KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Economics and
Applied Economics

Department of Economics

One Market, One Number?: a Composite Indicator Assessment of EU Internal Market Dynamics

by

Laurens CHERCHYE
C.A.Knox LOVELL
Wim MOESEN
Tom VANPUYENBROECK

Public Economics

Center for Economic Studies
Discussions Paper Series (DPS) 05.13
http://www.econ.kuleuven.be/ces/discussionpapers/default.htm

June 2005

DISCUSSION
PAPER

ONE MARKET, ONE NUMBER? A COMPOSITE INDICATOR ASSESSMENT OF EU INTERNAL MARKET DYNAMICS

by

Laurens CHERCHYE
C.A. Knox LOVELL
Wim MOESEN
Tom VAN PUYENBROECK

Public Economics

Center for Economic Studies
Discussion Paper Series DPS 05.13
http://www.econ.kuleuven.ac.be/ew/admin/default.htm

June 2005

D/2005/2020/15

One Market, One Number?

A Composite Indicator Assessment of EU Internal Market Dynamics[*]

Laurens Cherchye [a, b]

C.A. Knox Lovell [c]

Wim Moesen [a]

Tom Van Puyenbroeck [a, d]

JUNE 2005

ABSTRACT

We consider the lack of consensus about an appropriate theoretical framework linking sub-indicators as a defining characteristic of composite indicators. This intrinsic feature implies uncertainties about the appropriate normalisation and aggregation of the raw data. The two are related: index theory offers some valuable guidelines about their connection. Yet these do not fully solve the basic problem of expert disagreement. We embed such (residual) disagreement in the aggregation method itself. Specifically, we apply an impartial benefit-of-the-doubt weighting procedure, where weight restrictions incorporate the available information on experts' opinions. We apply this procedure to the dynamic performance assessment of EU Internal Market effects, thereby highlighting its capacity to disaggregate member states' observed performance shifts into changes *relative to* benchmarks and performance changes *of* the benchmarks (i.e. catching up versus genuine progress). Our results indicate that the latter factor is more important in explaining the observed progress.

Keywords: composite indicators, aggregation, weighting, Internal Market

JEL Classification Nrs: E13, E60, F02, O47

[*] This paper is an offshoot of the KEI-project (contract n° 502529) that is part of priority 8 of the policy orientated research under the European Commission's Sixth Framework Programme (see http://kei.publicstatistics.net/). We thank Stefano Tarantola and his colleagues at the Joint Research Centre, and the participants of the 2004 Workshop on Composite Indicators at the OECD for valuable comments on previous versions of this paper. Laurens Cherchye thanks the Fund for Scientific Research-Flanders for his postdoctoral fellowship.

[a] Centre for Economic Studies, KU Leuven
[b] Katholieke Universiteit Leuven Campus Kortrijk
[c] Terry College of Business, University of Georgia
[d] European University College Brussels (EHSAL).
Corresponding author: Tom.vanpuyenbroeck@ehsal.be

One Market, One Number?

A Composite Indicator Assessment of EU Internal Market Dynamics

> *"Composite indicators risk becoming exercises in measurement without a theoretical underpinning."*
>
> (Freudenberg, 2003, p. 29)

1. Introduction

Policy makers and their watchdogs rely on numerous data to know where they are and how they got there. The past century has seen massive efforts in statistical capacity building, which indisputably enlarged the informational basis of policy decisions. However, there are some indications that this basis has reached its limits. Organisations such as the UN, the OECD, or the European Commission, which traditionally act upon the presumption that better knowledge is a necessary condition for better public decision-making, nowadays appear to have mixed feelings about the overabundance of statistics. At least, one may infer this reluctance from the fact that such organisations, which have a long history as providers of many excellent and detailed statistics, have recently either singled out some "key indicators" or have constructed "composite indicators" in which several single indicators are aggregated into one index. This paper looks at the latter type of composite indicators, which comprise the UN's Human Development Indices, the OECD's Composite Leading Indicators, the WHO's Health System Performance Index, the World Economic Forum's Competitiveness index, etc. An extensive list of such indicators and methodological discussions can be found on the European Commission and OECD's composite indicator information server (http://farmweb.jrc.cec.eu.int/ci/).

Strictly speaking, composite indicators are far from new measures. Well-known indicators such as GDP, the CPI, the Gini coefficient, and so on, also merge information about different markets or agents into a single number. Each of these is firmly entrenched as a policy instrument, despite the fact that they continue to be criticized as inadequate measures of the underlying phenomena they purport to quantify. To mention but a few familiar criticisms: GDP is not an adequate indicator of a country's economic activity (let alone of its citizens' well-being) if only because it neglects the underground economy per definition; the CPI is at best just a rudimentary estimate of changes in the true cost-of-living; the Gini coefficient is rooted in a rather distinct welfarist framework when comparing individual incomes, etc.

Nevertheless, it is safe to say that such traditional aggregates are presently far less controversial than their more recent cognates. Many of their creators even present the new brand of composite indicators in an apologetic style. Likewise, the European Commision's generic definition – "*composite indicators are based on sub-indicators that have no common meaningful measurement unit and there is no obvious way of weighting these sub-indicators*" (as e.g. found on the aforementioned website) – alludes to their contentious nature. And in fact, once introduced, they often stir critical analyses of various elements that underlie their construction.[1]

To some extent, it is ironic that the new composite indicators are presented with a list of pros and cons attached (see e.g. Saisana *et al.,* 2005) that is in large measure appropriate for their older counterparts as well. Still, while aggregation is a concern in many economic contexts, the precise nature of the aggregation problem for composite indicators is different from the one relating to 'traditional' aggregates. Indeed, one can say that for aggregates such as GDP the predominant issue is not *how* to aggregate – aggregation is taken to be linear–, but rather *what* variables to include; only market goods and services are included, and non-market goods and services such as the underground economy, work in the home, environmental impacts, etc., are excluded (on the environmental discussion, see e.g. Nordhaus and Kokkelenberg, 1999). Quite distinct from that, the economic theory of production and consumption, based on optimising behaviour, duality theory and the theory of separability, imposes strict conditions for commodity aggregation on functions such as production, cost, profit, utility (examples include food in a utility function, labour in a production function, the wage rate in a cost function, etc.). Except under extreme conditions on technology or preferences, these aggregator functions are not linear. The issue in that particular area is, hence, one of functional form. Finally, for the composite indicators problem the predominant focus is also on *how* to aggregate, but economic theory based on optimisation, duality and separability provides no guidance. Specifically, whereas an essential feature of the other examples is that *exogenous* market prices act as natural weights, composite indicators are distinguished by the absence of market prices and the need to search for an alternative weighting system. For example, one has to rely on expert judgement, but then the further issue is raised that experts can and do disagree about such weights.

[1] The Human Development Index is a well-known case in point. See e.g., among many articles, the recent ones of Chakravarty (2003), Lind (2004), or Chatterjee (2005).

So, there is undeniably a particular epistemological sense in which the new composite indicators differ from their nowadays less contested precursors: at heart, the newcomers lack a sufficient degree of scientific consensus about an appropriate theoretical model that should, in principle, provide a precise insight into how the sub-indicators contribute to the underlying composite phenomenon.[2] Indeed, one often observes that an agreement emerges about the choice of key sub-indicators (though we concede that it may take time to reach such an agreement). But while there is a broad consent that all these single indicators can be 'associated with' the comprehensive phenomenon at hand, the hard question remains how and by how much.

In this paper, we therefore consider this lack of consensus as a *defining property* of (the new) composite indicators. As one important purpose of such indicators is to serve as a basis of comparison, with other geographical entities or over time, the natural question that follows is to what extent one can coherently employ them for such purpose. To clarify our general position on this issue further, we note that the uncertainty about the appropriate aggregation of sub-indicators carries over to the underlying phenomenon itself, as captured by a composite measure. And as far as the latter is concerned, we fully concur with the recommendation that "if a concept has some basic ambiguity (…), then a *precise* representation of that ambiguous concept must *preserve* that ambiguity, rather than try to remove it through some arbitrary complete ordering." (Foster and Sen, 1997, p. 121, italics in original). Urging for a representation which is precise and still preserves ambiguity is only seemingly contradictory: in some cases, including –in our opinion- the one in the current paper, one may use a well-defined aggregation *method* that incorporates the doubts intertwined with the aggregation *problem*.

For sure, there is no recipe for building composite indicators that is at the same time universally applicable and sufficiently detailed. The nature and quality of the underlying raw data, the availability and the heterogeneity of expert opinion, the specific purposes for which such an index is intended, etc., all (should) feed back into the construction of a composite indicator. In this sense, we think that the method discussed in this paper is particularly suited for cases (a) in which the underlying purpose is to get some idea of the composite 'performance dynamics' of, say, a country relative to other countries, and, (b) where expert

[2] Thus, whereas the traditional indices are not totally free of criticism either, the scientific assessment of their merits and demerits as composite indicators notably goes back to an analysis of their underlying *well-defined* theoretical framework. Similarly, one could say that the public acceptance of such indices is fostered by the broad (scientific) acceptance of the theories from which they are derived.

opinion about the proper weights for merging the sub-indicators is available but disparate. As we explain in section 2, the European Commission's efforts to monitor the development of its Internal Market provide a good illustration of such a context.

Notwithstanding their context-dependency, composite indicators ideally meet some minimal conditions of index theory. The primary concern here is the aforementioned issue that the original data may have no common meaningful measurement unit. We discuss this issue in section 3. However, as we also explain in that section, index theory alone does not usually suffice to provide a complete characterization of the composite indicator. One still has to face the weighting issue. We explain why equal weighting –a normal practice in composite indicator construction– is fundamentally flawed, and show how similar points can be raised against 'exact' weighting schemes that are uniformly applied to all observations.

In section 4 we propose to aggregate the data with an *endogenous weighting* procedure. In brief, this procedure generates weights that comply with the limited expert consensus, but deal with the remaining uncertainty using an impartial 'benefit-of-the-doubt' approach (after Melyn and Moesen, 1991). Adhering to such an approach entails that discussions about 'arbitrary orderings' or 'imposed value judgements', e.g. by the concerned country's policy makers, are likely to be minimized.

The proposed method has been used for composite indicators previously (see e.g. Cherchye *et al*., 2004, and the references cited therein), but the focus in these applications was on cross-section benchmarking of countries. Here we want to highlight its usefulness for panel data. Specifically, in section 5 we demonstrate how the method may be used to check to what extent a country's better performance over time is, relative to other countries in the sample, a result of genuine progress rather than a catching-up effect. We illustrate this type of analysis with the data the European Commission uses to track the effects of its Internal Market Policy. The latter is specifically geared towards *improving* economic performance*;* hence dynamically oriented performance evaluation seems particularly suited in this context.

Section 6 gathers the main points. In addition, it offers some concluding remarks and suggests avenues for further research.

2. The European Commission's Internal Market Index

Free movement of people, goods, services and capital within the European Union has been on the agenda ever since the Treaty of Rome, but the actual implementation of that principle is an ongoing process. A major actor in this domain is the European Commission's Internal Market Directorate-General, which is above all charged with removing (legal) barriers to free movement. In addition, it also informs citizens and businesses about the rights and benefits they are entitled to in the Single Market. One of its information vehicles is the Internal Market Scoreboard. Twice a year, this Scoreboard offers a picture of the current state of the Single Market, and gauges the degree to which member states, the Council and the Commission are meeting various Internal Market targets (e.g. the pace of transposition of European directives to national legislation, development of harmonized standards, etc.).

