

OECD Statistics Working Papers 2005/03

Handbook on Constructing Composite Indicators: Methodology and User Guide

Michela Nardo,
Michaela Saisana,
Andrea Saltelli,
Stefano Tarantola,
Anders Hoffmann,
Enrico Giovannini

https://dx.doi.org/10.1787/533411815016

OECD

Unclassified **STD/DOC(2005)3**

Organisation de Coopération et de Développement Economiques
Organisation for Economic Co-operation and Development **09-Aug-2005**

English - Or. English

STATISTICS DIRECTORATE

STD/DOC(2005)3
Unclassified

HANDBOOK ON CONSTRUCTING COMPOSITE INDICATORS: METHODOLOGY AND USER GUIDE

OECD Statistics Working Paper

by Michela Nardo, Michaela Saisana, Andrea Saltelli and Stefano Tarantola (EC/JRC)
Anders Hoffman and Enrico Giovannini (OECD)

JT00188147

Document complet disponible sur OLIS dans son format d'origine
Complete document available on OLIS in its original format

English - Or. English

OECD STATISTICS WORKING PAPER SERIES

The OECD Statistics Working Paper Series - managed by the OECD Statistics Directorate - is designed to make available in a timely fashion and to a wider readership selected studies prepared by staff in the Secretariat or by outside consultants working on OECD projects. The papers included are of a technical, methodological or statistical policy nature and relate to statistical work relevant to the organisation. The Working Papers are generally available only in their original language - English or French - with a summary in the other.

Comments on the papers are welcome and should be communicated to the authors or to the OECD Statistics Directorate, 2 rue André Pascal, 75775 Paris Cedex 16, France.

The opinions expressed in these papers are the sole responsibility of the authors and do not necessarily reflect those of the OECD or of the governments of its Member countries.

---

http://www.oecd.org/std/research

ABSTRACT

This Handbook aims to provide a guide for constructing and using composite indicators for policy makers, academics, the media and other interested parties. While there are several types of composite indicators, this Handbook is concerned with those which compare and rank country performance in areas such as industrial competitiveness, sustainable development, globalisation and innovation. The Handbook aims to contribute to a better understanding of the complexity of composite indicators and to an improvement of the techniques currently used to build them. In particular, it contains a set of technical guidelines that can help constructors of composite indicators to improve the quality of their outputs.

It has been prepared jointly by the OECD (the Statistics Directorate and the Directorate for Science, Technology and Industry) and the Applied Statistics and Econometrics Unit of the Joint Research Centre of the European Commission in Ispra, Italy. Primary authors from the JRC are Michela Nardo, Michaela Saisana, Andrea Saltelli and Stefano Tarantola. Primary authors from the OECD are Anders Hoffmann and Enrico Giovannini. Editorial assistance was provided by Candice Stevens, Gunseli Baygan and Karsten Olsen.

The research is partly funded by the European Commission, Research Directorate, under the project KEI (Knowledge Economy Indicators), Contract FP6 No. 502529. In the OECD context, the work has benefitted from a grant from the Danish government. The views expressed are those of the authors and should not be regarded as stating an official position of either the European Commission or the OECD.

Ce Manuel a pour objectif de procurer aux responsables politiques, universitaires, médias et autres parties concernées un guide sur la façon d'élaborer et d'utiliser des indicateurs composites. Si il existe plusieurs types d'indicateurs composites, ce Manuel intéresse ceux qui comparent et classent la performance d'un pays dans des domaines comme la compétitivité industrielle, le développement durable, la mondialisation et les innovations. Le Manuel a pour objectif de contribuer à une meilleure compréhension de la complexité des indicateurs composites et à une amélioration des techniques actuellement utilisées pour les élaborer. En particulier, il contient une série de lignes directrices techniques qui peuvent aider les concepteurs d'indicateurs composites à améliorer la qualité de leurs productions.

Il a été conjointement préparé par l'OCDE (la Direction des statistiques et la Direction de la science, de la technologie et de l'industrie) et la cellule des Statistiques appliquées et de l'économétrie du Centre commun de recherche (CCR) de la Commission européenne à Ispra en Italie. Les auteurs originaux du CCR sont Michela Nardo, Michaela Saisana, Andrea Saltelli et Stefano Tarantola. Les auteurs originaux de l'OCDE sont Anders Hoffmann et Enrico Giovannini. L'assistance éditoriale a été assurée par Candice Stevens, Gunseli Baygan et Karsten Olsen.

Les recherches sont partiellement financées par la Direction des recherches de la Commission européenne, pour le projet KEI (Knowledge Economy Indicators), Contrat FP6 no. 502529. Pour ce qui est de l'OCDE, le travail a bénéficié d'une subvention du gouvernement danois. Les points de vue exprimés sont ceux des auteurs et ils ne doivent pas être considérés comme l'expression d'une position officielle de la Commission européenne ou de l'OCDE.

TABLE OF CONTENTS

INTRODUCTION 8

I. CONSTRUCTING A COMPOSITE INDICATOR 12

Step 1. Developing a theoretical framework 12
Step 2. Selecting variables 13
Step 3. Multivariate analysis 14
Step 4. Imputation of missing data 16
Step 5. Normalisation of data 17
Step 6. Weighting and aggregation 21
Step 7. Robustness and sensitivity 23
Step 8. Links to other variables 24
Step 9. Back to the details 26
Step 10. Presentation and dissemination 28

II. QUALITY FRAMEWORK FOR COMPOSITE INDICATORS 31

III. TOOLBOX FOR CONSTRUCTORS 37

MULTIVARIATE ANALYSIS 37

Principal components analysis (PCA) 37
Factor analysis 43
Cronbach Coefficient Alpha 45
Cluster analysis 46

IMPUTATION OF MISSING DATA 52

Single imputation 52
Unconditional mean imputation 53
Regression imputation 53
Expected maximisation imputation 54
Multiple imputation 55

NORMALISATION 59

Scale transformation prior to normalisation 59
Standardisation (or z-scores) 60
Re-scaling 61
Distance to a reference 61
Indicators above or below the mean 62
Methods for cyclical indicators 62
Percentage of annual differences over consecutive years 62

WEIGHTING AND AGGREGATION 64
Weights based on statistical models 64
Data envelopment analysis (DEA) 66
Benefit of the doubt approach (BOD) 67
Unobserved components model (UCM) 69
Budget allocation (BAL) 70
Public opinion 70
Analytic hierarchy process (AHP) 70
Conjoint analysis (CA) 72
Performance of the different weighting methods 73
Additive aggregation methods 75
Non-compensatory multicriteria approach (MCA) 76
Geometric aggregation 79
UNCERTAINTY AND SENSITIVITY ANALYSIS 81
General framework 82
Uncertainty analysis (UA) 82
Sensitivity analysis using variance-based techniques 85
REFERENCES 94
APPENDIX: TECHNOLOGY ACHIEVEMENT INDEX 103
ENDNOTES 106

LIST OF TABLES

Table 1. Strength and weaknesses of multivariate analysis
Table 2. Normalisation methods
Table 3. Compatibility between aggregation and weighting methods
Table 4. Quality dimensions of composite indicators
Table 5. Correlation matrix for TAI sub-indicators
Table 6. Eigenvalues of TAI sub-indicators
Table 7. Component loadings for TAI sub-indicators
Table 8. Rotated factor loadings for TAI sub-indicators (method 1)
Table 9. Rotated factor loadings for TAI sub-indicators (method 2)
Table 10. Cronbach coefficient alpha results for TAI sub-indicators
Table 11. Distance measures for TAI sub-indicators
Table 12. K-means clustering for TAI countries
Table 13. Normalisation based on interval scales
Table 14. Examples of normalisation techniques using TAI data
Table 15. Eigenvalues of TAI dataset
Table 16. Factor loadings of TAI based on principal components
Table 17. Factor loadings of TAI based on maximum likelihood
Table 18. Data envelopment analysis (DEA) performance frontier
Table 19. Benefit of the doubt (BOD) approach applied to TAI
Table 20. Comparison matrix of eight TAI sub-indicators
Table 21. Comparison matrix of three TAI sub-indicators
Table 22. TAI weights based on different methods
Table 23. TAI country rankings based on different weighting methods
Table 24. Advantages and disadvantages of different weighting methods
Table 25. Impact matrix for TAI (five countries)
Table 26. Outranking impact matrix for TAI (five countries)
Table 27. Permutations obtained from the outranking matrix for TAI and associated score
Table 28. TAI country rankings by different aggregation methods
Table 29. Sobol' sensitivity measures of first order and total effect on TAI results
Table 30. Sobol' sensitivity measures and average shift in TAI rankings

LIST OF FIGURES

Figure 1. Link between TAI and GDP per capita, 2000
Figure 2. Example of bar chart decomposition presentation
Figure 3. Example of leader/laggard decomposition presentation
Figure 4. Example of spider diagram decomposition presentation
Figure 5. Example of traffic light decomposition presentation
Figure 6. Example of tabular presentation of composite indicator
Figure 7. Example of bar chart presentation of composite indicator
Figure 8. Example of line chart presentation of composite indicator
Figure 9. Example of trend diagram of composite indicator
Figure 10. Eigenvalues for TAI sub-indicators
Figure 11. Country clusters for TAI sub-indicators
Figure 12. Linkage distance vs fusion step in TAI hierarchical cluster
Figure 13. Means plot for TAI clusters
Figure 14. Logic of multiple imputation
Figure 15. Markov Chain Monte Carlo imputation method
Figure 16. Data envelopment analysis (DEA) performance frontier
Figure 17. Analytic hierarchy process (AHP) weighting of TAI
Figure 18. Uncertainty analysis of TAI country rankings
Figure 19. Sobol' sensitivity measures of first orderTAI results
Figure 20. Sobol' sensitivity measures of TAI total effect indices
Figure 21. Netherlands ranking by aggregation and weighting systems
Figure 22. Uncertainty analysis for TAI output variable
Figure 23. Average shift in TAI country rankings by aggregation and weighting combinations
Figure 24. Uncertainty analysis of TAI country rankings

INTRODUCTION

Composite indicators (CIs) which compare country performance are increasingly recognised as a useful tool in policy analysis and public communication. They provide simple comparisons of countries that can be used to illustrate complex and sometimes elusive issues in wide ranging fields, *e.g.*, environment, economy, society or technological development. These indicators often seem easier to interpret by the general public than finding a common trend in many separate indicators and have proven useful in benchmarking country performance. However, composite indicators can send misleading policy messages if they are poorly constructed or misinterpreted. Their "big picture" results may invite users (especially policy makers) to draw simplistic analytical or policy conclusions. Instead, composite indicators must be seen as a starting point for initiating discussion and attracting public interest. Their relevance should be gauged with respect to constituencies affected by the composite index.

Pros and cons of composite indicators

In general terms, an indicator is a quantitative or a qualitative measure derived from a series of observed facts that can reveal relative positions (*e.g.*, of a country) in a given area. When evaluated at regular intervals, an indicator can point out the direction of change across different units and through time. In the context of policy analysis, indicators are useful in identifying trends and drawing attention to particular issues. They can also be helpful in setting policy priorities and in benchmarking or monitoring performance. A composite indicator is formed when individual indicators are compiled into a single index on the basis of an underlying model. The composite indicator should ideally measure multi-dimensional concepts which cannot be captured by a single indicator alone, *e.g.*, competitiveness, industrialisation, sustainability, single market integration, knowledge-based society, etc. The main pros and cons of using composite indicators are the following (**Box 1**) (Saisana and Tarantola, 2002):

Box 1. Pros and Cons of Composite Indicators

Pros	Cons
• Can summarise complex or multi-dimensional issues in view of supporting decision-makers. • Easier to interpret than trying to find a trend in many separate indicators. • Facilitate the task of ranking countries on complex issues in a benchmarking exercise. • Can assess progress of countries over time on complex issues. • Reduce the size of a set of indicators or include more information within the existing size limit. • Place issues of country performance and progress at the centre of the policy arena. • Facilitate communication with general public (i.e. citizens, media, etc.) and promote accountability.	• May send misleading policy messages if they are poorly constructed or misinterpreted. • May invite simplistic policy conclusions. • May be misused, e.g., to support a desired policy, if the construction process is not transparent and lacks sound statistical or conceptual principles. • The selection of indicators and weights could be the target of political challenge. • May disguise serious failings in some dimensions and increase the difficulty of identifying proper remedial action. • May lead to inappropriate policies if dimensions of performance that are difficult to measure are ignored.

Composite indicators are much like mathematical or computational models. As such, their construction owes more to the craftsmanship of the modeller than to universally accepted scientific rules for encoding. As for models, the justification for a composite indicator lays in its fitness to the intended purpose and the acceptance of peers acceptance (Rosen, 1991). On the dispute whether composite indicators are good or bad as such , it has been noted:

"The aggregators believe there are two major reasons that there is value in combining indicators in some manner to produce a bottom line. They believe that such a summary statistic can indeed capture reality and is meaningful, and that stressing the bottom line is extremely useful in garnering media interest and hence the attention of policy makers. The second school, the non-aggregators, believe one should stop once an appropriate set of indicators has been created and not go the further step of producing a composite index. Their key objection to aggregation is what they see as the arbitrary nature of the weighting process by which the variables are combined." (Sharpe, 2004)

According to other commentators:

"[…] it is hard to imagine that debate on the use of composite indicators will ever be settled […] official statisticians may tend to resent composite indicators, whereby a lot of work in data collection and editing is "wasted" or "hidden" behind a single number of dubious significance. On the other hand, the temptation of stakeholders and practitioners to summarise complex and sometime elusive processes (e.g. sustainability, single market policy, etc.) into a single figure to benchmark country performance for policy consumption seems likewise irresistible." (Saisana *et al*., 2005)

Aim of the Handbook

This Handbook does not aim to resolve this debate, but only to contribute to a better understanding of the complexity of composite indicators and to an improvement of the techniques currently used to build composite indicators. In particular, it contains a set of technical guidelines that can help constructors of composite indicators to improve the quality of their outputs.

The proposal to develop a Handbook was launched at the end of a workshop on composite indicators jointly organised by the JRC and OECD in Spring 2003 which demonstrated:

- the growing interest in composite indicators from academic circles, media and policy makers;
- the existence of a wide range of methodological approaches to composite indicators; and
- the need, clearly expressed by the participants in the workshop, to have international guidelines in this domain.

Therefore, the JRC and OECD launched a project, open to other institutions, to develop the present Handbook. Key elements of the Handbook were then presented during a second workshop, held in Paris in February 2004, while the aims and the outline of the Handbook were presented at the OECD Committee on Statistics in June 2004.

The main aim of the Handbook is to provide builders of composite indicators with a set of recommendations on how to design, develop and disseminate a composite indicator. In fact, methodological issues need to be addressed transparently prior to the construction and use of composite indicators to avoid data manipulation and misrepresentation. In particular, to guide constructors and users, highlighting the technical problems and common pitfalls to be avoided, the first part of the Handbook discusses the following steps in the construction of composite indicators:

- *Theoretical framework* - A theoretical framework should be developed to provide the basis for the selection and combination of single indicators into a meaningful composite indicator under a fitness-for-purpose principle.

- *Data selection* - Indicators should be selected on the basis of their analytical soundness, measurability, country coverage, relevance to the phenomenon being measured and relationship to each other. The use of proxy variables should be considered when data are scarce.

- *Multivariate analysis* – An exploratory analysis should investigate the overall structure of the indicators, assess the suitability of the data set and explain the methodological choices, e.g., weighting, aggregation.

- *Imputation of missing data* - Consideration should be given to different approaches for imputing missing values. Extreme values should be examined as they can become unintended benchmarks.

- *Normalisation* - Indicators should be normalised to render them comparable.

- *Weighting and aggregation* – Indicators should be aggregated and weighted according to the underlying theoretical framework.

- *Robustness and sensitivity* – Analysis should be undertaken to assess the robustness of the composite indicator in terms of e.g., the mechanism for including or excluding single indicators, the normalisation scheme, the imputation of missing data and the choice of weights.

- *Links to other variables* – Attempts should be made to correlate the composite indicator with other published indicators as well as to identify linkages through regressions.

- *Visualisation* – Composite indicators can be visualised or presented in a number of different ways, which can influence their interpretation.

- *Back to the real data* - Composite indicators should be transparent and be able to be decomposed into their underlying indicators or values.

The second part of the Handbook presents a quality framework for composite indicators, where the relationships between methodologies used to construct and disseminate composite indicators and different quality dimensions are discussed. Finally, in the third part, methodologies to be used in the various steps are presented and discussed in more detail in the *Toolbox for Constructors*.

In order to help the reader in better understanding the content of the Handbook, a concrete example (the Technology Achievement Index - TAI) is used to explain the various steps in the construction of a composite indicator and highlight problems that may arise (Box 2). The TAI is a composite indicator developed by the United Nations for the Human Development Report (UN, 2001). It is composed of a relatively small number of sub-indicators, which renders it suitable for the didactic purposes of this Handbook. Moreover, the TAI is well documented by its developers and the underlying data are freely available on the Internet. For explanatory purposes, only the first 23 of the 72 original countries measured by the TAI are considered here. Further details are given in the Appendix.

The following notations are adopted throughout the Handbook:

$x_{q,c}^{t}$: raw value of sub-indicator q for country c at time t, with $q=1,...,Q$ and $c=1,...,M$

$I_{q,c}^{t}$: normalised value of sub-indicator

$w_{r,q}$: weight associated to sub-indicator q, with $r=1,...,R$ denoting the weighting method

CI_c^t: value of the composite indicator for country *c* at time *t*.

For reasons of clarity, the time suffix has been normally omitted and is present only in certain sections. When no time indication is present, the reader should consider that all variables have the same time dimension.

Box 2. Case study: Technology Achievement Index (TAI)

The TAI focuses on four dimensions of technological capacity (Table A.1**)**:

- **Creation of technology**. Two sub-indicators are used to capture the level of innovation in a society: 1) the number of patents granted per capita (to reflect the current level of innovative activities) and 2) receipts from royalty and license fees from abroad per capita (to reflect the stock of successful innovations that are still useful and hence have market value).
- **Diffusion of recent innovations**. Diffusion is measured by two sub-indicators: 1) diffusion of the Internet (indispensable to participation) and 2) exports of high-and medium-technology products as a share of all exports.
- **Diffusion of old innovations**. Two sub-indicators are included: telephones and electricity. These are needed to use newer technologies and have wide-ranging applications. Both indicators are expressed as logarithms, as they 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 technology achievement.
- **Human skills.** Two sub-indicators are used to reflect the human skills needed to create and absorb innovations: 1) mean years of schooling and 2) gross enrolment ratio of tertiary students enrolled in science, mathematics and engineering.

Limits of the Handbook

The literature on composite indicators is huge and almost every month new proposals on specific methodological aspects potentially relevant for the development of composite indicators are published. In this Handbook, taking into account its potential audience, we have preferred to make reference to relatively well established methodologies and procedures, avoiding the inclusion of some interesting, but still experimental, approaches. However, the Handbook should be seen as a "live" product, with successive editions, as long as new developments take place. On the other hand, this first version of the Handbook does not cover "composite leading indicators" normally used to identify cyclical movements of the economic activity. Although the OECD has a longstanding tradition and experience in this field, we have preferred to exclude them, because they are based on statistical and econometric approaches quite different from those relevant for other types of composite indicators.

I. CONSTRUCTING A COMPOSITE INDICATOR

Introduction

As already explained, the Handbook presents its recommendations following an "ideal sequence" of ten steps, from the development of a theoretical framework to the analysis of detailed data, once the indicator is built. Each step is extremely important, but the coherence of the whole process is equally important. Choices made in one step can have important implications for other steps: therefore, the composite indicator builder has not only to make the most appropriate methodological choices in each step, but also to identify if they fit well together.

Composite indicators builders have to face a relevant degree of scepticism among statisticians, economists and other groups of users. This scepticism is partially due to the lack of transparency of some existing indicators, especially as far as methodologies and basic data are concerned. To avoid these risks, the Handbook puts special emphasis on documentation and metadata. In particular, the Handbook recommends the preparation of relevant documentation at the end of each phase, both to ensure the coherence of the whole process and to prepare in advance the methodological notes that will be disseminated together with the numeric results.

This chapter provides an overview of individual steps in the construction of composite indicators. More detailed information about tools to be used in each phase is presented in Chapter III.

Step 1. Developing a theoretical framework

What is badly defined is likely to be badly measured …

A sound theoretical framework is the starting point in constructing composite indicators. The framework should clearly define the phenomenon to be measured and its sub-components and select individual indicators and weights that reflect their relative importance and the dimensions of the overall composite. Ideally, this process would be based on what is desirable to measure and not which indicators are available.

For example, gross domestic product (GDP) measures the total value of goods and services produced in a given country, where the weights are estimated based on economic theory and reflect the relative price of goods and services. The theoretical and statistical frameworks to measure GDP have been developed over the last 50 years and a revision of the 1993 System of National Accounts is currently being undertaken by the major international organisations. However, not all multi-dimensional concepts have such solid theoretical and empirical underpinnings. Composite indicators in newly emerging policy areas, *e.g.*, competitiveness, sustainable development, e-business readiness, etc., could be very subjective as the economic research in these fields is still being developed. Transparency thus is essential in constructing credible indicators. This requires:

- Defining the concept. The definition should give the reader a clear sense of what is being measured by the composite indicator. It should refer to the theoretical framework, linking various sub-groups and the underlying indicators. For example, the Growth Competitiveness Index (GCI) developed by the World Economic Forum is founded on the idea "that the process of economic growth can be analysed within three important broad categories: the macroeconomic environment, the quality of public institutions, and technology." Some complex concepts,

however, are difficult to define and measure precisely or could be subject to controversy among stakeholders. Ultimately, the users of composite indicators should assess their quality and relevance.

- Determining sub-groups. Multi-dimensional concepts can be divided into several sub-groups. These sub-groups need not be (statistically) independent of each other, and existing linkages should be described theoretically or empirically to the extent possible. The Technology Achievement Index, for example, is conceptually divided into four groups of technological capacity: creation of technology, diffusion of recent innovations, diffusion of old innovations and human skills. Such a nested structure improves the users' understanding of the driving forces behind the composite indicator. It could also make it easier to determine the relative weights across different factors. This step, and also the next one, should involve experts and stakeholders as much as possible so that multiple viewpoints are acknowledged and the conceptual framework and the set of indicators gain in robustness.

- Identifying the selection criteria for the underlying indicators. The selection criteria should work as a guide for whether an indicator should be included or not in the overall composite index. It should be as precise as possible and describe the phenomenon that is being measured, i.e., input, output or process. Too often composite indicators include both input and output measures. For example, an Innovation Index could combine R&D expenditures (inputs) and the number of new products and services (outputs) in order to measure the scope of innovative activity in a given country. However, only the latter set of output indicators should be included (or expressed in terms of output per unit of input) if the index is intended to measure innovation performance.

After Step 1. the constructor should have…

- A clear understanding and definition of the multidimensional phenomenon to be measured.
- A nested structure of the various sub-groups of the phenomenon.
- A list of selection criteria for the underlying variables, *e.g.*, input, output, process.
- Clear documentation of the above.

Step 2. Selecting variables

A composite indicator is above all the sum of its parts…

The strengths and weaknesses of composite indicators largely derive from the quality of the underlying variables. Ideally, variables should be selected on the basis of their relevance, analytical soundness, timeliness, accessibility, etc. Criteria for assuring the quality of the basic data set for composite indicators are discussed in the section of this Handbook on the *Quality Framework for Composite Indicators*. While the choice of indicators must be guided by the theoretical framework for the composite, the data selection process can be quite subjective as there may be no single definitive set of indicators. The lack of relevant data also limits the constructor's ability to build sound composite indicators. Given a scarcity of internationally comparable quantitative (hard) data, composite indicators often include qualitative (soft) data from surveys or policy reviews.

Proxy measures can be used when the desired data is unavailable or when cross-country comparability is limited. For example, data on the number of employees that use computers might not be available. Instead, the number of employees who have access to computers could be used as a proxy. As in the case of soft data, caution must be given to the utilisation of proxy indicators. To the extent data permit, the accuracy of proxy measures should be checked through correlation and sensitivity analysis. The constructor should also pay close attention to whether the indicator in question is dependent on GDP or other size-related factors. To have an objective comparison across small and large countries, scaling of

variables by an appropriate size measure, *e.g.*, population, income, trade volume, and populated land area, etc. is required. Finally, one has to make sure the type of the selected variables -- input, output or process indicators -- match the definition of the intended composite indicator.

The quality and accuracy of composite indicators should evolve in parallel with improvements in data collection and indicator development. The current trend towards constructing composite indicators of country performance in a range of policy areas may provide further impetus to improving data collection, identifying new data sources and enhancing the international comparability of statistics.

