Journal of the Operational Research Society (2008) 59, 239–251 0160-5682/08 $30.00

www.palgrave-journals.com/jors

Creating composite indicators with DEA and robustness analysis: the case of the Technology Achievement Index

L Cherchye[1]*, W Moesen[1], N Rogge[1], T Van Puyenbroeck[1], M Saisana[2], A Saltelli[2], R Liska[2] and S Tarantola[2]

[1]Catholic University of Leuven, Belgium; and [2]Joint Research Centre, European Commission, Italy

Composite indicators (CIs) are often used for benchmarking countries' performance, but they frequently stir controversies about the unavoidable subjectivity in their construction. Data Envelopment Analysis helps to overcome some key limitations, as it does not need any prior information on either the normalization of sub-indicators or on an agreed unique set of weights. Still, subjective decisions remain, and such modelling uncertainty propagates onto countries' CI scores and rankings. Uncertainty and sensitivity analysis are therefore needed to assess the robustness of the final outcome and to analyse how much each source of uncertainty contributes to the output variance. The current paper reports on these issues, using the Technology Achievement Index as an illustration.

Journal of the Operational Research Society (2008) **59,** 239–251. doi:10.1057/palgrave.jors.2602445
Published online 27 June 2007

Keywords: Data Envelopment Analysis; composite indicator construction; robustness analysis

1. Introduction

Organizations such as the United Nations, the European Commission, and others have developed and used composite indicators (CIs) in which sub-indicators are aggregated into one number, with a view to provide comparisons of countries in complex and sometimes elusive policy issues. These measures are increasingly recognized as a tool for policy making and, especially, public communications on countries' relative performance in wide ranging fields from environment, economy to technological development. More discussions on various CIs can be found in the information server: http://farmweb.jrc.cec.eu.int/ci/ provided by the Joint Research Centre of European Commission.

A CI is much like a mathematical or computational model. Just as for models, the justification for a CI lays in its fitness to the intended purpose and peer acceptance, while its construction owes more to craftsmanship than to universally accepted scientific rules for encoding. The construction of CIs involves stages that need subjective judgments: the selection of sub-indicators, the treatment of missing values, the choice of the aggregation model, the weights of the sub-indicators, and so on. These choices can, however, be used to manipulate the results. It is, thus, important to identify the sources of subjective assessment and data errors and use uncertainty and sensitivity analysis to gain insight during the process of CI building, including an appraisal of the reliability of countries' ranking. These considerations are a central theme of the current paper.

The construction methodology that is used in the present paper is rooted in data envelopment analysis (DEA). The original question in the DEA literature is how one could measure each decision-making unit's relative efficiency, given observations on input and output quantities in a sample of peers and, often, no reliable information on prices (Charnes and Cooper, 1985). One immediately appreciates the conceptual similarity between that original problem and the one of constructing CIs. In the latter case, quantitative sub-indicators for overall benchmarking are available, but as a rule there is only disparate expert opinion available about the appropriate weights to be used in an aggregation function. On the other hand, there are differences between the two settings; for example, a notable difference is that CIs typically look at 'achievements' without taking into account the input-side, though there are some interesting exceptions, for example the work of the European Commission (2005) on the Summary Innovation Index.

A well-known feature of DEA is that it looks for endogenous (possibly constrained) weights, which maximize the overall score for each decision-making unit given a set of other observations. This quality explains a major part of the appeal of DEA-based CIs in real policy-related settings. For example, several European policy issues entail an intricate balancing act between supra-national concerns and the country-specific policy priorities of member states. If one

*Correspondence: L Cherchye, Catholic University of Leuven, Campus Kortrijk, E. Sabbelaan 53; B-8500 Kortrijk, Belgium.
E-mail: Laurens.Cherchye@kuleuven-kortrijk.be

opts to compare the multi-dimensional performance of EU member states by subjecting them to a fixed set of weights, this may prevent acceptance of the entire exercise. To take an example: with reference to European social inclusion policy, Atkinson *et al* (2002) remark that 'in the context of the EU, there are evident difficulties in reaching agreement on such weights, given that each member state has its own national specificity.' As the essence of DEA is that it yields most favourable, country-specific weights, it may help to counteract such problems. However, the standard DEA set-up that only requires the non-negativity of weights does not suffice to guarantee peer acceptance. Often, expert opinion about the most appropriate weights is available, and such information should be considered in the DEA model.

DEA-based CIs have *inter alia* been used to assess labour market policy (Storrie and Bjurek, 2000), social inclusion policy (Cherchye *et al*, 2004), and internal market policy (Cherchye *et al*, 2007). A similar model has been tested to assess progress towards achieving the so-called Lisbon objectives (European Commission, 2004, pp. 376–378). Similarly, some authors have proposed a DEA-based model for the well-known Human Development Index (Mahlberg and Obersteiner, 2001; Despotis, 2005).

In this paper, we will illustrate our approach using the Technology Achievement Index, which together with the Human Development Index, was developed by the United Nations and included in the 2001 Human Development Report (United Nations, 2001). A further incentive to use the TAI is that it was extensively studied in the JRC-OECD *Handbook on Constructing Composite Indicators* (Nardo *et al*, 2005a,b). We will complement the handbook's results by providing a more in-depth application of the DEA approach. We will start in Section 2 by briefly discussing the TAI as well as the available information on possible sets of weights obtained from a panel of informed individuals. Section 3 presents the basic model and discusses its similarities and differences from conventional DEA models. We then address uncertainty and sensitivity analysis associated with our DEA model in Section 4. The current mainstream literature on sensitivity analysis for DEA-models is primarily concerned with the sensitivity of (in)efficiency scores induced by data perturbations for a *given* selection of inputs and outputs (Cooper *et al*, 2004). In the case of CIs, however, one is further concerned with the impact on the results of adding or deleting sub-indicators, altering expert information, and so on. Such issues have been addressed rather infrequently in the DEA literature (Valdmanis, 1992; Wilson, 1995; Banker *et al*, 1996; Simar and Wilson, 1998; Simar, 2003). Even here the parallel between CIs and mathematical models is useful. In mathematical models of natural or man-made systems uncertainty and sensitivity analysis related to modelling assumptions or scenarios have been studied (see Saltelli *et al*, 2004, for a review). The methodology that we present in Section 4 may therefore be valuable for a broader DEA audience as well. Section 5 concludes and offers some final remarks.

2. The Technology Achievement Index and expert opinion

The United Nations introduced the TAI to capture how well a country is creating and diffusing new or existent technologies and building a human skill base for technology creation, with the intention of helping policy-makers to define technology strategies (United Nations, 2001). As explained by Desai *et al* (2002), these dimensions are captured by eight achievement indicators: (I) the number of *patents* granted per 1 000 000 people, (II) the receipt of *royalties* (in US$, per 1000 inhabitants), (III) the number of *internet* hosts per 1000 people, (IV) *exports* of high and medium technology products (as a share of total goods exports), (V) the number of *telephone* lines per 1000 people (in logs), (VI) *electricity* consumption per capita (in logged kWh), (VII) the mean years of *schooling*, and (VIII) the gross *enrolment* ratio of tertiary students in science, mathematics and engineering.

This list exhibits a typical feature of most CIs, that is that the sub-indicators are displayed in various measurement units. The TAI authors deal with this problem by *normalizing* the original data on 0-1 scale using the formula: (original value − observed minimum value)/(observed maximum value − observed minimum value). Prior to this stage, the logarithms of the raw data for *telephone* and *electricity* are considered, as these sub-indicators are important at the earlier stages of technological advance but not at the most advanced stages. Expressing the measure in logarithms ensures that as the level increases, it contributes less to the overall index. The normalized sub-indicators are next *weighted* and *added*. Specifically, the UN takes the simple average of the eight sub-indicators.