In 2001, the Internal Market Scoreboard introduced the 'Internal Market Index' (IMI). From its inception, the basic purpose of the IMI has been to track progress on the Internal Market strategy, by looking at a combination of several outcome-variables. After the first IMI was presented with a word of caution, a substantially revised and improved IMI was presented in the November 2002 special issue of the Internal Market Scoreboard. Its basic structure has been unaffected since then. We provide a short account of its construction here, and refer to two methodological reports of the European Commission's Joint Research Centre for an in-depth treatment (Tarantola *et al.*, 2002, Tarantola *et al.*, 2004)

To begin with, in view of our remarks in the Introduction it is noteworthy that the substantive content of the IMI has always been downplayed by its authors. They have ever been the first to stress that the reality of the Internal Market is too complex to be summarized in a single number, and even after the release of the improved version, their modest claim was that it "should be seen more as a reality check than as a precise scientific exercise" (European Commision-Internal Market Directorate General, 2002, p. 38)**.**

That statement surely must be qualified, as one can definitely not say that the IMI is constructed haphazardly. On the contrary, throughout its creation process the IMI's authors built on statistical, economic and analytical expertise, and an extensive peer review backup by stakeholders such as Eurostat and the Internal Market Advisory Committee (IMAC, i.e. the group of Member State officials which the Commission consults on Internal Market matters; see also below).

The IMI's sub-indicators all capture different aspects through which the effects of the Internal Market are taken to materialize. In general terms, these are gauges for the elimination of barriers to free movement of production factors and final goods, and for its alleged downward effects on prices in some key markets. All together, twelve sub-indicators were selected. The raw data are expressed in different measurement units such as GDP-shares, prices in euro, population percentages, etc. Specifically, these twelve sub-indicators are (i) sectoral and ad hoc state aid (% of GDP), (ii) the share of published public procurement (% of GDP), (iii) telecommunication costs (in euro), (iv) electricity prices (in euro), (v) gas prices (in euro), (vi) countries' relative price level (PPP/market exchange rate), (vii) Intra-EU foreign direct investment (% of GDP), (viii) Intra-EU trade (% of GDP), (ix) the ratio of retail lending and savings interest rates, (x) the share of a country's active population coming from other member states, (xi) postal tariffs (in euro), and (xii) the value of pension funds assets (% of GDP).[3] Table 1 provides some summary statistics for the two years we will consider in more detail further on. The table also indicates whether a rise in the concerned indicator is taken to be good (+) or bad (-) in terms of the desirable internal market effects. An important feature of Table 1 is the disparate trends in the sub-indicators. All favourable indicators increase, but to very different degrees. Some unfavourable indicators decline, while others increase. This disparity in the performance of sub-indicators highlights the importance of attaching appropriate weights to the sub-indicators when constructing a composite indicator.

Table 1: IMI's sub-indicators, summary statistics

	StA	PuP	Tel	Elec	Gas	RPL	FDI	Trade	Pop	Pens	Intrst	Post
	(-)	(+)	(-)	(-)	(-)	(-)	(+)	(+)	(+)	(+)	(-)	(-)
1994												
EU-15 average	1,67	1,28	16,61	147793	157549	100	0,83	13,80	0,0137	18,90	2,56	0,45
Member States : sample std.dev	0,81	1,23	8,59	27454	55889	17,02	1,17	10,02	0,01	23,35	2,26	0,42
2000												
EU-15 average	0,78	2,41	5,14	129164	210368	100	7,97	17,80	0,0145	28,70	3,61	0,53
Member States : sample std.dev	0,49	0,92	1,86	26021,33	36889,52	16,01	5,32	13,87	0,01	36,50	3,03	0,44

Note: StA: sectoral and ad hoc state aid, PuP: value of published public procurement, Tel: telecommunication costs, Elec: electricity prices, Gas: gas prices, RPL: countries' relative price level, FDI: Intra-EU Foreign Direct Investment, Trade: Intra-EU trade, Pop: share of a country's active population coming from other member states, Pens: value of pension funds assets, Intrst: ratio of retail lending and savings interest rates, Post: postal tariffs.

The next steps in the IMI's construction are (i) a rescaling of the raw data so that they are all expressed in a common measurement unit, and (ii) the actual construction of the composite value, which is a weighted sum of the rescaled sub-indicators. Both features will be taken up in more detail in section 3 and section 4 respectively. Yet it is instructive to note at this point that the IMI is *not* a sum of *equally weighted* (rescaled) sub-indicators. In fact,

[3] See Tarantola *et al.* (2004) for an exact definition, data sources, and an explanation of the way in which these indicators capture policy targets and effects of market integration. Note that we take the interest rate indicator (Intrst) as it was defined in the 2002 report; in 2004 one switched to taking the *difference* between lending and savings rates. The original IMI was based on 20 sub-indicators, but the number was reduced after consultation of Internal Market experts and following a quality control check according to Eurostat guidelines. Tarantola *et al.* (2002) provide an account of the statistical quality of the original data and a principal component analysis to justify the use of these twelve indicators.

these weights are derived from a survey in which individual members of the Internal Market Advisory Committee were asked to forward their weighting scheme. Table 2, taken from Tarantola *et al.* (2002), displays the weights as they were provided by IMAC-members along with some summary statistics. The actual IMI's weights are the averages shown in bold. A quick glance at Table 2 reveals that the Member States' experts differ quite substantially on how they view the most appropriate weighting scheme.[4] The table thus clearly illustrates our central theme that differing expert opinions are a key constituent of many composite indicators.

Table 2: IMAC-members' proposed weights for the sub-indicators of IMI

	StA	PuP	Tel	Elec	Gas	RPL	FDI	Trade	Pop	Pens	Intrst	Post
AT	10	0	15	15	15	20	15	10	0	0	0	0
BE	0	25	10	20	15	0	10	10	10	0	0	0
DE	20	20	0	5	0	0	20	20	5	0	10	0
DK	30	0	10	10	10	15	0	0	0	0	15	10
ES	15	10	10	15	10	0	15	25	0	0	0	0
FI	10	0	10	20	10	0	15	20	0	0	15	0
FR	0	10	15	20	10	25	0	0	0	0	10	10
GR	20	18	8	4	0	20	15	15	0	0	0	0
IE	20	10	15	15	10	0	10	20	0	0	0	0
IT	30	35	10	15	0	0	10	0	0	0	0	0
NL	10	20	0	0	2	8	20	20	20	0	0	0
PT	0	15	5	15	10	15	20	20	0	0	0	0
SE	15	15	15	15	0	15	0	15	10	0	0	0
UK	10	10	10	0	0	20	20	20	0	10	0	0
Average	**13,6**	**13,4**	**9,5**	**12,1**	**6,6**	**9,9**	**12,1**	**13,9**	**3,2**	**0,7**	**3,6**	**1,4**
Max	30	35	15	20	15	25	20	25	20	10	15	10
Min	0	0	0	0	0	0	0	0	0	0	0	0
# 0	3	3	2	2	5	6	3	3	10	13	10	12
Median	12,5	12,5	10	15	10	11,5	15	17,5	0	0	0	0
Stdev	9,5	9,6	4,8	6,8	5,7	9,3	7,2	8,3	5,9	2,6	5,8	3,5
Var-coef	0,7	0,7	0,5	0,6	0,9	0,9	0,6	0,6	1,8	3,6	1,6	2,4

Note: StA: sectoral and ad hoc state aid, PuP: value of published public procurement, Tel: telecommunication costs, Elec: electricity prices, Gas: gas prices, RPL: countries' relative price level, FDI: Intra-EU Foreign Direct Investment, Trade: Intra-EU trade, Pop: share of a country's active population coming from other member states, Pens: value of pension funds assets, Intrst: difference between retail lending and savings interest rates, Post: postal tariffs. The weights provided by each IMAC-member sum to 100. Source: Tarantola *et al.*, 2002, p. 14.

[4] The expert weights that are closest to the average, as measured by the sum of squared residuals, are those of the Irish IMAC-member. Table 2 reveals that this can hardly be considered as a strong agreement. Some expert weights are similar in terms of correlation (the maximal value being 0.93 for Spain and Ireland), but there are many cases where this weight correlation is negative as well (the minimum is -0.47, between France and the Netherlands). The maximum and minimum *rank* correlations are 0.83 and -0.50. In 70% of all pairwise comparisons, the hypothesis that there is no significant relationship between sub-indicator rankings cannot be rejected at the 5%-level.

Eventually, the IMI is typically published as a line graph connecting 'EU-15 point values' for each year. The first, base-year value is set at 100, so implicitly advancing the message that these genuinely are index numbers in the way commonly understood. In fact, the companion methodological note is very specific about this interpretation, stating that "if a country has an IMI value of 120 for a given year, this implies that the country performs 20% better than its own state in [the base year]" (Tarantola *et al.,* 2002, p. 3).

Two final remarks are in order. First, similar *country-specific* IMI values can be (and have been) calculated, tracking a country's progress over time, where the country itself serves as its proper yardstick. In sections 4 and 5, we also focus on country-specific internal market performance, including inter-temporal performance shifts, but our indices will be based on panel data analysis. Second, the 'point values' have actually been checked with sensitivity and uncertainty analysis. For example, the weights in Table 2 were used in a Monte Carlo analysis to construct confidence intervals for the point estimates per year (Tarantola *et al.,* 2002, pp. 20-25). Sensitivity and uncertainty analysis will not be the central focus in this paper. This does not mean that we consider either of them as unimportant for composite indicators. Quite the reverse: it follows from our discussion that composite indicator construction is entirely permeated with uncertainties. However, it is a truism that one must have some specific base model before robustness assessments can be sensibly conducted. From this perspective, the point to be developed in the next sections is that such a specific model can already capture much of the uncertainty surrounding composite indicators. We will return to sensitivity and uncertainty issues in the concluding section.

3. Measuring Internal Market performance

In his discussion of the Human Development Index, Desai (1994, p. 34-35) defines the measurement problem in economics as one of "reducing a vector of variables to a scalar" by means of a weighted sum. It is however clear that several other aggregator functions exist to perform such a reduction. A more general description of the generic measurement problem is therefore that it addresses the interdependency of quantitatively meaningful representations of "raw data" on the one hand, and the precise method of aggregating these representations into a scalar on the other hand (see e.g. Aczél, 1988).

One important building block in such a setting is the scale (ordinal, cardinal, …) used as a numerical representation for each of the *individual* sub-indicators. Each scale is associated with a set of admissible transformations, which in turn define what kind of numerical statements are meaningful. For example, if one observes that the price for sending a standard letter is twice as high in France as in Germany, then this remains true regardless of the currency in which both prices are denominated. The other sub-indicators may be classified according to their (possibly different) measurement scales as well. The point is then to examine what kind of aggregator function can be applied to a given set of sub-indicators, and what kind of meaningful statements can be associated with the 'aggregate values' this function produces. Of course, the reasoning also holds the reverse way: particular aggregation methods presuppose particular measure-theoretic qualities of the original data in order to be meaningful. This measurement perspective is important for composite indicator construction as well (Munda and Nardo, 2003; Ebert and Welsch, 2004): one can expect a strong link between the 'normalisation' problem ("sub-indicators have no common meaningful measurement unit") and the 'aggregation problem' ("there is no obvious way of weighting these sub-indicators") with which such indices have to cope.