After Step 2. the constructor should have…

- Checked the quality of the available indicators.
- Discussed the strengths and weaknesses of each selected indicator.
- Made scale adjustments, if necessary.
- Created a summary table on data characteristics, *e.g.*, availability (across country, time), source, type (hard, soft or input, output, process)

Step 3. Multivariate analysis

Analysing the underlying structure of the data is still an art …

Over the last few decades, there has been an increase in the number of composite indicators developed by various national and international agencies. Unfortunately, individual indicators are sometimes selected in an arbitrary manner with little attention paid to the interrelationships between them. This can lead to indices which overwhelm, confuse and mislead decision-makers and the general public. Some analysts characterise this environment as 'indicator rich but information poor'. The underlying nature of the data needs to be carefully analysed before the construction of a composite indicator. This preliminary step is helpful in assessing the suitability of the data set and will provide an understanding of the implications of the methodological choices, e.g., weighting and aggregation, during the construction phase of the composite indicator. Information can be grouped and analysed along at least two dimensions of the dataset: sub-indicators and countries.

- **Grouping information on sub-indicators**. The analyst must first decide whether the nested structure of the composite indicator is well-defined (see Step 1) and if the set of available sub-indicators is sufficient or appropriate to describe the phenomenon (see Step 2). This decision can be based on expert opinion and the statistical structure of the data set. Different analytical approaches, such as principal components analysis, can be used to explore whether the dimensions of the phenomenon are statistically well-balanced in the composite indicator. If not, a revision of the sub-indicators might be needed.

  The goal of principal components analysis (PCA) is to reveal how different variables change in relation to each other and how they are associated. This is achieved by transforming correlated variables into a new set of uncorrelated variables using a covariance matrix or its standardised form – the correlation matrix. Factor analysis (FA) is similar to PCA, however it is based on a particular statistical model. An alternative way to investigate the degree of correlation among a set of variables is to use the Cronbach coefficient alpha (c-alpha), which is the most common estimate of internal consistency of items in a model or survey. These multivariate analysis techniques are useful for gaining insight into the structure of the data set of the composite. However, it is important to avoid carrying out multivariate analysis if the sample is small compared to the number of indicators since results will not have known statistical properties.

- **Grouping information on countries**. Cluster analysis is another tool for classifying large amounts of information into manageable sets. It has been applied to a wide variety of research problems and fields from medicine to psychiatry and archaeology. Cluster analysis is also used in developing composite indicators to group information on countries based on their similarity on different sub-indicators. Cluster analysis serves as: i) a purely statistical method of aggregation of the indicators, ii) a diagnostic tool for exploring the impact of the methodological choices made during the construction phase of the composite indicator, iii) a method of disseminating information on the composite indicator without losing that on the dimensions of the sub-indicators, and iv) a method for selecting groups of countries to impute missing data with a view to decreasing the variance of the imputed values.

  When the number of variables is large or when is it believed that some of these do not contribute to identify the clustering structure in the data set, continuous and discrete models can be applied sequentially. Researchers frequently carry out a PCA and then apply a clustering algorithm on the object scores on the first few components, called "tandem analysis". However, caution is required as PCA or FA may identify dimensions that do not necessarily contribute to revealing the clustering structure in the data and may mask the taxonomic information (Table 1).

  Various alternative methods combining cluster analysis and the search for a low-dimensional representation have been proposed and focus on multidimensional scaling or unfolding analysis. Factorial k-means analysis combines k-means cluster analysis with aspects of FA and PCA. A discrete clustering model together with a continuous factorial one are fitted simultaneously to two-way data to identify the best partition of the objects, described by the best orthogonal linear combinations of the variables (factors) according to the least-squares criterion. This has a wide range of applications since it reaches a double objective: data reduction and synthesis, simultaneously in the direction of objects and variables. Originally applied to short-term macroeconomic data, factorial k-means analysis has a fast alternating least-squares algorithm that extends its application to large data sets. This methodology can be recommended as an alternative to the widely used tandem analysis.

After Step 3, the constructor should have…

- Checked the underlying structure of the data along various dimensions, *i.e.*, sub-indicators, countries.
- Applied the suitable multivariate methodology, *e.g.*, PCA, FA, cluster analysis.
- Identified sub-groups of indicators or groups of countries that are statistically “similar”.
- Analysed the structure of the data set and compared this to the theoretical framework.
- Documented the results of the multivariate analysis and the interpretation of components and factors.

Table 1. Strength and weaknesses of multivariate analysis

	Strengths	Weaknesses
Principal Components/ /Factor Analysis	Can summarise a set of sub-indicators while preserving the maximum possible proportion of the total variation in the original data set. Largest factor loadings are assigned to the sub-indicators that have the largest variation across countries, a desirable property for cross-country comparisons, as sub-indicators that are similar across countries are of little interest and cannot possibly explain differences in performance.	Correlations do not necessarily represent the real influence of the sub-indicators on the phenomenon being measured. Sensitive to modifications in the basic data: data revisions and updates, *e.g.*, new countries. Sensitive to the presence of outliers, which may introduce a spurious variability in the data. Sensitive to small-sample problems, which are particularly relevant when the focus is on a limited set of countries. Minimisation of the contribution of sub-indicators which do not move with other sub-indicators.
Cronbach Coefficient Alpha	Measures the internal consistency in the set of sub-indicators, *i.e.*, how well they describe a unidimensional construct. Thus it is useful to cluster similar objects.	Correlations do not necessarily represent the real influence of the sub-indicators on the phenomenon expressed by the composite indicator. Meaningful only when the composite indicator is computed as a 'scale' (i.e. as the sum of the sub-indicators).
Cluster Analysis	Offers a different way to group countries; gives some insight into the structure of the data set.	Purely a descriptive tool; may not be transparent if the methodological choices made during the analysis are not motivated and clearly explained.

Step 4. Imputation of missing data

The idea of imputation could be both seductive and dangerous …

Missing data often hinder the development of robust composite indicators. Data can be missing in a random or non-random fashion. The missing patterns could be:

1. *Missing completely at random* (MCAR)**.** Missing values do not depend on the variable of interest or any other observed variable in the data set. For example, the missing values in variable income would be of the MCAR type if (i) people who do not report their income have, on average, the same income as people who do report income; and if (ii) each of the other variables in the dataset would have to be the same, on average, for the people who did not report the income and the people who did report their income.

2. *Missing at random* (MAR)**.** Missing values do not depend on the variable of interest, but they are conditional on other variables in the data set. For example, the missing values in income would be MAR, if the probability of missing data on income depends on marital status but, within each category of marital status, the probability of missing income is unrelated to the value of income. Missing by design, *e.g.*, if survey question 1 is answered yes, then survey question 2 is not to be answered, are also MAR as missingness depends on the covariates.

3. *Not missing at random* (NMAR)**.** Missing values depend on the values themselves. For example, high income households are less likely to report their income.

However, there is no statistical test for NMAR and often no basis to judge whether data are missing at random or systematically, while most of the methods that impute missing values require a missing at

random mechanism, *i.e.*, MCAR or a MAR. When there are reasons to assume a non-random missing pattern (NMAR), the pattern must be explicitly modelled and included in the analysis. This could be very difficult and could imply ad hoc assumptions that are likely to influence the result of the entire exercise.

There are three general methods for dealing with missing data: i) case deletion, ii) single imputation or iii) multiple imputation. The first one, also called complete case analysis, simply omits the missing records from the analysis. However, this approach ignores possible systematic differences between complete and in-complete samples and produces unbiased estimates only if deleted records are a random sub-sample of the original sample (MCAR assumption). Furthermore, standard errors will in general be larger in a reduced sample given that less information is used. As a rule of thumb, if a variable has more than 5% missing values, cases are not deleted (Little and Rubin, 2002).

The other two approaches consider the missing data as part of the analysis and try to impute values through either single imputation, *e.g.*, mean/median/mode substitution, regression imputation, hot- and cold-deck imputation, expectation-maximisation imputation, or multiple imputation, *e.g.*, Markov Chain Monte Carlo algorithm. Data imputation could lead to the minimisation of bias and the use of 'expensive to collect' data that would otherwise be discarded by case deletion. However, it can also allow data to influence the type of imputation. In the words of Dempster and Rubin (1983), "*The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to real and imputed data have substantial bias.* "

The uncertainty in the imputed data should be reflected by variance estimates. This allows taking into account the effects of imputation in the course of the analysis. However, single imputation is known to underestimate the variance, because it reflects partially the imputation uncertainty. The multiple imputation method, which provides several values for each missing value, can more effectively represent the uncertainty due to imputation.

No imputation model is free of assumptions and the imputation results should hence be thoroughly checked for their statistical properties such as distributional characteristics as well as heuristically for their meaningfulness, *e.g.*, whether negative imputed values are possible.

After Step 4, the constructor should have …

- A complete data set without missing values.
- A measure of the reliability of each imputed value so as to explore the impact of imputation on the composite indicator.
- Documented and explained the selected imputation procedure and the results.

Step 5. Normalisation of data

Avoid adding up apples and oranges …

Normalisation is required prior to any data aggregation as the indicators in a data set often have different measurement units. There exist a number of normalisation methods (Table 2) (Freudenberg, 2003; Jacobs et al., 2004):

1. *Ranking* is the simplest normalisation technique. This method is not affected by outliers and allows the performance of countries to be followed over time in terms of relative positions (rankings). Country performance in absolute terms however cannot be evaluated as information on levels are lost. Some examples that use ranking include: the Information and Communications

Technology index (Fagerberg, 2001) and the Medicare study on healthcare performance across the United States (Jencks *et al*., 2003).

2. *Standardisation* (or z-scores) converts indicators to a common scale with a mean of zero and standard deviation of one. Indicators with extreme values thus have a greater effect on the composite indicator. This might be desirable if the intention is to reward exceptional behaviour. That is, if an extremely good result on a few indicators is thought to be better than a lot of average scores. This effect can be corrected in the aggregation methodology, *e.g*., by excluding the best and worst sub-indicator scores from inclusion in the index or by assigning differential weights based on the "desirability" of the sub-indicator scores.

3. *Re-scaling* normalises indicators to have an identical range (0; 1). Extreme values/or outliers however, could distort the transformed indicator. On the other hand, re-scaling could widen the range of indicators lying within a small interval increasing the effect on the composite indicator, more than they would using the z-scores transformation.

4. *Distance to a reference* measures the relative position of a given indicator vis-à-vis a reference point. This could be a target to be reached in a given time frame. For example, the Kyoto Protocol has established an 8% reduction target for CO2 emissions by 2010 for European Union members. The reference could also be an external benchmark country. For example, the United States and Japan are often used as benchmarks for the composite indicators built in the framework of the EU Lisbon agenda. Alternatively, the reference country could be the average country of the group and would be given 1, while other countries receive scores depending on their distance from the average. Hence, standardised indicators that are higher than 1 indicate countries with above-average performance. The reference country could also be the group leader where the leading country receives 1 and the others are given percentage points away from the leader. This approach, however, is based on extreme values which could be unreliable outliers.

5. *Categorical scale* assigns a score for each indicator. Categories can be numerical, such as one, two or three stars, or qualitative, such as 'fully achieved', 'partly achieved' or 'not achieved'. Often, the scores are based on the percentiles of the distribution of the indicator across countries. For example, the top 5% receive a score of 100, the units between the 85th and 95th percentiles receive 80 points, the 65th and the 85th percentiles receive 60 points, all the way to 0 points, rewarding the best performing countries and penalising the worst. Since the same percentile transformation is used for different years, any change in the definition of the indicator over time will not affect the transformed variable. However, it is difficult to follow improvements over time. Categorical scales exclude large amounts of information about the variance of the transformed indicators. Besides, when there is little variation within the original scores, the percentile bands force the categorisation on the data, irrespective of the underlying distribution. A possible solution is to adjust the percentile brackets across the individual indicators in order to obtain transformed categorical variables with almost normal distributions.

6. *Indicators* above or below the mean are transformed such that values around the mean receive 0, whereas the ones above/or below a certain threshold receive 1, and -1 respectively, e.g., the Summary Innovation Index (EC, 2001a). This normalisation method is simple and not affected by outliers. However, the arbitrariness of the threshold level and the omission of absolute level information are usually criticised. For example, if the value of a given indicator for country A is 3 times (300%) above the mean, and the value for country B is 25% above the mean, both countries would be counted as 'above average' with a threshold of 20% around the mean.

7. *Methods for cyclical indicators*. The results of business tendency surveys are usually combined into composite indicators to reduce the risk of false signals, and to better forecast cycles in economic activities (Nilsson, 2000). See, for example, the OECD composite leading indicator, and the EU economic sentiment indicators. The latter is a balance of opinions as managers of firms from different sectors and sizes are asked to express their opinion on their firm's performance. This method gives implicitly less weight to the more irregular series in the cyclical movement of the composite indicator, unless some prior ad-hoc smoothing is performed.

8. *Percentage of annual differences over consecutive years* represents the percentage growth with respect to the previous year instead of the absolute level. The transformation can be used only when the indicators are available for a number of years, e.g., Internal Market Index (EC, 2001).

The selection of a suitable method however, is not trivial and deserves special attention (Ebert and Welsh, 2004). The normalisation method should take into account the data properties, as well as the objectives of the composite indicator. Different normalisation methods will yield different results. Robustness tests might be needed to assess their impact on the outcomes.

After Step 5, the constructor should have …

- Selected the appropriate normalisation procedure(s) with reference to the theoretical framework and to the properties of the data.
- Documented and explained the selected normalisation procedure and the results.

Table 2. Normalisation methods

Method	Equation
Ranking	$I_{qc}^{t} = Rank(x_{qc}^{t})$
Standardisation (or z-scores)	$I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t}}{\sigma_{qc=\bar{c}}^{t}}$
Re-scaling	$I_{qc}^{t} = \frac{x_{qc}^{t} - \min_{c}(x_{q}^{t_0})}{\max_{c}(x_{q}^{t_0}) - \min_{c}(x_{q}^{t_0})}.$
Distance to a reference country	$I_{qc}^{t} = \frac{x_{qc}^{t}}{x_{qc=\bar{c}}^{t_0}}$ or $I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t_0}}{x_{qc=\bar{c}}^{t_0}}$
Categorical scales	$I_{qc}^{t} = \begin{cases} 25 & \text{if } x_{qc}^{t} \in \{p^{25th}\}\text{percentile} \\ 50 & \text{if } x_{qc}^{t} \in \{p^{50th} - p^{25th}\}\text{percentile} \\ 75 & \text{if } x_{qc}^{t} \in \{p^{75th} - p^{50th}\}\text{percentile} \\ 100 & \text{if } x_{qc}^{t} \in \{p^{100th} - p^{75th}\}\text{percentile} \end{cases}$
Indicators above or below the mean	$I_{qc}^{t} = \begin{cases} 1 & \text{if } w > (1+p) \\ 0 & \text{if } (1-p) \le w \le (1+p) \\ -1 & \text{if } w < (1+p) \end{cases}$ where $w = x_{qc}^{t} / x_{qc=\bar{c}}^{t_0}$
Cyclical indicators (OECD)	$I_{qc}^{t} = \frac{x_{qc}^{t} - E_{t}(x_{qc}^{t})}{E_{t}(\lvert x_{qc}^{t} - E_{t}(x_{qc}^{t}) \rvert)}$
Balance of opinions (EC)	$I_{qc}^{t} = \frac{100}{N_e} \sum_{e}^{N_e} \text{sgn}_{e}(x_{qc}^{t} - x_{qc}^{t-1})$
Percentage of annual differences over consecutive years	$I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc}^{t-1}}{x_{qc}^{t}}$

Note: x_{ic}^{t} is the value of indicator for country *c* at time *t*. $\bar{c}$ is the reference country. The operator *sgn* gives the sign of the argument (i.e. +1 if the argument is positive, -1 if the argument is negative). N_e is the total number of experts surveyed.

Step 6. Weighting and aggregation

The relative importance of the indicators is a source of contention …

When used in a benchmarking framework, weights can have a significant effect on the overall composite indicator and the country rankings. A number of weighting techniques exists (**Table 3**). Some are derived from statistical models, such as factor analysis, data envelopment analysis and unobserved components models (UCM) or from participatory methods like budget allocation (BAL), analytic hierarchy processes (AHP) and conjoint analysis (CA). Unobserved components and conjoint analysis approaches are explained in the Toolbox for Constructors. No matter which method is used, weights are essentially value judgements. While some analysts might choose weights based only on statistical methods, others might reward (punish) the components that are deemed more (less) influential depending on expert opinion to better reflect the policy priorities or theoretical factors.

Most composite indicators rely on equal weighting (EW), *i.e.*, all variables are given the same weight. This could correspond to the case in which all variables are "worth" the same in the composite but also it could disguise the absence of statistical or empirical basis, *e.g.* when there is insufficient knowledge of causal relationships or a lack of consensus on the alternative. In any case, equal weighting does not mean "no weights", but implicitly implies the weights are equal. Moreover, if variables are grouped into components and those further aggregated into the composite, then applying equal weighting to the variables may imply an unequal weighting of the component (the components grouping the larger number of variables will have higher weight). This could result in an unbalanced structure of the composite index.

Weights may also be chosen to reflect the statistical quality of the data. Higher weights could be assigned to statistically reliable data with broad coverage. However, this method could be biased towards the readily available indicators, penalising the information that is statistically more problematic to identify and measure

When using equal weights, it may happen that - by combining variables with high degree of correlation - one may introduce an element of double counting into the index: if two collinear indicators are included in the composite index with a weight of w1 and w2, than the unique dimension that the two indicators measure would have weight (w1+w2) in the composite. The response has often been testing indicators for statistical correlation - for example with the Pearson correlation coefficient (Manly, 1994) - and choosing only indicators exhibiting a low degree of correlation or adjusting weights correspondingly, e.g. giving less weight to correlated indicators. Furthermore, minimizing the number of variables in the index may be desirable on other grounds such as transparency and parsimony.

Notice that there will almost always be some positive correlation between different measures of the same aggregate. Thus, a rule of thumb should be introduced to define a threshold beyond which the correlation is a symptom of double counting. On the other hand relating correlation analysis to weighting could be dangerous when motivated by apparent redundancy. For example, in the CI of e-business readiness the indicator I1 "Percentage of firms using Internet" and indicator I2 "The percentage of enterprises that have a web site" display a correlation of 0.88 in 2003: are we allowed to give less weight to the pair (I1, I2) given the high correlation or shall we consider the two indicators as measuring different aspects of innovation and communication technologies adoption and give them equal weight in constructing the composite indicator? If weights should ideally reflect the contribution of each indicator to the composite, double counting should not only be determined by statistical analysis but also by the analysis of the indicator itself *vis à vis* the rest of indicators and the phenomenon they all aim to picture.

Ideally, weights should reflect the contribution of each indicator to the overall composite. Statistical models such as principal components analysis (PCA) or factor analysis (FA) could be used to group sub-

indicators. These methods account for the highest variation in the data set, using the smallest possible number of factors that reflect the underlying "statistical" dimension of the data set. Weights, however, cannot be estimated if no correlation exists between indicators. Other statistical methods, such as the "benefit of the doubt" (BOD) approach is extremely parsimonious about weighting assumptions as it lets the data decide on the weights and is sensitive to national priorities. However, weights are country specific and have a number of estimation problems.

Alternatively, participatory methods that incorporate various stakeholders -- experts, citizens and politicians -- can be used to assign weights. This approach is feasible when there is a well-defined basis for a national policy. For international comparisons, such references are often not available, or they deliver contradictory results. In the budget allocation approach, experts are given a "budget" of *N* points, to be distributed over a number of sub-indicators, "paying" more for those indicators whose importance they want to stress (Jesinghaus in Moldan and Billharz, 1997). The budget allocation is optimal for a maximum of 10-12 indicators. If too many indicators are involved, this method can give serious cognitive stress to the experts who are asked to allocate the budget. Public opinion polls have been extensively used over the years as they are easy and inexpensive to carry out (Parker, 1991).

The analytic hierarchy process (AHP) (pair wise comparison of attributes) and conjoint analysis (comparison of attributes on different levels) are also widely used techniques for multi-attribute decision making, since they enable the derivation of overall attribute (*i.e*. sub-indicator) importance based on a number of rotating attribute comparisons, as opposed to simply assigning arbitrarily given weights. The resulting weights are less sensitive to errors of judgement. However, since the AHP is based on comparisons of indicator pairs, it is applicable only to low numbers of indicators.

Aggregation methods also vary. While the linear aggregation method is useful when all sub-indicators have the same measurement unit, geometric aggregations are better suited if non-comparable and strictly positive sub-indicators are expressed in different ratio-scales. The absence of synergy or conflict across the indicators is useful in applying either linear or geometric aggregation, however difficult to achieve. Furthermore, linear aggregations reward base-indicators proportionally to the weights, while geometric aggregations reward those countries with higher scores.

In both linear and geometric aggregations, weights express trade-offs between indicators. A shortcoming in one dimension thus can be offset (compensated) by a surplus in another. This implies an inconsistency between how weights are conceived (usually they measure the importance of the associated variable) and the actual meaning when geometric or linear aggregations are used. In a linear aggregation, the compensability is constant, while with geometric aggregations compensability is lower for the composite indicators with low values. In terms of policy, if compensability is admitted (as in the case of pure economic indicators) a country with low scores on one indicator will need a much higher score on the others to improve its situation, when geometric aggregation is used. Thus in benchmarking exercises, countries with low scores prefer a linear rather than a geometric aggregation. On the other hand, the marginal utility from an increase in low absolute score would be much higher than in a high absolute score under geometric aggregation. Consequently, a country would be more interested in increasing those sectors/activities/alternatives with the lowest score in order to have the highest chance to improve its position in the ranking if the aggregation is geometric rather than linear.

If one wants to assure that weights remain a measure of importance, other aggregation methods should be used, in particular methods that do not allow compensability. Moreover if different goals are equally legitimate and important, a non-compensatory logic might be necessary. This is usually the case when highly different dimensions are aggregated in the composite, as in the case of environmental indices that include physical, social and economic data. If the analyst decides that an increase in economic performance cannot compensate a loss in social cohesion, or a worsening in environmental sustainability,

then neither the linear nor the geometric aggregation is suitable. A non-compensatory multi-criteria approach (MCA) could assure non-compensability by finding a compromise between two or more legitimate goals. In its basic form, this approach does not reward outliers, as it keeps only ordinal information, *i.e.* those countries having a greater advantage (disadvantage) in sub-indicators. This method, however, could be computationally costly when the number of countries is high, as the number of permutations to calculate increases exponentially (Munda, 2005).

With regard to the time element, keeping weights unchanged across time might be justified if the researcher is willing to analyse the evolution of a certain number of variables, as in the case of the evolution of the EC internal market index from 1992 to 2002. Also in the MCA, weights do not change being associated to the intrinsic value of the indicators to explain the phenomenon. If, instead, the objective of the analysis is that of defining best practices or that of setting priorities, then weights should necessarily change over time.

The absence of an "objective" way of determining weights and aggregation methods does not necessarily lead to rejection of the validity of composite indicators, as long as the entire process is transparent. The modeller's objectives must be clearly stated at the outset, and the chosen model must be checked to see to what extent it fulfils the modeller's goal.

Table 3. Compatibility between aggregation and weighting methods

Weighting methods	Aggregation methods		
	Linear[4]	Geometric[4]	Multi-criteria
EW	Yes	Yes	Yes
PCA/FA	Yes	Yes	Yes
BOD	Yes[1]	No[2]	No[2]
UCM	Yes	No[2]	No[2]
BAL	Yes	Yes	Yes
AHP	Yes	Yes	No[3]
CA	Yes	Yes	No[3]

1 normalized with the maximin method.
2 BOD requires additive aggregation, similar arguments apply to UCM
3 At least with the multi-criteria methods requiring weights as importance coefficients.
4 With both linear and geometric aggregations weights need to trade-offs and not "importance" coefficients

After Step 6, the constructor should have …

- Selected the appropriate weighting and aggregation procedure(s) with reference to the theoretical framework.
- Considered the possibility of using multiple procedures
- Documented and explained the weighting and aggregation procedures selected.

Step 7. Robustness and sensitivity

Sensitivity analysis can be used to assess the robustness of composite indicators …

Several judgment calls have to be made when constructing composite indicators, *e.g.* on the selection of indicators, data normalisation, weights and aggregation methods, etc. The robustness of the composite indicators and the underlying policy messages may thus be contested. A combination of uncertainty and sensitivity analyses can help gauge the robustness of the composite indicator and improve transparency.

Uncertainty analysis focuses on how uncertainty in the input factors propagates through the structure of the composite indicator and affects the composite indicator values. Sensitivity analysis assesses the contribution of the individual source of uncertainty to the output variance. While uncertainty analysis is used more often than sensitivity analysis and they are almost always treated separately, the iterative use of

uncertainty and sensitivity analyses during the development of a composite indicator could improve its structure (Saisana *et al*., 2005*a*; Tarantola *et al.*, 2000). Ideally, all potential sources of uncertainty should be tackled: selection of sub-indicators, data quality, normalisation, weighting, aggregation method, etc. The approach taken to assess uncertainties could include the following steps:

1. inclusion and exclusion of sub-indicators.
2. modelling data error based on the available information on variance estimation.
3. using alternative editing schemes, e.g. single or multiple imputation.
4. using alternative data normalisation schemes, such as re-scaling, standardisation, use of rankings.
5. using different weighting schemes, e.g., methods from the participatory family (budget allocation, analytic hierarchy process) and endogenous weighting (benefit of the doubt).
6. using different aggregation systems, e.g., linear, geometric mean of un-scaled variable, and multi-criteria ordering.
7. using different plausible values for the weights.