We will now depart from the UN approach. One reason for doing so was previously mentioned in the introduction: applying DEA-based weights may help to foster acceptance of the eventual results by the national stakeholders. A second reason is that there exists some information on the weights for the eight sub-indicators, stemming from an internal JRC survey conducted on 21 interviewed individuals. The weights were obtained using the Budget Allocation method, in which individuals were requested to allocate points to the eight sub-indicators paying more for those sub-indicators whose importance they wanted to stress. Summary statistics about the budget allocation weights for the TAI sub-indicators is provided in Table 1.

There are considerable differences in the proposed weights; for example, there is not a single pair of individuals who suggest similar sets of weights. Limited consensus emerges from the panel on both the magnitudes and the relative importance of the sub-indicators. Interestingly, unanimity is achieved in judging that the *telephone* and *electricity* sub-indicators are the least important. Furthermore, although equal weights (of 1/8) fall within the upper and lower bounds over the sample of experts, no member of the panel proposed to weigh all sub-indicators equally, in contrast with the UN-TAI. If a

Table 1 Summary statistics on the eight sub-indicator weights provided by a panel of informed individuals

Weights	*Patents*	*Royalties*	*Internet*	*Exports*	*Telephone*	*Electricity*	*Schooling*	*Enrolment*
Mode	0.10	0.05	0.10	0.20	0.10	0.05	0.20	0.20
Average	0.11	0.11	0.11	0.18	0.10	0.06	0.15	0.18
St. dev.	0.05	0.07	0.05	0.07	0.05	0.04	0.06	0.08
Min	0.05	0.00	0.02	0.09	0.00	0.00	0.05	0.00
5th percentile	0.05	0.01	0.04	0.10	0.02	0.00	0.05	0.03
10th percentile	0.05	0.05	0.05	0.10	0.05	0.00	0.05	0.10
90th percentile	0.20	0.20	0.20	0.30	0.15	0.12	0.20	0.30
95th percentile	0.20	0.26	0.20	0.31	0.17	0.15	0.24	0.30
Max	0.20	0.30	0.20	0.33	0.20	0.15	0.25	0.30
Ranked (*)								
• on top	2	2	1	8	0	0	5	9
• at bottom	4	8	6	1	5	15	3	3

(*): Entries provide the number of times a sub-indicator figures on top (or at the bottom) of experts' preference. The horizontal sum exceeds the number of experts due to tied preferences.

single set of weights was to be used to represent *this particular* panel of individuals, the mode (first line, Table 1) would be a proper choice. However, much information on the weights would be left unused. To this end, we suggest an approach that incorporates this type of (disagreement) information in the calculation of the CI, and further assesses the impact of uncertainties on the country CI scores.

Before doing so, one more remark is in order. Weights in a linear aggregation $\Sigma y_{ij} w_i$ have the meaning of trade-offs. Hence, what matters in a linear aggregation are the relative weights (which directly refer to the substitutability of the different dimensions) instead of the absolute weights. It has, however, been observed (eg by Munda and Nardo, 2003) that experts usually interpret weights, such as those stemming from a Budget Allocation method, as 'importance coefficients' (cf Freudenberg, 2003, p 10: 'Greater weight should be given to components which are considered to be more significant in the context of the particular composite indicator'). In fact, the 21 experts were literally asked to assign more points to a sub-indicator 'the more important this indicator is'. We will consequently adhere to such an interpretation of the above weights in what follows. An alternative would have been to use results on the relative importance of the TAI sub-indicators, which derived from the Analytical Hierarchy Process (Nardo *et al*, 2005a,b). Owing to space limitations, we focus on the budget allocation results. However, as demonstrated in Cooper *et al* (2000, pp. 169–174), information deriving from Analytic Hierarchy Process can also be appended to DEA models by creating assurance regions for the weights.

3. A simple DEA-model

To construct a DEA-based CI, we consider a set of m sub-indicators for n countries, with y_{ij} the value of sub-indicator i in country j, where higher values indicate better performance. Our objective is to merge the sub-indicator values per country into a single number, calculated as the weighted average of the m sub-indicators; we use w_i to represent the weight of the ith sub-indicator. In the absence of reliable information about the true weights, we endogenously select those weights that maximize the CI score for the country under consideration. This gives the following linear programming problem for each country j:

$$CI_j = \max_{wij} \sum_{i=1}^{m} y_{ij} w_{ij}, \qquad j = 1, \ldots, n, \quad i = 1, \ldots, m$$

s.t.

$$\sum_{i=1}^{m} y_{ij} w_{ij} \leqslant 1 \qquad \textit{(bounding constraint)}$$

$$w_{ij} \geqslant 0 \qquad \textit{(non-negativity constraint)}$$

In this basic programming problem, the weights are non-negative and a country's score is between 0 (worst) and 1 (best). As pointed out by Despotis (2005), this model is formally equivalent to the original output maximizing multiplier DEA model with multiple outputs and constant inputs presented by Charnes *et al* (1978), in which the sub-indicators represent the different outputs and a single 'dummy input' with value unity is assigned to each country.

A DEA-based CI meets the important property of 'units invariance', which makes the normalization stage redundant. This is particularly convenient for practical reasons; see, for example the discussion in Freudenberg (2003) on the sensitivity of composite indicators results with respect to the specific normalization scheme that is used.

The non-negativity restriction on the weights, however, allows for extreme scenarios. If a country's value in a given sub-indicator dominates those of other countries, this country would always obtain a score of 1.0 even if it has very low values in many other sub-indicators. Furthermore, it may lead to a situation where a large number of countries are on the frontier (score = 1.0), rendering a further assessment impossible. Therefore, some additional constraints on the weights need to be introduced. In case of available information on the relative importance of the sub-indicators, as in the present case, this needs to be considered in the development of the DEA model.

The issue of imposing additional *a priori* weight restrictions has attracted considerable attention in the DEA literature (see Thanassoulis *et al* (2004) for a survey). In the present context, restrictions regarding the pie shares are particularly interesting as these (i) do not depend on measurement units and (ii) directly reveal how the respective pie shares contribute to a CI score. Formally, the ith pie share for country j is given as the product $y_{ij} w_{ij}$. Clearly, the sum of the pie shares equals the CI_j. In what follows, we focus on *pie share constraints* (for each sub-indicator i) of the type

$$L_i \leqslant \frac{y_{ij} w_{ij}}{\sum_{i=1}^{m} y_{ij} w_{ij}} \leqslant U_i \qquad \textit{(pie share constraint)}$$

with L_i and U_i the respective lower and upper bounds (Wong and Beasley, 1990).

The resulting CI_j remains invariant to the measurement units. An alternative would have been to bound the weights across countries by imposing that weights cannot vary (too much) over different country observations (Kao and Tung, 2005; Cherchye and Kuosmanen, 2006). We will refrain from pursuing this further in this paper, and instead build explicitly on the information provided by the panel. The available (budget allocation) information on the pie shares in the specific TAI case is consistent with formulating such upper and lower bounds.

We end this section by presenting the results for the TAI obtained using the constrained DEA model applied to 23 countries and eight sub-indicators. *Raw* data are used, since normalization is redundant in DEA models. We further refrain from using logarithms for the *telephone* and *electricity* values—as was originally used by the UN authors- and instead build the model on the raw data for these two indicators. However, we will return to this issue later in our analysis. For the ‘baseline scenario’, we additionally append pie share constraints that reflect the variation in the weighting sets stated by the panel. Specifically, we require that the relative pie share of each indicator should not lie outside the minimum and maximum bounds reported in Table 1 (ie the pie share of *patents* is between 5 and 20% of the aggregate score, the pie share of *royalties* and *enrolment* between 0 and 30%, etc).