Turning to the IMI-case, what does one want to measure if one is interested in the aggregate performance dynamics of a multi-dimensional phenomenon such as the Internal Market? We take it that part of the answer is contained in the quote that "if a country has an IMI value of 120 for a given year, this implies that the country performs 20% better than its own state in [the base year]".   Put differently, we take it that the purpose of the IMI is to convey a reasonable statement about the "average performance growth" of the set of sub-indicators, just as e.g. the CPI's purpose is to depict the (appropriately defined) average growth of the price level between a base year and another moment in time. This immediately adds some structure to both the normalisation and the aggregation problem, as we discuss next.

A. Normalisation: the units of measurement issue

First, the previous remarks do suggest an appropriate normalisation of the raw data, viz. the ratio scale transformation

$$\frac{y_i^1}{y_i^O} \qquad \text{(1a)},$$

where $y_i^O \neq 0$ is the base value for the i-th sub-indicator and y_i^1 its value at time '1'. Such normalisation is indeed an admissible transformation of our raw data: it is meaningful to say, for example, that Portugal's Intra EU-Trade has increased by a factor 1.99 between 1994 and 2000, that the share of Denmark's non-Danish (EU citizens) active population has almost doubled between 1994 and 2001, etc. Moreover, (and with a little extra notation we provide in section 5), a normalization such as (1a) is also meaningful for comparisons between countries as well as for 'panel' comparisons ('Germany's 2000 public procurement share of GDP was in fact less than one fifth of Greece's corresponding value in 1994'). Note that higher ratios are taken to represent beneficial internal market effects in each of these examples. Yet this does not hold for all sub-indicators (e.g. the postal tariffs discussed above). In such cases however, we apply the inverse normalisation

$$\frac{y_i^0}{y_i^1} \qquad \text{(1b),}$$

to convey statements such as "sending a standard letter in Denmark in 1992 was almost four times cheaper than sending the same letter in Austria in 2000". In our IMI-application below, both (1a) and (1b) are therefore used as normalised versions of the raw data.

B. Aggregation: the weighting issue

There are also some formal desiderata that one would like to see incorporated by the aggregator, i.e. the composite indicator used to gauge 'average growth'. For instance, one would indeed like to see the IMI increase by *x* percent if, ceteris paribus, all the sub-indicators have increased by *x* percent; that a base-year value of 100 has a clear meaning, etc. Specifically, we propose to focus on gauges that are similar to output distance functions. The latters' axiomatic properties have been well-documented in the literature (e.g. Shephard, 1970; Balk, 1998), and some of these properties are indeed desirable for the composite indicator problem at hand. For example, an output distance function is weakly monotonic, homogeneous of degree +1 and convex in sub-indicators, and bounded above by unity, which corresponds to best practice.

At this point, we may paraphrase the opening quote of this paper by insisting that the act of measurement itself has a theoretical underpinning that cannot be neglected with impunity when constructing composite indicators. But it is also clear that measurement (or index number) theory alone is not sufficient to identify 'the' suitable composite indicator. For composite indicators in particular, one still has to cope with the inherent difficulty that experts' opinions about the relative importance of each sub-indicator in such an index are usually quite disparate.

In fact, some parallels can be drawn with the illustrious 'measurement without theory' discussion in index number theory. Recall that price (or quantity) index theory is developed from three different angles. First, the axiomatic approach forwards desirable properties upon which suitable index numbers are derived. Second, the stochastic approach starts from the idea that each good *i*'s individual price ratio $p_{i,t+1} / p_{i,t}$ is a random variable, i.e. an estimate of inflation which can itself be found by averaging over these ratios (e.g. Selvanathan and Rao, 1994). Finally, the so-called economic perspective, introduced by Könus (1924), traditionally disavowed the former two strands as 'measurement without theory'.[5]

To briefly explain the parallel, assume everyone would agree that expression (1a) is a fitting way to measure the performance dynamics of each separate sub-indicator. Then let us take a stochastic perspective by stating that all the numbers so derived are in fact estimates of the common internal market performance growth rate γ. To take two particular examples, one could start from the logarithmic specification $\log(y_i^1 / y_i^O) = \log\gamma + \varepsilon_i$, or from the linear specification $(y_i^1 / y_i^O) = \gamma + \varepsilon_i$, in both cases with the error terms independently and symmetrically distributed. The former assumption eventually leads to the 'Jevonian' γ-estimator (2a), whereas the latter variant leads to a 'Carli'-type index (2b)[6]:

$$\hat{\gamma} = \prod_{i=1}^{m} \sqrt[m]{\frac{y_i^1}{y_i^O}} \tag{2a}$$

$$\hat{\gamma} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_i^1}{y_i^O} . \tag{2b}$$

The key criticism that has historically been raised against indices (2a) and (2b) is that all observed sub-indicator growth rates are taken to be equally important in calculating the underlying common growth rate.[7] In particular, the economic approach to price index numbers has carried this argument further by insisting that prices and their weights (i.e. consumed or produced quantities) in a price index number are in fact connected *via* the underlying (hypothesized model of) optimizing behaviour of the economic agent(s) concerned.

A quick glance at existing composite indicators reveals that a large majority of them are of the equal weighting type. Somewhat surprisingly then, the weighting scheme disavowed for economic indexes such as (2a) and (2b) is quite common for composite indicators. Among other authors, Babbie (1995, p. 171) goes so far as to recommend it *qua*

[5] Further references can e.g. be found in the contributions of Diewert collected in Diewert and Nakamura (1993).

[6] If one replaces the sub-indicators y_i by prices, expression (2a) is the price-index as proposed by Jevons (1865) while (2b) is the Carli (1804) price index.

[7] Even from a purely stochastic perspective, this point makes sense: if, by way of example, the sub-indicators are output prices (as in a CPI), and if we were to draw sub-indicators at random to provide an estimate of the common inflation rate, clearly not all of them have an equal chance of being selected. As argued by Theil (1967), the probability in this specific case is rather given by the revenue share of the good concerned. One particular expenditure-weighted version of expression (2a) is indeed known as the Theil-Törnqvist index.

standard, stating that "items be weighted equally unless there are compelling reasons for differential weighting. That is, the burden of the proof should be on differential weighting; equal weighting should be the norm".

This obviously provides no substantive justification for equal weighting. Neither, we think, does the appeal to Occam's razor by Hopkins (1991, p. 1471): "Since it is probably impossible to obtain agreement on weights, the simplest arrangement is the best choice." The reason why the principle of parsimony provides no guidance here is important: opting for equal weighting does not imply choosing from a set of *otherwise equivalent* models of a given phenomenon. In fact, as exemplified by Table 2, the problem is rather that there are, at best, conflicting opinions available. Hence, equal weighting is not even an adequate *description* of the core debate in composite indicator construction.[8]

The IMI's construction illustrates how differential weighting is often introduced: a group of experts is consulted and the weight information they provide is aggregated, usually by averaging. This means that one uses the experts' weight information to arrive at an exogenously weighted quantity index, e.g. of the form:[9]

$$\hat{\gamma}_{ex} = \frac{\sum_{i=1}^{m} \overline{w_i} y_i^1}{\sum_{i=1}^{m} \overline{w_i} y_i^0},$$

where the weight $\overline{w_i}$ for a sub-indicator is the sample average of the experts' proposed valuations. Thus, the observed opinions are considered as a particular sample to which the laws of probability can be applied. However, if the experts are truly experts, i.e. when each one of them is endowed with a profound knowledge of the phenomena under consideration, one has to come to grips with the deep problem that "decision makers who wish to base choices on the advice of the panel have no way to objectively assign probabilities to the alternatives" (Woodward and Bishop, 1997, p. 494). Or, with reference to the IMI-application: can the Commission consider the weights given to postal tariffs by the Danish

[8] Moreover, as already touched upon, equal weighting as a rule interferes with the specific preliminary normalization process. For examples, see e.g. Panigrami and Sivramkrishna (2002) or Cherchye *et al.* (2004). This means that, at best, equal weighting is only apparently the simplest arrangement.

[9] The closest analogue to this formula in standard index theory is the 'Lowe quantity index' (after Lowe, 1823): prices for some base year *b* are selected, held fixed, and applied as weights to quantity vectors y^0, y^1. The corresponding Lowe price index (with a fixed quantity basket) is often used in applied calculations, e.g. of monthly price indices. Note however that in such cases the fixed basket is ultimately taken from direct market *observations*, whereas in the composite indicator case the fixed weights stem from some aggregation over individual expert *opinions*.

and French experts as ‘less *probable*’, just because these experts hold a minority position?[10] This issue is likely to gain importance in cases where opinions may be disparate (think of Table 2).

Finally, even if experts themselves perform the mental act of providing the right objective, *explanatory* weights, at least part of the purpose of the composite indicator is *normative*: it does eventually determine the countries’ benchmarks. Clearly, when even experts disagree, there is a flavour of strong value judgements present if one sees countries’ vaguely describable reality about ‘the’ internal market effects (or another composite phenomenon) being rigidly weighted and transformed into an exact number.

In short, we think that agreement on *any* particular common set of weights, whether that is the equal weight set or another one, is usually a mirage, and to depict it as such is therefore *intrinsically* problematic. However, this still need not imply that building reasonable composite indicators is impossible. We address the weighting issue in the following section. The proposed methodology will be used afterwards to construct an index of performance growth.

4. Benefit-of-the-doubt weighting

Foster and Sen (1997, p. 206) asserted that “[w]hile 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.” The aggregation method we employ in this section can be taken as an illustration of this idea. Basically, the idea is to apply the experts’ stated weight vectors as constraints in a weight optimization problem that seeks to maximize aggregate performance for each particular observation. Hence, for each observation, the weights leading up to its index value are to some degree *endogenous*, the degree of endogeneity depending on the extent of disagreement within the expert panel. Since for each observation the problem is formulated such as to yield the highest possible index value (given the weight constraints), this method has alternatively been labeled ‘benefit-of-the-doubt’-weighting (e.g. by Melyn and Moesen, 1991).

[10] According to Woodward and Bishop, the answer is unambiguously negative (1997, p. 494): “Since experts’ opinions vary because of underlying theories, in many circumstances the relative number of experts that hold a particular position tells us little about the likelihood that that perspective will be correct […] If the opinion of each expert is highly respected, then the inclusion of additional experts with opinions already represented on the panel should not be important to the decision process.”

The origins of this procedure are found in so-called non-parametric performance analysis, or 'data envelopment analysis' (DEA) (see e.g. Cooper *et al.*, 2004, for a recent overview of the vast DEA literature). The original question in that literature was how one could measure each firm's efficiency, given observations on input and output quantities in a sample of firms 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 in this paper, in which quantitative sub-indicators are available but weights are not. Indeed, and unsurprisingly, the scope of DEA has broadened considerably over the last two decades, including *inter alia* '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. The European Commission itself uses the method in its *EU Economy Review 2004*; see European Commission , 2004, p. 376-378).

