.10% (23)
Slovenia	0.4580 (23)	55.34% (17)	36.38% (19)	38.70% (19)
Slovakia	0.4470 (24)	29.59% (23)	13.15% (26)	15.20% (25)
Portugal	0.4190 (25)	10.44% (28)	8.63% (27)	9.91% (27)
Poland	0.4070 (26)	30.03% (22)	13.63% (25)	14.72% (26)
Mexico	0.3890 (27)	1.95% (31)	1.51% (30)	1.65% (30)
Argentina	0.3810 (28)	12.19% (27)	4.23% (28)	4.52% (28)
Romania	0.3710 (29)	26.42% (24)	19.98% (22)	23.88% (22)
Uruguay	0.3430 (30)	3.71% (29)	1.14% (31)	1.21% (32)
Thailand	0.3370 (31)	1.92% (32)	1.10% (32)	1.34% (31)
Brazil	0.3110 (32)	3.65% (30)	2.78% (29)	3.17% (29)
China	0.2990 (33)	1.62% (34)	0.50% (34)	0.72% (34)
Colombia	0.2740 (34)	1.88% (33)	0.90% (33)	1.02% (33)

NOTES

[1] Equal weighting has been justified by referring to Occam's razor: "Since it is probably impossible to obtain agreement on weights, the simplest arrangement [equal weighting] is the best choice". Hopkins (1991, p. 1471). However, Occam's razor (the principle of parsimony) refers to choosing the simplest among 'otherwise equivalent models'. Given that outcomes (ranks) *depend* on the weighting scheme, this prerequisite clearly doesn't hold. As for composite indicators, our own opinion regarding Babbie's statement is, hence, the other way around: the burden of the proof should be on equal weighting whereas the norm should be differential (benefit of the doubt) weighting.

[2] In fact, such subjective judgements can also enter at the normalization stage. In the actual TAI, for example, the telephone and electricity figures are not normalized in exactly the same way as the other six indicators, since in the former case one normalizes logged values rather than original values. (The option to take logs is often taken to reflect the idea of diminishing marginal importance of an indicator). Some authors refer to this phenomenon as the 'implicit' weighting of sub-indicators. To the extent that it is preferable to render composite indicators transparent, such 'implicit weighting' (which clearly contains value judgements about the relative worth of each component) should best be avoided, e.g., by skipping the normalization stage when possible, or by only resorting to 'explicit' weighting.

[3] See e.g. Brandolini (2002): "For the sake of simplicity – but the observation carries over to more complicated formulations – suppose that the summary index equals the arithmetic mean of the selected indicators. In adopting such an index, we are implicitly assuming that one unit more of indicator A can be substituted for one unit less of indicator B or vice versa. If A is the unemployment rate and B the proportion of people failing to reach 65, our summary index would suggest that the valuation of the social situation is unchanged when the unemployment rate is reduced by 1% point at the same time as the proportion of people dying before 65 is raised by 1% point. I do not think that this conclusion is acceptable, nor is it likely to gain wide acceptance."

[4] As pointed out by a referee, it may be desirable at this point to make the explicit distinction between policy priorities and societal priorities, as a one-to-one relationship between the two may not be taken for granted. Upholding this distinction, and recognizing that composite indicators are as a rule used for relative performance assessment (and hence, implicitly, as a device for monitoring policies), the benefit-of-the-doubt idea hence boils down to looking for plausible (national) *policy* priority weights. Precisely because information about 'true' policy weights is lacking, one 'lets the data speak for themselves' to assign the weights. By contrast, it seems far more difficult to defend that relative policy performance evaluation enables to discern *societal* preferences. This does however not mean that the latter type of preferences – if such information were available – would be of no value when constructing a benefit-of-the-doubt indicator. In Section 3 we discuss several methods through which 'value judgments' can be appended to a benefit-of-the-doubt evaluation exercise.