Figure 1 gives a graphical presentation of the results for top-ranked Finland and low-ranked Singapore in the baseline scenario. The difference in the total CI score is indicated by the size of the pies, while the importance of the indicators by the pie shares. Table 2 presents the pie shares (measured in *absolute* terms) for a few selected countries including Finland and Singapore. The sum of a country’s pie shares equals the country’s CI score. All pie shares are in accordance with our starting point of granting leeway to each country when assigning the shares, while not violating the (relative) upper and lower bounds. The pie shares can be quite diverse in terms of their *relative importance.* For example, Finland assigns 1/4 of its total score to schooling, while the same dimension accounts for no more than 1/14 of Singapore’s CI score. On the other hand, Finland assigns 16/100 to *royalties* whereas Singapore actually maximizes its (duly constrained) score by completely neglecting that indicator. Different pie share patterns occur for countries with similar CI scores (eg Belgium and New Zealand). Note that assigning zero weights to some indicators (*royalties, telephone, electricity, enrolment*) is effectively consistent with the idea of respecting the lower bounds provided by the panel (eg three experts recommended to discard the electricity indicator). Interestingly, most countries’ pie shares equal the lower or upper bound for at least five sub-indicators.

Table 3 reports on the CI scores and ranks of this baseline scenario and those provided by the simple unrestricted DEA model where only the non-negativity constraint is in place. Countries are ordered according to their original rank (United Nations, 2001). The simple DEA model allows for extreme scenarios, resulting in zero weight values in 63.5% of all 184 cases (=23 countries × 8 dimensions). As a direct consequence, there are eight countries—Finland, USA, Sweden, Japan, Korea, Australia, Singapore and Norway—that reach a top score of 1. On the other hand, only four countries—Finland, USA, Sweden and Japan—score 1 in the baseline scenario. It is evident that the additional constraints on the pie shares allowed for a better assessment of countries’ technological achievement. It is worth noting that Singapore

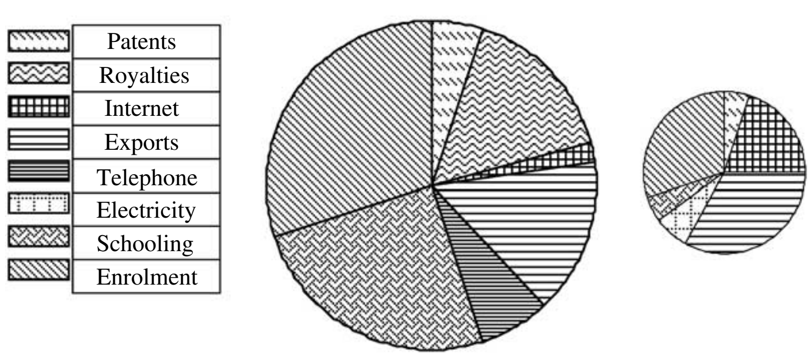

Figure 1 TAI for Finland (100%) and Singapore (14.3%) in baseline DEA scenario

Table 2 Pie shares (absolute terms) and their ('composite') sum for a few countries

Pie shares	*Patents*	*Royalties*	*Internet*	*Exports*	*Telephone*	*Electricity*	*Schooling*	*Enrolment*	Σ
Finland	0.05^L	0.16	0.02^L	0.15	0.07	0.00^L	0.25^U	0.30^U	**1.000**
Japan	0.20^U	0.00^L	0.02^L	0.33^U	0.20^U	0.12	0.09	0.04	**1.000**
Belgium	0.03^L	0.18^U	0.01^L	0.07	0.01^L	0.00^L	0.15^U	0.15	**0.616**
New Zealand	0.03^L	0.00^L	0.12^U	0.06^L	0.07	0.09^U	0.15^U	0.09	**0.614**
Italy	0.01^L	0.00^L	0.00^L	0.04	0.04^U	0.00^L	0.05^U	0.06^U	**0.204**
Singapore	0.01^L	0.00^L	0.03^U	0.05^U	0.00^L	0.01	0.01^L	0.04^U	**0.143**

Superscript '*L*' (or '*U*') indicates that this value equals the lower (or upper) bound of the relative pie share constraint associated with this indicator.

Table 3 DEA-based scores and ranks for the Technology Achievement Index

	DEA-based scores		*DEA-based ranks*	
Countries	*Unconstrained*	*Baseline*	*Unconstrained*	*Baseline*
Finland	1.000	1.000	1	1
US	1.000	1.000	1	1
Sweden	1.000	1.000	1	1
Japan	1.000	1.000	1	1
Rep. of Korea	1.000	0.625	1	12
Netherlands	0.994	0.901	9	5
UK	0.976	0.750	11	7
Canada	0.982	0.435	10	19
Australia	1.000	0.618	1	13
Singapore	1.000	0.143	1	23
Germany	0.921	0.818	13	6
Norway	1.000	0.732	1	10
Ireland	0.831	0.735	16	9
Belgium	0.802	0.616	20	14
New Zealand	0.975	0.614	12	15
Austria	0.820	0.729	18	11
France	0.849	0.736	15	8
Israel	0.813	0.565	19	16
Spain	0.756	0.436	22	18
Italy	0.822	0.204	17	22
Czech Republic	0.792	0.331	21	20
Hungary	0.856	0.320	14	21
Slovenia	0.684	0.553	23	17

Countries are ordered according to the original UN-TAI rank.

scores highest (1.000) under the simple unconstrained DEA model, while it scores lowest (0.143) under the baseline DEA model. Such a maximum score produced by the simple model is a direct consequence of the absence of any pie share constraints, which allows Singapore to assign no less than 85% of its pie to the *exports* indicator, and effectively neglect five other dimensions (viz those for which it reaches the lower bound in Table 2).

We have thus far analysed the technological achievement in several countries and discussed the advantages of using constrained DEA models, combined with information from experts, instead of the simple DEA approaches where only the non-negativity constraint is considered. We will next study the uncertainties that could be present in any DEA model used to develop a CI and analyse them under a sensitivity analysis framework.

4. Uncertainty and sensitivity analysis

4.1. Uncertainty analysis

Uncertainties in the development of a CI can be attributed to a number of factors induced by modelling choices, namely (see Nardo *et al*, 2005a, for a detailed discussion):

(a) The model chosen for estimating the measurement error in the data (eg based on the available information regarding variance estimation).
(b) The mechanism for including or excluding sub-indicators in the CI.
(c) The transformation and/or trimming of sub-indicators (eg aimed at removing the impact of outliers).
(d) The type of normalization scheme applied to remove scale effects from the sub-indicators.

(e) The amount of missing data and the choice of a particular imputation algorithm for replacing these missing data.
(f) The choice of the weighting scheme (eg equal weights, weights derived from a DEA-based approach, etc).
(g) The level of aggregation, if more than one level is used (eg at the indicator level or at the sub-indicator level).
(h) The choice of the aggregation scheme (eg additive aggregation, multiplicative aggregation, or aggregation based on multi-criteria analysis).

All these choices may influence the countries' CI scores and should be taken into account before attempting any interpretation of the results. Saisana *et al* (2005) studied the uncertainties in the Technology Achievement Index, focusing on the type of normalization for the sub-indicators, the way of eliciting information from the experts on the weight issue and the sub-indicators' weights.

In the previous section, we discussed the use of a suitable DEA model that incorporates expert opinion on the relative importance of the different dimensions of a CI. Still, this baseline scenario is effectively characterized by specific modelling choices. We focus on two points (see point (c) and (f) in the list above) that could introduce uncertainty in the countries' CI scores:

- the consideration of logarithms for *telephone* and *electricity* indicators, as applied in the original version of the TAI, and
- the weights (pie shares in our case) provided by experts, together with the corresponding pie share constraints for the DEA model.

The remaining types of uncertainty listed above are either non-applicable or non-relevant in our case. There is no information about the measurement error in the data (point (a)), the type of normalization has no impact on the results due to the units' invariance feature of DEA models (point (d)), there are no missing data in the dataset (point (e)), and finally, the UN authors selected one level of aggregation for the indicators in TAI (point (g)). The inclusion/exclusion of an indicator (point (b)) is already inherent in our second source of uncertainty, since several experts have suggested eliminating some of the indicators from the dataset.