The method is captured formally in expressions (3a) – (3c)[11]. Note that (3a) is an aggregate of normalised variables of type (1a) only, i.e. we divide each y_i^j, observation *j*'s value of sub-indicator *i*, by the corresponding base value y_i^O for that sub-indicator. This is however only to facilitate presentation; type (1b) normalisations have been used wherever appropriate. (As regards the base observation, we set y_i^O at the *i*-th sub-indicator's 1992 average value for the EU-15.)

As (3a) reveals, the denominator of the index value, i.e. the benchmark observation value, is itself obtained from an optimization problem. It is in fact the observation that, by employing the 'most favorable weights' for the *evaluated* observation, yields the maximal weighted sum of all observations in the sample. Consequently, this benchmark is endogenous too. Literally, it is either an observation that demonstrably outperforms the evaluated observation in terms of the latter's most flattering weighting scheme or, if such a superior observation does not exist, the evaluated observation serves as its proper benchmark[12]. Clearly then, the benefit-of-the-doubt character of the comparisons extends to the choice of the benchmark. The full set of index values is found by repeating this optimization procedure for all *n* observations.

[11] Note that this is a benefit-of-the-doubt weighted counterpart of the Jevonian quantity index (2a). Essentially the same idea can be applied to the Carli-index (2b), starting from a log-transformation of the original data (compare with Charnes *et al* (1983), and Banker and Maindiratta (1986) in the original DEA-context).

[12] In the latter case, the index value is trivially set at 100.

There are 12 sub-indicators that comprise the IMI, hence *m*=12 in (3a)-(3c). Next, we take *n*=30 in the empirical illustrations of this section. Specifically, we use data for the 14 Member States listed in Table 2 plus the average data for the EU-15. We confine ourselves here to observations for two years, *viz*. 1994 and 2000. Note further that in this section we pool all observations, regardless of the time period. This implies, for instance, that country *X*'s index for 1994 can be calculated with country *Y*'s aggregate performance in 2000 figuring in the denominator of (3a). Specifically, pooling implies an unaltered environment in both years and therefore the best practices are also assumed the same in both periods. This approach will be altered when explicitly dealing with dynamical aspects in section 5. Our primary attention is here on the weighting issue, as captured in (3b) and (3c).

Formally, for country/year j under evaluation, the benefit-of-the-doubt weighting problem is to select weights w_i that solve the problem

$$\max_{w_i} \frac{\sum_{i=1}^{m} w_i \left(\dfrac{y_i^j}{y_i^0} \right)}{\max\limits_{y^k} \sum_{i=1}^{m} w_i \left(\dfrac{y_i^k}{y_i^0} \right)} \tag{3a}$$

subject to

$$\sum_{i=1}^{m} w_i \left(\frac{y_i^k}{y_i^0} \right) \leq 100 \qquad \forall k : 1,...,n \tag{3b}$$

$$(w_1,..,w_i,..,w_m) \in \mathbf{W} \subseteq \mathbf{R}_{+}^{\mathbf{m}} \tag{3c}$$

As problem (3a) – (3c) is a weight selection problem for an individual observation *j*, and since we have 30 observations (14 countries plus a mean country, each for two years), the problem will be solved 30 times. Accordingly, we get 30 scores, one for each country in each year, and one for the EU-15 mean in each year.

Expression (3b) reveals that the weighted sum of the normalised sub-indicators is constrained to be at most 100, and, in fact, this implies that the endogenous benchmark value is normalised at that '100%'-value. Note that this goes against a common practice in composite indicator construction, where as a rule the weights themselves are directly restricted to add up to one. However, the latter approach is superfluous, and may even be misleading: since (3b) is a linear value function, any ordering of *n* different *m*-vectors of sub-indicators is unique up to a similarity transformation of the weights (i.e. $\sum_{i=1}^{m} w_i (y_i^k / y_i^0)$ and $\sum_{i=1}^{m} \omega_i (y_i^k / y_i^0)$ convey the same meaning iff $\omega_i = \lambda w_i, \forall \lambda > 0$). Stated differently: what matters in the linear composite are the relative weights (i.e. the $-w_i / w_j$, which directly refer to the substitutability of the different dimensions) rather than the absolute weights. Of course, by the same rule, nothing prohibits presenting the optimal endogenous weights *eventually* in the form $w_i^* /(\sum w_i^*)$, and we will adhere to that presentation of the weights below. But this should not obscure the essential idea that the resulting benefit-of-the-doubt 'number' is expressed *relative* to a base value of 100%. To recall, this base value is associated with the (best practice) bechmark for the evaluated observation. While eventually this is a matter of normalisation, the attractiveness of this particular choice in a setting characterised by uncertainty should be clear. In point of fact, given the limited information one can (and does) incorporate, the very concept of a best practice is itself necessarily relative and observation-specific. (Indeed, if there would be universal agreement about what constitutes best practice, then one *has* unanimous agreement about a base value, and country indices could be calculated by measuring the distance to this 'absolute' reference point).

Finally, the weight restrictions as defined in (3c) serve the objective of limiting the variation in the relative importance of the sub-indicators reflected in the experts' stated opinions. In general, there are many different ways in which 'value judgments' can be appended to the benefit-of-the-doubt procedure (see e.g. Thanassoulis *et al.*, 2004, for an overview). The alternatives we single out below do certainly not exhaust the full range of possibilities and should therefore be seen as illustrative examples of possible "agreed judgments", tailored to the IMI example.

A. Ordinal weight bounds on mutually agreed categories.

Looking back at table 2, one sees that several sub-indicators can be grouped under more general headings. For instance: it seems reasonable to say that the first two indicators relate to 'governmental barriers', that several indicators are concerned with specific market prices, and so on. Specifically, let us assume by way of example that experts commonly agree that the full set of 12 indicators can be subdivided in the following 5 categories (the explanation for the abbreviations is found under table 2):

I. 'Governmental barriers': {StA, Pup}

II. 'Prices': {Tel, Elec, Gas, Intrst, Post}

III. 'Free movement': {FDI, Trade, Pop}

IV. Countries' general price level, relative to the EU-average {RPL}

V. Pension funds assets {Pens}

From table 3, which is based on the accordingly categorized data of table 2, one can infer that the average 'group weights' I-III are close to each other. Also, although some individuals' judgments are still relatively distinct, the coefficients of variation for the so constructed groups are low in comparison with the original partition (IV and V are of course the same as columns RPL and Pens in table 2).

Table 3: Summary statistics on grouped sub-indicators of Internal Market Performance

	Categories				
	I	II	III	IV	V
Average	**27,0**	**33,1**	**29,3**	**9,9**	**0,7**
Max	65	65	60	25	10
Min	10	2	0	0	0
# 0	0	0	2	6	13
Median	27,5	32,5	30	11,5	0
Stdev	14,2	18,3	16,2	9,3	2,6
Var-coef	0,5	0,6	0,6	0,9	3,6

Let us next assume that the experts, after agreeing on the composition of the different categories, all support the following judgment:

> (O): *"In calculating an IMI for a specific country at a specific time, categories I, II and III are of equal relative importance. Their relative importance for a composite internal market index is not lower than that of category IV, which in turn is at least as important as V."*

In formal terms, this implies that expression (3c) can be specialized to the following set of weight restrictions: $\sum_{i\in I} w_i = \sum_{i\in II} w_i = \sum_{i\in III} w_i \geq \sum_{i\in IV} w_i \geq \sum_{i\in V} w_i \geq 0$. Or, in a lengthier notation:

$$(w_{StA} + w_{PuP}) = (w_{Tel} + w_{Elec} + w_{Gas} + w_{Intrst} + w_{Post}) = (w_{FDI} + w_{Trade} + w_{Pop}) \geq w_{RPL} \geq w_{Pens} \geq 0.$$

Imputing these constraints and solving (3a) – (3c) for our pooled sample leads to the results displayed partly in table 4. (These are results for the 1994 observations only. The index values for 2000 are at least as high for all countries, a feature that we will discuss in more detail further on.) Table 4 thus shows that 'limited expert consensus', as defined above, can still lead to composite index values: the remaining uncertainties are in fact taken at face value as the 'absence of further restrictions'. The remaining leeway granted to the evaluated observation is filled in by the impartial benefit-of-the-doubt weighting procedure. Indeed, the weights in table 4 all comply with the preceding broad value judgment, but one sees that it can be satisfied differently in order to arrive at a composite index.[13]

As regards the specific index values, we confine ourselves in this (a-temporal) exercise to single out two notable cases. First, it may be surprising that one finds both Sweden and Greece to achieve the highest possible index value. Sweden's result is less of a surprise: it performs good or excellent in several categories (e.g. published procurement, the intra EU-trade and human mobility variable, interest rates,…). However, Greece's favourable result is largely due to one sub-indicator, viz. its share of published public procurement (5.27% of GDP). Relative to the corresponding values of other observations, this can be regarded as an outlier (the second highest value being 3.81% for the UK in 2000). Here we take the position that this outlier value indeed reveals (observed) best practice in the concerned dimension and therefore conveys valuable information for relative performance analysis. We return briefly to the correction for outliers (notably when they can be due to measurement error) in the concluding section. Second, for sake of clarity we recall that the value for EU-15 in table 4 is not the average of individual country scores. Rather, it is the score of the average of the individual country category values.

Note that the discriminatory power of a benefit-of-the-doubt approach is rather high, despite the very modest nature of the weight restrictions associated with judgement 'O'. Indeed, if we were to drop these restrictions, a 'full' benefit-of-the-doubt approach would effectively allow for considerably more leeway in the selection of observation-specific weights and, a fortiori, in the calculation of an index value. To show this, we added in the rightmost column the index values that would result if the only extra constraint added to (3a)-(3b) were that all sub-indicator weights have to be equal or larger than zero. The statement $w_1,..,w_i,..,w_m \geq 0$ would then capture the most limited kind of possible consensus among the experts, viz. that each of the 12 sub-indicators *may* be used to construct a composite indicator to assess a country's internal market performance. Thus, when countries are given unlimited freedom to choose their own (nonnegative) weights, their choice of relatively idiosyncratic weights generates uniformly high unconstrained index values, and discriminatory power is sacrificed. This illustrates the trade-off between freedom to choose and discriminatory power.

[13] Judgment 'O' is defined on categories rather than at the level of the sub-indicators. Hence, even countries that share the same weights in table 3 may in fact mutually differ as regards the choice of weights *within* a category. For instance, the *sum* of the weights in category I equals 0.25 for seven countries in 1994, but at the more disaggregated level these values range from (0.01, 0.24) for Austria and France to (0.06, 0.19) for Belgium. Similar 'permissible disagreements' are found within categories II and III.