The consideration of the uncertainty inherent in the development of a composite indicator is cited in very few studies. The Human Development Index produced annually since 1990 by the United Nations Development Programme has encouraged improvement in the indicators used in its formulation. "No index can be better than the data it uses. But this is an argument for improving the data, not abandoning the index." (UN, 1992). The results of the robustness analysis are generally reported as country rankings with their related uncertainty bounds, which are due to the uncertainties at play. This would enable communicating to the user the plausible range of the composite indicator values for each country. The sensitivity analysis results are generally shown in terms of the sensitivity measure for each input source of uncertainty. These sensitivity measures represent how much the uncertainty in the composite indicator for a country would be reduced if that particular input source of uncertainty were removed. The results of a sensitivity analysis are often also shown as scatter-plots with the values of the composite indicator for a country on the vertical axis and each input source of uncertainty on the horizontal axis. Scatter-plots are useful to see patterns in the input-output relationships.

After Step 7, the constructor should have…

- Identified the sources of uncertainty in the development of the composite indicator .
- Assessed the impact of the uncertainties/assumptions on the final result.
- Conducted sensitivity analysis of the inference, *e.g.* to show what sources of uncertainty are more influential in determining the relative ranking of two entities.
- Documented and explained the sensitivity analyses and the results.Step 8. Links to other variables

Step 8. Links to other variables

Composite indicators can be linked to other variables and measures ….

Composite indicators often measure concepts that are linked to well-known and measurable phenomena, *e.g.*, productivity growth, entry of new firms. These links can be used to test the explanatory power of a composite. Simple cross-plots are often the best way to illustrate such links. An indicator measuring the environment for business start-ups, for example, could be linked to entry rates of new firms, where good performance on the composite indicator of business environment would be expected to yield higher entry rates.

For example, the Technology Achievement Index (TAI) helps to assess the position of a country relative to others concerning technology achievements. Higher technology achievement should lead to higher wealth, that is, countries with a high TAI would be expected to have high GDP per capita. Correlating TAI with GDP per capita shows this link (**Figure 1**). Most countries are close to the trend line. Only Norway and Korea are clear outliers. Norway is an outlier due to revenues from oil reserves, while

Korea has long prioritised technology development as an industrial strategy to catch-up with high-income countries.

High correlation suggests high quality of the composite indicator, although the correlation analysis should not be mistaken with causality analysis. Correlation simply indicates that the variation in the two data sets is similar. A change in the indicator does not necessarily lead to a change in the composite indicator and vice versa. Countries with high GDP might invest more in technology or more technology might lead to higher GDP. The causality remains unclear in the correlation analysis. More detailed econometric analyses can be used to determine causality, *e.g.* the Granger-causality test. However, causality tests require time series for all variables which are often not available.

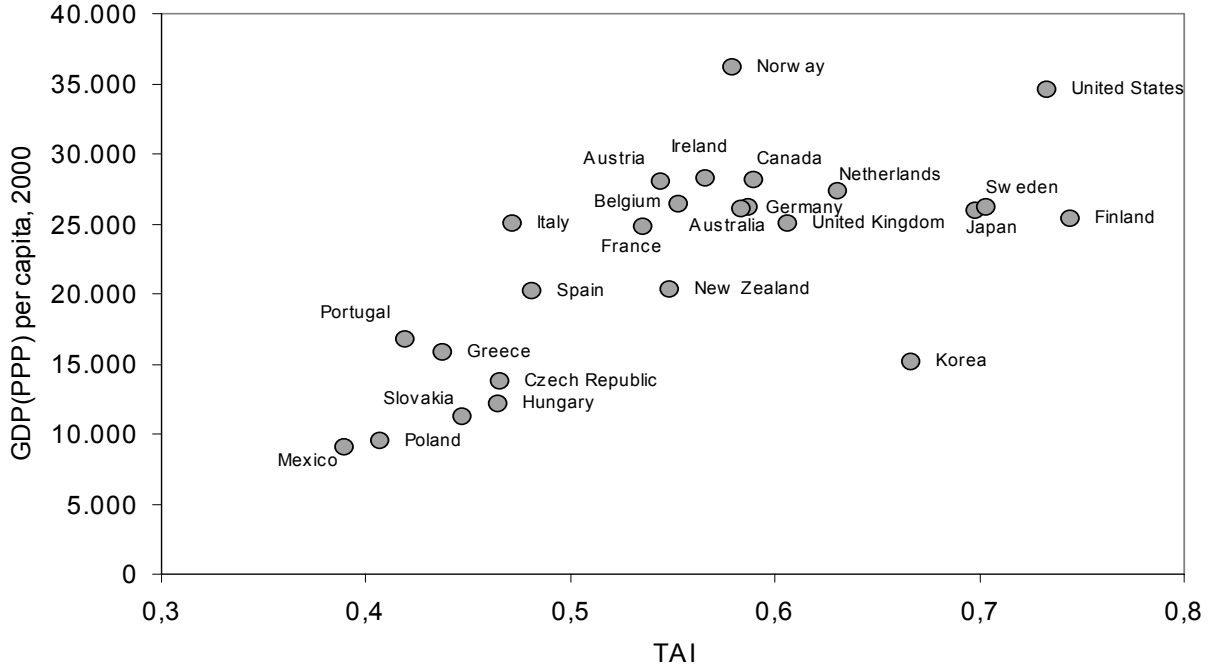
The correlation can be tested with different normalisation techniques and weights. The weights for the underlying indicators can, for example, be allowed to vary between 0 and 1 for each indicator and the calculations repeated 10 000 times. The correlation analysis can then be repeated for the 10 000 indicators and the highest, median and lowest possible correlation determined. Alternatively, the correlation between the composite indicator and the measurable phenomenon can be maximised or minimised by allowing weights to vary.

It should be noted that composite indicators often include some of the indicators with which they are being correlated leading to double counting. For example, most composite indicators of sustainable development include some measure of GDP as a sub-component. In such cases, the GDP measure should be removed from the composite indicator before running any correlation.

After Step 8, the constructor should have…

- Correlated the composite indicator with related measurable phenomena.
- Tested the links with variations of the composite indicator as determined through sensitivity analysis.
- Performed econometric analysis of the links, data permitting.
- Documented and explained the correlations and the results.

Figure 1. Link between TAI and GDP per capita, 2000



Note: The correlation is significantly different from zero at the 1% level and r^2 equals 0.47. Only OECD countries are included in the correlation, as correlation with very heterogeneous groups tends to be misleading.

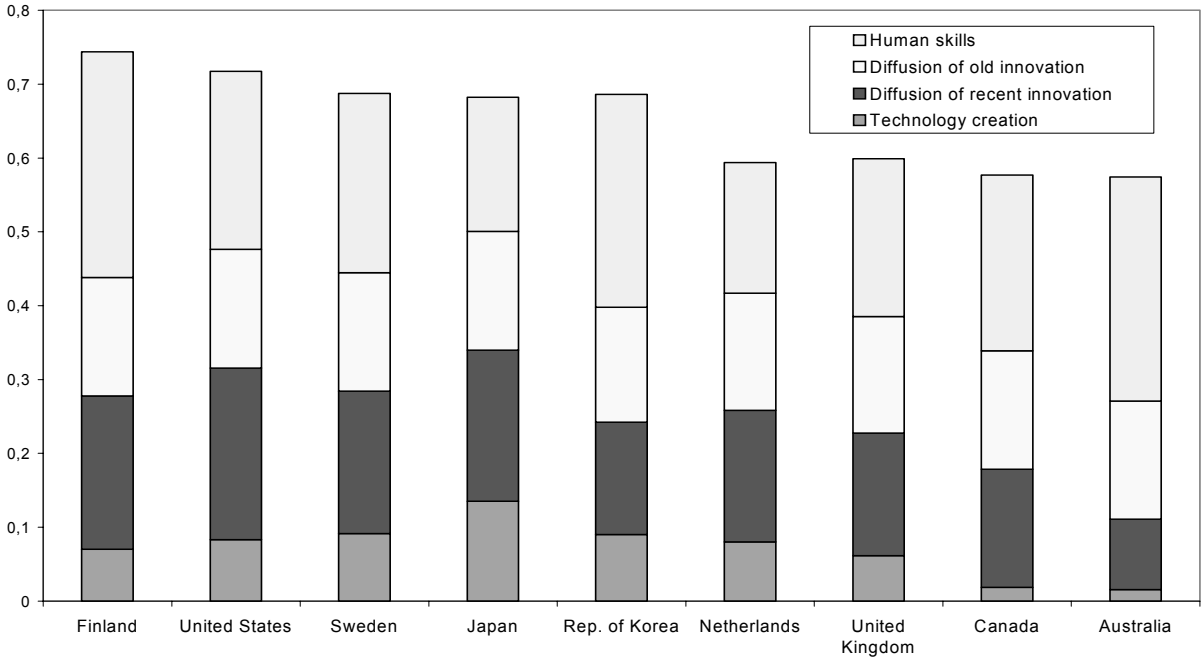
Step 9. Back to the details

De-constructing composite indicators can help extend the analysis …

Composite indicators provide a starting point for analysis. While they can be used as summary indicators to guide policy and data work, they can also be decomposed such that the contribution of sub-components and individual indicators can be identified and the analysis of country performance can be extended.

For example, the TAI index has four sub-components, which contribute differently to the aggregated composite indicator and country rankings (**Figure 2)**. This shows that a country like Finland is very strong in human skills and diffusion of recent innovations, while Japan is strong in technology creation but weaker in human skills. The decomposition of the composite indicator can thus shed light on the overall performance of a given country.

Figure 2. Example of bar chart decomposition presentation



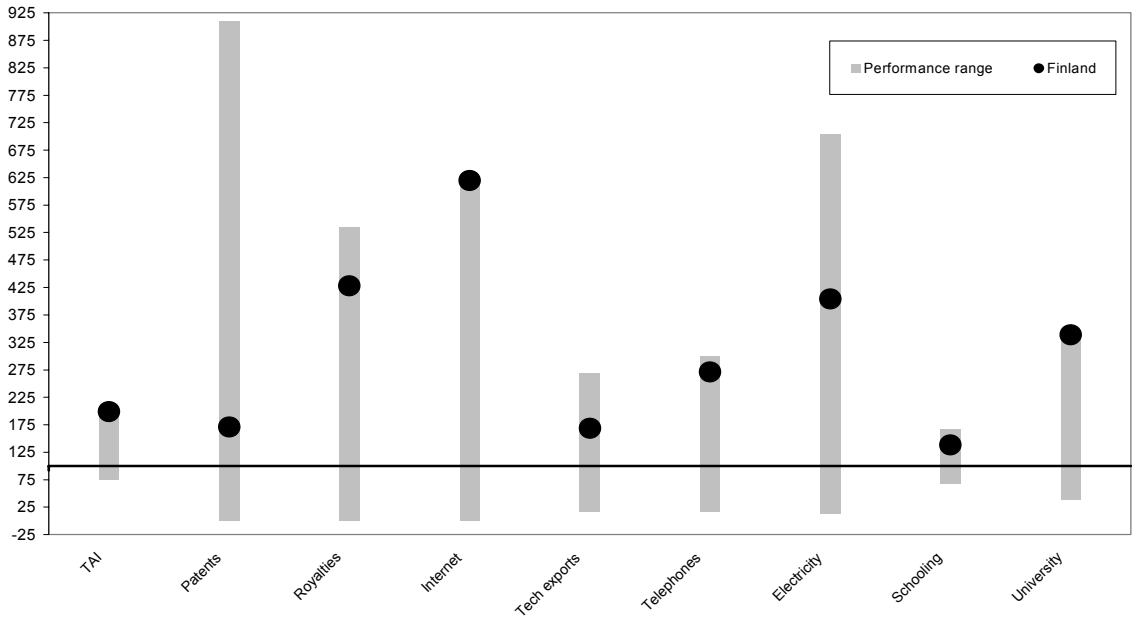
Note: Contribution of components to overall Technology Achievement Index (TAI) composite indicator. The figure is constructed by showing the standardised value of the sub-components multiplied with their individual weights. The sum of these four components equals the overall TAI index.

To profile national innovation performance**,** each sub-component of the index has been further disaggregated. The individual indicators are then used to show strengths and weaknesses. There is no optimal way of presenting individual indicators and country profiles can be presented in various ways. The following discusses three examples: 1) leaders and laggards, 2) spider diagrams and 3) traffic light presentations.

In the first example, performance on each indicator can be compared to the leader, the laggard and average performance (**Figure 3)**. Finland's top ranking is primarily based on having the highest values for the indicators relating to the Internet and universities, while the country's only weakness relates to the patent indicator.

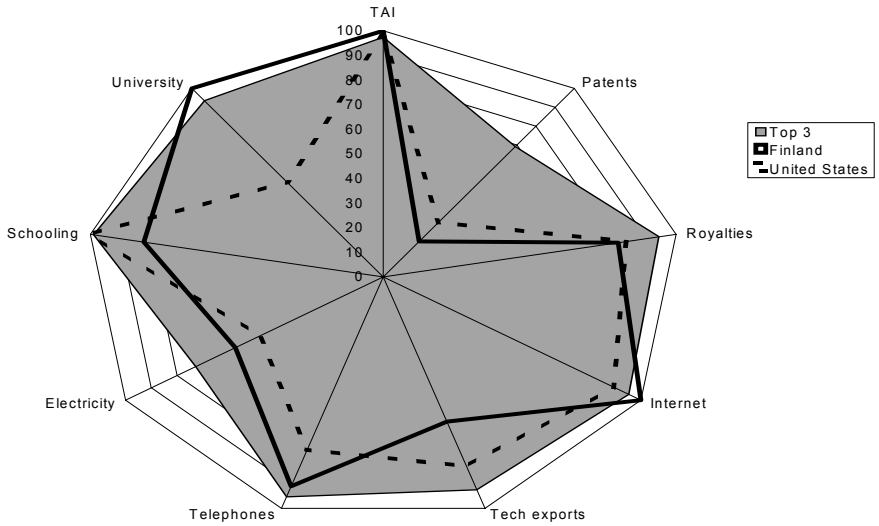
Another way of illustrating country performance is to use spider diagrams or radar charts (**Figure 4**). Here, Finland is compared to the best three countries on each indicator and another country, here the United States.

Figure 3. Example of leader/laggard decomposition presentation



Note: Technology Achievement Index (TAI). Finland (the dot) is used as an example. The figure is constructed based on the standardised indicators (distance to the mean is used). The grey area shows the range of values for that particular indicator. The average of all countries is illustrated by the 100-line.

Figure 4. Example of spider diagram decomposition presentation



Note: Technology Achievement Index (TAI). Finland is compared to the top three TAI performers and the United States. The best performing country for each indicator has the value 100 and the worst performing 0.

Finally, one can use a traffic light approach, where each indicator gets the colour green, yellow or red according to the relative performance of the country. This approach is useful when many indicators are used in the composite. For example, **Figure 5** shows that Finland has only one indicator in red (patents) and one in yellow (electricity). Japan has three in red and one in yellow.

After Step 9, the constructor should have…

- Decomposed the composite into its individual parts.
- Profiled country performance at the indicator level to reveal what is driving the aggregate results.

Documented and explained the relative

Figure 5. Example of traffic light decomposition presentation

		Well below average (under 20)	Below average (20-40)	Average (40-60)	Above average (60-80)	Well above average (over 80)
Finland	Index value					
TAI	100					
Patents	19	X				
Royalties	80					X
Internet	100					
Tech exports	63				X	
Telephones	90					X
Electricity	57			X		
Schooling	82					X
University	100					
Japan						
TAI	93					X
Patents	100					
Royalties	41			X		
Internet	24		X			
Tech exports	100					
Telephones	76				X	
Electricity	30		X			
Schooling	77				X	
University	36		X			

Note: Technology Achievement Index (TAI). Several ways exist for assigning colours. In the chosen format five colours are used but this can easily be reduced to 3. For example, green might be given to all indicators with values above the 66% quintile, yellow above 33% but below 67%.

Step 10. Presentation and dissemination

A well-designed graph can speak louder than words …

The way composite indicators are presented is not a trivial issue. Composite indicators must be able to communicate a picture to decision-makers and other end-users quickly and accurately. In particular, graphical representation of composite indicators should provide clear messages, without obscuring individual data points. On the other hand, visual presentations of composite indicators can provide signals extremely delicate from the user perspective, *e.g.,* problematic areas that require policy intervention. There are interesting ways to display and visualise composite indicators from simple tabular tools to more complicated multi-dimensional graphics and interactive software. Some examples are given below.

A tabular format is the simplest presentation where the composite indicator is presented for each country as a table of values. Usually countries are displayed in descending ranking order. Rankings can be used to track changes in country performance over time as *e.g.*, the Growth Competitiveness Index which shows the rankings of countries for two consecutive years (**Figure 6**). While tables are a comprehensive approach for displaying results, it may not be visually appealing and too detailed. However, it can be adapted to show targeted information for sets of countries grouped by geographic location, GDP, etc.

Composite indicators can be expressed via a simple bar chart (**Figure 7**). The countries are on the vertical axis and the values of the composite on the horizontal axis. The top bar indicates the average performance of all countries and enables the reader to identify how a country is performing *vis-à-vis* the average. The underlying sub-indicators can also be displayed on a bar chart. The use of colours can make

the graph more visually appealing and highlight the countries performing well or not so well, growing or not growing, etc. The top bar can be thought as a target to be reached by countries.

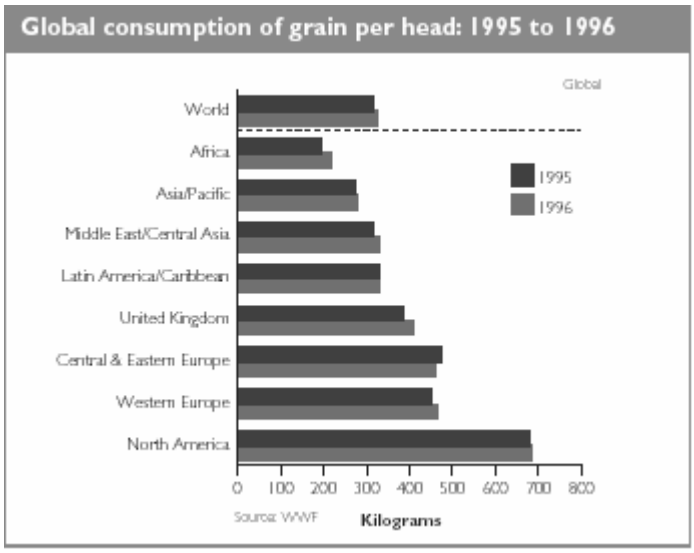
Figure 6. Example of tabular presentation of composite indicator

GROWTH COMPETITIVENESS INDEX RANKINGS

Country	Growth Competitiveness ranking 2003	Growth Competitiveness ranking 2003 among GCR 2002 countries	Growth Competitiveness ranking 2002*
Finland	1	1	1
United States	2	2	2
Sweden	3	3	3
Denmark	4	4	4
Taiwan	5	5	6
Singapore	6	6	7
Switzerland	7	7	5
Iceland	8	8	12
Norway	9	9	8
Australia	10	10	10
Japan	11	11	16
Netherlands	12	12	13
Germany	13	13	14
New Zealand	14	14	15
United Kingdom	15	15	11
Canada	16	16	9
Austria	17	17	18
Korea	18	18	25
Malta	19	—	—
Israel	20	19	17
Luxembourg	21	—	—
Estonia	22	20	27

Source: WEF, 2004.

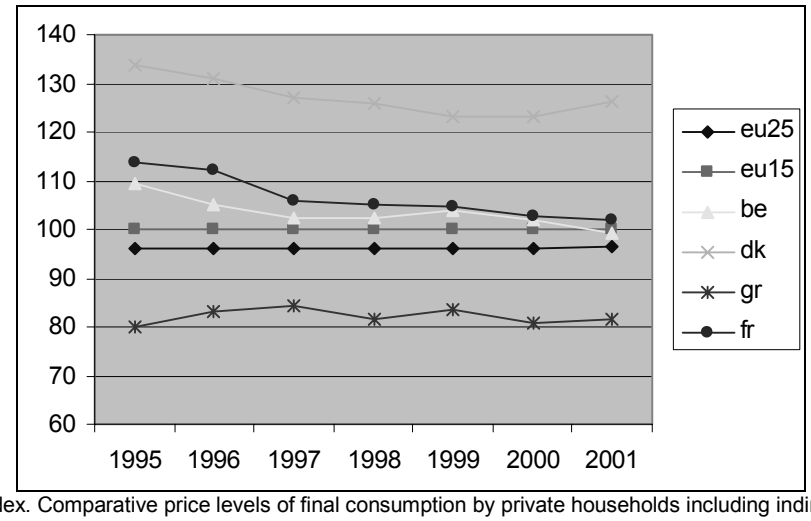
Figure 7. Example of bar chart presentation of composite indicator



Source : UK, 2004.

Line charts can be used to show performance across time. A number of lines are usually superimposed in the same chart to allow comparison between countries. Performance can be displayed *e.g.*, using a) absolute levels, b) absolute growth rates, *e.g.*, in percentage points with respect to the previous year or a number of past years, c) indexed levels and d) indexed growth rates. When indexed, the values of the indicator are linearly transformed so that their indexed value at a given year is 100. For instance, the price level index shows values such that EU15=100 at each year, with more expensive countries having values larger than 100 and less expensive countries having values lower than 100 (**Figure 8**).

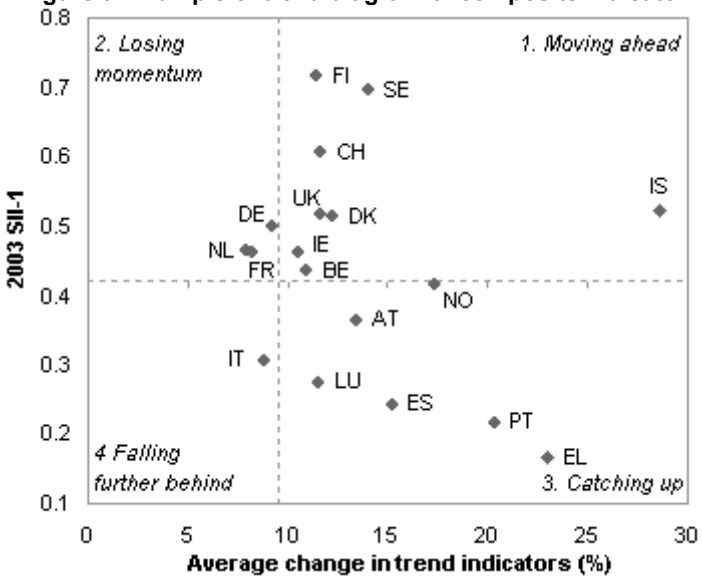
Figure 8. Example of line chart presentation of composite indicator



Note: EU price level index. Comparative price levels of final consumption by private households including indirect taxes (EU-15=100).
Source: Eurostat, 2004.

Trends in country performance as revealed through a composite indicator can be presented through trend diagrams. When a composite indicator is available for a set of countries for at least two different time points, changes or growth rates can be depicted. For example, the EU Summary Innovation Index is used to track relative performance of European countries on innovation indicators (**Figure 9**). Overall country trends are reported on the X-axis and levels are given on the Y-axis. The horizontal axis gives the EU average value and the vertical axis gives the EU trend. The two axes divide the area into four quadrants. Countries in the upper quadrant are "moving ahead", because both their value and their trend are above the EU average. Countries in the bottom left quadrant are "falling further behind" because they are below the EU average for both variables.

Figure 9. Example of trend diagram of composite indicator



Note: EU Summary Innovation Index
Source: EC, 2004.

After Step 10, the constructor should have…

- Identified a coherent set of presentational tools for the targeted audience.
- Selected the visualisation technique which communicates the most information.
- Visualised the results of the composite indicator in a clear and accurate manner.

II. QUALITY FRAMEWORK FOR COMPOSITE INDICATORS

Quality profile for composite indicators

The development of a quality framework for composite indicators is not an easy task. In fact, the overall quality of the composite indicator depends on several aspects, related both to the quality of elementary data used to build the indicator and the quality of procedures used to do it. Quality is usually defined as "fitness for use" in terms of user needs. As far as statistics are concerned, this definition is broader than has been used in the past when quality was equated with accuracy. It is now generally recognised that there are other important dimensions. Even if data are accurate, they cannot be said to be of good quality if they are produced too late to be useful, cannot be easily accessed, or appear to conflict with other data. Thus, quality is viewed as a multi-faceted concept. The most important quality characteristics depend on user perspectives, needs and priorities, which vary across groups of users.

Several organisations (*e.g.*, Statistics Canada, Statistics Sweden, Eurostat, International Monetary Fund) have been working towards the identification of various dimensions of quality for statistical products. Particularly important are the frameworks developed by the International Monetary Fund (IMF) and Eurostat. The IMF framework views quality through a prism that covers governance of statistical systems, core statistical processes and observable features of the outputs. To assess the overall quality of statistics produced by its member countries, the IMF has developed the "Data Quality Assurance Framework (DQAF)", which addresses a broad range of questions that are captured through the i) prerequisites of quality and ii) five quality dimensions.