The uncertain input factors in our analysis are described in Table 4. The triggers X_1 to X_{21} decide whether a particular individual should be included in the panel. The members of the panel are sampled independently. Next, factor X_{22} determines the type of the pie share constraint that is used in the DEA model, be it either the min–max values of the set of weights provided by the panel, or the 5th–95th percentiles, or the 10th–90th percentiles. Finally, trigger X_{23} determines whether to use logarithms or raw data for the *telephone* and *electricity* indicators. Note that in the $K = 23$ dimensional space of uncertainties there are $2^{21} \times 3 \times 2 = 12\,582\,912$ possible combinations of the input factor values. Given that we cannot afford a full design with so many simulations, we need a representative sampling of the space of uncertainties. We use an **LP**-τ (quasi random) sampling scheme (Sobol', 1967) of size $N = 24\,576$ for the purposes of the sensitivity analysis to be discussed in detail in Section 4.2.

At this point, it is worth stressing that we do not build the panel by eliminating one individual at a time, but by including a *group* of them randomly. To illustrate this point, Figure 2 shows the probability distribution function (pdf) of the number of individuals included in the simulations. With **LP**-τ we guarantee that all simulations consider at least three individuals and a maximum of 21. On average, the panel is composed of 11 individuals. We undertook this approach to avoid that the results of our analysis depend on the number of individuals interviewed.

These uncertainties are translated into a set of N combinations of scalar input factors, which are sampled from their discrete distributions (three levels for X_{22}, two levels for the remaining factors) in a Monte Carlo simulation framework. The CI is then evaluated N times. These N CI scores are non-linear functions of (the corresponding draws of) the uncertain input factors, and the estimation of their pdf is the purpose of the uncertainty analysis.

Figure 3 shows the results of the uncertainty analysis for the technological achievement: whisker plots present the median value, the 5th and 95th percentiles of the empirical CI

Table 4 The 23 uncertain input factors for the analysis

Input factor	*Definition*	*Alternatives*
X_1	Consideration of Individual #1	• Included • Excluded
X_2	Consideration of Individual #2	• Included • Excluded
…	…	…
X_{21}	Consideration of Individual #21	• Included • Excluded
X_{22}	Pie share constaint	• Min-max • 5th–95th percentile • 10th–90th percentile
X_{23}	Data transformation	• Raw data • Logarithms for *telephone* and *electricity*

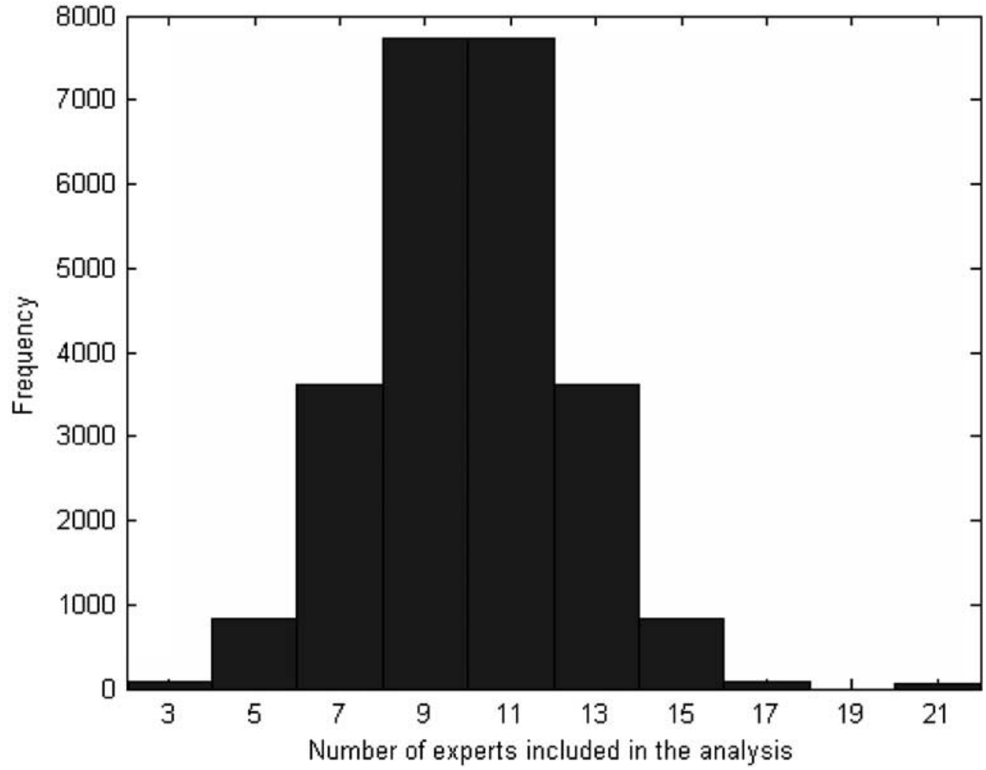

Figure 2 Probability distribution function of number of experts included in the analysis.

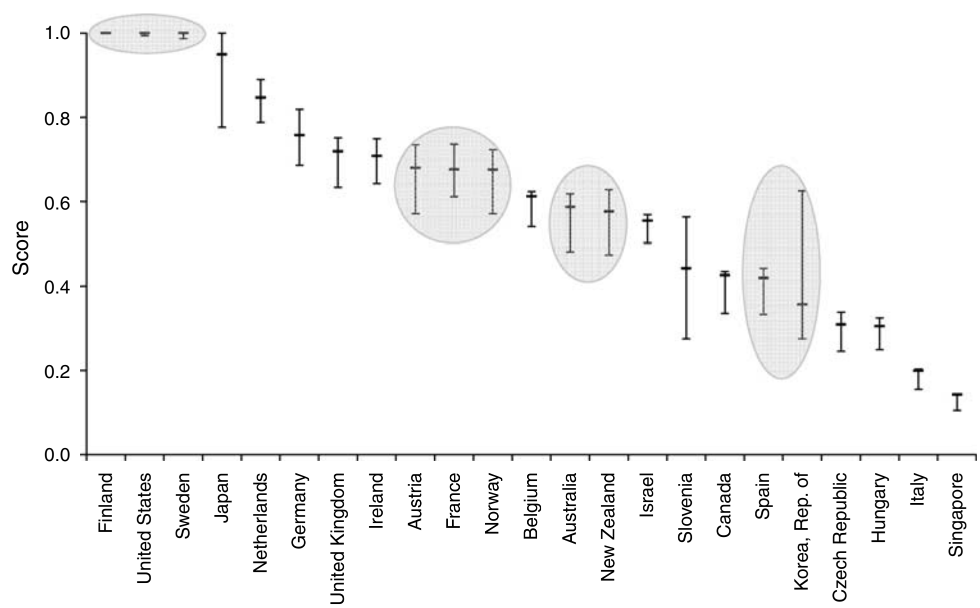

Figure 3 Results of Uncertainty analysis—Countries' scores using the DEA model. *Note*: Median composite indicator scores (black mark), 5th and 95th percentiles (bounds). Countries are ordered according to the median of the pdf of the corresponding ranks.

scores distribution. The graph should be read 'horizontally': if whisker plots partially overlap then this indicates that the ranks of the corresponding countries can interchange and the countries have similar performance. Conversely, if two countries are not overlapping, then this means that the policy inference (ie 'one country performs better than another') is robust, and it is independent of the uncertainties. Finland, USA, and Sweden are unarguably the countries with the highest technological achievement. There are, however, several countries whose relative performance is strongly influenced