Table 4: IMIs for 1994, complying with experts' common judgment 'O'

	Index	Weights per category					*(unconstrained*
	Value	I	II	III	IV	V	*Index Value)*
AT	**62.7**	25.0	25.0	25.0	25.0	0.0	*100*
BE	**75.0**	25.0	25.0	25.0	25.0	0.0	*96.5*
DE	**84.2**	30.7	30.7	30.7	7.9	0.0	*100*
DK	**76.9**	29.8	29.8	29.8	10.5	0.0	*90.4*
ES	**62.5**	24.0	24.0	24.0	24.0	4.0	*100*
FI	**54.4**	25.0	25.0	25.0	25.0	0.0	*87.3*
FR	**56.0**	25.0	25.0	25.0	25.0	0.0	*99.5*
GR	**100**	33.3	33.3	33.3	0.0	0.0	*100*
IE	**85.4**	24.0	24.0	24.0	24.0	4.0	*100*
IT	**64.1**	25.0	25.0	25.0	25.0	0.0	*100*
NL	**76.8**	24.4	24.4	24.4	2.4	0.0	*100*
PT	**71.8**	25.0	25.0	25.0	25.0	0.0	*100*
SE	**100**	33.3	33.3	33.3	0.0	0.0	*100*
UK	**88.4**	24.4	24.4	24.4	2.4	0.0	*100*
EU-15	**60.4**	25.0	25.0	25.0	25.0	0.0	*95.8*

B. Relative weight bounds on mutually agreed categories.

As discussed earlier, relative weights rather than absolute weights are important for the method described in equations (3a)-(3c).[14] In principle, this means that each expert's weight vector can be transformed into a corresponding $m \times m$ relative weight matrix, and that one may consequently append to (3a)-(3b) constraints of the form $e_{ij}^{L} \leq w_i / w_j \leq e_{ij}^{U}$. For each pair ($i,j$), the lower and upper bounds e_{ij}^{L}, e_{ij}^{U} would then be given by the corresponding minimum and maximum values over all experts. The eventual result would be benefit-of-the-doubt index values of which the associated relative weights are, by construction, never set outside the limits as provided by the expert panel.

In reality, these limits can be quite extreme, and the specific IMI weight set displayed in table 2 is a notable example of this. For the IMI, every sub-indicator is granted a zero weight by at least one of the experts, and therefore applying this procedure would not lead to a truly 'restricted' optimization problem.[15] Zero minimum weights are (slightly) less of a

[14] For composite indicator construction in general, this fact has not gone unnoticed. In particular, the Analytic Hierarchy Process (Saaty, 1980), which typically builds on relative weights, has been suggested as one method for deriving weights (see e.g. the website referred to in the main text). However, contrary to the IMI case, AHP weights are typically picked from a 1 to 9 semantic pair-wise comparison scale (e.g. "1" means that the two attributes are "equally important", "3" that one is "weakly more important" than the other, etc.).

[15] Strictly speaking, relative weights are undefined if the denominator is zero, but in the IMI example one could sidestep this by letting the trade-off value go to infinity in such cases. This option would effectively imply that the weight set **W** in (3c) coincides with the non-negative orthant of $\mathbf{R}^{\mathbf{m}}$. Trivial as it may seem, such a model at least captures the limited consensus that *is* undeniably present in the IMI-case, *viz.* the unanimous agreement that

problem at the level of the categories defined above (cf. table 3). The Appendix shows in more detail how the just outlined procedure can be applied at the category level.

The second example we present can be conceived of as being intermediate between judgement "(O)" and the 'min-max'-approach just discussed: it loosens the strict indifference between the three first categories, and appends numerical upper and lower trade-off bounds between these and the other categories. Suppose that experts would agree on the following broad judgment, which we describe first in qualitative terms, but which is clearly inspired by the quantitative information contained in table 3:

> (B): *"In calculating an IMI for a specific country at a specific time, categories I, II and III are 'approximately' equally important. A decrease in any of the variables in these three categories can be counterbalanced by better performance in category IV, but the required trade-off is 'roughly somewhere between three and four to one'. Finally, better performance as regards pensions (category V) can compensate for worse performance in a country's general price level (category IV), but this requires a 'substantial' compensation."*

Obviously, these qualitative statements need to be translated into numerical weight bounds in order to solve (3a)-(3c). Suppose, more explicitly, that one agrees to capture them in the weight bounds:

$$0.85 \leq \frac{\sum_{i \in X} w_i}{\sum_{i \in Y} w_i} \leq 1.15 \qquad (X, Y \text{ representing categories I,II, and III,}) \qquad (4a)$$

$$0.25 \leq \frac{\sum_{i \in IV} w_i}{\sum_{i \in X} w_i} = \frac{w_{RPL}}{\sum_{i \in X} w_i} \leq 0.35 \;\; (X \text{ representing categories I,II, and III}), \qquad (4b)$$

$$0.06 \leq \frac{\sum_{i \in V} w_i}{\sum_{i \in IV} w_i} = \frac{w_{Pens}}{w_{RPL}} \leq 0.1. \qquad (4c),$$

and (recalling the original formulation in (3c)),

$w \geq 0$. Of course, problems such as these vanish if stated weights are all strictly positive. Interestingly, as the preceeding footnote makes clear, individual expert weights are non-negative by construction in AHP. In point of fact, experts' AHP weights have been used to create restrictions such as those discussed in the main text in a benefit-of-the-doubt optimization problem (see e.g. Cooper, Seiford and Tone, 2000, p. 169-174).

$$w_1,..,w_i,..,w_m \geq 0 \tag{4d}$$

Obviously, these restrictions implicitly define weight bounds between categories I-III and category V as well. Appending them to (3a)-(3b) leads to the results summarized in Table 5. Note that, in comparison with the results in table 4, the 1994 IMI values do not change very much, with some increasing and others decreasing, and discriminatory power remains high. The most outspoken differences are those for Italy and Portugal. One can see further that this is essentially due to the impact of (4b), i.e. the value judgment concerning trade-offs between the three 'most important' categories and category IV. Of course, the particular bound values (4a)-(4c) that *we* propose may well deviate from those the IMAC members themselves would deem appropriate. Since such information is lacking, the index values and endogenous weights shown in table 5 are unlikely to be the correct ones as far as the IMI is concerned. But the table illustrates the essence of the Foster and Sen argument, particularly since a mutual agreement on weight bounds is more likely to come about than agreement on specific weight values.

Table 5: IMIs, complying with experts' common judgment 'B' (as detailed further in restrictions (4a)-(4d)

	Index Value (1994)	Weights per category (1994)					**Index Value (2000)**
		I	II	III	IV	V	
AT	**65.0**	26.9	31.6	31.6	9.4	0.6	**98.2**
BE	**74.6**	26.9	31.6	31.6	9.4	0.6	**100**
DE	**90.1**	27.3	32.1	32.1	8.0	0.5	**88.9**
DK	**77.4**	28.2	33.2	28.2	9.9	0.6	**92.7**
ES	**62.5**	26.8	31.5	31.5	9.4	0.9	**88.0**
FI	**53.0**	26.9	31.6	31.6	9.4	0.6	**81.5**
FR	**54.6**	26.9	31.6	31.6	9.4	0.6	**78.7**
GR	**100**	32.1	32.1	27.3	8.0	0.5	**100**
IE	**82.4**	28.1	33.0	28.1	9.8	1.0	**99.4**
IT	**60.5**	28.2	33.2	28.2	9.9	0.6	**65.8**
NL	**75.5**	26.8	31.5	31.5	9.4	0.9	**100**
PT	**66.3**	28.2	33.2	28.2	9.9	0.6	**100**
SE	**100**	28.7	33.2	28.7	8.4	0.5	**100**
UK	**86.7**	28.2	33.2	28.2	9.9	0.6	**100**
EU-15	**59.8**	26.9	31.6	31.6	9.4	0.6	**77.6**

A similar analysis can be carried out for the 2000 data. In the last column of table 5 we confine ourselves to reporting the index values for that year. One can see that, except for Germany, no country has a lower aggregate performance index value in 2000. Indeed, many of them have a value close to or equal to the maximum value as set in equation (3b). To recall, this means that they are not demonstrably outperformed by other observations in the sample. Clearly, this indicates that countries moved forward in terms of beneficial internal market effects.[16] However, there are better ways to focus the analysis on dynamical aspects than working with a pooled, 'a-temporal' data set. We will address these in the following section.

5. Decomposing Internal Market performance dynamics

This section applies the benefit-of-the-doubt weighting method to the dynamic performance assessment of EU Internal Market effects. As we will show, there is a specific sense in which this approach may actually deliver more information than one normally retrieves from the output distance functions we employed in section 3. To see this, recall that the benefit-of-the-doubt approach endogenises the identification of a country's best practice observation as well. Clearly, such best practices may themselves alter over time. Thus, apart from the mere measurement of performance shifts, we also highlight the capacity of the method to disaggregate member states' observed performance shifts into changes *relative to* the benchmarks and performance changes *of* the benchmarks (i.e. catching up vis-à-vis the best practice versus genuine progress of the best practice itself). The aggregate performance index we propose is in fact strongly related to the Malmquist (1953) (output) quantity index, which is a ratio of two output distance functions, one using base period data and the other using comparison period data. The axiomatic properties of an output quantity index are e.g. described by Balk (1998, pp. 90-91). Färe *et al.* (1994) popularised the methodology (including the decomposition into catching up and best practice progress) for analyzing productivity changes; our following discussion adapts those ideas to the assessment of policy performance, and extends them by means of endogenously weighted composite indicators.

We specifically consider performance changes between a period t and a subsequent period t+1; in our application t stands for the year 1994 and t+1 for 2000. Because we shift from an atemporal to an intertemporal analysis, our notation will deviate somewhat from that previously used: we make an explicit distinction between normalised sub-indicators and

[16] In fact, this can be discerned from the original disaggregated data: the move forward is indeed primarily due to improvements in sub-indicator values rather than e.g. to changes in weights. Only price variables show, on average, a mixed picture in this respect. Germany's modest overall decline is largely due to its deterioration in intra-EU FDI (dropping back to slightly over one third of the 1994 value of 0.5% of GDP). This cannot be compensated by the (marginally weaker) performance in its population variable nor by the increase in intra-EU trade (since, in fact, a country as Sweden still outperforms Germany on that latter dimension).

weights pertaining to the base period *t* versus those for period *t*+1. For any country *j*, the *i*-th normalised sub-indicator for each period *l* (*l* = *t* or *t*+1) is presented as[17]

$$\left(\frac{y_i^{j,l}}{y_i^O}\right) \quad (l = t, t+1) \qquad \text{(5a)},$$

while the corresponding weights are denoted as

$$w_i^l \quad (l = t,\ t+1) \qquad \text{(5b)}.$$

In the following, we first discuss the measurement of overall performance change. Next, we decompose performance change into catching up and environmental change (which is reflected in different best practices, or "genuine progress"). To facilitate the presentation, we first take the aggregation weights (see (5b)) for periods *t* and *t*+1 as given; and we address computational issues following from endogenizing those weights afterwards. After the methodological part, we present the results of our application.

A. Measuring internal market performance change

Measuring performance change between periods *t* and *t*+1 essentially boils down to comparing the aggregate (or weighted) sub-indicator performance in the two periods. Given this, there essentially are two possibilities for aggregating the different (normalised) sub-indicators: one may use either the period *t* weights or the period *t*+1 weights. (Recall that, for simplicity, we first assume that these weights are known.) Using the period *t* weights yields the performance shift measure

$$PC^{j,t} = \frac{\sum_{i=1}^{m} w_i^t \left(\frac{y_i^{j,t+1}}{y_i^0}\right)}{\sum_{i=1}^{m} w_i^t \left(\frac{y_i^{j,t}}{y_i^0}\right)} \qquad \text{(6a)},$$

while the period *t*+1 weights yield

$$PC^{j,t+1} = \frac{\sum_{i=1}^{m} w_i^{t+1} \left(\frac{y_i^{j,t+1}}{y_i^0}\right)}{\sum_{i=1}^{m} w_i^{t+1} \left(\frac{y_i^{j,t}}{y_i^0}\right)} \qquad \text{(6b)}.$$

[17] For each sub-indicator *i* the same base value y_i^O will be used in periods *t* and *t*+1. To recall, we take y_i^O as the EU-15 mean for 1992.