With regard to the prerequisites of quality, the DQAF assesses how the quality of statistics is affected by the legal and institutional environment and available resources and whether there exists quality awareness in managing statistical activities. Therefore, an evaluation of the way in which a national statistical office (or system) performs its task is carried out through a detailed questionnaire to identify the degree of scientific independence of statistical agencies, the autonomy given to statistical agencies, etc.

The five quality dimensions used by the IMF are the following:

1. *Assurance of integrity*: What are the features that support firm adherence to objectivity in the production of statistics, so as to maintain users' confidence?

2. *Methodological soundness*: How do the current practices relate to the internationally agreed methodological practices for specific statistical activities?

3. *Accuracy and reliability*: Are the source data, statistical techniques, etc. adequate to portray the reality to be captured?

4. *Serviceability*: How are users' needs met in terms of timeliness of the statistical products, their frequency, consistency, and their revision cycle?

5. *Accessibility*: Are effective data and metadata easily available to data users and is there assistance to users?

The Eurostat framework focuses on statistical outputs as viewed by users and works its way back to the underlying processes only where the outputs do not yield a direct measurement. It is based on seven dimensions, which try to answer to the following questions:

1. *Relevance*: are the data what the user expects?

2. *Accuracy*: are the figures reliable?

3. *Comparability*: are the data in all necessary respects comparable across countries?

4. *Completeness:* are domains for which statistics are available reflecting the needs expressed by users?

5. *Coherence*: are the data coherent with other data?

6. *Timeliness and punctuality*: does the user receive the data in time and according to pre-established dates?

7. *Accessibility and clarity*: is the figure accessible and understandable?

Given the institutional set-up of the European Statistical System, the main aim of the Eurostat quality approach is to ensure that certain standards are met in various aspects of statistical production processes carried out by national statistical agencies and by Eurostat itself. In addition, it largely aims to use quantifiable measures, such as measurement errors or days (or months) of publication delay after the reference period.

There are several areas of commonalities between the two approaches, but, notwithstanding the effort made over the last few years to further harmonise them, they are quite different in scope. For example, the IMF approach focuses on process-oriented indicators, is mainly based on qualitative assessments and was designed with national sources in mind, while the Eurostat approach focuses on output-oriented indicators, aims to provide, to the extent possible, quantitative measures and can be applied both to national and European data.

More recently, the OECD developed and published the first version of its "Quality Framework and Guidelines for OECD Statistics" (OECD, 2003). It relies heavily on the results achieved by the international statistical community, adapting them to the OECD context. In fact, for an international organisation, the quality of statistics disseminated depends on two aspects: i) the quality of national statistics received, and ii) the quality of internal processes for collection, processing, analysis and dissemination of data and metadata. From this point of view, there are some similarities between what the OECD has done in the development of its own quality framework and the characteristics of composite indicators, whose overall quality depends on two aspects: i) the quality of basic data, and ii) the quality of procedures used to build and disseminate the composite indicator.

Both elements are equally important: the application of the most advanced approaches to the development of composite indicators based on inaccurate or incoherent data would not produce high quality results. Similarly, a composite indicator which combines very good basic data but uses poor procedures would produce unreliable and unstable results. Finally, composite indicators disseminated without appropriate metadata could easily be misinterpreted. Therefore, the quality framework for composite indicators must consider all these aspects. In the following section, each are considered separately.

Quality dimensions for basic data

The selection of basic data should maximise the overall quality of the final result. In particular, in selecting these data the following dimensions are to be considered:

Relevance

The relevance of data is a qualitative assessment of the value contributed by these data. Value is characterised by the degree to which the data serves to address the purposes for which they are sought by users. It depends upon both the coverage of the required topics and the use of appropriate concepts.

In the context of composite indicators, relevance has to be evaluated considering the overall purpose of the indicator. A careful evaluation and selection of basic data have to be carried out to ensure that the right range of domains is covered in a balanced way. Given the actual availability of data, "proxy" series are often used, but in this case some evidence about their relationships with "target" series should be produced whenever possible.

Accuracy

The accuracy of basic data is the degree to which they correctly estimate or describe the quantities or characteristics that they are designed to measure. Accuracy refers to the closeness between the values provided and the (unknown) true values. Accuracy has many attributes, and in practical terms there is no single aggregate or overall measure of it. Of necessity, these attributes are typically measured or described in terms of the error, or the potential significance of error, introduced through individual major sources of error.

In the case of sample survey-based estimates, the major sources of error include coverage, sampling, non-response, response, processing, and problems in dissemination. For derived estimates, such as for national accounts or balance of payments, sources of error arise from the surveys and censuses that provide source data; from the fact that source data do not fully meet the requirements of the accounts in terms of coverage, timing, and valuation and that the techniques used to compensate can only partially succeed; from seasonal adjustment; and from separation of price and quantity in the preparation of volume measures.

An aspect of accuracy is the closeness of the initially released value(s) to the subsequent value(s) of estimates. In light of the policy and media attention given to first estimates, a key point of interest is how close a preliminary value is to subsequent estimates. In this context it useful to consider the sources of revision, which include (1) replacement of preliminary source data with later data, (2) replacement of judgmental projections with source data, (3) changes in definitions or estimating procedures, and (4) updating of the base year for constant-price estimates. Smaller and fewer revisions is an aim; however, the absence of revisions does not necessarily mean that the data are accurate.

In the context of composite indicators, accuracy of basic data is extremely important. Here the issue of credibility of the source becomes crucial. The credibility of data products refers to confidence that users place in those products based simply on their image of the data producer, i.e., the brand image. One important aspect is trust in the objectivity of the data. This implies that the data are perceived to be produced professionally in accordance with appropriate statistical standards and policies and that practices are transparent (for example, data are not manipulated, nor their release timed in response to political pressure). Other things equal, data produced by "official sources" (*e.g.* national statistical offices or other public bodies working under national statistical regulations or codes of conduct) should be preferred to other sources

Timeliness

The timeliness of data products reflects the length of time between their availability and the event or phenomenon they describe, but considered in the context of the time period that permits the information to be of value and still acted upon. The concept applies equally to short-term or structural data; the only difference is the timeframe. Closely related to the dimension of timeliness, the punctuality of data products is also very important, both for national and international data providers. Punctuality implies the existence of a publication schedule and reflects the degree to which data are released in accordance with it.

In the context of composite indicators, timeliness is especially important to minimise the need for estimating missing data and for revisions of previously published data. As individual basic data sources establish their optimal trade-off between accuracy and timeliness taking into account institutional, organisational and resource constraints, often data covering different domains are released at different points of time. Therefore, special attention must be paid to the overall coherence of vintages of data used to build composite indicators (see also coherence).

Accessibility

The accessibility of data products reflects how readily the data can be located and accessed from original sources. The range of different users leads to such considerations as multiple dissemination formats and selective presentation of metadata. Thus, accessibility includes the suitability of the form in which the data are available, the media of dissemination, and the availability of metadata and user support services. It also includes the affordability of the data to users in relation to its value to them and whether the user has a reasonable opportunity to know that the data are available and how to access them.

In the context of composite indicators, accessibility of basic data can affect the overall cost of production and updating of the indicator over time. It can also influence the credibility of the composite indicator if poor accessibility of basic data makes it difficult for third parties to replicate the results of the composite indicators. In this respect, given improvements in electronic access to databases released by various sources, the issue of coherence across data sets can become relevant. Therefore, the selection of the source should not always give preference to the most accessible source, but also look at other quality dimensions.

Interpretability

The interpretability of data products reflects the ease with which the user may understand and properly use and analyse the data. The adequacy of the definitions of concepts, target populations, variables and terminology underlying the data, and information describing the limitations of the data, if any, largely determines the degree of interpretability. The range of different users leads to such considerations as metadata presentation in layers of increasing detail. Definitional and procedural metadata assist in interpretability: thus, the coherence of these metadata is an aspect of interpretability.

In the context of composite indicators, the wide range of data used to build them and the difficulties due to the aggregation procedure require the full interpretability of basic data. The availability of definitions and classifications used to produce basic data is essential to assess the comparability of data over time and across countries (see coherence): for example, series breaks need to be assessed when composite indicators are built to compare performances over time. Therefore, the availability of adequate metadata is an important element to assess the overall quality of basic data.

Coherence

The coherence of data products reflects the degree to which they are logically connected and mutually consistent. Coherence implies that the same term should not be used without explanation for different concepts or data items; that different terms should not be used without explanation for the same concept or

data item; and that variations in methodology that might affect data values should not be made without explanation. Coherence in its loosest sense implies the data are "at least reconcilable". For example, if two data series purporting to cover the same phenomena differ, the differences in time of recording, valuation, and coverage should be identified so that the series can be reconciled.

In the context of composite indicators, two aspects of coherence are especially important: coherence over time and across countries. *Coherence over time* implies that the data are based on common concepts, definitions, and methodology over time, or that any differences are explained and can be allowed for. Incoherence over time refers to breaks in a series resulting from changes in concepts, definitions, or methodology. *Coherence across countries* implies that from country to country the data are based on common concepts, definitions, classifications and methodology, or that any differences are explained and can be allowed for.

Quality dimensions for procedures to build and disseminate composite indicators

Each phase of the composite indicator building process is important and has to be carried out with quality concerns in mind. For example, the design of the theoretical framework can affect the relevance of the indicator; the multivariate analysis is important to increase its reliability; the imputation of missing data, as well as the normalisation and the aggregation, can affect its accuracy, etc. In the following matrix, the most important links between each phase of the building process and quality dimensions are identified, using the seven dimensions of the OECD Quality Framework (**Table** 4).

The proper definition of the theoretical framework affects the relevance of the composite indicator, but also its credibility and interpretability. The relevance of a composite indicator is usually evaluated taking into account analytical and policy needs, but also its theoretical foundation. From this point of view, several composite indicators are quite weak and such weakness is often quoted to criticise the overall idea of composite indicators.

The quality of basic data chosen to build the composite indicator strongly affects its accuracy and credibility. Also timeliness can be largely influenced by the choice of appropriate data. The use of multivariate analysis to identify the data structure can increase both the accuracy and the interpretability of final results. This step is also very important to identify redundancies among selected phenomena and evaluate possible gaps in basic data.

The imputation of missing data affects the accuracy of the composite indicator and its credibility. Furthermore, too much use of imputation techniques can undermine the overall quality of the indicator and its relevance, even if it can improve the dimension of timeliness. The normalisation phase is crucial both for the accuracy and the coherence of final results. A inappropriate normalisation procedure can bring about unreliable or biased results. On the other hand, the interpretability of the composite indicator heavily relies on the correctness of the approach followed in the normalisation phase.

One of the key issues in the construction of composite indicators is the choice of the weighting and aggregation model. Almost all quality dimensions are affected by this choice, especially accuracy, coherence and interpretability. This is also one of the most criticised characteristics of composite indicators: therefore, the indicator builder has to pay special attention to avoid internal contradictions and mistakes when dealing with weighting and aggregating individual indicators.

To minimise the risks of producing meaningless composite indicators, sensitivity and robustness analyses are needed. Analysis of this type can improve the accuracy, credibility and interpretability of the final results. Given public and media interest in country rankings, sensitivity checks can help distinguish significant and insignificant differences, minimising the risk of misinterpretation and misuse.

The comparison between the composite indicator and other well known and "classical" measures of relevant phenomena can be very useful to evaluate the capacity of the former to produce meaningful and relevant results. Therefore, relevance and interpretability of the results can be strongly reinforced by such comparison. In addition, the credibility of the indicator can benefit by its capacity to produce results which are highly correlated with the reference data.

The presentation of composite indicators and their visualisation affects both relevance and interpretability of the results. Given the complexity of composite indicators, the general public (media, citizens, etc.), as well as policy makers, will not generally read methodological notes and "caveats". Therefore, their comprehension of the results will be largely based on the "messages" given through summary tables or charts.

As highlighted in this Handbook, composite indicators provide a starting point for analysis, which has to be deepened going back to the detail. Therefore, this analytical phase can affect the relevance of the indicator and also its interpretability. Moreover, if the way in which the indicator is built or disseminated does not allow users and analysts to go into the details, the overall credibility of the exercise can be affected.

Finally, the dissemination phase is crucial to assure the relevance of the indicator, its credibility, accessibility and interpretability. Too often statisticians do not pay enough attention to this fundamental phase, thus limiting the audience for their products and their overall impact. The OECD has recently developed a Handbook for data and metadata presentation which contains useful practices to improve the dissemination of statistical products.

Table 4. Quality dimensions of composite indicators

CONSTRUCTION PHASE	**QUALITY DIMENSIONS**						
	Relevance	*Accuracy*	*Credibility*	*Timeliness*	*Accessibility*	*Interpretability*	*Coherence*
Theoretical framework	X		X			X	
Data selection		X	X	X			
Multivariate analysis		X				X	X
Imputation of missing data	X	X	X	X			
Normalisation		X				X	X
Weighting and aggregation	X	X	X			X	X
Robustness and sensitivity		X	X			X	
Links to other variables	X		X			X	X
Visualisation	X					X	
Back to the data	X		X			X	
Dissemination	X		X		X	X	

1. For explanatory purposes, only the first 23 of the 72 original countries measured by the TAI are considered here. Further details are given in the Appendix.

III. TOOLBOX FOR CONSTRUCTORS

A number of statistical methods are discussed here in detail to provide constructors the necessary tools for building sound composite indicators, focusing on the practical implementation of the steps previously outlined. The need for multivariate analysis prior to the aggregation of the individual indicators is stressed. Also discussed are the problem of missing data and the techniques used to standardise indicators of a very different nature into a common unit. Different methodologies for weighting and aggregating indicators into a composite are explored as well as the need to test the robustness of the composite using uncertainty and sensitivity analysis. The example of the Technology Achievement Index (TAI) (see Appendix) is used as a baseline case to illustrate differences across different methods and to highlight potential pitfalls.

MULTIVARIATE ANALYSIS

Multivariate data analysis techniques which have found use inthe construction or analysis of composite indicators are described in this section.

Principal components analysis

The objective[1] is to explain the variance of the observed data through a few linear combinations of the original data. Even though there are Q variables, $x_1, x_2, ...x_Q$, much of the data's variation can often be accounted for by a small number of variables – principal components, or linear relations of the original data, $Z_1, Z_2, ...Z_Q$ that are uncorrelated. At this point there are still Q principal components, *i.e.*, as many as there are variables. The next step is to select the first, say P<Q principal components that preserve a "high" amount of the cumulative variance of the original data.

$$\begin{aligned} Z_1 &= a_{11}x_1 + a_{12}x_2 + ... + a_{1Q}x_Q \\ Z_2 &= a_{21}x_1 + a_{22}x_2 + ... + a_{2Q}x_Q \\ &... \\ Z_Q &= a_{Q1}x_1 + a_{Q2}x_2 + ... + a_{QQ}x_Q \end{aligned} \qquad \textbf{(1)}$$

The lack of correlation in the principal components is a useful property. It indicates that the principal components are measuring different "statistical dimensions" in the data. When the objective of the analysis is to present a huge dataset using a few variables, some degree of economy can be achieved by applying Principal Components Analysis (PCA), if the variation in the Q original x variables can be accounted for by a small number of Z variables. It must be stressed that PCA cannot always reduce a large number of original variables to a small number of transformed variables. Indeed, if the original variables are uncorrelated, then the analysis is of no value. On the other hand, a significant reduction is obtained when the original variables are highly correlated--positively or negatively.

The weights a_{ij} (also called component or factor loadings) applied to the variables x_j in Equation (1) are chosen so that the principal components Z_i satisfy the following conditions:

(i) they are uncorrelated (orthogonal),

(ii) the first principal component accounts for the maximum possible proportion of the variance of the set of xs, the second principal component accounts for the maximum of the remaining variance and so on until the last of the principal component absorbs all the remaining variance not accounted for by the preceding components, and [2]

$$\alpha_{i1}^2 + \alpha_{i2}^2 + ... + \alpha_{iQ}^2 = 1, i = 1,2,...,Q$$

PCA involves finding the *eigenvalues* λ_j , $j=1,...,Q$, of the sample covariance matrix *CM*,

$$CM = \begin{bmatrix} cm_{11}\ cm_{12}\ ...cm_{1Q} \\ cm_{21}\ cm_{22}\ ...cm_{2Q} \\ ... \\ cm_{Q1}\ cm_{Q2}\ ...cm_{QQ} \end{bmatrix} \quad \textbf{(2)}$$

where, the diagonal element cm_{ii} is the variance of x_i and cm_{ij} is the covariance of variables x_i and x_j. The eigenvalues of the matrix *CM* are the variances of the principal components and can be found by solving the characteristic equation. $|CM - \lambda I| = 0$ where *I* is the identity matrix with the same order as CM, and λ is the vector of eigenvalues. This is possible, however, only if Q is small. If there are too many variables solving for λ is non-trivial and other methods exist (see i.e. Gentle, Härdle & Mori, 2004; and Golub & van der Vorst, 2000)[3]. There are Q eigenvalues, some of which may be negligible. Negative eigenvalues are not possible for a covariance matrix. An important property of the eigenvalues is that they add up to the sum of the diagonal elements of CM. That is, the sum of the variances of the principal components is equal to the sum of the variances of the original variables:

$$\lambda_1 + \lambda_2 + ... + \lambda_Q = cm_{11} + cm_{22} + ... + cm_{QQ} \quad (3)$$

In order to avoid one variable having an undue influence on the principal components, it is common to standardise the variables -- xs-- to have zero means and unit variances at the start of the analysis. The co-variance matrix CM then takes the form of the correlation matrix (**Table 5**). For the TAI example, the highest correlation is found between the sub-indicators ELECTRICITY & INTERNET with a coefficient of 0.84.

Table 5. Correlation matrix for TAI sub-indicators

	PATENTS	ROYALTIES	INTERNET	EXPORTS	TELEPHONE	ELECTRICITY	SCHOOLING	ENROLMENT
PATENTS	1.00	0.13	-0.09	**0.45**	0.28	0.03	0.22	0.08
ROYALTIES		1.00	**0.46**	0.25	**0.56**	0.32	0.30	0.06
INTERNET			1.00	**-0.45**	**0.56**	**0.84**	**0.63**	0.27
EXPORTS				1.00	0.00	-0.36	-0.35	-0.03
TELEPHONE					1.00	**0.64**	0.30	0.33
ELECTRICITY						1.00	**0.65**	0.26
SCHOOLING							1.00	0.08
ENROLMENT								1.00

Note: n=23. Marked correlations are statistically significant at $p < 0.05$.

Table 6 gives the eigenvalues of the correlation matrix of the eight sub-indicators (standardised values) that compose TAI. Note that the sum of the eigenvalues is equal to the number of sub-indicators ($Q = 8$).

Figure 0a is a graphical presentation of the eigenvalues in descending order. Given that the correlation matrix rather than the covariance matrix is used in the PCA, all 8 sub-indicators are assigned equal weights in forming the principal components (Chatfield and Collins, 1980). The first Principal Component explains the maximum variance in all the sub-indicators – eigenvalue of 3.3. The second principal component explains the maximum amount of the remaining variance – a variance of 1.7. The third and fourth principal components have an eigenvalue close to 1. The last four principal components explain the remaining 12.8% of the variance in the dataset.

Table 6. Eigenvalues of TAI sub-indicators

	Eigenvalue	% of variance	Cumulative %
1	3.3	41.9	41.9
2	1.7	21.8	63.7
3	1.0	12.3	76.0
4	0.9	11.1	87.2
5	0.5	6.0	93.2
6	0.3	3.7	96.9
7	0.2	2.2	99.1
8	0.1	0.9	100.0

Note: Extraction method: Principal Components Analysis, n=23.

A drawback of the conventional PCA is that it does not allow for inference on the properties of the general population. Traditionally, drawing such inferences requires certain distributional assumptions to be made regarding the population characteristics, which the PCA techniques are not based upon There are several assumptions made in the application of PCA/FA which are discussed in **Box 3**. These assumptions are mentioned in almost all textbooks, yet they are often neglected when composite indicators are developed.

Furthermore, in a traditional PCA framework, there is no estimation of the statistical precision of the results, which is essential for relatively small sample sizes, as in the present case of the TAI example. Therefore, the bootstrap method has been utilised in conjunction with PCA to make inferences about the population (Efron and Tibshirani, 1991, 1993). Bootstrap refers to the process of randomly re-sampling the original data set to generate new data sets. Estimates of the relevant statistics are made for each bootstrap sample. A very large number of bootstrap samples will give satisfactory results but the computation may be cumbersome. Various values have been suggested, ranging from 25 (Efron and Tibshirani, 1991) to as high as 1000 (Efron, 1987; Mehlman *et al*., 1995).

Box 3. Assumptions in Principal Components Analysis and Factor Analysis

Enough number of cases. The question of how many cases (or countries) are necessary to do PCA/FA has no scientific answer and methodologists' opinions differ. Alternative arbitrary rules of thumb in descending order of popularity include those below.

Rule of 10. There should be at least 10 cases for each variable.

3:1 ratio. The cases-to-variables ratio should be no lower than 3 (Grossman et al. 1991).

5:1 ratio. The cases-to-variables ratio should be no lower than 5 (Bryant and Yarnold, 1995; Nunnaly 1978, Gorsuch 1983).

Rule of 100: The number of cases should be the larger between (5 × number of variables), and 100. (Hatcher, 1994).

Rule of 150: Hutcheson and Sofroniou (1999) recommend at least 150 - 300 cases, more toward 150 when there are a few highly correlated variables.

Rule of 200. There should be at least 200 cases, regardless of the cases-to-variables ratio (Gorsuch, 1983).

Significance rule. There should be 51 more cases than the number of variables, to support chi-square testing (Lawley and Maxwell, 1971)

These rules are not mutually exclusive. Bryant and Yarnold (1995), for instance, endorse both the cases-to-variables ratio and the Rule of 200. In the TAI example, there are 23:8 cases-to-variables, therefore the first and the second rule are satisfied.

No bias in selecting sub-indicators. The exclusion of relevant sub-indicators and the inclusion of irrelevant sub-indicators in the correlation matrix being factored will affect, often substantially, the factors which are uncovered. Although social scientists may be attracted to factor analysis as a way of exploring data whose structure is unknown, knowing the factorial structure in advance helps select the sub-indicators to be included and yields the best analysis of factors. This dilemma creates a chicken-and-egg problem. Note this is not just a matter of including all relevant sub-indicators. Also, if one deletes sub-indicators arbitrarily in order to have a "cleaner" factorial solution, erroneous conclusions about the factor structure will result (Kim and Mueller, 1978a).

No outliers. As with most techniques, the presence of outliers can affect interpretations arising from PCA/FA. One may use Mahalanobis distance to identify cases, which are multivariate outliers and remove them prior to the analysis. Alternatively, one can also create a dummy variable set to 1 for cases with high Mahalanobis distance, then regress this dummy on all other variables. If this regression is non-significant (or simply has a low R-squared for large samples) then the outliers are judged to be at random and there is less danger in retaining them. The ratio of the regression coefficients indicates which variables are most associated with the outlier cases.

Assumption of interval data. Kim and Mueller (1978b) note that ordinal data may be used if it is thought that the assignment of ordinal categories to the data does not seriously distort the underlying metric scaling. Likewise, the use of dichotomous data is allowed, if the underlying metric correlation between the variables are thought to be moderate (.7) or lower. The result of using ordinal data is that the factors may be much harder to interpret. Note that categorical variables with similar splits will necessarily tend to correlate with each other, regardless of their content (see Gorsuch, 1983). This is particularly apt to occur when dichotomies are used. The correlation will reflect similarity of "difficulty" for items in a testing context; hence such correlated variables are called *difficulty factors*. The researcher should examine the factor loadings of categorical variables with care to assess whether common loading reflects a difficulty factor or substantive correlation.

Linearity. Principal components factor analysis (PFA), which is the most common variant of FA, is a linear procedure. Of course, as with multiple linear regression, nonlinear transformation of selected variables may be a pre-processing step, but this is not common. The smaller the sample size, the more important it is to screen data for linearity.

Multivariate normality of data is required for related significance tests. PCA and PFA have no distributional assumptions. Note, however, that a variant of factor analysis, maximum likelihood factor analysis, does assume

multivariate normality. The smaller the sample size, the more important it is to screen data for normality. Moreover, as factor analysis is based on correlation (or sometimes covariance), both correlation and covariance will be attenuated when variables come from different underlying distributions (ex., a normal vs. a bimodal variable will correlate less than 1.0 even when both series are perfectly co-ordered).

Underlying dimensions shared by clusters of sub-indicators are assumed. If this assumption is not met, the "garbage in, garbage out" principle applies. Factor analysis cannot create valid dimensions (factors) if none exist in the input data. In such cases, factors generated by the factor analysis algorithm will not be comprehensible. Likewise, the inclusion of multiple definitionally-similar sub-indicators representing essentially the same data will lead to tautological results.

Strong intercorrelations are not mathematically required, but applying factor analysis to a correlation matrix with only low intercorrelations will require for solution nearly as many factors as there are original variables, thereby defeating the data reduction purposes of factor analysis. On the other hand, too high inter-correlations may indicate a multi-collinearity problem and collinear terms should be combined or otherwise eliminated prior to factor analysis.