[5] For completeness, we should stress that the benefit-of-the-doubt approach effectively allows one to impose a *common* (endogenously selected) weight scheme for assessing the performance of each evaluated country. More generally, it is possible to reduce (or even eliminate) dispersion of weight values *over the countries*. This type of weight restrictions could seem appropriate, e.g., if one is dealing with countries that are (a priori) taken to be 'similar' to some extent (equal circumstances, same policy issues, etc). For an application of this idea, see Cherchye and Kuosmanen (2006). It must be stressed that in such cases one needs to introduce a 'meta-objective', serving to link the various optimization programs that otherwise could be run in isolation for each country (see particularly expressions (4)–(5) below). For example, the (common) weights could be chosen such that the sum of the eventual CIs is maximal, or the minimal value over all countries is maximal, etc. (see Kao and Hung, 2005). In this respect, a

difficulty from a practical point of view is that it is often not clear which of such meta-objectives is to be preferred.

[6] Note that this does not necessarily imply that a country will assign all weight to the one sub-indicator dimension in which it performs relatively best (compared to its relative performance on the other sub-indicators), and assign zero weights to all other dimensions. By way of example, we append a third country to the dataset of Table I, with values 140 and 65 for patents and royalties respectively. Note that this new country will be outperformed by both the US and Finland. Upon applying the benefit-of-the-doubt model, the best weights for this new country are 0.002653 (patents) and 0.001795 (royalties), which demonstrates our point. Its score then amounts to $(0.002653 \times 140) + (0.001795 \times 65) = 0.488$. Indeed, if we arbitrarily fix maximal weight, say at 100%, and if the new country would have assigned this maximal weight to patents, then expression (4) yields a *lower* performance score of $140/289 = 0.484$. Similarly, putting all weight on royalties and applying (4) would yield the even lower score of $65/156.6 = 0.415$. The same finding holds for our TAI-example: for the model without any further weight restrictions but the non-negativity requirement (5b), we find e.g. that Slovenia assigns non-zero weights to no less than five dimensions (Exports, Telephones, Electricity, Schooling and Enrollment).

[7] The benefit of the doubt weights can be connected to a game-theoretic set-up: they can be conceived of as Nash equilibria in an evaluation game between a regulator and an organization. See e.g. Semple (1996).

[8] Equations (4) and (5a) clearly reveal that, when evaluating country c, the optimal weights for c are applied to all countries $j = \{1, \ldots, c, \ldots, n\}$. To illustrate, we recapture the example discussed in footnote 4. For that example, we have that the third country's optimal weights, multiplied by the values of the other two countries yields a performance score of 100% (e.g., for Finland we have $(0.002653 \times 271) + (0.001795 \times 156.6) = 1$)

[9] The model of Zaim et al. (2001) has some noteworthy additional features. First, it focuses on a sub-vector of performance indicators only, keeping other outputs and inputs fixed when assessing potential improvement. Second, their 'achievement index' is in fact the *ratio* of two DEA-scores; one for the evaluated country and (in the denominator) one for a fixed reference country. This slightly different normalization option entails a slightly different presentation of a country's score (viz., relative to this fixed reference country), but the underlying model is of course the same as ours. Third, on a more technical level, they use the dual ('envelopment') formulation of our primal ('multiplier') model (4)–(5a)–(5b). Both approaches are fully equivalent from a mathematical point of view. However, it may be argued that the multiplier/benefit-of-the-doubt formulation is easier to convey to CI-practitioners (e.g., given their frequent use of the form (1)).

[10] In other cases (e.g. switching from Fahrenheit to Celsius, which are interval scales, or with ordinal sub-indicators), other models in the DEA-class can still be useful. See Halme et al. (2002) for the first case. The case of ordinal data is briefly taken up in Section 4.

[11] In the DEA literature, this concept is usually labelled a 'virtual output' ('virtual input'). See especially Thanassoulis et al. (2004) for a discussion of virtual outputs (or pure weights, or exogenous benchmarks) as means to include value judgments in DEA.

[12] The TAI-case itself provides an example: only the absolute bounds $\alpha_i = 0$ and $\beta_i = 1$ (and thus full flexibility) allow the computation of composite indicators. All tighter absolute restrictions lead to infeasibilities.

[13] For example, in footnote 5 we suggest the possibility of reducing the dispersion of weight values *across* countries.