The interpretation is clear: values above unity for the measures (6a) and (6b) indicate performance progress, whereas values below unity have the opposite interpretation.

Still, given that the aggregation weights will generally be different for the two time periods, the two performance change measures will usually have different values. It may even occur that the two measures yield conflicting conclusions; e.g., the measure (6a) may indicate performance progress while the alternative (6b) suggests performance regress. To avoid an arbitrary base of comparison, we suggest a performance change measure that is calculated as the geometric mean of the measures (6a) and (6b):

$$PC^{j} = \left(\frac{\sum_{i=1}^{m} w_i^{t} \left(\frac{y_i^{j,t+1}}{y_i^{0}} \right)}{\sum_{i=1}^{m} w_i^{t} \left(\frac{y_i^{j,t}}{y_i^{0}} \right)} \right)^{1/2} \left(\frac{\sum_{i=1}^{m} w_i^{t+1} \left(\frac{y_i^{j,t+1}}{y_i^{0}} \right)}{\sum_{i=1}^{m} w_i^{t+1} \left(\frac{y_i^{j,t}}{y_i^{0}} \right)} \right)^{1/2} \qquad (7).$$

Obviously, this ("Fisher ideal") metric preserves the intuitive interpretation of the above two measures: a value above unity indicates performance progress of country *j* between *t* and *t*+1; and a value below unity indicates performance regress.

B. Decomposing internal market performance change

The performance shift metric (7) is instrumental to know whether a country has advanced or not, but is silent on the question whether this performance change is mainly idiosyncratic rather than a result of generally changed circumstances, which, intuitively, would be revealed by changes in best practices. However, (7) can precisely be decomposed into (i) a part that is attributable to the country-specific better (resp. worse) performance relative to the best practice benchmark in period *t*+1 as compared to period *t,* and (ii) a part that is attributable to overall better (resp. worse) practice in period *t*+1 as compared to period *t*, including a better (resp. worse) performance of the best practice benchmark itself. In the following, we label part (i) as 'catching up' vis-à-vis the best practice: it captures the better (resp. worse) performance that is effectively due to country *j*'s catching up (resp. losing ground) relative to the best possible performance. Next, we label part (ii) as 'environmental change'. At heart, it exactly reflects genuine progress, viz. changes of the best practices in period *t*+1 as compared to the base period *t.* These changes essentially indicate different performance possibilities following from a more (resp. less) favorable environment, which defines the scope for policy-making.

To obtain the decomposition, we first multiply the numerator and the denominator of (7) by the geometric mean of the benchmark performances in period *t* and period *t*+1:

$$\left(\max_{y^{k,t}} \sum\nolimits_{i=1}^{m} w_i^t \left(\frac{y_i^{k,t}}{y_i^0}\right)\right)^{1/2} \left(\max_{y^{k,t+1}} \sum\nolimits_{i=1}^{m} w_i^{t+1} \left(\frac{y_k^{k,t+1}}{y_i^0}\right)\right)^{1/2} \qquad (8).$$

After rearranging, this multiplication yields

$$PC^j = CU^j \times EC^j , \qquad (9a),$$

where

$$CU^j = \left[\frac{\sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{j,t+1}}{y_i^0}\right) / \max\limits_{y^{k,t+1}} \sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{k,t+1}}{y_i^0}\right)}{\sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{j,t}}{y_i^0}\right) / \max\limits_{y^{k,t}} \sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{k,t}}{y_i^0}\right)}\right] \qquad (9b)$$

and

$$EC^j = \left[\left(\frac{\sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{j,t+1}}{y_i^0}\right) / \max\limits_{y^{k,t}} \sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{k,t}}{y_i^0}\right)}{\sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{j,t+1}}{y_i^0}\right) / \max\limits_{y^{k,t+1}} \sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{k,t+1}}{y_i^0}\right)}\right)^{1/2} \left(\frac{\sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{j,t}}{y_i^0}\right) / \max\limits_{y^{k,t}} \sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{k,t}}{y_i^0}\right)}{\sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{j,t}}{y_i^0}\right) / \max\limits_{y^{k,t+1}} \sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{k,t+1}}{y_i^0}\right)}\right)^{1/2}\right] \qquad (9c).$$

The expression (9a) presents the performance shift measure PC^j as the product of a 'catching up' component CU^j and an 'environmental change' component EC^j. The first component captures the extent to which the evaluated country *j* gets closer to its best practice benchmark in period *t*+1 as compared to period *t*; see (9b). The second component measures shifts in the best possible performance between periods *t* and *t*+1, which in turn reflects a more favorable policy environment. To see this interpretation of EC^j, consider the first factor of the geometric mean in (9c). This term will be above unity if

$$\left(\frac{\max\limits_{y^{k,t+1}} \sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{k,t+1}}{y_i^0}\right)}{\sum\nolimits_{i=1}^{m} w_i^{t+1}\left(\frac{y_i^{j,t+1}}{y_i^0}\right)}\right) > \left(\frac{\max\limits_{y^{k,t}} \sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{k,t}}{y_i^0}\right)}{\sum\nolimits_{i=1}^{m} w_i^{t}\left(\frac{y_i^{j,t+1}}{y_i^0}\right)}\right),$$

i.e., the distance from country *j*'s performance in period *t*+1 to the best practice in period *t*+1 exceeds that to the best practice in period *t*. If this inequality holds, then this suggests that better practice has become possible in period *t*+1 as compared to period *t*. A similar interpretation applies for the second factor of the geometric mean in (9c), but now the distances relate to country *j*'s performance in period *t*. Like before, taking the geometric mean avoids choosing an arbitrary base of comparison (i.e., country *j*'s performance in period *t* or in period *t*+1).

Obviously, both catching up and environmental change components can take values above and below unity. CU^j and EC^j values above unity indicate performance progress, while the opposite interpretation holds for values below unity. Note that the two components may move in opposite directions. For example, performance progress may occur because of a more favorable environment while the relative distance from the best possible practice deteriorates (i.e., the catching up value is below unity), or *vice versa*. Of course, it is possible that a country obtains a 100% value for the first component (e.g. because in each of the two periods it acts as a 'contemporaneous' best practice), and still has a value above/below 100% for the second factor (as it its performance improved/worsened over time, judged by cross-comparison with the observations of the other time period). The above decomposition of performance change is intuitive as both catching up and environmental change have a clear impact on the perception of overall performance shifts.

A final point of attention concerns the computation of the metrics PC^j and its components CU^j and EC^j when exact weighting information is not available. Consistent with our previous discussion, we suggest a benefit-of-the-doubt approach in such a case. In this respect, we note that the constituent components of (9b) and (9c) have the same formal structure as the performance ratio in (3a). Specifically, they include four ratios of the following form (for l_1, l_2 = t, t+1):

$$\frac{\sum_{i=1}^{m} w_i^{l_2}\left(\frac{y_i^{j,l_1}}{y_i^0}\right)}{\max\limits_{y^{k,l_2}} \sum_{i=1}^{m} w_i^{l_2}\left(\frac{y_i^{k,l_2}}{y_i^0}\right)}.$$

Under endogenous weights, each such ratio can be computed as

$$\max_{w_i^{l_2}} \frac{\sum_{i=1}^{m} w_i^{l_2}\left(\frac{y_i^{j,l_2}}{y_i^0}\right)}{\max\limits_{y^{k,l_2}} \sum_{i=1}^{m} w_i\left(\frac{y_i^{k,l_2}}{y_i^0}\right)} \qquad (10a)$$

subject to

$$\sum_{i=1}^{m} w_i^{l_2}\left(\frac{y_i^{k,l_2}}{y_i^0}\right) \leq 100 \qquad \forall k:1,...,n^{l_2} \qquad (10b)$$

$$(w_1,..,w_i,..,w_m) \in \mathbf{W} \subseteq \mathbf{R}_+^{\mathbf{m}} \qquad (10c)$$

In this problem, the weight set **W** is constructed in the same way as before (see our discussion of (4a)-(4d)). The value n^{l_2} refers to the number of countries observed in the time period l_2 (= t or t+1). (For example, in our application we have $n^t = n^{t+1} = 15$.) Like (3a)-(3c), this non-linear programming problem may be converted into a linear program; the reasoning is directly adapted from Cherchye *et al*. (2004, p. 934).

C. Empirical application

Table 6 reports the changes in IMI performance between 1994 and 2000; these results are obtained under the relative weight restrictions (4a)-(4d). A first observation is that internal market performance has generally improved; the average improvement amounts to almost 67% (i.e., the difference between 166.89% and the *status quo* value 100%). In fact, performance progress for individual countries is often substantial: for example, Spain, Finland, France, The Netherlands, Portugal, Sweden and the UK improved performance by more than 50%. Finally, if the 'average' EU country (see EU-15) did exist, it would have increased its overall performance by slightly more than 100% (recall again that 'EU-15' is the score of the average of the individual country category values).

Next, when focusing on the individual components of these overall performance shifts, an interesting finding is that the environmental change values systematically (and largely) exceed the catching up values. This suggests that the overall better internal market performance is almost exclusively due to a more favorable environment (i.e. an enlarged scope for policy performance in the EU). In fact, the environmental change value is everywhere above 100%, and in some cases even above 200% (see Finland, France, the Netherlands, Portugal and the UK). Stated differently, as judged by the composite index, Internal Market Policy may well have led to better best practices: one interpretation of the results is indeed that European policy-makers have succeeded in creating an environment that has fostered integration. This finding is underscored further by the observation that *all* individual observations' 2000-values lie above the 1994 best practice frontier.[18] On the other hand, none of the 1994 observations comes out as best practice when compared with the 2000 frontier.

Specific country results provide further intuition. In that respect, it is interesting to compare the environmental change values for countries with catching up components of 100%, reflecting that they are best practice in 1994 as well as 2000. For compactness, we focus on the specific example of Sweden and Greece (but a straightforwardly similar reasoning applies for other comparisons that include Belgium, The Netherlands, Portugal or the UK). The environmental change component amounts to 104.89% for Greece (i.e. the minimum value in our sample), while it equals as much as 157.00% for Sweden. In view of the formal expression (9c), this finding can be given a specific interpretation in terms of evaluating the beneficial effects of the more favorable policy environment in 2000 at different (best practice) 'policy mixes' (i.e., the country-specific configurations of the single-dimensional performances). For the current example, it means that the environmental change has contributed to better performance much more substantially at the more equilibrated Swedish policy mix than at the strongly specialized Greek one, which is in turn reflected in the overall performance change results for these two countries.

Another insightful example is Finland. When judged in the contemporaneous time frame, Finland's results for 1994 and 2000 are 100% and 81.53%, respectively. Recalling expression (9b), the latter two values explain the value of Finland's 'catching up' (or better: 'falling behind') score in table 6. Next, evaluating the 1994 performance with respect to the 2000 best possible practice yields a value of 53.04%, while assessing the 2000 performance by the 1994 environment obtains 203.17%. Combining this with the first two index values yields the environmental change component of $\sqrt{203.17\% / 81.53\%}\sqrt{100\% / 53.04\%}$ = 216.73%; see (9c).