(a) The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy is a statistics for comparing the magnitudes of the observed correlation coefficients to the magnitudes of the partial correlation coefficients. The concept is that the partial correlations should not be very large if one is to expect distinct factors to emerge from factor analysis (Hutcheson and Sofroniou, 1999). A KMO statistic is computed for each individual sub-indicator, and their sum is the KMO overall statistic. KMO varies from 0 to 1.0. A KMO overall should be .60 or higher to proceed with factor analysis (Kaiser and Rice, 1974), though realistically it should exceed 0.80 if the results of the principal components analysis are to be reliable. If not, it is recommended to drop the sub-indicators with the lowest individual KMO statistic values, until KMO overall rises above .60.

(b) Variance-inflation factor (VIF) is simply the reciprocal of tolerance. A VIF value greater than 4.0 is an arbitrary but common cut-off criterion for suggesting that there is a multi-collinearity problem. Some researchers use the more lenient cutoff VIF value of 5.0.

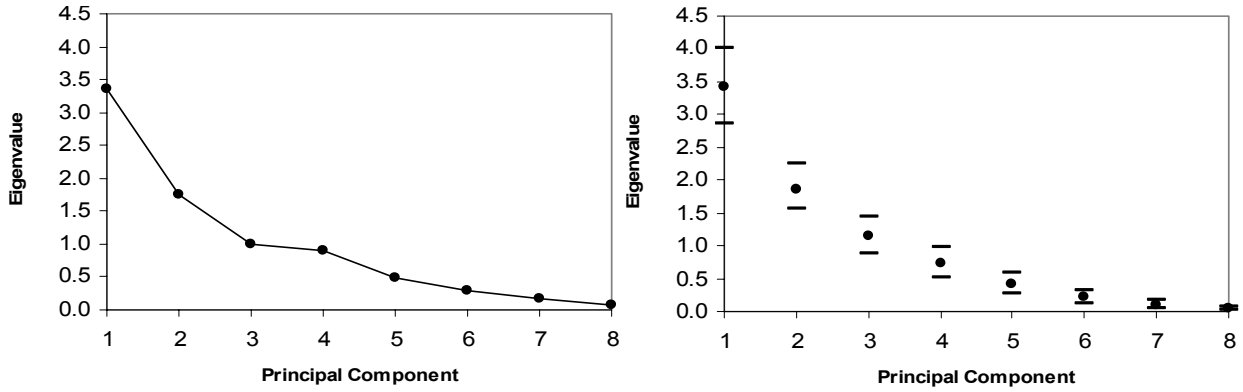
(c) The Bartlett's test of sphericity is used to test the null hypothesis that the sub-indicators in a correlation matrix are uncorrelated, that is to say that the correlation matrix is an identity matrix. The statistic is based on a chi-squared transformation of the determinant of the correlation matrix. However, as Bartlett's test is highly sensitive to sample size (Knapp and Swoyer 1967), Tabachnick and Fidell (1989) suggest implementing it with the KMO measure.

An important issue however, is whether the TAI dataset for the 23 countries can be viewed as a 'random' sample of the entire population, as required by the bootstrap procedures (Efron 1987; Efron and Tibshirani 1993). Several points can be made regarding the issues of randomness and representativeness of the data. First, it is often difficult to obtain complete information for a dataset in the social sciences, as controlled experiments are not always possible, unlike in natural sciences.. As Efron and Tibshirani (1993) state: 'in practice the selection process is seldom this neat […], but the conceptual framework of random sampling is still useful for understanding statistical inferences.' Second, the countries included in the restricted set show no apparent pattern as to whether or not they are predominately developed or developing countries. In addition, the countries of varying sizes span all the major continents of the world, ensuring a wide representation of the global state of technological development. Consequently, the restricted set could be considered as representative of the total population. A third point on the data quality is that a certain amount of measurement error is likely to exist. While such measurement error can only be controlled at the data collection stage, rather than at the analytical stage, it is argued that the data represent the best estimates currently available (UN, 2001).

Figure 10b demonstrates graphically the relationship between the eigenvalues from the deterministic PCA, their bootstrapped confidence intervals (5^{th} and 95^{th} percentiles) and the ranked principal components. These confidence intervals allow one to generalise the conclusions concerning the small set of the sub-indicators (23 countries) to the entire population (e.g. of 72 countries or even more general), rather than confining the conclusions only to the sample set being analysed. Bootstrapping has been performed

for 1000 sample sets of size 23 (random sampling with replacement). It is shown that the values of the eigenvalues drop sharply at the beginning and then gradually approach zero after a certain point.

Figure 10a & b. Eigenvalues for TAI sub-indicators



Note: (a) Eigenvalues from traditional Principal Components Analysis (Scree plot); (b) Bootstrapped eigenvalues (1000 samples randomly selected with replacement).

The correlation coefficients between the principal components Z and the variables x are called *component loadings*, $r(Z_j, x_i)$. In case of uncorrelated variables *x*, the loadings are equal to the weights a_{ij} given in equation (1). Analogous to Pearson's r, the squared loading is the percent of variance in that variable explained by the principal component. The *component scores* are the scores of each case (country in our example) on each principal component. The component score for a given case for a principal component is calculated by taking the case's standardised value on each variable, multiplying by the corresponding loading of the variable for the given principal component factor, and summing these products.

Table 7 presents the component loadings for the TAI sub-indicators. High and moderate loadings (>0.50) indicate how the sub-indicators are related to the principal components. It can be seen that with the exception of PATENTS and ROYALTIES, all the other sub-indicators are entirely accounted for by one principal component alone and that the high and moderate loadings are all found in the first four principal components. An undesirable property of these components is that two sub-indicators are related strongly to two principal components.

Table 7. Component loadings for TAI sub-indicators

	1	2	3	4	5	6	7	8
PATENTS	-0.11	**-0.75**	0.13	**0.60**	-0.10	-0.12	-0.17	0.05
ROYALTIES	**-0.56**	-0.48	0.22	**-0.54**	0.27	-0.17	-0.04	0.10
INTERNET	**-0.92**	0.21	0.02	-0.10	0.04	0.11	-0.27	-0.13
EXPORTS	0.35	**-0.85**	0.01	-0.13	0.11	0.35	0.06	-0.08
TELEPHONES	**-0.76**	-0.39	-0.16	-0.16	-0.41	-0.16	0.16	-0.09
ELECTRICITY	**-0.91**	0.13	0.01	0.07	-0.19	0.30	0.04	0.16
SCHOOLING	**-0.74**	0.11	0.37	0.39	0.33	-0.02	0.20	-0.07
ENROLMENT	-0.36	-0.12	**-0.87**	0.15	0.26	-0.03	0.02	0.02

Note: Extraction method: PCA. Loadings greater than 0.5 (absolute values) are highlighted, n=23 countries.

The question of how many principal components should be retained in the analysis without losing too much information and how the interpretation of the components might be improved are addressed in the following section on Factor Analysis.

Factor analysis

Factor analysis (FA) is similar to PCA. It aims to describe a set of Q variables $x_1, x_2, ..., x_Q$ in terms of a smaller number of m factors, and highlight the relationship between these variables. However, whereas the PCA simply is based on linear data combinations, the FA is based on a rather special model (Spearman, 1904). Contrary to the PCA, the FA model assumes that the data is based on the underlying factors of the model, and that the data variance can be decomposed into that accounted for by common and unique factors. The model is given by:

$$\begin{aligned} x_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + ... + \alpha_{1m}F_m + e_1 \\ x_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + ... + \alpha_{2m}F_m + e_2 \\ &... \\ x_Q &= \alpha_{Q1}F_1 + \alpha_{Q2}F_2 + ... + \alpha_{Qm}F_m + e_Q \end{aligned} \qquad \textbf{(4)}$$

where x_i (i=1,…,Q) represents the original variables but standardized with zero mean and unit variance; $\alpha_{i1}, \alpha_{i2}, ..., \alpha_{im}$ are the factor loadings related to the variable X_i; $F_1, F_2,...,F_m$ are m uncorrelated common factors, each with zero mean and unit variance; and e_i are the Q specific factors supposed independently and identically distributed with zero mean. There are several approaches to deal with the model given in equation (4), e.g. communalities, maximum likelihood factors, centroid method, principal axis method, etc. The most common is the use of PCA to extract the first m principal components and consider them as factors and neglect the remaining. Principal components factor analysis is most preferred in the development of composite indicators, e.g., Product Market Regulation Index (Nicoletti et al. 2000), as it has the virtue of simplicity and allows the construction of weights representing the information content of sub-indicators. Notice however that different extraction methods supply different values for the factors thus for the weights, influencing the score of the composite and the corresponding country ranking.

On the issue of how factors should be retained in the analysis without losing too much information, methodologists' opinions differ. The decision of when to stop extracting factors basically depends on when there is only very little "random" variability left, and it is rather arbitrary. However, various guidelines ("stopping rules") have been developed, roughly in the order of frequency of their use in social science (Dunteman, 1989: 22-3) (**Box 4**).

Box 4. A sample of "stopping rules"

Kaiser criterion. Drop all factors with eigenvalues below 1.0. The simplest justification to this rule is that it doesn't make sense to add a factor that explains less variance than is contained in one sub-indicator. According to this rule, 3 factors should be retained in the analysis of the TAI example, although the 4th factor follows closely with an eigenvalues of 0.90.

Scree plot. This method proposed by Cattell plots the successive eigenvalues, which drop off sharply and then tend to level off. It suggests retaining all eigenvalues in the sharp descent before the first one on the line where they start to level off. This approach would result in retaining 3 factors in the TAI example (

Figure a).

Variance explained criteria. Some researchers simply use the rule of keeping enough factors to account for 90% (sometimes 80%) of the variation. The first 4 factors account for 87.2% of the total variance.

Joliffe criterion. Drop all factors with eigenvalues under 0.70. This rule may result in twice as many factors as the Kaiser criterion, and it is less often used. In the present case study, this criterion would have lead to the selection of 4 factors.

Comprehensibility. Though not a strictly mathematical criterion, there is much to be said for limiting the number of factors to those whose dimension of meaning is readily comprehensible. Often this is the first two or three.

A relatively recent method for deciding on the number of factors to retain combines the bootstrapped eigenvalues and eigenvectors (Jackson, 1993; Yu *et al*., 1998). Based on a combination of the Kaiser

criterion and the bootstrapped eigenvalues, we should consider the first 4 factors in the TAI example. In light of the above analysis, we retain the first four principal components as identified by the bootstrap eigenvalue approach combined with the Kaiser criterion. This choice implies a greater willingness to overstate the significance of the fourth component and be in line with the idea that there are four main categories of technology achievement indicators.

After choosing the number of factors to keep, rotation is a standard step performed to enhance the interpretability of the results (Kline, 1994). The sum of eigenvalues is not affected by rotation, but changing the axes, will alter the eigenvalues of particular factors and will change the factor loadings. There are various rotational strategies that have been proposed. The goal of all of these strategies is to obtain a clear pattern of loadings. However, different rotations imply different loadings, and thus different meanings of principal components - a problem some cite as a drawback to the method. The most common rotation method is the "varimax rotation".

Table 8 presents the factor loadings for the first factors in the TAI example. Note that the eigenvalues have been affected by the rotation. The variance accounted for by the rotated components is spread more evenly than for the unrotated components. The first four factors account now for 87% of the total variance and are not sorted into descending order according to the amount of the original's dataset variance explained. The first factor has high positive coefficients (loadings) with Internet (0.79), electricity (0.82) and schooling (0.88). Factor 2 is mainly dominated by patents and exports, whilst enrolment is exclusively loaded on Factor 3. Finally, Factor 4 is formed by royalties and telephones. Yet, despite the rotation of factors, the sub-indicator of exports has sizeable loadings in both Factor 1 (negative loading) and Factor 2 (positive loading). A meaningful interpretation of the factors is not straightforward. Furthermore, the statistical treatment of the eight sub-indicators results in different groups (factors) than the conceptual ones (see Table A.1 in Appendix).

Table 8. Rotated factor loadings for TAI sub-indicators (method 1)

	Factor 1	Factor 2	Factor 3	Factor 4	Communality
PATENTS	0.07	**0.97**	0.06	0.06	0.95
ROYALTIES	0.13	0.07	-0.07	**0.93**	0.89
INTERNET	**0.79**	-0.21	0.21	0.42	0.89
EXPORTS	-0.64	**0.56**	-0.04	0.36	0.86
TELEPHONES	0.37	0.17	0.38	**0.68**	0.77
ELECTRICITY	**0.82**	-0.04	0.25	0.35	0.85
SCHOOLING	**0.88**	0.23	-0.09	0.09	0.85
ENROLMENT	0.08	0.04	**0.96**	0.04	0.93
Explained variance	2.64	1.39	1.19	1.76	
Cumulative (%)	33	50	65	87	

Note: Extraction method: principal components, varimax normalised rotation. Positive loadings greater than 0.5 are highlighted.

Another method of extracting factors that deals with the uncorrelation issue of the specific factors would have given different results. **Table 9** presents the rotated factor loadings of the four factors for the TAI case study (extraction method: principal factors maximum likelihood). For instance, electricity and schooling are not loaded any more both on F1, but electricity is loaded on F4 and schooling on F3. There is 76% variance that is common in the sub-indicators set and expressed by the four rotated common factors. In contrast, the total variance explained in the previous analysis by the four rotated principal components was much higher (87%). The commonalties for seven sub-indicators are greater than 0.66, with the exception of enrolment for which the communality is only 0.15, which indicates that enrolment does not move with the other sub-indicators in the dataset, and therefore it is not well-represented by the four common factors. This conclusion does not depend on the factor analysis method applied, as it has been confirmed by different methods (centroid method, principal axis method).

Table 9. Rotated factor loadings for TAI sub-indicators (method 2)

	Factor 1	Factor 2	Factor 3	Factor 4	Communality
PATENTS	0.01	0.11	**0.88**	0.13	0.80
ROYALTIES	**0.96**	0.14	0.09	0.18	0.99
INTERNET	0.31	**0.56**	-0.29	**0.60**	0.86
EXPORTS	0.29	-0.45	**0.58**	-0.14	0.65
TELEPHONES	0.41	0.13	0.18	**0.73**	0.75
ELECTRICITY	0.13	0.57	-0.13	**0.73**	0.89
SCHOOLING	0.14	**0.95**	0.10	0.14	0.95
ENROLMENT	-0.01	0.03	0.03	0.39	0.15
Explained Variance	1.31	1.80	1.27	1.67	
Cumulative (%)	16	39	55	76	

Note: Extraction method: principal factors maximum likelihood, varimax normalised rotation.

To sum up the steps of PCA/FA as exploratory analysis method:

1. Calculate the covariance/correlation matrix: if the correlation between sub-indicators is small, it is unlikely that they share common factors.

2. Identify the number of factors that are necessary to represent the data and the method for calculating them.

3. Rotate factors to enhance their interpretability (by maximising loading of sub-indicators individual factors).

Cronbach Coefficient Alpha

The Cronbach Coefficient Alpha, c-alpha henceforth, (Cronbach, 1951) is the most common estimate of internal consistency of items in a model or survey – Reliability/Item Analysis (e.g. Boscarino et al., 2004; Raykov, 1998; Cortina, 1993; Feldt et al., 1987; Green et al., 1977; Hattie, 1985; Miller, 1995). It assesses how well a set of items (in our terminology sub-indicators) measures a single unidimensional object (e.g. attitude, phenomenon etc.).

Cronbach's Coefficient Alpha can be defined as:

$$\alpha_c = \left(\frac{Q}{Q-1}\right)\frac{\sum_{i\neq j} cov(x_i, x_j)}{var(x_o)} = \left(\frac{Q}{Q-1}\right)\left(1 - \frac{\sum_j var(x_j)}{var(x_o)}\right) \quad c = 1,..,M;\ i,j = 1,..,Q \tag{5}$$

where M indicates the number of countries considered, Q the number of sub-indicators available, and $x_o = \sum_{q=1}^{Q} x_j$ is the sum of all sub-indicators. C-alpha measures the portion of total variability of the sample of sub-indicators due to the correlation of indicators. It increases with the number of sub-indicators and with the covariance of each pair. If no correlation exists and sub-indicators are independent then C-alpha is equal to zero, while if sub-indicators are perfectly correlated the C-alpha is equal to one.

C-alpha is not a statistical test, but a coefficient of reliability based on the correlation between sub-indicators. That is if the correlation is high, then there is evidence that the sub-indicators are measuring the same underlying construct. Therefore a high c-alpha, or equivalently a high "reliability", indicates that the

sub-indicators measure well the latent phenomenon. Although widely interpreted as such, strictly speaking c-alpha is *not a measure of unidimensionality*. A set of sub-indicators can have a high alpha and still be multidimensional. This happens when there are separate clusters of sub-indicators (separate dimensions) which inter-correlate highly, even though the clusters themselves are not highly correlated. An issue is how large the c-alpha must be. Nunnally (1978) suggests 0.7 as an acceptable reliability threshold. Yet, some authors use .75 or .80 as cut-off value, while others are as lenient as .60. In general this varies by discipline.

If the variances of the sub-indicators vary widely, like in our test case, a standard practice is to standardise the sub-indicators to a standard deviation of 1 before computing the coefficient alpha. In our notation this would mean substituting x_i with I_i. The c-alpha is .70 for the dataset of the 23 countries, which is equal to the Nunnally's cutoff value. An interesting exercise is to determine how the c-alpha varies with the deletion of each sub-indicator at a time. This helps to detect the existence of clusters of sub-indicators, thus it is useful to determine the nested structure of the composite. If the reliability coefficient increases after deleting a sub-indicator from the scale, one can assume that the sub-indicator is not correlated highly with other sub-indicators in the scale.

Table 10 presents the values for the Cronbach coefficient alpha and the correlation with the total after deleting one sub-indicator at-a-time. Telephones has the highest variable-total correlation and if deleted the coefficient alpha would be as low as 0.60. If exports were to be deleted from the set then the value of standardised coefficient alpha will increase from the current .70 to .77. Note that the same sub-indicator has the lowest variable-total correlation value (-.108). This indicates that exports is not measuring the same construct as the rest of the sub-indicators are measuring. Note also, that the factor analysis in the previous section had indicated enrolment as the sub-indicator that shares the least amount of common variance with the other sub-indicators. Although both factor analysis and the Cronbach coefficient alpha are based on correlations among sub-indicators, their conceptual framework is different.

Table10. Cronbach coefficient alpha results for TAI sub-indicators

Deleted sub-indicator	Correlation with total	Cronbach coefficient alpha
PATENTS	0.261	0.704
ROYALTIES	0.527	0.645
INTERNET	0.566	0.636
EXPORTS	-0.108	0.774
TELEPHONES	0.701	0.603
ELECTRICITY	0.614	0.624
SCHOOLING	0.451	0.662
ENROLMENT	0.249	0.706

Note: Cronbach coefficient alpha results for the 23 countries after deleting one sub-indicator (standardised values) at a time.

Cluster analysis

Cluster analysis (CLA) is a collection of algorithms to classify objects such as countries, species, and individuals. (Anderberg 1973, Massart and Kaufman 1983). The classification aims to reduce the dimensionality of a dataset by exploiting the similarities/dissimilarities between cases. CLA techniques can be hierarchical, if the classification has an increasing number of nested classes, e.g., *tree clustering*; or non-hierarchical when the number of clusters is decided ex ante e.g., the *k-means clustering*. However, care should be given that classes are meaningful, and are not arbitrary or artificial.

Homogeneous and distinct groups could be delineated based upon assessment of distances or as in the case of Ward's method, an F-test (Davis, 1986). A distance measure is an appraisal of the degree of

similarity or dissimilarity between cases in the set. A small distance is equivalent to a large similarity. It can be based on a single dimension or on multiple dimensions, for example countries in TAI example can be evaluated according to the TAI composite indicator or they can be evaluated according to all single sub-indicators. Some of the most common distance measures are listed in **Table 11** including Euclidean and non-Euclidean distances[4].

Table 11. Distance measures for TAI sub-indicators

Euclidean	$D(x,y) = \left(\frac{\sum_{i=1}^{N_d} (x_i - y_i)^2}{N_d} \right)^{1/2}$	This is the geometric distance in a multidimensional space and is usually computed from raw data (prior to any normalisation). This measure is not affected by the addition of new objects such as outliers. However, it is highly affected by the difference in scale, e.g., whether the same object is measured in centimetres or in meters the *D(x,y)*
Squared Euclidean	$D(x,y) = \frac{\sum_{i=1}^{N_d} (x_i - y_i)^2}{N_d}$	This measure places progressively greater weight on objects that are further apart. Usually this is computed from raw data and shares the same advantages and disadvantages of the Euclidean distance.
City-block[5] (Manhattan)	$D(x,y) = \frac{\sum_{i=1}^{N_d} \lvert x_i - y_i \rvert}{N_d}$	This distance is the average of distances across dimensions and it yields similar results to the Euclidean distance. In this measure the effect of outliers is less pronounced, since it is not squared.
Chebychev	$D(x,y) = Max \lvert x_i - y_i \rvert$	This measure is mostly used when one wants to define objects as "different" if they are different in any one of the dimensions.
Power	$D(x,y) = \left(\frac{\sum_{i=1}^{N_d} (x_i - y_i)^p}{N_d} \right)^{1/r}$	This distance measure is useful when one wants to increase or decrease the progressive weight placed on one dimension, for which the respective objects are very different. The parameters *r* and *p* are user-defined, such that *p* controls the progressive weights placed on differences on individual dimensions, and *r* controls the progressive weight placed on larger differences between objects. Note that for *p* = *r* = 2 corresponds to the Euclidean distance.
Percent disagreement	$D(x,y) = \frac{number\ of\ x_i \neq y_i}{N_d}$	This measure is useful if the data are categorical in nature.

The next step is to choose the clustering algorithm, i.e. the rules, which govern how distances are measured between clusters. There are many methods available. The selection criteria could differ and hence different classifications may be obtained for the same data, even using the same distance measure. The most common linkage rules are (Spath, 1980):

- **Single linkage** (nearest neighbour). The distance between two clusters is determined by the distance between the two closest elements in the different clusters. This rule called also single linkage, produces clusters chained together by single objects.

- **Complete linkage** (farthest neighbour). The distance between two clusters is determined by the greatest distance between any two objects belonging to different clusters. This method usually performs well when objects naturally form distinct groups.

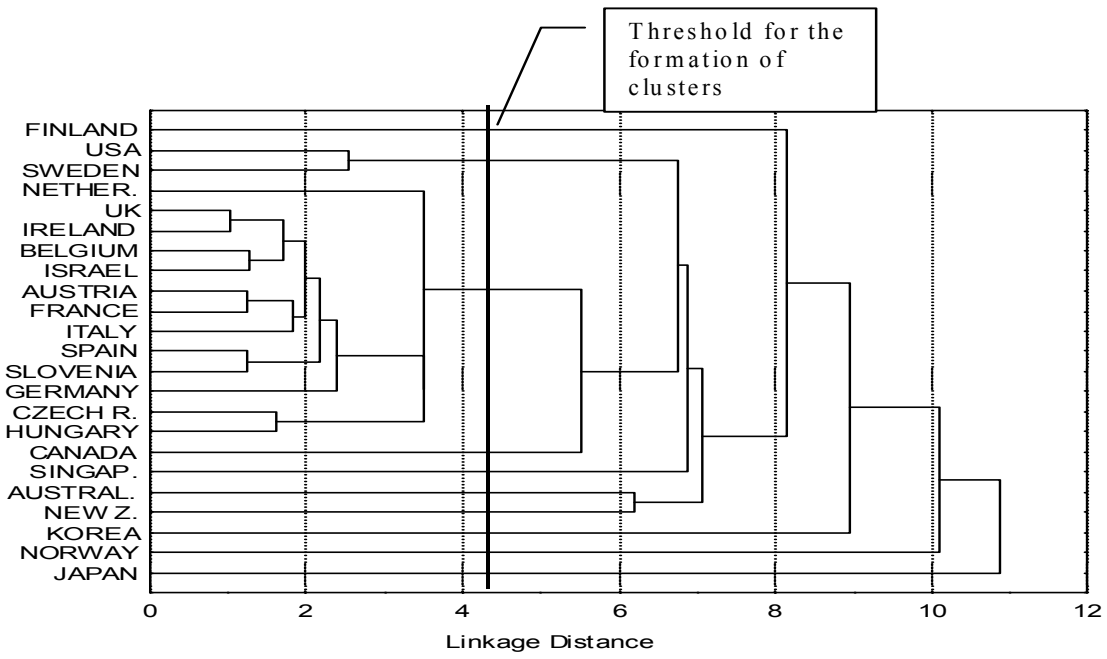
- **Unweighted pair-group average**. The distance between two clusters is calculated as the average distance between all pairs of objects in the two clusters. This method usually performs well when objects naturally form distinct groups. A variation of this method is using the centroid of a cluster--the distance is the average point in the multidimensional space defined by the dimensions.

- **Weighted pair-group average**. Similar to the unweighted pair-group average (centroid included) except that the size of the cluster, i.e. the number of objects contained, is used as weight for the average distance. This method is useful when cluster sizes are very different.

- **Ward's method** (Ward, 1963). Cluster membership is determined by calculating the variance of elements, i.e., the sum of the squared deviations from the mean of the cluster. An element will belong to the cluster if it produces the smallest possible increase in the variance.