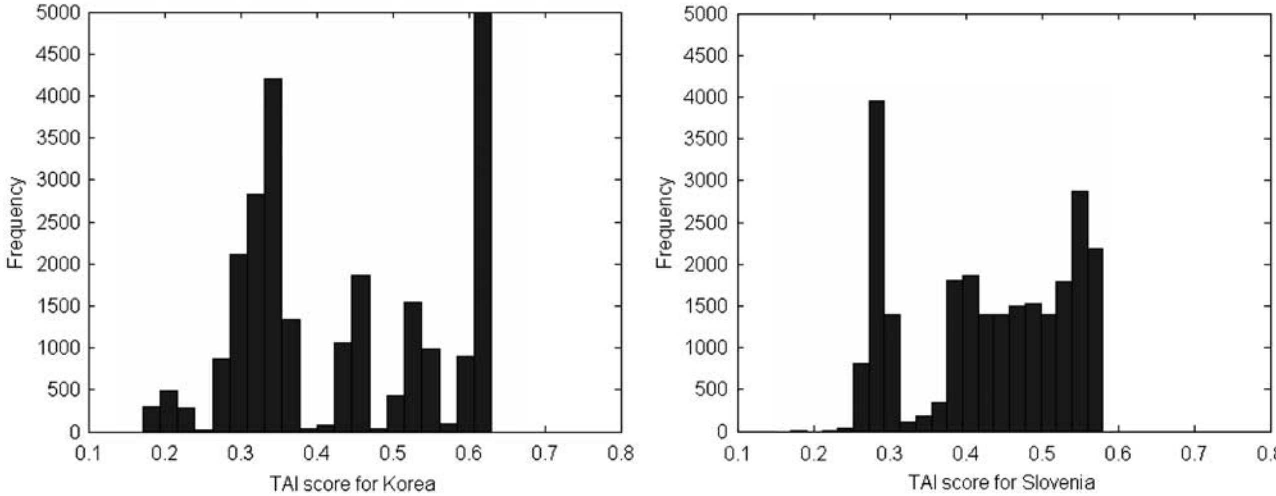

Figure 4 Uncertainty analysis of the DEA-derived TAI scores for Korea and Slovenia.

by the assumptions in the DEA model, with the most notable being Korea and Slovenia. The distributions of their empirical CI scores are plotted in Figure 4. Korea's score can range between 0.276 (5th percentile) and 0.625 (95th percentile), while for Slovenia the performance is estimated between 0.275 and 0.564.

Going back to the overlap in the countries' CI scores, as shown in Figure 3, an evident question is: which countries have *significantly* different performance in technological development? Can we argue that France (median score = 0.676) performs better than Norway (0.675), or that Canada's level of technological achievement (0.426) is superior to that of Spain (0.420)? An hypothesis test could provide the answer. We applied the Wilcoxon signed rank test on the median differences in paired TAI scores. The test, also known as the Wilcoxon matched pairs test, is the non-parametric equivalent of the paired t-test (see Conover, 1980). The assumption that the differences follow a normal distribution (as needed for the t-test) was not confirmed in our case. In contrast, the only assumption required for the Wilcoxon test, that the distribution of the differences is symmetric, was confirmed. Applying this test we identify four groups of countries for which no distinction can be made in terms of their technological achievement level. The groups are shaded in grey in Figure 3. The first group contains the top three performing countries Finland (1.000), USA (1.000) and Sweden (1.000). The second group is composed of Austria (0.680), France (0.676) and Norway (0.675). Australia (0.587) and New Zealand (0.576) belong to the third group. Finally, Slovenia (0.443) and Canada (0.426) belong to the fourth group. Note that, although the median score for Spain is 0.420, which is very close to that of Canada (0.426), the performance of the two countries can be clearly distinguished.

We next complement our uncertainty analysis with a sensitivity analysis. We first investigate sensitivity of the countries' scores with respect to outlier countries that define the frontier of the DEA model. We then use variance-based techniques to apportion the variance in the countries' scores to the two major sources of uncertainty in our analysis.

4.2. Sensitivity analysis due to outlier countries

Procedures that randomly omit some observations (in our case countries; for example one randomly excludes one country at a time) have been suggested in the DEA literature as a way to correct for the impact of outlier observations (eg Wilson, 1995; Cazals *et al*, 2002; Simar, 2003). To assess such an impact on the countries' scores, we have repeated the Monte Carlo approach described above, after eliminating one country at a time from the set of 23 countries. The $N = 24\,576$ CI scores are estimated for each group of 22 countries. The corresponding box plots are presented in Figure 5. The leftmost box plot in each graph represents the CI score distribution that is obtained for the given country when using the full sample of countries (compare with Figure 3); the following box plots (left to right) represent the CI distribution obtained after eliminating one country from the sample, starting from Finland (second box plot), USA (third box plot) to France (penultimate box plot) and Israel (final, rightmost box plot). The two countries that have the greatest impact on the countries' scores are Japan and Finland. When Japan is eliminated from the set, the countries that improve their score are: Finland, United Kingdom, Australia, Ireland, Spain, Czech Republic, Canada, Singapore, Norway, Belgium, Israel, Italy and Hungary. When Finland is eliminated from the set, the countries that improve their score are: Sweden, United States, Korea, Germany, France, and to a lesser degree Japan. Eliminating any of the remaining countries does not have a notable impact on the countries' scores. This result suggests that, for this application, the DEA model is quite robust with respect to outlier observations. Consequently, we do not explicitly account for outlier countries in the remaining analysis.

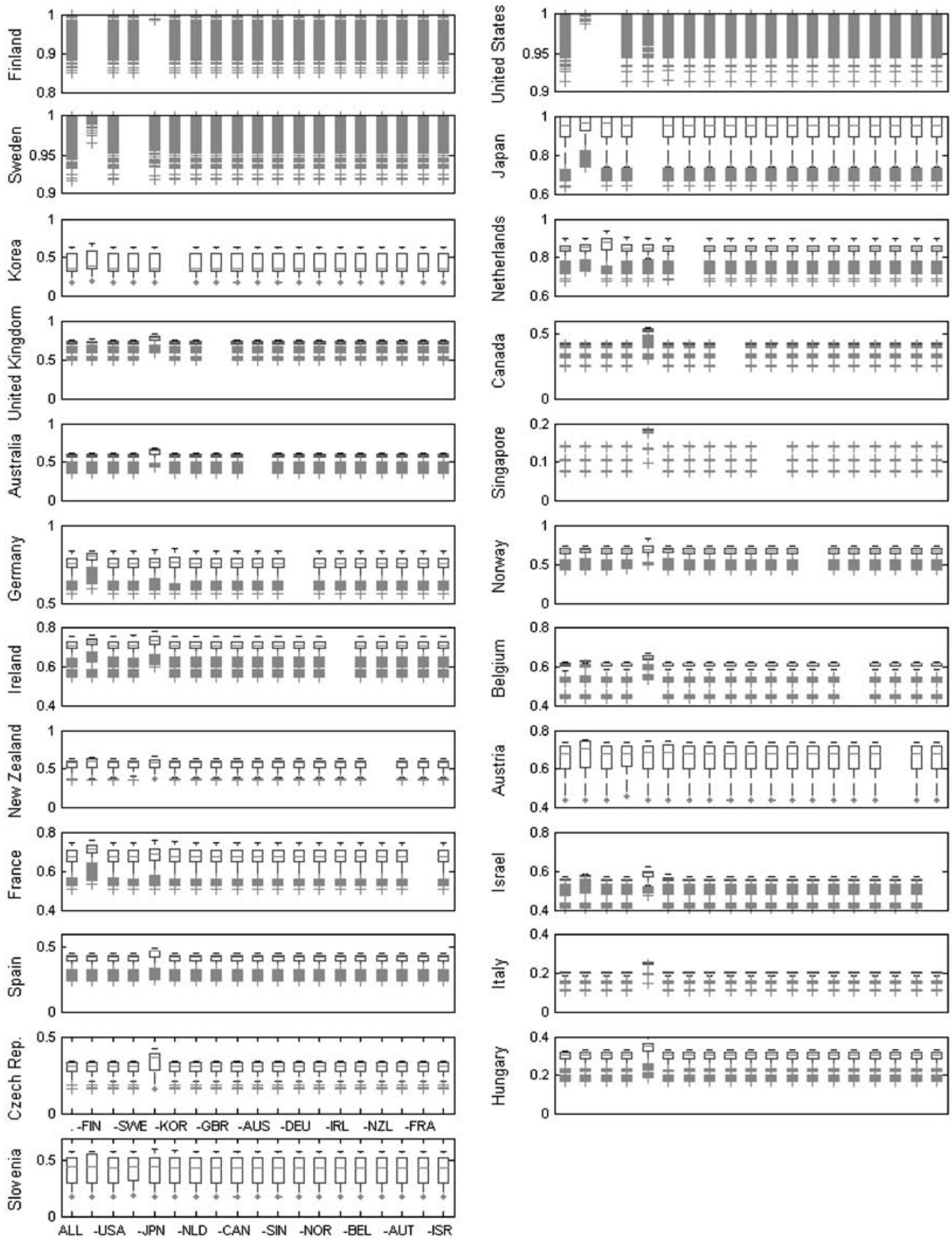

Figure 5 Boxplots of DEA-based countries' scores when eliminating a country at a time. *Note*: The box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers (+) are data with values beyond the ends of the whiskers. If there are no data outside the whisker, a dot is placed at the bottom whisker.