While genuine progress is common to all countries, some countries improved less rapidly than others, a feature which is reflected by their falling behind (i.e. column 1). Again, one can take the above Finnish example (81.53%/100%) as an illustration. Only a single country (Spain) obtains a catching up value that strictly exceeds 100%, as it moves from a dominated position in 1994 to one of the best practices in the 2000 subsample. The

[18] This feature can not directly be inferred from table 6; it relates to the numerator of the first factor in expression (9c), which lies above 100% for each observation (ranging from 127.93% to no less than 516.49%, with an average value of 248.20%).

predominant impact of the more favorable policy environment also appears from the summarizing statistics at the bottom of Table 6.

Table 6: Performance change and its components, complying with experts' common judgment 'B' (as detailed further in equations (4a)-(4d))

	Catching up	Environmental change	Overall performance change
AT	98.46%	143.79%	141.57%
BE	100.00%	136.19%	136.19%
DE	90.59%	130.50%	118.21%
DK	95.71%	137.09%	131.20%
ES	116.23%	143.76%	167.10%
FI	81.53%	216.73%	176.71%
FR	84.13%	235.98%	198.54%
GR	100.00%	104.89%	104.89%
IE	99.44%	124.93%	124.23%
IT	96.83%	138.39%	134.01%
NL	100.00%	228.33%	228.33%
PT	100.00%	239.78%	239.78%
SE	100.00%	157.00%	157.00%
UK	100.00%	243.68%	243.68%
EU-15	98.35%	205.35%	201.97%
Average	97.42%	172.43%	166.89%
Max	116.23%	243.68%	243.68%
Min	81.53%	104.89%	104.89%
Median	99.44%	143.79%	157.00%

Our approach should be contrasted with the way the *actual* IMI is constructed, since in the latter case one measures progress with a fixed weighting scheme and, therefore, one implicitly evaluates each country relative to some (exogenous) 'average' point of reference. By contrast, the approach presented above is primarily concerned with identifying *best* practices *within* the set of observations, with shifts of these best practices over time, and with gauging country performance relative to these best practices. This means that it is problematical to compare the magnitude of our indices with findings such as "the index for the EU as a whole improved by 60 points in the period 1994-2002" (Tarantola *et al.*, 2004, p.

11).[19] Even though such a strict comparison is somewhat tricky, it is still interesting to note that our average performance improvement value (of almost 67%) is close in magnitude to the 'fixed weight'-IMI estimate reported by Tarantola *et al.* (2004).

Of course, like before, this application mainly serves illustrative purposes. As a result, the above findings are at best interpreted as indicative rather than conclusive. Still, in our opinion, they do highlight the potential of the outlined procedure for inter-temporal policy performance analysis. Specifically within the European context, it seeks to address the crucial question whether internal market performance progress is mainly due to favorable policy-environmental changes, or rather to country-specific catching-up effects. Our analysis suggests quite clearly that the main driver of internal market performance progress has been the establishment of a global policy environment that has led to improved best practice.

6. Concluding remarks

Booysen (2002, p.131) summarized the debate on composite indicators by noting that "not one single element of the methodology of composite indexing is above criticism". As indicated above, we think this lack of consensus is actually a defining feature of composite indicators, and in fact constitutes the unifying thread that links several critical issues in this area.

Before offering some concluding observations on the specific methodology reviewed we re-iterate that an alternative –or rather: complementary– approach to assess the uncertainties surrounding the construction of composite indicators is to present the calculations with extensive uncertainty and sensitivity analysis. We have refrained from doing so in our illustrative example, but we fully agree with Saisana *et al.* (2005) that for practical applications such analyses are warranted. In fact, the specific type of global analysis used by these authors can readily be applied to benefit-of-the-doubt models. As the indices (3) and (10) hinge on the selection of best practice benchmarks, sensitivity to outliers may be a concern here perhaps more than with other composite indicators. A notable example in our own sample was Greece, with its exceptional performance in terms of published public procurement, which we treated as a reliable observation in our analysis. If one casts doubt on

[19] Direct comparisons with the numerical values of the actual IMI are further complicated by the fact that its authors use a different normalisation procedure (to wit, a z-score transformation) of the original sub-indicator values.

the quality of the underlying values, one could always perform a robustness assessment of the results (see, e.g., Cazals *et al.*, 2002)

One way to interpret Booysen's remark is that the normalisation issue should not be regarded as an isolated stage in the construction methodology; it too pops up because ultimately there is no consensus on the proper underlying scientific model for aggregating 'apples and oranges' in a composite indicator.[20] Regarding normalisation, one well-known feature of (3a)-(3b) is that the eventual value of the composite gauge is invariant to ratio-scale transformations of the original sub-indicators in the 'unrestricted' case where $w_i \geq 0$ for all i=1,...m. (see e.g. Cooper *et al.*, 2000, p. 24). In general, this is no longer ascertained when, as in our approach, weight restrictions are introduced. In this respect, however, we point out that (a) the (individually meaningful) preliminary ratio-scale transformations in fact guarantee independence of measurement units and, as a result, (b) that the combination of such pre-normalised sub-indicators with relative weight restrictions is in fact equivalent to restricting the ratios of the (dimensionless) 'virtual outputs' associated with the base observation y^0.[21] Evidently, this also holds for the Malmquist-type of performance index discussed in the previous section.

We recall that the so resulting index has some desirable axiomatic properties (e.g., weak monotonicity, independence of units of measurement, proportionality,...) but also that other traditional desiderata are not, or are only partially, met. Specifically, one may feel unease with the fact that endogenous weighting (and the concomittant endogenous choice of a best practice observation) ultimately prevents a conventional ordering (i.e. an ordering on the basis of a common, fixed objective function). In fact, we think this is actually an attractive way to "preserve the ambiguity", and that uniform weighting is an instance of "trying to remove it through some arbitrary complete ordering", to recall again the Foster and Sen (1997)

[20] The aggregation of apples and oranges (or of apples and scientific journals) is a rather uncontroversial problem when constructing GDP, even if goods and services are strictly speaking not really commensurable. The trick to render them so is of course by multiplying with market prices, i.e. to work with monetary values. Trivial as this example may be, it proves the point made above, and also by Ebert and Welsch (2004, p. 271) that "arbitrary choices of measurement units can be accommodated on the basis of known scientific relationships". In the GDP-example, the 'known scientific relationship' requires a sufficient consensus that prices are sensible weights (e.g. because they are taken to represent relative factor productivities or marginal utilities).

[21] To see this, note that expressions of the form $\sum_{i=1}^{m} w_i \left(y_i^k / y_i^0 \right)$ can be rewritten as $\sum_{i=1}^{m} \tilde{w}_{i0} y_i^k$, i.e. with $\tilde{w}_{i0} y_i^0 = w_i, \forall i$. Consequently, relative weight restrictions of the kind we used are of the form $\omega_{i,j}^{-} \leq (\tilde{w}_{i0} y_i^0)/(\tilde{w}_{j0} y_j^0) \leq \omega_{i,j}^{+}$, constituting a special –actually simplified– instance of what Pedraja-Chaparro *et al.* (1997) call 'contingent weight restrictions'. Moreover, expressions such as e.g. (6a) could consequently have been written as $(\sum_{i=1}^{m} \tilde{w}_{i0}^{t} y_i^{j,t+1})/(\sum_{i=1}^{m} \tilde{w}_{i0}^{t} y_i^{j,t})$. As in a Lowe quantity index, such hybrid weights combine the weights for observation *j* with the sub-indicator values for the base observation y^0.

recommendations. But others may disagree on this, especially in view of the fact that composite indicators are sometimes used to provide a country ranking. On this account however, the proposed methodology is in fact sufficiently flexible: *if* the experts' judgments point in such direction, one can incorporate the restriction that endogenous weights should be 'similar' or even 'the same' for all observations. See e.g. Cherchye and Kuosmanen (2004), Kuosmanen et al. (2005), Despotis (2004).

With regard to the issue of weight flexibility, it should additionally be noted that our findings regarding performance dynamics are not primarily a consequence of the particular weight restrictions we imposed, but rather of our use of a theoretical construct which decomposes dynamics. To recall: expressions (9a)-(9c) allow to check the primacy of environmental change over catching up also if explicit weights would have been available for the respective time periods. Yet evidently, common agreement on sets of weights, one set for each time-period considered, is even more unlikely than in an atemporal context. Thus the (constrained) benefit-of-the-doubt approach to resolve remaining disagreements surely retains its attractive character.

A quite different possible criticism is that we have taken inspiration from the literature on (Malmquist) *quantity* indices, without accounting for the real possibility that better ('output') performance may have its price either in terms of diminished performance in other areas or in terms of the inputs that are needed in order to improve 'output' performance. Partly, this is due to the fact that many composite indicators only look at one side of the equation. In any case, the approach we presented is readily amenable to address issues of *productivity* change, for cases in which sub-indicator data can be categorized as outputs or inputs: as we have indicated in section 5, the idea of decomposing a (Malmquist) index into a catching up component and a best practice shift in fact originates from *productivity* indices rather than *quantity* indices (see Diewert and Nakamura, 2003, for a recent summary of appropriate productivity indices). Nonetheless, the Malmquist decomposition yields policy-relevant information even when focusing on outcome variables only. Our empirical analysis illustrates this, as it strongly suggests that 'Europe' created an environment conducive to global performance improvement, and that the overwhelming majority of EU-15 members have exploited the opportunity, albeit to different degrees. In fact, our decomposition suggests that many countries could have done an even better job in this respect. Such a finding cannot be delivered by the Internal Market Index, which is incapable of identifying the sources of the observed performance improvement.

Finally, and more generally, we stress again that the particular benefit-of-the-doubt approach we discussed is not a universal panacea for building composite indicators. As in other areas, it holds that the 'best' index is contingent on the specific context at hand. The prime feature it possesses is that, unlike the Internal Market Index itself, it respects diversity of expert judgement by incorporating ambiguity into a specific method used to create a composite indicator. Whatever the specific details of the eventual composite indicator, we are firmly convinced that such a flexible methodology is warranted.

References

Aczél, J. (1988), "'Cheaper by the Dozen': Twelve Functional Equations and their Applications to the 'Laws of Science' and to Measurement in Economics'", in W. Eichorn (ed.), *Measurement in Economics: Theory and Applications of Economic Indices*, Physica-Verlag, Heidelberg, pp. 3-17.

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

Balk, B. M. (1998), *Industrial Price, Quantity, and Productivity Indexes*. Kluwer Academic Publishers, Boston.

Banker, R., A. Maindiratta (1986), "Piecewise Loglinear Estimation of Efficient Production Surfaces", *Management Science* **32**, 126-135.

Booysen, F. (2002), "An Overview and Evaluation of Composite Indices of Development", *Social Indicators Research* **59**, 115-151.

Carli, G.-R. (1804), "Del valore e della proporzione de' metalli monetati", in Custodi, P. (ed.) *Scrittori classici italiani di economia politica,* Tomo 13, 297-366.

Cazals, C., J.P. Florens, L. Simar (2002), "Nonparametric Frontier Estimation: A Robust Approach", *Journal of Econometrics* **106**, 1-25.

Chakravarty, S.R. (2003), "A Generalized Human Development Index", *Review of Development Economics* **7**, 99-114.