Figure 11 shows the country clusters based on the technology achievement sub-indicators using tree clustering (hierarchical) with single linkage and squared Euclidean distances. Similarity between countries belonging to the same cluster decreases as the linkage distance increases. One of the biggest problems with CLA is identifying the optimum number of clusters. As the amalgamation process continues increasingly dissimilar clusters must be fused, i.e. the classification becomes increasingly artificial. Deciding upon the optimum number of clusters is largely subjective, although looking at the plot of linkage distance across fusion steps may help (Milligan and Cooper, 1985).

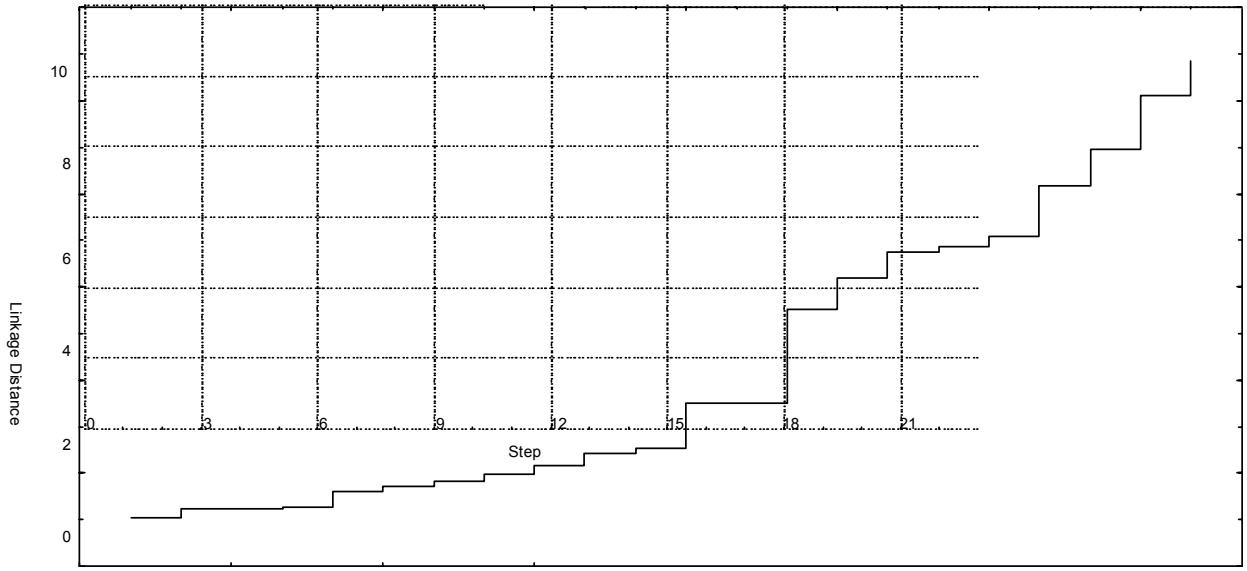
Figure 11. Country clusters for TAI sub-indicators



Note: Standardised data. Type: Hierarchical, single linkage, squared Euclidean distances.

Sudden jumps in the level of similarity (abscissa) could indicate that dissimilar groups or outliers are fused. Such a plot is presented in **Figure 12**, where the greatest dissimilarity among the 23 countries in the TAI example is found at a linkage distance close to 4.0, which indicates that the data are best represented by ten clusters: Finland; Sweden and USA; the group of countries located between the Netherlands and Hungary; Canada; Singapore; Australia; New Zealand; Korea; Norway; and Japan. Note that the most dissimilar are Korea, Norway and Japan, which are aggregated only at the very end of the analysis. Besides, this result does not fully correspond to the division in laggard, average and leading countries resulting from the standard aggregation methods. Japan, in fact, would be in the group of leading countries, together with Finland, Sweden, USA, while Hungary, Czech Republic, Slovenia and Italy would be the laggards, far away from the Netherlands, USA or Sweden (see Table 30).

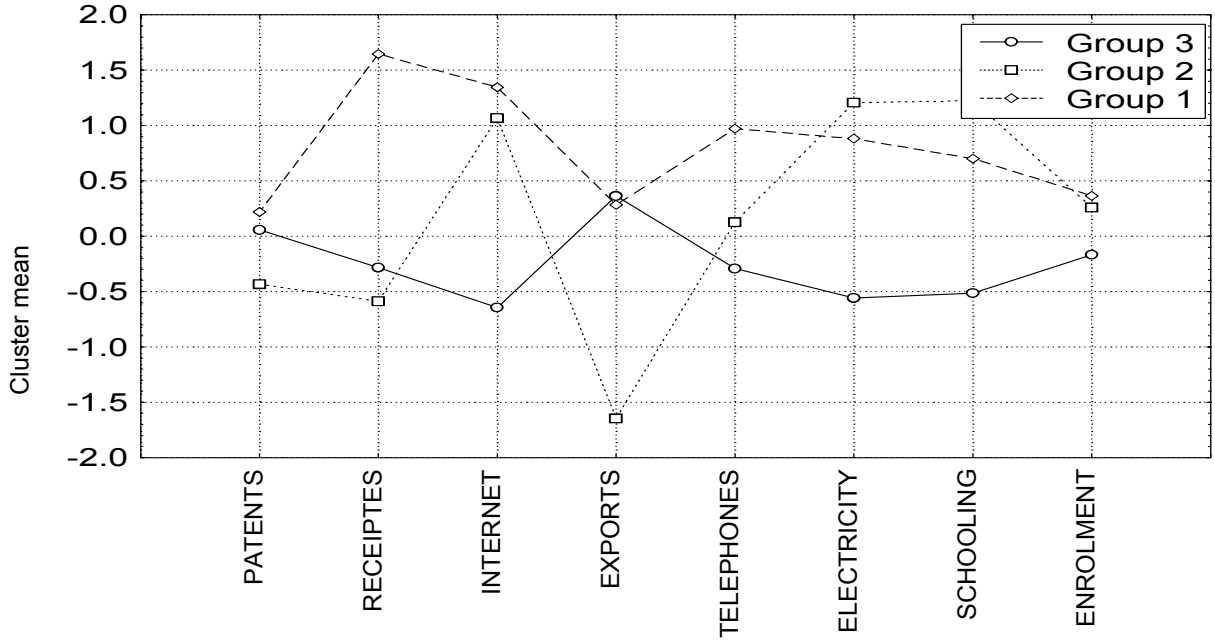
Figure 12. Linkage distance vs fusion step in TAI hierarchical cluster



A non-hierarchical method of clustering, different from the *joining* or *tree* clustering shown below, is the k-means clustering (Hartigan, 1975). This method is useful when the aim is to divide the sample in *k* clusters of greatest possible distinction. The parameter *k* is decided by the analyst, for example we may decide to cluster the 23 countries in the TAI example into 3 groups, e.g., leaders, potential leaders, and dynamic adopters. The k-means algorithm will supply 3 clusters, as distinct as possible, by analysing the variance of each cluster. This algorithm can be applied with continuous variables, yet it can be also modified to accommodate for other types of variables. The algorithm starts with *k* random clusters and moves the objects in and out the clusters with the aim of (i) minimising the variance of elements within the clusters, and (ii) maximise the variance of the elements outside the clusters.

A line graph of the means across clusters is displayed in **Figure 13**. This plot could be very useful in summarising the differences in the means between clusters. It is shown for example that the main difference between the *leaders* and the *potential leaders* (**Table 12**) is on receipts and exports. At the same time, the *dynamic adopters* are lagging behind the *potential leaders* due to their lower performance on Internet, electricity and schooling. They are, however, performing better on exports. Two of the sub-indicators, patents and enrolment, are not useful in distinguishing between these 3 groups, as the cluster means are very close.

Figure13. Means plot for TAI clusters



Note: Type: k-means clustering (standardised data).

Table 12. K-means clustering for TAI countries

Group1 (leaders)	**Group 2** (potential leaders)	**Group 3** (dynamic adopters)
Finland	Canada	Japan
USA	Australia	Korea
Sweden	Norway	UK
Netherlands	New Zealand	Singapore
		Germany
		Ireland
		Belgium
		Austria
		France
		Israel
		Spain
		Italy
		Czech Rep.
		Hungary
		Slovenia

Finally, expectation maximisation (EM) clustering extends the simple k-means clustering in two ways. First, instead of clustering the objects by maximising the differences in means for continuous variables, EM clusters membership on the basis of probability distributions--each observation will belong to each cluster with a certain probability. EM estimates mean and standard deviation of each cluster so as to maximises the overall likelihood of the data, given the final clusters (Binder, 1981). Second, unlike k-means, EM can be applied to both continuous and categorical data.

Ordinary significance tests are not valid for testing differences between clusters, as clusters are formed to be as much separated as possible. Thus the assumption of usual tests--parametric or non-parametric--is violated (see Hartigan 1975). As final remark a warning: CLA will always produce a grouping, this means that clusters may or may not prove useful for classifying objects depending upon the

objectives of the analysis. For example, if grouping zip code areas into categories based on age, gender, education and income discriminates between wine drinking behaviours, then this would be useful information only if the aim of the CLA was that of establishing a wine store in new areas. Furthermore, CLA methods are not clearly established, there are many options, all giving very different results (Everitt, 1979).

Various alternative methods combining cluster analysis and the search for a low-dimensional representation have been proposed, and focus on multidimensional scaling or unfolding analysis (e.g., Heiser, 1993, De Soete and Heiser, 1993). A method that combines k-means cluster analysis with aspects of Factor Analysis and PCA is presented by Vichi and Kiers (2001). A discrete clustering model together with a continuous factorial one are fitted simultaneously to two-way data, with the aim to identify the best partition of the objects, described by the best orthogonal linear combinations of the variables (factors) according to the least-squares criterion.

This methodology named factorial k-means analysis has a wide range of applications since it reaches a double objective: data reduction and synthesis simultaneously in direction of objects and variables. Originally applied to short-term macroeconomic data, factorial k-means analysis has a fast alternating least-squares algorithm that extends its application to large data sets, *i.e.*, multivariate data sets with >2 variables. The methodology can therefore be recommended as an alternative to the widely used tandem analysis that sequentially performs PCA and CLA..

IMPUTATION OF MISSING DATA

The literature on the analysis of missing data is extensive and in rapid development. This section covers the main methods. More comprehensive surveys can be found in Little and Rubin (2002), Little (1997) and Little and Schenker (1994).

Single imputation

Imputations are means or draws from a predictive distribution of the missing values (Little and Rubin, 2002). The predictive distribution must be generated by employing the observed data either through implicit or explicit modelling:

Implicit modelling. The focus is on an algorithm, with implicit underlying assumptions that need to be verified whether they are reasonable and fit to the issue under consideration. The danger of this type of modelling missing data is to consider the resulting data set as complete, and forget that an imputation has been done. Implicit modelling includes:

- **Hot deck imputation.** Fill in blanks cells with individual data, drawn from "similar" responding units, e.g. missing values for individual income may be replaced with the income of another respondent with similar characteristics, e.g., age, sex, race, place of residence, family relationships, job, etc.

- **Substitution.** Replace non-responding units with units not selected into the sample, e.g. if a household cannot be contacted, then a previously non-selected household in the same housing block is selected.

- **Cold deck imputation.** Replace the missing value with a value from an external source, e.g. from a previous realisation of the same survey.

Explicit modelling. The predictive distribution is based on a formal statistical model where the assumptions are made explicitly, such as:

- **Unconditional mean/median/mode imputation.** The sample mean (median, mode) of the recorded values for the given sub-indicator replaces the missing values.

- **Regression imputation**. Missing values are substituted by the predicted values obtained from regression. The dependent variable of the regression is the sub-indicator hosting the missing value, and the regressor(s) is (are) the sub-indicator(s), showing a strong relationship with the dependent variable, i.e., usually a high degree of correlation.

- **Expectation Maximisation (EM) imputation.** This model focuses on the interdependence between model parameters and the missing values. The missing values are substituted by estimates obtained through an iterative process. First, one predicts the missing values based on initial estimates of the model parameter values. These predictions are then used to update the parameter values, and the process is repeated. The sequence of parameters converges to maximum-likelihood estimates, and the time to convergence depends on the proportion of missing data and the flatness of the likelihood function.

If simplicity is its main appeal, an important limitation of the single imputation method is its systematic underestimation of the variance of the estimates (with some exceptions for the EM method where the bias depends on the algorithm used to estimate the variance). Therefore, this method does not fully assess the implications of imputation or the robustness of the composite index derived from the imputed dataset.

Unconditional mean imputation

Let X_q be the random variable associated to the sub-indicator q, with $q=1,...,Q$, and $x_{q,c}$ the observed value of X_q for country c, with $c=1,..,M$. Let m_q be the number of recorded or non-missing values on X_q, and M- m_q the number of missing values. The unconditional mean is then given by:

$$\bar{x}_q = \frac{1}{m_q} \sum_{recorded} x_{q,c} \qquad (6)$$

Similarly, the median[6] and the mode[7] of the distribution could be calculated on the available sample and to substitute missing values.[8] By "filling in" blank spaces with the sample mean, the imputed value becomes a biased estimator of the population mean, even in the case of MCAR mechanisms, and the sample variance underestimates true variance, underestimating the uncertainty on the composite due to the imputation.

Regression imputation

Suppose a set of h-1<Q fully observed sub-indicators ($x_1,...,x_{h-1}$) and a sub-indicator x_h only observed for r countries, but missing for the remaining M-r countries. Regression imputation computes the regression of x_h on ($x_1,...,x_{h-1}$) using r complete observations, and impute the missing values as prediction from the regression[9]:

$$\hat{x}_{ih} = \hat{\beta}_0 + \sum_{j=1}^{h-1} \hat{\beta}_j x_{ij} \quad i = 1,..,M-r \qquad (7)$$

In general, the strategy to define the 'best' regression is a two step procedure. First, all different subsets of predictors are adopted in a multiple regression manner. Then, the best subset(s) is determined using the following criteria:[10]

- the value of R2
- the value of the residual mean square RMS
- the value of Mallows' Ck
- stepwise regression

A variation of the regression approach is the stochastic regression approach that imputes a conditional draw instead of imputing the conditional mean:

$$\hat{x}_{ih} = \hat{\beta}_0 + \sum_{j=1}^{h-1} \hat{\beta}_j x_{ij} + \varepsilon_i \quad i = 1,..,M-r \qquad (7^*)$$

where, ε_i is a random variable $N(0,\hat{\sigma}^2)$ and $\hat{\sigma}^2$ is the residual variance from the regression of x_h on ($x_1,...,x_{h-1}$) based on the r complete cases.

A key problem for both approaches is again the underestimation of the standard errors, although stochastic regression ameliorates the distortions. Hence, the inference based on the entire dataset, including the imputed data, does not fully count for imputation uncertainty. The result is that *p-values* of tests are too small and confidence intervals too narrow. Replication methods and multiple imputation could correct the loss of precision of simple imputation.

What if the variable with missing observations is categorical? Regression imputation is still possible but adjustment using, *e.g.*, rounding of the predictions or a logistic, ordinal or multinomial logistic regression models, is required. For nominal variables, frequency statistics such as the mode or hot- and cold-deck imputation methods might be more appropriate.

Expected maximisation imputation

Suppose that *X* denotes the matrix of data. In the likelihood based estimation, the data is assumed to be generated by a model, described by a probability or density function $f(X \mid \theta)$, where θ is the unknown parameter vector lying in the parameter space Ω_θ (e.g. the real line for means, the positive real line for variances and the interval [0,1] for probabilities). The probability function captures the relationship between the data set and the parameter of the model, and describes the probability of observing a dataset for a given $\theta \in \Omega_\theta$. Since θ is unknown, while the data set is known, it makes sense to reverse the argument, and look for the probability, or the likelihood, of observing a certain θ given *X*. Therefore, given *X*, the likelihood function $L(\theta \mid X)$ is any function of $\theta \in \Omega_\theta$ proportional to $f(X \mid \theta)$:

$$L(\theta \mid X) = k(X) f(X \mid \theta) \tag{8}$$

where, $k(X) > 0$ is a function of *X* and not of θ. The log-likelihood is then the natural logarithm of the likelihood function. For *M* independent and identically distributed observations $X = (x_1,...,x_M)^T$ from a normal population with mean μ and variance σ^2 the joint density is

$$f(X \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-M/2} \exp\left(-\frac{1}{2} \sum_{c=1}^{M} \frac{(x_c - \mu)^2}{\sigma^2} \right) \tag{9}$$

For a given sample *X* the log-likelihood is (ignoring additive constants of function $f(\cdot)$) a function of (μ, σ^2):

$$\begin{aligned} l(\mu, \sigma^2 \mid X) &= \ln[L(\mu, \sigma^2 \mid X)] = \ln[k(X) f(X \mid \mu, \sigma^2)] \\ &= \ln k(X) - \frac{M}{2} \ln \sigma^2 - \frac{1}{2} \sum_{c=1}^{M} \frac{(x_c - \mu)^2}{\sigma^2} \end{aligned} \tag{10}$$

Maximising the likelihood function corresponds to the question of which value of $\theta \in \Omega_\theta$ is mostly supported by a given sampling realisation *X*. This implies solving the likelihood equation:

$$D_l(\theta \mid X_{obs}) \equiv \frac{\partial \ln L(\theta \mid X_{obs})}{\partial \theta} = 0 \tag{11}$$

When a closed-form solution of equation (11) cannot be found, iterative methods can be applied. The EM algorithm is one of these iterative methods. [11] The issue is that X contains both observable and missing values, i.e. $X = (X_{obs}, X_{mis})$. Thus one has to find both the unknown parameters and the unknown observations of the model.

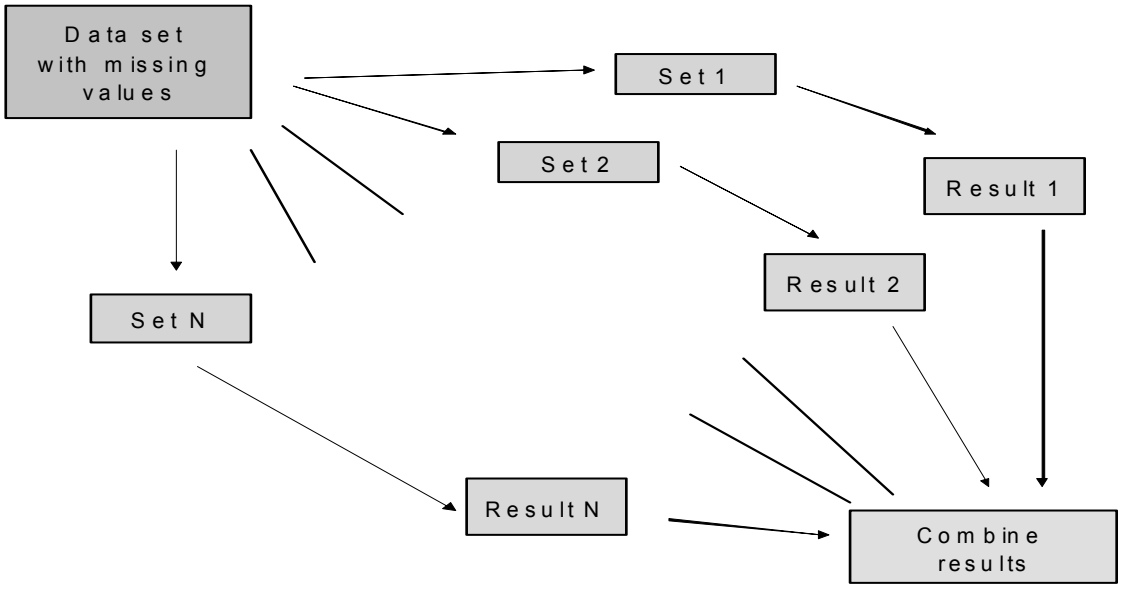
Assuming that missing data are MAR or MCAR[12], the EM consists of two components, the expectation (E) and maximisation (M) steps. Each step is completed once within each algorithm cycle. Cycles are repeated until a suitable convergence criterion is satisfied. The procedure is as follows. First (M) the parameter vector θ is estimated by applying maximum likelihood as if there was no missing data, and Second (E) the expected values of the missing variables are then calculated given the estimate of θ just obtained in the M step. This procedure is repeated until convergence (absence of changes in estimates and in the variance-covariance matrix). Effectively, this process maximises the expectation of the complete data log-likelihood in each cycle conditional on the observed data and parameter vector. To start the process, however, one needs an initial estimate of the missing data. This is done by running the first M step on the non-missing observations only, and then predict the missing variables by using the estimate on θ.

The advantage of the EM is its broadness. It can be used for a broad range of problems, e.g. variance component estimation or factor analysis. EM algorithm is also often easy to construct conceptually and practically. Besides, each step has a statistical interpretation and convergence is reliable. The main drawback however, is that convergence may be very slow, when a large fraction of information is missing (if there was no missing information, convergence would be immediate). The user should also be careful that the maximum found is indeed a global maximum and not a local one. To test this, different initial starting values for theta can be used.

Multiple imputation

Multiple imputation (MI) is a general approach that does not require a specification of parameterised likelihood for all data (**Figure 14)**. The imputation of missing data is performed with a random process that reflects uncertainty. Imputation is done N times, to create N "complete" data sets. On each data set the parameters of interest are estimated, together with their standard errors. Average (mean or median) estimates are combined using the N sets and between and within imputation variance is calculated.

Figure 14. Logic of multiple imputation

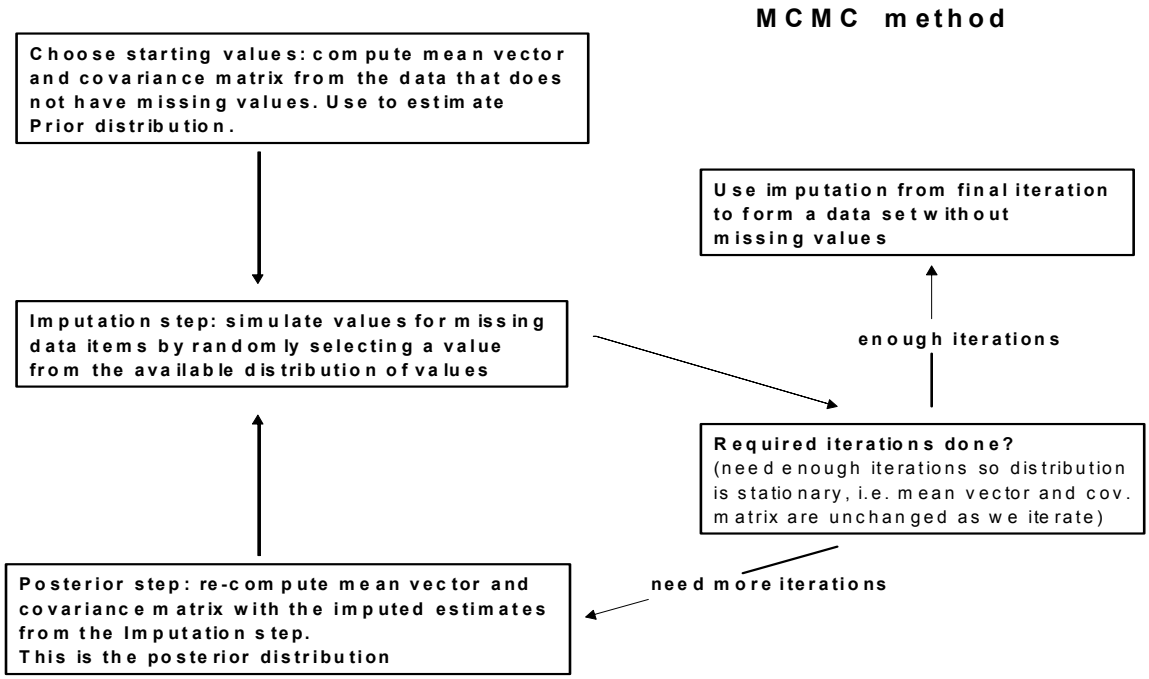


Any "proper" imputation method can be used in multiple imputation. For example, one could use regression imputation repeatedly, drawing *N* values of the regression parameters using the variance matrix of estimated coefficients. However, one of the most general models is the *Markov Chain Monte Carlo* (MCMC) method. MCMC is a sequence of random variables, in which the distribution of the actual element depends on the value of the previous one. It assumes that data are drawn from a multivariate Normal distribution and requires MAR or MCAR assumptions.

The theory of MCMC is most easily understood using Bayesian methodology (**Figure 15**). Denote the observed data as X_{obs}, and the complete dataset as $X=(X_{obs}, X_{mis})$, where X_{mis} is to be filled in via multiple imputation. If the distribution of X_{mis}, with parameter vector θ, were known then we could impute X_{mis} by drawing from the conditional distribution $f(X_{mis}|X_{obs}, \theta)$. However, since θ is unknown, we shall estimate it from the data, yielding $\hat{\theta}$, and use the distribution $f(X_{mis}|X_{obs}\hat{\theta})$. Since $\hat{\theta}$ is itself a random variable, we must also take its variability into account in drawing imputations.

The missing data generating process may also depend on additional parameters φ but if φ and θ are independent, the process is called ignorable and the analyst can concentrate on modelling the missing data given the observed data and θ**.** If the two processes are not independent, then we have a non-ignorable missing data generating process, which cannot be solved adequately without making assumptions on the functional form of the interdependency.