4.3. Sensitivity analysis using variance-based techniques

At this stage, sensitivity analysis is needed to quantify the impact of the K (=23; see Table 4) different uncertain input factors on the variance (uncertainty) of the countries' scores. Variance-based techniques are appealing and well suited to our case study. The discussion of their methodological formulation to compute sensitivity measures that account for

Table 5 Sensitivity measures of first order and total effect for the composite indicators scores

	X_1	X_2	X_3	X_4	X_5	X_7	X_{10}	X_{12}	X_{14}	X_{15}	X_{16}	X_{17}	X_{19}	X_{20}	X_{22}	X_{23}	*Sum*
First-order sensitivity measures, S_k																	
Finland	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.091
United States	0.00	0.00	0.00	0.00	0.09	0.06	0.00	0.00	0.00	0.00	0.14	0.02	0.00	0.00	0.00	0.00	0.317
Sweden	0.01	0.00	0.01	0.00	0.03	0.11	0.03	0.00	0.00	0.02	0.10	0.00	0.01	0.01	0.05	0.01	0.460
Japan	0.00	0.00	0.04	0.00	0.06	0.12	0.14	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.13	0.01	0.685
Korea	0.00	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.00	0.855
Netherlands	0.00	0.00	0.00	0.00	0.13	0.12	0.01	0.00	0.01	0.14	0.00	0.00	0.02	0.00	0.17	0.01	0.653
U. Kingdom	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.05	0.09	0.00	0.00	0.01	0.03	0.03	0.00	0.323
Canada	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.08	0.06	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.316
Australia	0.05	0.01	0.17	0.01	0.00	0.00	0.01	0.08	0.05	0.00	0.00	0.00	0.03	0.15	0.02	0.00	0.615
Singapore	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.06	0.03	0.00	0.00	0.00	0.07	0.00	0.00	0.332
Germany	0.00	0.00	0.05	0.00	0.00	0.00	0.04	0.00	0.01	0.31	0.00	0.00	0.00	0.01	0.13	0.07	0.627
Norway	0.02	0.00	0.12	0.01	0.00	0.01	0.00	0.09	0.03	0.00	0.00	0.00	0.01	0.19	0.09	0.02	0.621
Ireland	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.19	0.00	0.00	0.01	0.01	0.06	0.03	0.401
Belgium	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.07	0.02	0.00	0.00	0.00	0.05	0.00	0.04	0.315
New Zealand	0.01	0.01	0.48	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.01	0.18	0.02	0.05	0.808
Austria	0.00	0.00	0.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.02	0.01	0.748
France	0.00	0.00	0.06	0.00	0.00	0.00	0.03	0.00	0.02	0.39	0.00	0.00	0.00	0.00	0.12	0.02	0.646
Israel	0.05	0.00	0.00	0.02	0.00	0.00	0.00	0.05	0.10	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.287
Spain	0.05	0.01	0.28	0.01	0.00	0.00	0.00	0.02	0.06	0.01	0.00	0.00	0.00	0.14	0.00	0.02	0.583
Italy	0.06	0.00	0.04	0.00	0.00	0.00	0.00	0.06	0.07	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.347
Czech Rep.	0.02	0.01	0.43	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.17	0.00	0.01	0.698
Hungary	0.04	0.00	0.25	0.00	0.00	0.00	0.00	0.03	0.04	0.00	0.00	0.00	0.00	0.14	0.00	0.02	0.545
Slovenia	0.00	0.00	0.70	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.08	0.00	0.01	0.826
Total-effect sensitivity measures, S_{Tk}																	
Finland	0.65	0.00	0.00	0.43	0.00	0.01	0.00	0.66	0.57	0.01	0.01	0.00	0.10	0.69	0.00	0.00	
United States	0.02	0.48	0.08	0.02	0.56	0.56	0.13	0.02	0.02	0.16	0.51	0.14	0.00	0.02	0.09	0.00	
Sweden	0.16	0.22	0.02	0.04	0.35	0 56	0.11	0.04	0.14	0.29	0.26	0.08	0.09	0.05	0.10	0.07	
Japan	0.04	0.04	0.12	0.05	0.18	0.24	0.31	0.03	0.03	0.28	0.03	0.04	0.03	0.04	0.33	0.01	
Korea	0.03	0.01	0.15	0.02	0.01	0.03	0.03	0.01	0.02	0.78	0.02	0.01	0.00	0.04	0.03	0.00	
Netherlands	0.08	0.07	0.05	0.06	0.30	0.28	0.04	0.08	0.11	0.21	0.07	0.03	0.05	0.09	0.27	0.01	
U. Kingdom	0.46	0.01	0.01	0.10	0.01	0.02	0.02	0.42	0.43	0.12	0.01	0.01	0.02	0.42	0.07	0.00	
Canada	0.50	0.01	0.01	0.09	0.01	0.00	0.01	0.54	0.52	0.00	0.01	0.00	0.00	0.53	0.01	0.00	
Australia	0.26	0.02	0.21	0.06	0.02	0.04	0.02	0.36	0.33	0.00	0.02	0.01	0.06	0.39	0.06	0.00	
Singapore	0.54	0.01	0.00	0.08	0.01	0.00	0.01	0.55	0.54	0.00	0.01	0.00	0.00	0.55	0.00	0.00	
Germany	0.06	0.03	0.15	0.04	0.07	0.06	0.14	0.04	0.06	0.43	0.02	0.02	0.02	0.08	0.25	0.07	
Norway	0.15	0.02	0.17	0.05	0.03	0.07	0.01	0.30	0.22	0.01	0.03	0.02	0.05	0.38	0.19	0.02	
Ireland	0.34	0.02	0.03	0.10	0.02	0.06	0.05	0.27	0.31	0.24	0.02	0.01	0.03	0.26	0.12	0.03	
Belgium	0.50	0.01	0.00	0.13	0.00	0.01	0.01	0.50	0.50	0.02	0.01	0.00	0.01	0.47	0.00	0.05	
New Zealand	0.05	0.01	0.59	0.03	0.01	0.03	0.01	0.11	0.08	0.01	0.01	0.01	0.02	0.31	0.06	0.05	
Austria	0.01	0.01	0.75	0.01	0.02	0.01	0.02	0.01	0.02	0.02	0.01	0.01	0.01	0.29	0.07	0.01	
France	0.06	0.02	0.13	0.03	0.06	0.06	0.12	0.02	0.07	0.49	0.02	0.02	0.02	0.07	0.22	0.03	
Israel	0.44	0.01	0.01	0.14	0.02	0.02	0.02	0.46	0.50	0.02	0.01	0.00	0.00	0.41	0.06	0.01	
Spain	0.27	0.01	0.35	0.07	0.02	0.01	0.02	0.25	0.28	0.03	0.02	0.00	0.00	0.40	0.02	0.02	
Italy	0.49	0.01	0.05	0.08	0.01	0.00	0.01	0.49	0.50	0.00	0.01	0.00	0.00	0 54	0.00	0.00	
Czech Rep.	0.17	0.01	0.52	0.05	0.01	0.01	0.00	0.17	0.18	0.01	0.01	0.00	0.00	0.39	0.02	0.01	
Hungary	0.30	0.01	0.31	0.07	0.01	0.01	0.01	0.30	0.30	0.01	0.01	0.00	0.00	0.43	0.02	0.02	
Slovenia	0.01	0.00	0.81	0.01	0.01	0.01	0.01	0.01	0.01	0.06	0.01	0.01	0.00	0.18	0.03	0.01	

Marked values are in light shaded (>0.10), dark shaded (>0.30) and black (>0.50). S_k measures sensitivity of output to single factor X_k, S_{Tk} measures total effect (including interactions with otherfactors) of X_k.

the interaction between the input factors falls beyond the scope of the current study; we refer the interested reader to Saltelli *et al* (2000). We confine ourselves to indicating those additional properties of model-free variance-based techniques that render them suitable for the present analysis:

- they allow an exploration of the whole range of variation of the input factors, instead of just sampling factors over a limited number of values, as done for example in fractional factorial design (Box *et al*, 1978);
- they are quantitative, and can distinguish main effects (first order) from interaction effects (second and higher order);
- they are easy to interpret and to explain.