Charnes, A., W.W. Cooper, L. Seiford, J. Stutz (1983), "Invariant Multiplicative Efficiency and Piecewise Cobb-Douglas Envelopments", *Operations Research Letters* **2-3**, 101-103.

Chatterjee, S.K. (2005), "Measurement of Human Development: an Alternative Approach", *Journal of Human Development* **6**, 31-53.

Cherchye, L., T. Kuosmanen (2004), Benchmarking Sustainable Development: A Synthetic Meta-Index Approach, forthcoming in M. McGillivray (ed.), *Perspectives on Inequality, Poverty and Human Well-Being*, United Nations University Press.

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

Cooper, W.W., Seiford, L.M., Tone, K. (2000), *Data Envelopment Analysis*. Kluwer Academic Publishers, Boston.

Cooper W.W., Seiford L.M., Zhu J. (2004): *Handbook on data envelopment analysis,* Kluwer Academic Publishers, Boston.

Desai, M. (1994), "The Measurement Problem in Economics", *Scottish Journal of PoliticalEconomy* **41**, 34-42.

Despotis, DK (2004), A reassessment of the human development index via data enevelopment analysis, *Journal of the Operational Research Society,*1-12.

Diewert, W. E. (1981), "The Economic Theory of Index Numbers: A Survey", original version in A. Deaton, (ed.), *Essays in the Theory and Measurement of Consumer Behaviour in Honor of Sir Richard Stone*, Cambridge University Press.

Diewert, W. E. and A. O Nakamura (eds) (1993), *Essays in Index Number Theory, vol 1*, North-Holland, Amsterdam.

Diewert, W. E.and A. O Nakamura (2003), "Index Number Concepts, Measures and Decompositions of Productivity Growth", *Journal of Productivity Analysis* **19**, 127-159.

Ebert, U, H. Welsch, (2004), "Meaningful Environmental Indices: A Social Choice Approach", *Journal of Environmental Economics and Management* **47**, 270-283.

European Commission (2002), '*Structural Indicators'*, Communication from the Commission, COM (2002), 551 Final.

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

European Commision-Internal Market Directorate General (2002), "Internal Market Scoreboard No. 11", Brussels. (Scoreboards can be found at http://europa.eu.int/comm/internal_market/score/)

Färe, R., S. Grosskopf and M. Norris and Z. Zhang (1994), Productivity Growth, Technical Progress and Efficiency Change in Industrialized Countries, *American Economic Review* 84, 66-83.

Foster, J., A. Sen (1997), *On Economic Inequality*, (2nd, expanded edition), Clarendon Press.

Freudenberg, M. (2003), "Composite Indicators of Country Performance: a Critical Assessment", *STI Working Paper 2003/16*, OECD, Paris.

Hopkins, M. (1991),"Human Development Revisited: A New UNDP Report", *World Development* **19**, 1469-1473.

Jevons, W. S (1865), "The Variation of Prices and the Value of the Currency since 1782", *Journal of the Statistical Society of Londen* **28**, 294-320.

Könus, A.A. (1924) "The Problem of the True Cost of Living Index", translation published in *Econometrica* **7** (1939), 10-29.

Kumar, S., and R.R. Russell (2002), Technical Change, Technological Catch-Up, and Capital Deepening: Relative Contributions to Growth and Convergence, *American Economic Review* 92, 527-548.

Kuosmanen, T., Cherchye, L., and T. Spilaïnen (2005), "The law of one price in data envelopment analysis: restricting weight flexilbility across firms", forthcoming in *European Journal of Operational Research.*

Lind, N.(2004), "Values reflected in the Human Development Index", *Social Indicators Research* **66**, 283-293.

Lowe, J. (1823), *The Present State of England in Regard to Agriculture, Trade and Finance,* Longman, Hurst, Rees, Orme and Brown, London.

Malmquist, S. (1953), "Index Numbers and Indifference Surfaces," *Trabajos de Estadistica* 4, 209-42.

Melyn W., and Moesen W., (1991), Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available, Public Economic research Paper 17, CES, KU Leuven.

Munda, G., M. Nardo (2003), "*On the Construction of Composite Indicators for Ranking Countries*", mimeo, Universitat Autonoma de Barcelona.

Nordhaus, W. D., and Kokkelenberg, E. C. 1999, *Nature's Numbers: Expanding the National Economic Accounts to Include the Environment*, Report of the Panel on Integrated Environmental and Economic Accounting to the Committee on National Statistics of the National Research Council, National Academy Press, Washington, DC, USA.

Panigrahi R., Sivramkrishna, S. (2002) An adjusted Human Development Index: robust country rankings with respect to the choice of fixed maximum and minimum indicator values, *Journal of Human Development* **3**, 301-311.

Pedraja-Chaparro, F., J. Salinas-Jimenez, P. Smith (1997), "On the role of weight restrictions in data envelopment analysis", *Journal of Productivity Analysis* **8**, 215-230.

Saaty, T.L., 1980. *The Analytic Hierarchy Process*. McGraw-Hill, New York.

Saisana, M., A. Saltelli, 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**, 1-17.

Selvanathan, E.A, D.S. Prasada Rao (1994), *Index Numbers: A Stochastic Approach*, University of Michigan Press, Ann Arbor.

Shephard, R.W., 1970. *Theory of Cost and Production Functions.* Princeton: Princeton University Press.

Tarantola, S., M. Saisana, A. Saltelli (2002), "Internal Market Index 2002": Technical Details of the Methodology", Ispra: European Commission: Joint Research Centre-Applied Statistics Group. (available at http://europa.eu.int/comm/internal_market/score/docs/score11/im-index-2002_en.pdf)

Tarantola, S., R. Liska, A. Saltelli, N. Leapman, C. Grant (2004), "The Internal Market Index 2004", EUR 21274, European Commission: Joint Research Centre.

Thanassoulis, E., M. C. Portela, R. Allen (2004), "Incorporating value judgements in DEA" , in W.W. Cooper, L. Seiford, J. Zhu (eds.), *Handbook on Data Envelopment Analysis (International Series In Operations Research And Management Science 71)*, Kluwer Academic Publishers, Boston.

Theil, H. (1967), *Economics and Information Theory*, North-Holland, Amsterdam.

Woodward, R.T., R.C Bishop (1997), "How to Decide when Experts Disagree: Uncertainty-Based Choice Rules in Environmental Policy", *Land Economics* **73**, 492-507.

WHO, 2000. *The World Health Report 2000.* Health Systems: Improving Performance. World Health Organization, Geneva.

Appendix 1: individual expert opinions as direct upper and lower bounds on trade-offs.

In section 4 we briefly discussed appending weight bounds of the form $e_{ij}^{L} \leq w_i / w_j \leq e_{ij}^{U}$, where the lower and upper bounds are taken as the minimum and maximum relative weights for each pair (*i,j*) over the expert panel. We here illustrate this approach at the level of the five categories discerned in the main text. The category weights per IMAC member are listed in table A1.

Table A1

	AT	BE	DE	DK	ES	FI	FR	GR	IE	IT	NL	PT	SE	UK
Category I	10	25	40	30	25	10	10	38	30	65	30	15	30	20
II	45	45	15	55	35	55	65	12	40	25	2	30	30	10
III	25	30	45	0	40	35	0	30	30	10	60	40	25	40
IV	20	0	0	15	0	0	25	20	0	0	8	15	15	20
V	0	0	0	0	0	0	0	0	0	0	0	0	0	10

The next step is then to calculate relative weights per expert, after which minima and maxima are set as bounds. For example, the minimum $w_{\mathrm{II}}/\ w_{\mathrm{I}}$ in the table above is given by the Dutch IMAC-member (2/30), the maximum $w_{\mathrm{II}}/\ w_{\mathrm{I}}$ is France's 65/10, max $w_{\mathrm{III}}/\ w_{\mathrm{II}} = 4$ (UK), etc. Due to the many zeroes in the above table, we make the assumption that 0/0 = 0, $x/0 \to \infty$ for $x>0$. In fact, this implies that no weight constraints can be imposed directly (i.e. on the basis of the table) for the pairs (III,IV), (III,V) and (IV,V). (Recall though that expression 3c requires that all single-indicator weights are positive. This requirement continues to hold here). Proceeding as such eventually yields the following weight restrictions for problem (3a-b):

$$0.067 \leq \frac{\sum_{i\in II} w_i}{\sum_{i\in I} w_i} \leq 6.5; 0 \leq \frac{\sum_{i\in III} w_i}{\sum_{i\in I} w_i} \leq 2.5; 0 \leq \frac{\sum_{i\in IV} w_i}{\sum_{i\in I} w_i} \leq 3.5; 0 \leq \frac{\sum_{i\in V} w_i}{\sum_{i\in I} w_i} \leq 0.5;$$

$$0 \leq \frac{\sum_{i\in III} w_i}{\sum_{i\in II} w_i} \leq 4; 0 \leq \frac{\sum_{i\in IV} w_i}{\sum_{i\in II} w_i} \leq 30; 0 \leq \frac{\sum_{i\in V} w_i}{\sum_{i\in II} w_i} \leq 1;$$

$$w_1,..,w_i,..,w_m \geq 0$$

The results are given in table A2. The 'wide disagreement' that is captured by the broad restrictions is of course directly mirrored both by the large increase in the benefit-of-the-doubt index values and by the dispersion in the associated optimal weights. Comparison of tables A1 and A2 shows that there may be rather wide divergence between national expert's original scheme and the corresponding scheme for their country following the benefit-of-the-doubt approach (although the lack of importance for category V, i.e. the pension fund indicator, is a consistent and notable finding in both approaches). Whether this implies that the panel should favor A1 from the start and directly should apply each national expert's proposal exclusively to the corresponding member state, is questionable. First, because this is tantamount to adding further (in fact extremely narrow) restrictions to model (3a-c). Put differently: this means that country values will *ceteris paribus* never be higher than in table A2. Second, because in terms of Foster and Sen’s argument, a direct application of A1-weights on their respective countries means constructing indices in a context of ‘universal disagreement on values’. With an eye towards real EU decision making, this seems a non-starter. Conversely, A2 de facto implies ‘universal agreement on bounds’.

Table A2: IMI’s for 1994, complying with panel’s upper and lower bounds

	Index	Weights per category				
	Value	I	II	III	IV	V
AT	**95.4**	17.3	22.1	60.6	0.0	0.0
BE	**90.6**	7.4	48.1	25.9	18.5	0.0
DE	**100**	9.1	59.1	31.8	0.0	0.0
DK	**86.7**	29.8	29.8	29.8	10.5	0.0
ES	**100**	10.0	65.0	0.0	25.0	0.0
FI	**75.5**	8.5	55.3	15.0	21.3	0.0
FR	**87.1**	7.5	48.6	25.3	18.7	0.0
GR	**100**	43.3	56.7	0.0	0.0	0.0
IE	**100**	8.7	56.3	13.5	21.6	0.0
IT	**98.1**	10.0	65.0	0.0	25.0	0.0
NL	**100**	9,5	62,0	22.0	6.4	0.0
PT	**82.4**	19,0	17,5	16.0	47.6	0.0
SE	**100**	9,1	59,0	31.7	0.0	0.0
UK	**100**	10.0	65.0	0.0	25.0	0.0
EU-15	**85.3**	7,5	48,6	25.2	18.7	0.0