[14] Missing data are a recurring source of frustration when building CI's. The *Handbook* of Nardo et al. (2005) lists some statistical methods for data imputation, which evidently are valuable for a benefit of the doubt model as well. However, we note that for such models some specific treatments of missing data have been proposed. The standard DEA approach will

automatically assign a zero weight to missing data if the latter are set at zero. Consequently, sub-indicators with missing data are excluded from the performance evaluation of that specific country. Cherchye and Kuosmanen (2006) propose a way to account for missing data when defining weight restrictions, to avoid that missing data arbitrarily influence the final results.

[15] For other TAI weights, obtained by using Saaty's Analytical Hierarchy Process, see Nardo et al. (2005).

REFERENCES

Babbie, E.: 1995, The Practice of Social Research (Wadsworth Publishing Company, Belmont).

Booysen, F.: 2002, 'An Overview and Evaluation of Composite Indices of Development'. Social Indicators Research 59, pp. 115–151.

Brandolini, A.: 2002, 'Education and Employment Indicators for the EU Social Agenda'. Politica Economica 18, pp. 55–62.

Charnes, A., W. W. Cooper and E. Rhodes: 1978, 'Measuring the Efficiency of Decision Making Units'. European Journal of Operational Research 2, pp. 429–444.

Cherchye L., C. A. K. Lovell, W. Moesen and T. Van Puyenbroeck: 2005, 'One Market, One Number? A Composite Indicator Assessment of EU Internal Market Dynamics', forthcoming in *European Economic Review*.

Cherchye, L., W. Moesen and T. Van Puyenbroeck: 2004, 'Legitimately Diverse, Yet Comparable: On Synthesizing Social Inclusion Performance in the EU'. Journal of Common Market Studies 42, pp. 919–955.

Cherchye, L. and T. Kuosmanen: 2006, 'Benchmarking Sustainable Development: A Synthetic Meta-Index Approach' Chapter 7. in M. McGillivray and M. Clarke (eds), Perspectives on Human Development (United Nations University Press), to appear.

Cook, W. D.: 2004, 'Qualitative Data in DEA', in W. W. Cooper, L. Seiford and J. Zhu (eds), Handbook on Data Envelopment Analysis (Kluwer Academic Publishers, Dordrecht), pp. 75–97.

Cook, W. D. and M. Kress: 1991, 'A Multiple Criteria Decision Model with Ordinal Preference Data'. European Journal of Operations Research 54, pp. 191–193.

Cook, W. D. and M. Kress: 1994, 'A Multiple Criteria Composite Index Model for Quantitative and Qualitative Data'. European Journal of Operations Research 78, pp. 367–379.

Cooper, W. W., L. M. Seiford and K. Tone: 2000, Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software (Kluwer Academic Publishers, Dordrecht).

Cooper, W. W., L. M. Seiford and J. Zhu: 2004, Handbook on Data Envelopment Analysis (Kluwer Academic Publishers, Dordrecht).

Desai, M., S. Fukuda-Parr, C. Johansson and F. Sagasti: 2002, 'Measuring the Technology Achievement of Nations and the Capacity to Participate in the Networking Age'. Journal of Human Development 3(1), pp. 95–122.

Despotis, D. K.: 2005, 'A Reassessment of the Human Development Index via Data Envelopment Analysis', Journal of the Operational Research Society 56(8), pp. 969–980.

Ebert, U. and H. Welsch: 2004, 'Meaningful Environmental Indices: A Social Choice Approach'. Journal of Environmental Economics and Management 47, pp. 270–283.

European Commission: 2004, The EU Economy Review 2004, European Economy, Nr. 6 (Office for Official Publications of the EC, Luxembourg).

Foster, J. and A. Sen: 1997, On Economic Inequality, 2nd expanded edn (Clarendon Press, Oxford).

Freudenberg, M.: 2003, 'Composite Indicators of Country Performance: A Critical Assessment', OECD Science, Technology and Industry Working Papers 2003/16.