The starting point of such analysis is the variance decomposition of a model output Y: $V(Y) = V(E(Y|X_k)) + E(V(Y|X_k))$, where X_k is an uncertain input factor ($k = 1, \ldots, K$). (Note that $V(E(Y|X_k)) \equiv V_{X_k}(E_{\mathbf{X}_{-k}}(Y|X_k))$ where $\mathbf{X}_{-k}$ is the vector of all–but–k factors.) First-order sensitivity measures can be calculated as $S_k = V(E(Y|X_k))/V(Y)$ for each uncertain factor; the higher the S_k, the higher the sensitivity of the output Y to factor X_k. In the case of an additive (hence, linear) model, where no interactions between its uncertain factors occur, $\sum_{k=1}^{K} S_k = 1$. However, as several layers of uncertainty are often simultaneously activated, CIs should be regarded as non-linear, possibly non-additive models due to interactions between the uncertain input factors (Saisana *et al*, 2005). As argued by practitioners (Saltelli *et al*, 2000; EPA, 2004), robust, 'model-free' techniques for sensitivity analysis should be used for non-linear models.

For non-additive models, higher order sensitivity measures that capture interaction effects among sets of input factors could be computed to improve insight of the model structure. However, higher order measures are usually not estimated, as in a model with K factors the total number of sensitivity measures (including the first-order) that should be estimated is as high as $2^K - 1$. We will therefore confine ourselves to using a more compact sensitivity measure. This is the *total-effect measure* that concentrates in a single term all the interactions involving a given factor X_k (Homma and Saltelli, 1996). We use S_{Tk} to denote the total-effect sensitivity measure for the given factor X_k.

The extended variance-based methods that we used for this study are based on the work of Saltelli (2002) and are implemented in the freely distributed software SIMLAB (Saltelli *et al*, 2004). The pair (S_k, S_{Tk}) gives a fairly good description of the DEA model sensitivities at a reasonable cost, which amounts to $N(2K + 2)$ model evaluations. In our analysis, the base sample is of size $N = 512$ and, hence, evaluating the CI score for each country involves $512 \times (2 \times 23 + 2) = 24\,576$ DEA model runs. When using S_k for sensitivity analysis, one looks for important input factors that—if fixed singularly—would reduce the most the variance of the output variable (in our case a country's CI score). There exists no established threshold to determine when an input factor is important or not. In practice (Saisana *et al*, 2005), one usually considers that an input factor is important if it explains more than 10% of the variance of the output ($S_k > 0.10$). Furthermore, the greater the value of the difference $S_{Tk} - S_k$, the more that factor is involved in interactions with other factors.

The sensitivity measures S_k and S_{Tk} for all countries' scores are given in Table 5. We make a few notable remarks. About 70% of Slovenia's variance—recall that Slovenia was one of the two countries with large uncertainty bounds—is driven by a single factor, that is the set of weights provided by Individual #3. In total, all factors taken singularly explain 82.6% of the variance in Slovenia's score (see rightmost column of Table 5, upper part). The remaining 17.4% of the variance is, hence, due to interactions among the input factors. Similarly, the variance in the score of Korea is mostly attributed due to the inclusion in the panel of Expert 15 (73% variance explained).

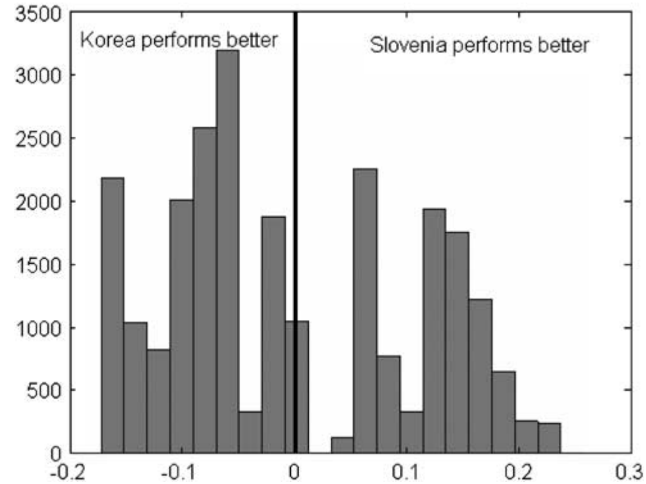

Figure 6 Uncertainty analysis for the DEA-based Technology Achievement Index for Slovenia *versus* Korea.

For the entire set of countries, five of the 21 individuals (#1, 3, 12, 14, and 20) have a remarkable impact on the variance of the countries' scores. Had the results been more robust to the panel opinions, a policy maker would have more confidence on the countries' relative performance in technological achievement. Given that in our case the results are sensitivity to the opinion of five individuals, an iterative process, for example a Delphi method, could be undertaken to assist in the convergence of the opinions of those five experts to the remaining opinions of the panel, or to replace them with other individuals. The assumption on the type of the pie shares constraint (X_{22}) is influential for very few countries: Japan, Netherlands, Germany, Norway, Ireland and France. Finally, the data transformation which consists in considering the logarithms for the *telephone* and *electricity* indicators is not influential to any country's variance. This immediately reveals that it would be futile to discuss the use of scale transformations for those two indicators, if the DEA model was selected as the proper aggregation method.

Sensitivity analysis may also be useful in cases where partial overlapping between two countries occur. As an illustration, we will look at the difference in the technological achievement levels between Slovenia (median score = 0.443) and Korea (0.357), which, as shown previously, present significant overlap in their scores. Figure 6 shows the pdf of the difference in the two countries' relative performance, while explicitly acknowledging the uncertainties inherent in our DEA approach. Note that, although Slovenia has a higher median score of technological performance than Korea, 61% of the differences fall in the left-hand region, where Korea outperforms Slovenia. An obvious issue for CI practitioners is, hence, which factors are mostly responsible for this ambiguity? The results of the sensitivity analysis for the difference in the two countries' scores are given in Table 6. Taken singularly, the input factors account for 91.3% of the

Table 6 Sensitivity measures of first-order and total effect for the difference between the composite indicator scores of Slovenia and Korea

	S_k	S_{Tk}
X_1	0.000	0.018
X_2	0.002	0.006
X_3	**0.163**	**0.209**
X_4	0.000	0.007
X_5	0.001	0.004
X_6	0.003	0.002
X_7	0.000	0.015
X_8	0.001	0.003
X_9	0.000	0.006
X_{10}	0.002	0.015
X_{11}	0.002	0.004
X_{12}	0.000	0.008
X_{13}	0.003	0.001
X_{14}	0.000	0.021
X_{15}	**0.705**	**0.731**
X_{16}	0.003	0.007
X_{17}	0.000	0.001
X_{18}	0.000	0.005
X_{19}	0.000	0.000
X_{20}	0.019	0.047
X_{21}	0.000	0.001
X_{22}	0.000	0.015
X_{23}	0.008	0.008
Sum	0.913	

variance in the difference between the two countries. Most of the variance is due to the consideration of the set of weights by Individual #15 (70.5%) and by Individual #3 (16.3%). The remaining small portion of the output variance, that is 8.7%, is explained by the interactions among the factors.