Figure 15. Markov Chain Monte Carlo iimputation method



Source: Rearranged from K. Chantala and C. Suchindran, (http://www.cpc.unc.edu/services/computer/presentations/mi_presentation2.pdf)

In Bayesian terms, θ is a random variable, whose distribution depends on the data. The first step for its estimation is to obtain the posterior distribution of θ from the data. Usually this posterior is approximated by a normal distribution. After formulating the posterior distribution of θ, the following imputation algorithm can be used:

- Draw θ^* from the posterior distribution of θ, f(θ|Y, Xobs) where Y denotes exogenous variables that may influence θ.
- Draw Xmis from f(Xmis|Y, Xobs, θ^*)
- Use the completed data X and the model to estimate the parameter of interest (e.g. the mean) β^* and its variance V(β^*) (within-imputation variance).

These steps are repeated independently N times, resulting in β_n^*, $V(\beta_n^*)$, $n=1,...,N$. Finally, the N imputations are combined. A possible *combination* is the mean of all individual estimates (but also the median can be used):

$$\beta^* = \frac{1}{N}\sum_{n=1}^{N}\beta_n^* \qquad (12)$$

This *combination* will be the value that fills in the blank space in the dataset. The total variance is obtained as a weighted sum of the *within-imputation* variance and the *between-imputations* variance:

$$V^* = \bar{V} + \frac{N+1}{N}B \qquad (13)$$

where the mean of the *within-imputation* variances is

$$\overline{V} = \frac{1}{N}\sum_{n=1}^{N} V\left(\beta_n^*\right) \tag{14}$$

and the *between-imputations* variance is given by

$$B = \frac{1}{N-1}\sum_{n=1}^{N}\left(\beta_n^* - \beta^*\right)\left(\beta_n^* - \beta^*\right)' \tag{15}$$

Confidence intervals are obtained by taking the overall estimate plus or minus a multiple of standard error, where that number is a quantile of Student's t-distribution with degrees of freedom:

$$df = (N-1)\left(1 + \frac{1}{r}\right)^2 \tag{16}$$

where r is the *between-to-within* ratio.

$$r = \left(1 + \frac{1}{N}\right)\frac{B}{\overline{V}} \tag{17}$$

Based on these variances, one can calculate approximate 95% confidence intervals.

Multiple Imputation method imputes several values (N) for each missing value (from the predictive distribution of the missing data), to represent the uncertainty about which values to impute. The N versions of completed data sets are analysed by standard complete data methods and the results are combined using simple rules to yield single combined estimates (e.g., MSE, regression coefficients), standard errors, p-values, that formally incorporate missing data uncertainty. The pooling of the results of the analyses performed on the multiple imputed data sets, implies that the resulting point estimates are averaged over the N completed sample points, and the resulting standard errors and p-values are adjusted according to the variance of the corresponding N completed sample point estimates. Thus, the '*between imputation variance*', provides a measure of the extra inferential uncertainty due to missing data, which is not reflected in single imputation).

NORMALISATION

The objective is to identify the most suitable normalization procedures to apply to the problem at hand, taking into account their properties with respect to the measurement units in which the indicators are expressed, and their robustness to possible outliers in the data (Ebert and Welsch, 2004). Different normalization methods will supply different results for the composite indicator. Therefore, overall robustness tests should be carried out to assess their impact on the outcomes

Scale transformation prior to normalisation

Certain normalisation procedures provide the same normalised value of the indicator irrespective of the measurement unit. Applying a normalisation procedure, which is not invariant to changes in the measurement unit, however could result in different outcomes. Below is a simple example with two indicators--temperature and humidity--for two hypothetical countries A and B, in 2003 and 2004. The raw data and normalised composites are given in **Table 13** where the temperature is first expressed in Celsius and then in Fahrenheit. Each indicator is divided by the value of the leading country and aggregated with equal weights. Using Celsius data normalised based on "distance to the best performer", the performance of Country A has increased over time. While the same normalisation and aggregation methods are used, the results in Fahrenheit show a completely different pattern. The composite indicator for country A now decreases over time.

Table 13. Normalisation based on interval scales

Temperature data in Celsius		
	2003	2004
Country A –Temperature (ºC)	35	35.9
Country A –Humidity (%)	75	70
Country B –Temperature (ºC)	39	40
Country B –Humidity (%)	50	45
Normalised data in Celsius		
Country A	0.94872	0.94875
Country B	0.83333	0.82143
Temperature data in Fahrenheit		
Country A –Temperature (F)	95	96.62
Country A –Humidity (%)	75	70
Country B –Temperature (F)	102.2	104
Country B –Humidity (%)	50	45
Normalised data in Fahrenheit		
Country A	0.964775	0.964519
Country B	0.83333	0.82143

The example illustrated so far is a case of *interval scale*, based on a transformation *f* defined as:

$$f: \quad x \to y = \alpha x + \beta; \alpha > 0, \beta \neq 0$$

where, the variable x is the temperature expressed in Celsius and y is the temperature expressed in Fahrenheit. Their relationship is given by:

$$y(\mathrm{F}) = \frac{9}{5}x(^{\circ}\mathrm{C}) + 32$$

Another common change of measurement unit is the so-called *ratio scale*, which is based on the transformation:

$$f: \quad x \to y = \alpha x; \alpha > 0 .$$

To give an example, a "length" might be expressed in centimetres (cm) or yards (yd). Their relationship is indeed: 1 yd = 91.44 cm. The normalisation by country leader, not invariant on the 'interval scale', is invariant on the 'ratio scale'. In general, all normalisation methods that are invariant on the 'interval scale', are also invariant on the 'ratio scale'.

Another transformation, which is often used to reduce the skewness of (positive) data, is the logarithmic transformation:

$$f: \quad x \to y = \log(x); x > 0 .$$

When the range of values for the indicator is wide, or it is positively skewed, the log transformation shrinks the right-hand side of the distribution. As values approach zero they are also penalised, given after transformation, they become largely negative. Expressing the weighted variables in a linear aggregation in logarithms is equivalent to the geometric aggregation of the variables without logarithms. The ratio between two weights indicates the percentage improvement in one indicator that would compensate for one percentage point decline in another indicator. This transformation leads to attributing higher weight for a one-unit improvement, starting from a low level of performance, compared to an identical improvement starting from a high level of performance.

The normalisation methods described below are all non-invariant to this type of scale transformation. The user may decide whether or not to use the log transformation before the normalisation, bearing in mind that the normalised data will be affected by the log transformation.

In some circumstances, outliers[13] can reflect the presence of unwanted information. An example is offered in the *Environmental Sustainability Index*, where the variable distributions outside the 2.5 and 97.5 percentile scores are trimmed to partially correct for outliers, as well as to avoid having extreme values overly dominate the aggregation algorithm. That is, any observed value greater than the 97.5 percentile is lowered to match the 97.5 percentile. Any observed value lower than the 2.5 percentile on is raised to the 2.5 percentile. It is advisable to first try to remove outliers, and consequently perform the normalisation, as this latter procedure can be more or less sensitive to outliers.

Standardisation (or z-scores)

For each sub-indicator x_{qc}^{t}, the average across countries $x_{qc=\bar{c}}^{t}$ and the standard deviation across countries $\sigma_{qc=\bar{c}}^{t}$ are calculated. The normalisation formula is: $I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t}}{\sigma_{qc=\bar{c}}^{t}}$, so that all the I_{qc}^{t} have similar dispersion across countries. The actual minima and maxima of the I_{qc}^{t} across countries depend on the sub-indicator. For time-dependent studies, in order to assess country performance across years, the

average across countries $x_{qc=\bar{c}}^{t_0}$ and the standard deviation across countries $\sigma_{qc=\bar{c}}^{t_0}$ are calculated for a reference year, usually the initial time point, t_0.

Re-scaling

Each indicator x_{qc}^t for a generic country c and time t is transformed in $I_{qc}^t = \frac{x_{qc}^t - min_c(x_q^t)}{max_c(x_q^t) - min_c(x_q^t)}$ where $min_c(x_q^t)$ and $max_c(x_q^t)$ are the minimum and the maximum value of x_{qc}^t across all the countries c at time t. In this way, the normalised indicators I_{qc} have values laying between 0 (laggard, $x_{qc}^t = min_c(x_q^t)$), and 1 (leader, $x_{qc}^t = max_c(x_q^t)$.

The expression $I_{qc}^t = \frac{x_{qc}^t - min_c(x_q^{t_0})}{max_c(x_q^{t_0}) - min_c(x_q^{t_0})}$ is sometimes used for time-dependent studies. However, if $x_{qc}^t > max_c(x_q^{t_0})$, the normalised indicator y_{qc}^t would be larger than 1.

Another variant of the re-scaling method is $I_{qc}^t = \frac{x_{qc}^t - min_{t\in T}\, min_c(x_q^t)}{max_{t\in T}\, max_c(x_q^t) - min_{t\in T}\, min_c(x_q^t)}$ where the minimum and maximum for each indicator is calculated across countries and time to take into account the evolution of indicators. The normalised indicators, I_{qc}^t, have values between 0 and 1. However, this transformation is not stable, when data for a new time point becomes available. This implies an adjustment of the analysis period T, which may, in turn, affect the minimum and the maximum for some sub-indicators and, hence the values of I_{qc}^t. To maintain comparability between the existing and the new data, the composite indicator for the existing data needs be recalculated.

Distance to a reference

This method takes the ratios of the indicator x_{qc}^t for a generic country c and time t with respect to the sub-indicator $x_{qc=\bar{c}}^{t_0}$ for the reference country at the initial time t_0.

$$I_{qc}^t = \frac{x_{qc}^t}{x_{qc=\bar{c}}^{t_0}}$$

Using the denominator $x_{qc=\bar{c}}^{t_0}$, the transformation takes into account the evolution of indicators across time; alternatively one can use the denominator $x_{qc=\bar{c}}^t$, with running time t.

A different approach is to consider the country itself as the reference country and calculate the distance in terms of the initial time point as

$$I_{qc}^t = \frac{x_{qc}^t}{x_{qc}^{t_0}}.$$

This approach is used in *Concern about environmental problems* (Parker, 1993) for measuring the concern of the public on certain environmental problems in three countries (Italy, France and the UK) and in the European Union. An alternative distance for the normalisation can be:

$$y_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc=\bar{c}}^{t_0}}{x_{qc=\bar{c}}^{t_0}}$$

which is essentially same as above. Instead of being centred on one, it is centred on zero. In the same way, the reference country can either be the average country, the group leader, or an external benchmark.

Indicators above or below the mean

This transformation considers the indicators that are above and below an arbitrarily defined threshold, *p,* around the mean:

$$I_{qc}^{t} = \begin{cases} 1 & \textbf{if} \quad w > (1+p) \\ 0 & \textbf{if} \quad (1-p) \le w \le (1+p) \\ -1 & \textbf{if} \quad w < (1+p) \end{cases}, \text{ where } w = x_{qc}^{t} \Big/ x_{qc=\bar{c}}^{t_0}$$

The threshold builds a neutral region around the mean, where the transformed indicator is zero. This reduces the sharp discontinuity from -1 to +1 that exists across the mean value, to two minor discontinuities from -1 to 0 and from 0 to +1 across the thresholds. A larger number of thresholds could be created at different distances from the mean value, which might overlap with the categorical scales. For time-dependent studies to assess country performance over time, the average across countries $x_{qc=\bar{c}}^{t_0}$ is calculated for a reference year (usually the initial time point t_0). An indicator that moves from significantly below the mean to significantly above the threshold in the consecutive year will have a positive effect on the composite.

Methods for cyclical indicators

When indicators are in the form of time series the transformation can be made by subtracting the mean over time $E_t\left(x_{qc}^{t}\right)$ and then by dividing by the mean of the absolute values of the difference from the mean. The normalised series are then converted into index form by adding 100.

$$I_{qc}^{t} = \frac{x_{qc}^{t} - E_t\left(x_{qc}^{t}\right)}{E_t\left(\left|x_{qc}^{t} - E_t\left(x_{qc}^{t}\right)\right|\right)}$$

Percentage of annual differences over consecutive years

Each indicator is transformed using the formula:

$$I_{qc}^{t} = \frac{x_{qc}^{t} - x_{qc}^{t-1}}{x_{qc}^{t}} * 100$$

The transformed indicator is dimension-less.

Examples of the above normalisation methods are shown in **Table 14** using the TAI data. The data are sensitive to the choice of the transformation and this might cause problems in terms of loss of the interval level of the information, sensitivity to outliers, arbitrary choice of categorical scores and sensitivity to weighting.

Table 14. Examples of normalisation techniques using TAI data

	Mean years of school	Rank	z-score	re-scaling	distance to reference country					Above/below the mean	Percentile	Categorical scale
	age 15 and above	high value = top in the list			ratio			difference				
Country					c=mean	c=best	c= worst	c=mean	c=worst	p= 20 %		
Finland	10	15	0.26	0.59	1.04	0.83	1.41	0.04	0.41	0	65.2	60
United States	12	23	1.52	1.00	1.25	1.00	1.69	0.25	0.69	1	100	100
Sweden	11.4	19	1.14	0.88	1.19	0.95	1.61	0.19	0.61	0	82.6	60
Japan	9.5	12	-0.06	0.49	0.99	0.79	1.34	-0.01	0.34	0	52.2	50
Korea, Rep. of	10.8	17	0.76	0.76	1.13	0.90	1.52	0.13	0.52	0	73.9	60
Netherlands	9.4	9	-0.12	0.47	0.98	0.78	1.32	-0.02	0.32	0	39.1	50
UK	9.4	9	-0.12	0.47	0.98	0.78	1.32	-0.02	0.32	0	39.1	50
Canada	11.6	20	1.27	0.92	1.21	0.97	1.63	0.21	0.63	1	87.0	80
Australia	10.9	18	0.83	0.78	1.14	0.91	1.54	0.14	0.54	0	78.3	60
Singapore	7.1	1	-1.58	0.00	0.74	0.59	1.00	-0.26	0.00	-1	4.3	0
Germany	10.2	16	0.38	0.63	1.06	0.85	1.44	0.06	0.44	0	69.6	60
Norway	11.9	22	1.46	0.98	1.24	0.99	1.68	0.24	0.68	1	95.7	100
Ireland	9.4	9	-0.12	0.47	0.98	0.78	1.32	-0.02	0.32	0	39.1	50
Belgium	9.3	8	-0.19	0.45	0.97	0.78	1.31	-0.03	0.31	0	34.8	40
New Zealand	11.7	21	1.33	0.94	1.22	0.98	1.65	0.22	0.65	1	91.3	80
Austria	8.4	6	-0.76	0.27	0.88	0.70	1.18	-0.12	0.18	0	26.1	40
France	7.9	5	-1.08	0.16	0.82	0.66	1.11	-0.18	0.11	0	21.7	40
Israel	9.6	14	0.00	0.51	1.00	0.80	1.35	0.00	0.35	0	60.9	50
Spain	7.3	4	-1.46	0.04	0.76	0.61	1.03	-0.24	0.03	-1	17.4	40
Italy	7.2	3	-1.52	0.02	0.75	0.60	1.01	-0.25	0.01	-1	13.0	20
Czech Republic	9.5	12	-0.06	0.49	0.99	0.79	1.34	-0.01	0.34	0	52.2	50
Hungary	9.1	7	-0.31	0.41	0.95	0.76	1.28	-0.05	0.28	0	30.4	40
Slovenia	7.1	1	-1.58	0.00	0.74	0.59	1.00	-0.26	0.00	-1	4.3	0

Sometimes, there is no need to normalise the indicators, if the indicators are already expressed with the same standard. See, for example, the case of the e-business readiness (Nardo *et al*., 2004), where all the indicators are expressed in terms of percentages of enterprises possessing a given infrastructure or using a given ICT tool. In such case the normalisation would rather obfuscate the issue, as one would loos the inherent information contained in the percentages.

WEIGHTING AND AGGREGATION

Weights based on statistical models

Principal components analysis, and more specifically factor analysis, group together sub-indicators that are collinear to form a composite indicator that captures as much of common information among sub-indicators as possible. Note that sub-indicators must have the same unit of measurement. Each factor (usually estimated using principal components analysis) reveals the set of indicators having the highest association with it. The idea under PCA/FA is to account for the highest possible variation in the indicators set using the smallest possible number of factors. Therefore, the composite no longer depends upon the dimensionality of the dataset but it is rather based on the "statistical" dimensions of the data.

According to PCA/FA, weighting only intervenes to correct for the overlapping information of two or more correlated indicators, and it is not a measure of theoretical importance of the associated indicator. If no correlation between indicators is found, then weights can not be obtained estimated with this method. This is the case for the new economic sentiment indicator, where factor and principal components analysis excluded the weighing of individual questions within a sub-component of the composite index (see the supplement B of the Business and Consumer Surveys Result N. 8/9 August/September 2001[14]). PCA/FA was excluded in the construction of an indicator of environmental sustainability when it was found that this procedure assigned negative weights to some sub-indicators (World Economic Forum, 2002).

The first step in FA is to check the correlation structure of the data, as explained in the section of multivariate analysis. If the correlation between the indicators is low, then it is unlikely that they share common factors. The second step is the identification of a certain number of latent factors, smaller than the number of sub-indicators, representing the data. Each factor depends on a set of coefficients (loadings), each coefficient measuring the correlation between the individual indicator and the latent factor. Principal component analysis is usually used to extract factors (Manly, 1994[15]). For a factor analysis only a subset of principal components are retained (m), the ones that account for the largest amount of the variance.

The standard practice is to choose factors that: (i) have associated eigenvalues larger than one; (ii) individually contribute to the explanation of overall variance by more than 10%; (iii) cumulatively contribute to the explanation of the overall variance by more than 60%. With the TAI reduced dataset (the one with 23 countries) the factors with eigenvalues close to the unity are the first four, as summarised in **Table 15**. Individually they explain more than 10% of the total variance and overall they count for about the 87% of variance.

Table 15. Eigenvalues of TAI dataset

Eigenvalues			
	Eigenval	% total Variance	Cumul. %
1	**3.3**	**41.9**	41.9
2	**1.7**	**21.8**	63.7
3	**1.0**	**12.3**	76.0
4	**0.9**	**11.1**	**87.2**
5	0.5	6.0	93.2
6	0.3	3.7	96.9
7	0.2	2.2	99.1
8	0.1	0.9	100.0

The third step deals with the rotation of factors (**Table 16**). The rotation (usually the *varimax rotation*) is used to minimise the number of sub-indicators that have a high loading on the same factor. The idea in transforming the factorial axes is to obtain a "simpler structure" of the factors (ideally a structure in which each indicator is loaded exclusively on one of the retained factors). Rotation is a standard step in factor analysis, it changes the factor loadings and hence the interpretation of the factors leaving unchanged the analytical solutions obtained *ex-ante* and *ex-post* the rotation.

Table 16. Factor loadings of TAI based on principal components

	Factor 1	Factor 2	Factor 3	Factor 4	*Factor 1*	*Factor 2*	*Factor 3*	*Factor 4*
Patents	0.07	0.97	0.06	0.06	*0.00*	***0.68***	*0.00*	*0.00*
Royalties	0.13	0.07	-0.07	0.93	*0.01*	*0.00*	*0.00*	***0.49***
Internet	0.79	-0.21	0.21	0.42	***0.24***	*0.03*	*0.04*	*0.10*
Tech exports	-0.64	0.56	-0.04	0.36	*0.16*	***0.23***	*0.00*	*0.07*
Telephones	0.37	0.17	0.38	0.68	*0.05*	*0.02*	*0.12*	***0.26***
Electricity	0.82	-0.04	0.25	0.35	***0.25***	*0.00*	*0.05*	*0.07*
Schooling	0.88	0.23	-0.09	0.09	***0.29***	*0.04*	*0.01*	*0.00*
University	0.08	0.04	0.96	0.04	*0.00*	*0.00*	***0.77***	*0.00*
Expl.Var	2.64	1.39	1.19	1.76				
Expl./Tot	*0.36*	*0.26*	*0.24*	*0.42*				

The last step deals with the construction of the weights from the matrix of factor loadings after rotation, given that the square of factor loadings represent the proportion of the total unit variance of the indicator which is explained by the factor. The approach used by Nicoletti G., Scarpetta S., Boylaud O. (2000) is that of grouping the sub-indicators with the highest factors loadings in *intermediate* composite indicators. With the TAI dataset the *intermediate* composites are 4 (Table 6.2). The first includes Internet (with a weight of 0.24), Electricity (weight 0.25) and Schooling (weight 0.29).[16] Likewise the second *intermediate* is formed by Patents and Technology Exports (worth 0.68 and 0.23 respectively), the third only by University (0.77) and the fourth by Royalties and Telephones (weighted with 0.49 and 0.26).

Then the four *intermediate* composites are aggregated by weighting each composite using the proportion of the explained variance in the dataset: 0.36 for the first (0.36 = 2.64/(2.64+1.39+1.19+1.76)), 0.26 for the second, 0.24 for the third and 0.42 for the fourth.[17] Notice that different methods for the extraction of principal components imply different weights, hence different scores for the composite (and possibly different country ranking). For example if Maximum Likelihood (ML) were to be used instead of Principal Component (PC) the weights obtained would be as given in **Table 17**.

Table 17. Factor loadings of TAI based on maximum likelihood

	M L	PCA
Patents	0.19	0.17
Royalties	0.20	0.20
Internet	0.07	0.08
Tech exports	0.07	0.06
Telephones	0.15	0.11
Electricity	0.11	0.09
Schooling	0.19	0.10
University	0.02	0.18

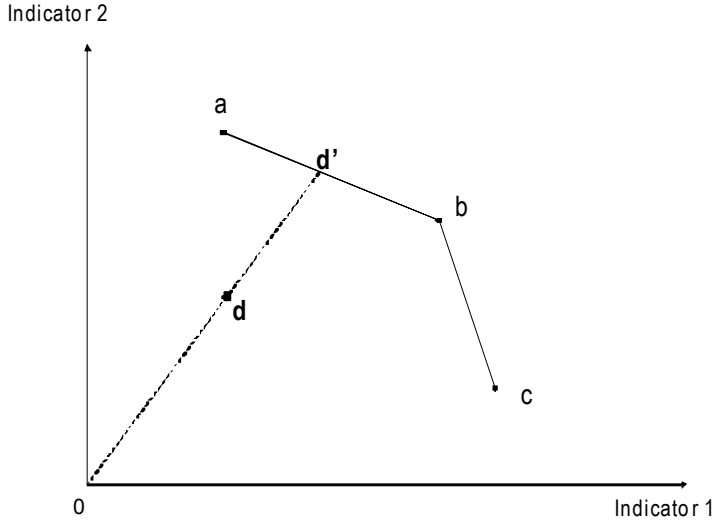
Data envelopment analysis (DEA)

Data Envelopment Analysis (DEA) employs linear programming tools to estimate an *efficiency frontier* that would be used as a benchmark to measure the relative performance of countries.[18] This requires construction of a benchmark (the *frontier*) and the measurement of the distance between countries in a multi-dimensional framework. The following assumptions are made for the benchmark:

(i) positive weights--the higher the value of a given sub-indicator, the better for the corresponding country;
(ii) non-discrimination of countries that are the best in any single dimension (sub-indicator), thus ranking them equally; and
(iii) a linear combination of the best performers is feasible, i.e., convexity of the frontier.

The distance of each country with respect to the benchmark is determined by the location of the country and its position relative to the frontier. Both issues are represented in **Figure 16**, for the simple case of four countries and two base indicators that are represented in the two axes. Countries (*a, b, c, d*) are ranked according to the score of the indicators. The line connecting countries *a*, *b* and *c* constitutes the performance frontier and the benchmark for country *d* which lies beyond the frontier. The countries supporting the frontier are classified as the *best performing*, while country *d* is the *worst performing*.

Figure 16. Data envelopment analysis (DEA) performance frontier



Source: Rearranged from Mahlberg and Obersteiner (2001).

The performance indicator is the ratio of the distance between the origin and the actual observed point and that of the projected point in the frontier: $\overline{0d} / \overline{0d'}$. The best performing countries will have a performance score of 1, and the least performing less than one. This ratio corresponds to $(w_{1d}I_{1d} + w_{2d}I_{2d})/(w_{1d}I^{*}_{1d} + w_{2d}I^{*}_{2d})$, where I^{*}_{id} is the frontier value of indicator *i*, *i=1,2*, and I_{id} is its actual value (see expression 6.1 for more than 2 indicators). The set of weights for each country therefore depends on its position with respect to the frontier, while the benchmark corresponds to the ideal point with a similar mix of indicators (*d'* in the example). The benchmark could also be determined by a hypothetical decision-maker (Korhonen et al. 2001), who locates the target in the efficiency frontier with the most preferred combination of sub-indicators. This is similar to the budget allocation method (see below) where experts are asked to assign weights (i.e. priorities) to sub-indicators.