Golany, B.: 1988, 'A Note on Including Ordinal Relations among Multipliers in Data Envelopment Analysis', Management Science 34(8), pp. 1029–1033.
Halme, M., T. Joro and M. Koivu: 2002, 'Dealing with Interval Scale Data in Data Envelopment Analysis'. European Journal of Operational Research 137, pp. 22–27.
Hopkins, M.: 1991, 'Human Development Revisited: A New UNDP Report'. World Development 19, pp. 1469–1473.
Kao, C. and H. T. Hung: 2005, 'Data Envelopment Analysis with Common Weights: The Compromise Solution Approach'. Journal of the Operational Research Society 56, pp. 1196–1203.
Kumar, S. and R. R. Russel: 2002, 'Technical Change, Technological Catch-Up, and Capital Deepening: Relative Contributions to Growth and Convergence'. American Economic Review 92, pp. 527–548.
Lovell, C. A. K, J. T. Pastor and J. A. Turner: 1995, 'Measuring Macroeconomic Performance in the OECD: A Comparison of European and Non-European Countries'. European Journal of Operational Research 87, pp. 507–518.
Mahlberg, B. and M. Obersteiner: 2001, 'Remeasuring the HDI by Data Envelopment Analysis', International Institute for Applied Systems Analysis Interim Report 01–069.
Melyn, W. and W. Moesen: 1991, 'Towards a Synthetic Indicator of Macroeconomic Performance: Unequal Weighting when Limited Information is Available', *Public Economics Research Paper 17, CES*, KU Leuven.
Micklewright, J.: 2001, 'Should the UK Government Measure Poverty and Social Exclusion with a Composite Index?', in CASE, Indicators of progress: A Discussion of Approaches to Monitor the Government's Strategy to Tackle Poverty and Social Exclusion, CASE Report 13, London School of Economics.
Munda, G. and M. Nardo: 2003, 'On the Methodological Foundations of Composite Indicators Used for Ranking Countries', *mimeo*, Universitat Autonoma de Barcelona.
Nardo, M., M. Saisana, A. Saltelli and S. Tarantola (EC/JRC) and A. Hoffman and E. Giovannini (OECD): 2005, Handbook on Constructing Composite Indicators: Methodology and User Guide, OECD Statistics Working Paper.
Pedraja-Chaparro, F., J. Salinas-Jimenez and P. Smith: 1997, 'On the Role of Weight Restrictions in Data Envelopment Analysis'. Journal of Productivity Analysis 8, pp. 215–230.
Saisana, M., A. Saltelli and S. Tarantola: 2005, 'Uncertainty and Sensitivity Analysis as Tools for the Quality Assessment of Composite Indicators'. Journal of the Royal Statistical Society Series A 168, pp. 1–17.
Semple, J.: 1996, 'Constrained Games for Evaluating Organizational Performance'. European Journal of Operational Research 96, pp. 103–112.
Storrie, D. and H. Bjurek: 2000, 'Benchmarking European Labour Market Performance with Efficiency Frontier Techniques', CELMS Discussion Paper, Göteborg University.
Thanassoulis, E., M. C. Portela and R. Allen: 2004, 'Incorporating Value Judgements in DEA', in W. W. Cooper, L. Seiford and J. Zhu (eds), Handbook on Data Envelopment Analysis (Kluwer Academic Publishers, Dordrecht), pp. 99–138.
Wong, Y. H. B. and J. E. Beasly: 1990, 'Restricting Weight Flexibility in Data Envelopment Analysis'. Journal of the Operational Research Society 41, pp. 829–835.
Zaim, O., R. Färe and S. Grosskopf: 2001, 'An Economic Approach to Achievement and Improvement Indexes'. Social Indicators Research 56, pp. 91–118.
Zhu, J.: 2003, Quantitative Models for Performance Evaluation and Benchmarking, International Series in Operations Research and Management Science (Kluwer Academic Publishers, Dordrecht).

Centre for Economic Studies L. Cherchye
Catholic University of Leuven W. Moesen
Naamsestraat 69 N. Rogge
3000, Leuven, Belgium Tom V. Puyenbroeck

European University College N. Rogge
Stormstraat 2 Tom V. Puyenbroeck
1000, Brussels, Belgium
E-mail: Nicky.Rogge@econ.kuleuven.be