This example also shows that the analysis of country differences may provide additional insight, relative to the sensitivity analysis of individual country scores. Indeed, we inferred from Table 5 that individual #15 was the main driver of Korea's score variance, and individual #3 for Slovenia. Yet, this does not reveal the degree by which these two experts determine the difference in scores between the two countries. However, given the results presented in Table 6, one sees that individual #15 is the main driver of the preference of Korea over Slovenia.

5. Conclusion

Media and policy-makers look with increasing interest at CIs as appealing tools to attract the attention of the community, build narratives and help focus policy debates. Methodological gaps or fragilities in their design and construction may invite politicians to draw simplistic conclusions or the press to communicate misleading information. That is why national and international organizations believe that it is important to focus on methodological issues in the design of CIs.

Here, we have illustrated a generalization of the DEA-model for the selection of weights using expert panel information and combined it with a variance-based sensitivity analysis. In addition, we have tested it on a practical case study from the CI field, where rarely robustness and sensitivity analysis are applied, despite the fact that many potential sources of uncertainty are present in the CI construction.

The focus of our sensitivity analysis was on quantifying the extent to which eventual results are influenced by expert opinion on the relative importance of the sub-indicators (and other 'modelling options'). This is not a coincidence, as this source of uncertainty often triggers most of the debate on the CI's ultimate intrinsic value. In our specific TAI-case, the results showed that five experts drive most of the variance in the countries' scores. In case the results are sensitive to certain experts' opinion, then CI practitioners could consider the nationality of the experts involved to avoid some kind of "national bias" in the scoring system. Unfortunately, we did not have this type of information. More generally, the sensitivity analysis results could guide the subsequent steps of the evaluation/construction process. A Delphi approach combined with sensitivity analysis could be instrumental in assisting to reach convergence between those panel members whose opinion has the highest impact on the results and the remaining panel members.

At the same time, it is important to stress again that a desirable feature of the DEA-approach to building CIs is precisely the flexibility of experts' opinions and the fact that no *unique* set of weights is sought. In fact, disagreement about such a set is the rule rather than the exception. A DEA-based model such as the one we presented here is particularly apt to deal with this challenging reality.

Acknowledgements— We thank two anonymous referees for insightful comments and suggestions. This paper is an offshoot of the KEI-project (contract no 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/). Laurens Cherchye thanks the Fund for Scientific Research-Flanders (FWO-Vlaanderen) for his postdoctoral fellowship.

References

Atkinson A, Cantillon B, Marlier E and Nolan B (2002). *Social Indicators. The EU and Social Inclusion*. Oxford University Press: Oxford.

Banker RD, Chang H and Cooper WW (1996). Simulation studies of efficiency, returns to scale and misspecification with nonlinear functions in DEA. *Ann Opns Res* **66**: 233–253.

Box G, Hunter W and Hunter J (1978). *Statistics for Experimenters*. John Wiley: New York.

Cazals C, Florens JP and Simar L (2002). Nonparametric frontier estimation: A robust approach. *J Econ* **106**: 1–25.

Charnes A and Cooper WW (1985). Preface to topics in data envelopment analysis. *Ann Opns Res* **2**: 59–94.

Charnes A, Cooper WW and Rhodes E (1978). Measuring the efficiency of decision making units. *Eur J Opl Res* **2**: 429–444.

Cherchye L and Kuosmanen T (2006). Benchmarking Sustainable Development: A Synthetic Meta-Index Approach. In: McGillivray

M and Clarke M (eds). *Understanding Human Well-being*. United Nations University Press: Tokyo, pp 139–168.

Cherchye L, Moesen W and Van Puyenbroeck T (2004). Legitimately diverse yet comparable: On synthesizing social inclusion performance in the EU. *J Common Market Stud* **42**: 919–955.

Cherchye L, Knox Lowell CA, Moesen W and Van Puyenbroeck T (2007). One market, one number? A composite indicator assessment of EU internal market dynamics. *Eur Econ Rev* **51**: 749–779.

Conover WJ (1980). *Practical Nonparametric Statistics*. John Wiley: New York.

Cooper WW, Seiford L and Tone K (2000). *Data Envelopment Analysis*. Kluwer Academic Publishers: Boston.

Cooper WW, Shanling L, Seiford LM and Zhu J (2004). Sensitivity analysis in DEA. In: Cooper WW, Seiford LM and Zhu J (eds). *Handbook on Data Envelopment Analysis*. Kluwer Academic Publishers: Boston, pp 75–97.

Desai M, Fukuda-Parr S, Johansson C and Sagasti F (2002). Measuring the technology achievement of nations and the capacity to participate in the network age. *J Human Dev* **3**: 95–122.

Despotis DK (2005). A reassessment of the human development index via data envelopment analysis. *J Opl Res Soc* **56**: 969–980.

EPA (2004). Council for Regulatory Environmental Modeling, 'Draft Guidance on the Development, Evaluation, and Application of Regulatory Environmental Models', http://www.epa.gov/osp/crem/library/CREM%20Guidance%20Draft%2012_03.pdf.

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

European Commission (2005). European innovation scoreboard 2005—Comparative analysis of innovation performance, http://trendchart.cordis.lu/scoreboards/scoreboard2005/pdf/EIS%202005.pdf.

Freudenberg M (2003). *Composite indicators of country performance: A critical assessment*, STI Working Paper 2003/16, OECD, Paris.

Homma T and Saltelli A (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliab Eng Syst Safety* **52**: 1–17.

Kao C and Tung HT (2005). Data envelopment analysis with common weights: The compromise solution approach. *J Opl Res Soc* **56**: 1196–1203.

Mahlberg B and Obersteiner M (2001). *Remeasuring the HDI by data envelopment analysis*. IIASA interim report IR-01-069, Luxemburg.

Munda G and Nardo M (2003). *On the methodological foundations of composite indicators used for ranking countries*. Mimeo, Universitat Autonoma de Barcelona.

Nardo M, Saisana M, Saltelli A and Tarantola S (2005a). *Tools for composite indicators building*. EUR 21682 EN, European Commission-Joint Research Centre.

Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A and Giovannini E (2005b). *Handbook on constructing composite indicators: methodology and user's guide*, OECD Statistics Working Paper, JT00188147.

Saisana M, Saltelli A and Tarantola S (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J Roy Stat Soc* **168**: 307–323.

Saltelli A (2002). Making best use of model evaluations to compute sensitivity indices. *Comput Phys Commu* **145**: 280–297.

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

Saltelli A, Tarantola S, Campolongo F and Ratto M (2004). *Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models*. Wiley: England.

Simar L (2003). Detecting outliers in frontier models: A simple approach. *J Prod Anal* **20**: 391–424.

Simar L and Wilson P (1998). Sensitivity of efficiency scores: How to bootstrap in Nonparametric frontier models. *Mngt Sci* **44**: 49–61.

Sobol IM (1967). On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput Math Phys* **7**: 86–112.

Storrie D and Bjurek H (2000). *Benchmarking European labour market performance with efficiency frontier techniques*, CELMS Discussion paper, Göteborg University.

Thanassoulis E, Portela MC and Allen R (2004). Incorporating value judgments in DEA. In: Cooper WW, Seiford LM and Zhu J (eds). *Handbook on Data Envelopment Analysis*. Kluwer Academic Publishers: Dordrecht, pp 99–138.

United Nations (2001). Annex 2.1: The technology achievement index: A new measure of countries' ability to participate in the network age. *Human Development Report*, Oxford University Press: United Kingdom, http://www.undp.org.

Valdmanis V (1992). Sensitivity analysis for DEA models: An empirical example using public vs. NFP hospitals. *J Public Econ* **48**: 185–205.

Wilson PW (1995). Detecting influential observations in data envelopment analysis. *J Prod Anal* **6**: 27–46.

Wong Y-HB and Beasley JE (1990). Restricting weight flexibility in data envelopment analysis. *J Opl Res Soc* **47**: 136–150.

Received January 2006;
accepted April 2007 after three revisions