Benefit of the doubt approach (BOD)

The application of the DEA to the field of composite indicators is known as the benefit of the doubt approach (BOD) and it was originally proposed to evaluate macroeconomic performance (Melyn and Moesen, 1991)[19]. In the BOD approach, the composite indicator is defined as the ratio of a country's actual performance over its benchmark performance:

$$CI_c = \frac{\sum_{c=1}^{M} I_{qc} w_{qc}}{\sum_{c=1}^{M} I_{qc}^{*} w_{qc}} \tag{18}$$

where I_{qc} is the normalised (with the max-min method) score of qth sub-indicator (q=1,…,Q) for country c (c=1,…M) and wqc the corresponding weight. Cherchye et al. (2004) who firstly implemented this method suggested obtaining the benchmark as solution of a maximisation problem, although external benchmarks are also possible:

$$I^{*} = I^{*}(w\ ) = \underset{I_k, k \in \{1,\ldots,M\}}{arg\,max} \left(\sum_{q=1}^{Q} I_{qk} w_q \right) \tag{19}$$

I* is the score of the hypothetical country that maximises the overall performance (defined as the weighted average), given the (unknown) set of weights w. Note that (i) weights are country specific: different sets of weights may lead to choose different countries as far as there is no country having the highest score in all sub-indicators; (ii) the benchmark would in general be country-dependent, so no unique benchmark would exist (unless, as before, a country is better in all sub-indicators), (iii) sub-indicators must be comparable, i.e. have the same unit of measurement.

The second step is the specification of the set of weights for each country. The optimal set of weighs--if it exists--guarantees the best position for the associated country *vis-à-vis* all other countries in the sample. With any other weighting profile, the relative position of that country would have been worse. Optimal weights are obtained by solving the following constrained optimisation:

$$CI_c^{*} = \underset{w_{qc}, q=1,\ldots,Q}{arg\,max} \frac{\sum_{q=1}^{Q} I_{qc} w_{qc}}{\underset{I_k, k \in \{1,\ldots,M\}}{max} \left(\sum_{q=1}^{Q} I_{qk} w_{qc} \right)} \quad \text{for } c=1,..,M \tag{20}$$

subject to non negativity constraints on weights. [20]

The resulting composite index will range between zero (lowest possible performance) and 1 (the benchmark). Operationally, equation (20) can be reduced to a linear programming problem (21) by multiplying all the weights with a common factor that does not alter the index value and then solved using optimisation algorithms

$$CI_c^* = \arg\max_{wqc} \sum_{q=1}^{Q} I_{qc} w_{qc}$$

$$s.t. \qquad (21)$$

$$\sum_{q=1}^{Q} I_{qk} w_{qk} \leq 1$$

$$w_{qk} \geq 0$$

$$\forall k = 1,...,M; \forall q = 1,...,Q$$

The result of BOD approach applied to the TAI example can be seen in **Table 19**. Weights are given in the first eight columns, while the last column contains the composite indicator values. Finland, the United States and Sweden have a composite index value of one, i.e. they have the top score in the ranking. This however hides a problem of multiple equilibria. In Figure 18, any point between country *a* (e.g., Finland) and country *b* (e.g., USA) can be an optimal solution for these countries. Thus weights are not uniquely determined. Notice also that the multiplicity of solutions is likely to depend upon the set of constraints imposed to the weights of the maximisation problem in (21) – the wider is the range of variation of weights, the lower is the possibility of obtaining a unique solution[21]. Second, the set of weights for each country as calculated by the above algorithm[22] does not sum up to one, making the comparison between countries and with other methods (like FA or EW) impossible.

Table 19. Benefit of the doubt (BOD) approach applied to TAI

	Patents	Royalties	Internet	Tech. Export	Telephones	Electricity	Schooling	University	CI
Finland	0.15	0.17	0.17	0.16	0.19	0.17	0.17	0.19	**1**
United States	0.20	0.20	0.17	0.21	0.15	0.15	0.21	0.14	**1**
Sweden	0.18	0.21	0.15	0.19	0.19	0.16	0.20	0.14	**1**
Japan	0.22	0.15	0.15	0.22	0.22	0.16	0.21	0.15	**0.87**
Korea	0.22	0.14	0.14	0.22	0.14	0.14	0.22	0.22	**0.80**
Netherlands	0.22	0.22	0.14	0.22	0.22	0.14	0.14	0.14	**0.75**
United Kingdom	0.14	0.21	0.14	0.21	0.21	0.14	0.20	0.15	**0.71**
Canada	0.14	0.14	0.14	0.21	0.21	0.21	0.21	0.14	**0.73**
Australia	0.13	0.13	0.20	0.13	0.13	0.20	0.20	0.20	**0.66**
Singapore	0.14	0.14	0.14	0.20	0.20	0.20	0.14	0.20	**0.62**
Germany	0.22	0.15	0.15	0.22	0.21	0.15	0.22	0.15	**0.62**
Norway	0.14	0.14	0.20	0.14	0.20	0.20	0.20	0.14	**0.86**
Ireland	0.14	0.21	0.14	0.21	0.21	0.14	0.20	0.15	**0.60**
Belgium	0.14	0.16	0.14	0.21	0.19	0.21	0.21	0.14	**0.54**
New Zealand	0.21	0.14	0.21	0.14	0.14	0.21	0.21	0.14	**0.58**
Austria	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.52**
France	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.51**
Israel	0.21	0.15	0.15	0.22	0.22	0.15	0.22	0.15	**0.49**
Spain	0.21	0.14	0.14	0.21	0.21	0.14	0.14	0.21	**0.34**
Italy	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.38**
Czech Rep.	0.22	0.15	0.15	0.22	0.15	0.22	0.22	0.15	**0.31**
Hungary	0.22	0.14	0.21	0.22	0.14	0.14	0.22	0.15	**0.27**
Slovenia	0.22	0.14	0.14	0.22	0.22	0.22	0.14	0.14	**0.28**

Note: Columns 1 to 8 : weights, column 9: composite indicator for a given country, n=23 countries.

Unobserved components model (UCM)

In the unobserved components model (UCM), sub-indicators are assumed to depend on an unobserved variable plus an error term, e.g. the "percentage of firms using internet in country *j*" depends upon the (unknown) propensity to adopt new information and communication technologies plus an error term, accounting, for example, for the error in the sampling of firms. Therefore, estimating the unknown component sheds some light on the relationship between the composite and its components. The weight obtained will be set to minimise the error in the composite. This method resembles the well known regression analysis. The main difference resides in the dependent variable, which is unknown under UCM.

Let *ph(c)* be the unknown phenomenon to be measured. The observed data consist on a cluster of *q=1,...,Q(c)* indicators, each measuring an aspect of *ph(c)*. Let *c=1,...M(q)* the countries covered by indicator *q*. The observed score of country *c* on indicator *q*, *I(c,q)*, can be written as a linear function of the unobserved phenomenon and an error term, $\varepsilon(c,q)$:

$$I(c,q) = \alpha(q) + \beta(q)[\,ph(c) + \varepsilon(c,q)\,] \tag{22}$$

$\alpha(q)$ and $\beta(q)$ are unknown parameters mapping *ph(c)* on *I(c,q)*.

The error term captures two sources of uncertainty. First, the phenomenon can be imperfectly measured or observed in each country (e.g. errors of measurement). Second, the relationship between *ph(c)* and *I(c,q)* can be imperfect (e.g. *I(c,q)* may only be a noisy indicator of the phenomenon, if there are differences among countries about the indicator). The error term $\varepsilon(c,q)$ is assumed to have a zero mean, $E(\varepsilon(c,q)) = 0$, and the same variance across countries within a given indicator, but a different variance across indicators, $E(\varepsilon(c,q)^2) = \sigma_q^2$; it also holds $E(\varepsilon(c,q)\varepsilon(i,h)) = 0$ for $c \neq i$ or $q \neq h$.

The error term is assumed to be independent across indicators, given each sub-indicator should ideally measure a particular aspect of the phenomenon independent of others. Furthermore, it is usually assumed that *ph(c)* is a random variable with mean zero and unit variance, and the indicators are re-scaled to take values between zero and one. The assumption that both *ph(c)* and $\varepsilon(c,q)$ are jointly normally distributed simplifies the estimation of the level of *ph(c)* in country *c*. This is done by using the mean of the conditional distribution of the unobserved component, once the observed scores are appropriately re-scaled:

$$E[\,ph(c) / I(c,1),...,I(c,Q(c))] = \sum_{q=1}^{Q(c)} w(c,q)\frac{I(c,q) - \alpha(q)}{\beta(q)} \tag{23}$$

The weights are equal to:

$$w(c,q) = \frac{\sigma_q^{-2}}{1 + \sum_{q=1}^{Q(c)} \sigma_q^{-2}} \tag{24}$$

w(c,q) is a decreasing function of the variance of indicator *q*, and an increasing function of the variance of the other indicators. The weight, *w(c,q),* depends on the variance of indicator *q* (numerator) and on the sum of the variances of the all the other sub-indicators, including *q* (denominator). However, since not all countries have data on all sub-indicators, the denominator of *w(c,q)* could be country-specific. This may produce non-comparability of country values for the composite as in BOD. Obviously whenever the set of

indicators is equal for all countries, weights will be no longer country specific and comparability will be assured. The variance of the conditional distribution is given by:

$$var[\, ph(c) / I(c,1),...,I(c,Q(c))] = [\, 1 + \sum_{q=1}^{Q(c)} \sigma_q^{-2} \,]^{-1} \qquad (25)$$

and can be used as a measure of the precision of the composite. The variance decreases in the number of indicators for each country, and increases in the variance of the disturbance term for each indicator. The estimation of the model could be simplified by the assumption of normality for *ph(c)* and $\varepsilon(c,q)$. The likelihood function of the observed data is maximised with respect to the unknown parameters, $\alpha(q)s$, $\beta(q)s$, and $\sigma_q^2 s$, and their estimated values substituted in equation (23) to obtain the composite indicator and the weights.

Budget allocation (BAL)

It is essential to bring together experts that have a wide spectrum of knowledge, and experience to ensure that a proper weighting system is established. Special care should be given in the identification of the population of experts from which to draw a sample, stratified or otherwise[23]. The budget allocation method (BAL) has four different phases:

- Selection of experts for the valuation;
- Allocation of budget to the sub-indicators;
- Calculation of the weights;
- Iteration of the budget allocation until convergence is reached (optional).

Public opinion

From a methodological point of view, opinion polls focus on the notion of "concern". That is people are asked to express the degree of concern (e.g., much or little) on issues, measured by the base indicators. As with expert assessments, the budget allocation method could also be applied in public opinion polls. However it is more difficult to ask the public to allocate a hundred points to several sub-indicators than to express a degree of concern about a given problem.

Analytic hierarchy process (AHP)

The Analytic Hierarchy Process (AHP) is a widely used technique for multi-attribute decision making (Saaty, 1987). It enables the decomposition of a problem into hierarchy and assures that both qualitative and quantitative aspects of a problem are incorporated in the evaluation process, during which opinions are systematically extracted by means of pairwise comparisons. According to Forman et al. (1983): "*AHP is a compensatory decision methodology because alternatives that are efficient with respect to one or more objectives can compensate by their performance with respect to other objectives. AHP allows for the application of data, experience, insight, and intuition in a logical and thorough way within a hierarchy as a whole. In particular, AHP as weighting method enables decision-maker to derive weights as opposed to arbitrarily assign them.*"

Weights represent the trade-off across indicators. They measure the willingness to forego a given variable in exchange for another. Hence, they are not importance coefficients. This could create a misunderstanding, if AHP weights are interpreted as importance coefficients (see Ülengin et al. 2001).

The core of AHP is an ordinal pair-wise comparison of attributes. For a given objective, the comparisons are made per pairs of sub-indicators: Which of the two is the more important? And, by how much? The preference is expressed on a semantic scale of 1 to 9. A preference of 1 indicates equality between two sub-indicators, while a preference of 9 indicates that the sub-indicator is 9 times more important than the other one. The results are represented in a comparison matrix (**Table 20**), where $A_{ii} = 1$ and $A_{ij} = 1 / A_{ji}$.

Table 20. Comparison matrix of eight TAI sub-indicators

Objective	Patents	Royalties	Internet	Tech exports	Telephone	Electricity	Schooling	University
Patents	1	2	3	2	5	5	1	3
Royalties	1/2	1	2	1/2	4	4	½	3
Internet	1/3	½	1	1/4	2	2	1/5	1/2
Tech. exports	1/2	2	4	1	4	4	1/2	3
Telephones	1/5	1/4	1/2	1/4	1	1	1/5	1/2
Electricity	1/5	¼	1/2	1/4	1	1	1/5	1/2
Schooling	1	2	5	2	5	5	1	4
University	1/3	1/3	2	1/3	2	2	1/4	1

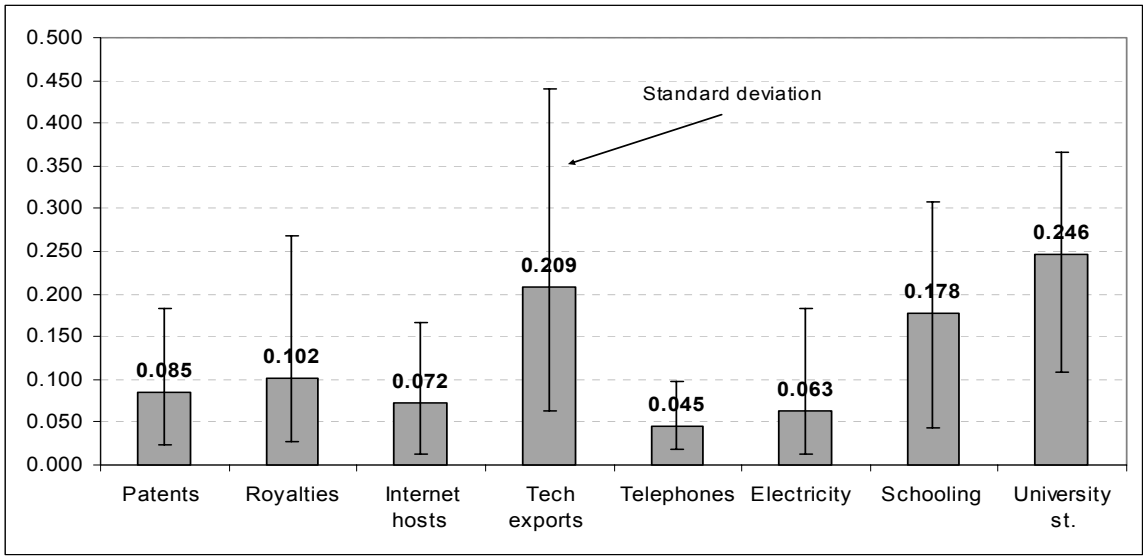
For the example, *patents* is three times more important than *Internet*. Each judgement reflects the perception of the relative contributions (weights) of the two sub-indicators to the overall objective (**Table 21**).

Table 21. Comparison matrix of three TAI sub-indicators

Objective	Patents	Royalties	Internet
Patents	w_P/w_P	w_P/w_{ROY}	w_P/w_I
Royalties	w_{ROY}/w_P	w_{ROY}/w_{ROY}	w_{ROY}/w_I
Internet	w_I/w_P	w_I/w_{ROY}	w_I/w_I

The relative weights of the sub-indicators are calculated using an eigenvector. This method enables to check the consistency of the comparison matrix through the calculation of the eigenvalues. **Figure 17** shows the results of the evaluation process and the weights, together with the corresponding standard deviation[24].

Figure 17. Analytic hierarchy process (AHP) weighting of TAI



Note. Average weight (bold) and standard deviation.

People's beliefs however are not always consistent. For example, if one claims that A is much more important than B, B slightly more important than C, and C slightly more important than A, his/her judgement is inconsistent and the results are less trustworthy. Inconsistency, however, is part of the human nature. Therefore it could be enough to measure the degree of inconsistency, such that results could be acceptable in the public eye. For a matrix of size Q × Q, only Q–1 comparisons are required to establish weights for Q indicators. The actual number of comparisons performed in AHP is Q(Q–1)/2. This is computationally costly, but results in a set of weights that is less sensitive to errors of judgement. In addition, the redundancy allows for a measure of judgement errors, an inconsistency ratio. Small inconsistency ratios--the suggested rule-of-thumb is less than 0.1, although 0.2 is often cited--do no drastically affect the weights (Saaty, 1980; Karlsson, 1998).

Conjoint analysis (CA)

Merely asking respondents how much importance they attach to a sub-indicator is unlikely to yield effective "willingness to pay" valuations. These can be inferred by using conjoint analysis (CA) from respondents' ranking of alternative scenarios (Hair et al. 1995). The conjoint analysis is a decompositional multivariate data analysis technique frequently used in marketing (McDaniel and Gates, 1998) and consumer research (Green and Srinivasan, 1978). If AHP derives the "worth" of an alternative, *summing up* the "worth" of the individual sub-indicators, the CA does the opposite, i.e. it disaggregates preferences.

This method asks for an evaluation (a preference) over a set of alternative scenarios. A scenario can be a given set of values for the sub-indicators. The preference is then decomposed by relating the single components (the known values of sub-indicators of that scenario) to the evaluation. Although this methodology uses statistical analysis to treat data, it relies on the opinion of people, *e.g.* experts, politicians, citizens, who are asked to choose which set of sub-indicators they prefer, with each person presented with different choice sets to evaluate.

The absolute value (or level) of sub-indicators could be varied both within the choice sets presented to the same individual and across individuals. A preference function would be estimated using the information coming from the different scenarios. Therefore a probability of the preference could be estimated as a function of the levels of the sub-indicators defining the alternative scenarios:

$$pref_c = P(I_{1c}, I_{2c}, ..., I_{Qc}) \tag{26}$$

where I_{qc} is the level of sub-indicator $q=1,...,Q,$ for country $c=1,...,M$. After estimating this probability (often using discrete choice models), the derivatives with respect to the sub-indicators of the preference function can be used as weights to aggregate the sub-indicators in a composite index:

$$CI_c = \sum_{q=1}^{Q} \frac{\partial P}{\partial I_{qc}} I_{qc} \tag{27}$$

The idea is to calculate the total differential of the function P at the point of indifference between alternative states of nature. Solving for the sub-indicator $q,$ one obtains the marginal rate of substitution of I_{qc}. Therefore $\partial P / \partial I_{qc}$ (thus the weight) indicates a trade-off--how the preference changes with the change of the indicator. This implies compensability among indicators, i.e. the possibility of offsetting the lack in some dimension with an outstanding performance in another dimension. This is an important feature of this method, and should be carefully evaluated vis-à-vis the objectives of the whole analysis. For example, compensability might not be desirable when dealing with environmental issues.

Performance of the different weighting methods

The weights for the TAI example are calculated using different weighting methods -- equal weighting (EW), factor analysis (FA), budget allocation (BAL), and analytical hierarchy process (AHP) (**Table 22**). The diversity in the resulting weights from applying different methods is noteworthy. Clearly with each method, different sub-indicators are evaluated in a different way. Patents, for example, are worth 17% of the weight according to the FA, but only 9% according to the AHP. This deeply influences the variability of each country's ranking (**Table 23**). For example, Korea ranks second with the AHP, but only fifth, when EW or FA are used. AHP assigns high weights (more than 20%) to two indicators, *High tech exports* and *University enrolment ratio*, for which Korea has higher scores for one or both indicators, compared to the United States, Sweden or Japan.The role of the variability in the weights and their influence in the value of the composite are discussed in the section on sensitivity analysis.

Table 22. TAI weights based on different methods

Equal weighting (EW), factor analysis (FA), budget allocation (BAL), and analytic hierarchy process (AHP)

Methods/ Sub-indicators	Patents	Royalties	Internet	Tech exports	Telephones	Electricity	Schooling	University
EW	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
FA	0.17	0.15	0.11	0.06	0.08	0.13	0.13	0.17
BAL	0.11	0.11	0.11	0.18	0.10	0.06	0.15	0.18
AHP	0.09	0.10	0.07	0.21	0.05	0.06	0.18	0.25

Table 23. TAI country rankings based on different weighting methods

Weighting method/Country	**EW**	**FA**	**BOD**	**BAL**	**AHP**
Finland	1	1	1	1	1
United States	2	2	1	2	3
Sweden	3	3	1	3	4
Japan	4	4	4	5	5
Korea, Rep. of	5	5	6	4	2
Netherlands	6	6	7	8	11
United Kingdom	7	8	9	7	7
Singapore	8	11	12	6	6
Canada	9	10	8	10	10
Australia	10	7	10	11	9
Germany	11	12	11	9	8
Norway	12	9	5	13	16
Ireland	13	14	13	12	12
Belgium	14	15	15	14	13
New Zealand	15	13	14	17	18
Austria	16	16	16	15	15
France	17	17	17	16	14
Israel	18	18	18	18	17
Spain	19	19	20	19	19
Italy	20	20	19	21	21
Czech Republic	21	21	21	22	22
Hungary	22	23	23	20	20
Slovenia	23	22	22	23	23

Note: e.g., the United States ranks first according to BOD, second according to EW, FA, and BAL and third according to AHP.

Table 24. Advantages and disadvantages of different weighting methods

Advantages	Disadvantages
Benefit of the doubt (BOD) -- *e.g.*, Human Development Index (Mahlberg and Obersteiner, 2001); Sustainable Development (Cherchye and Kuosmanen, 2002); Social Inclusion (Cherchye, Moesen, Van Puyenbroeck, 2004); Macro-economic performance evaluation (Melyn and Moesen, 1991, and Cherchye 2001); Unemployment (Storrie and Bjurek, 1999, and 2000).	
The indicator is sensible to national policy priorities, in that the weights are endogenously determined by the observed performances (this is a useful second best approach whenever the first best – full information about true policy priorities- can not be attained). The benchmark is not based upon theoretical bounds, but a linear combination of observed best performances. It is useful in policy arena, since policy makers could not complain about unfair weighting: any other weighting scheme would have generated lower composite scores. Such an index could be "incentive generating" rather than "punishing" the countries lagging behind. Weights, by revealing information about the policy priorities, may help to define trade-offs, overcoming the difficulties of linear aggregations.	Weights are country specific, thus cross-country comparisons is not possible. Without imposing constraints on weights (except the non-negativity) the most likely solution is to have all countries with a composite equal to 1. It may happen that there exist a multiplicity of solutions making the optimal set of weights undetermined (this is likely to happen when the CI=1). Different normalisation of the scores is likely to give different weighting schemes. The index is likely to reward the status-quo, since for each country the maximisation problem gives higher weights to higher scores. Endogenous weighting has the risk of substituting open experts' opinions with the analyst's manipulation of weights (through the constraints). Transparency of the procedure would be lost. The value of the scoreboard depends on the benchmark performance. If this changes the composite will change as well as the set of weights (and the country ranking). Moreover if one country is "lazy" in all but one variable where it is the best performing, then it could be a benchmark in the frontier The best performer (the one with a composite equal to one) will not see its progress reflected in the composite (that will remain stacked to 1). This can be solved by imposing an external benchmark.
Unobserved Components Models -- *e.g.*, Governance indicators (see Kaufmann, Kraay and Zoid-Iobatón, 1999 and 2003)	
Weights do not depend on ad hoc restrictions. It can be used even if component indicators are not correlated.	Reliability and robustness of results depend on the availability of enough data. With highly correlated sub-indicators there could be identification problems. Thus the method is likely to work well with independent sub-indicators. The method rewards the absence of outliers, given that weights are a decreasing function of the variance of sub-indicators. If each country has a different number of sub-indicators; weights are hence country specific
Budget Allocation -- *e.g.*, Employment Outlook (OECD,1999); Composite Indicator on e-Business Readiness (EC-JRC, 2004b);National Health Care System Performance (King's Fund., 2001);Eco-indicator 99 (Pré-Consultants NL, 2000) (weights based on survey from experts); Overall Health System Attainment (WHO, 2000) (weights based on survey from experts).	
Weighting is based on experts' opinion and not on technical manipulations. Experts' opinions are likely to increase the legitimacy of the composite and create a forum of discussion around which to form a consensus for policy action.	Weighting reliability. Weights could reflect specific local conditions (e.g. in environmental problems), so expert weighting may not be transferable from one area to another. Allocating a certain budget over a too large number of indicators can give serious cognitive stress to the experts, as it implies circular thinking. The method is likely to produce inconsistencies for a number of indicators higher than 10. The weighting may not measure the importance of each sub-indicator but rather the urgency or need for political intervention in the dimension of the sub-indicator concerned (e.g. more weight on Ozone emissions if the expert feels that not enough has been made to abate them).
Public Opinion – *e.g.* **c**oncern about environmental problems Index (Parker, 1991)	

Deals with issues on the public agenda. Allows all stakeholders to express their preference, and creates a consensus for policy actions.	Implies the measurement of "concern" . The method could produce inconsistencies when dealing with high number of indicators.
Analytic Hierarchy Process -- *e.g.*, Index of Environmental Friendliness, (Puolamaa *et al.*, 1996).	
The method can be used both for qualitative and quantitative data. The transparency of the composite is higher	The method requires a high number of pairwise comparisons and thus it can be computationally costly. The results depends on the set of evaluators chosen and the setting of the experiment
Conjoint Analysis – *e.g.*, indicator of quality of life in the city of Istanbul (Ülengin et al. 2001); advocated by Kahn (1998) and Kahn and Maynard (1996) for environmental applications.	
Weights represent trade-offs across indicators It takes into account the socio-political context, and the values of respondents.	It needs a pre-specified utility function and it implies compensability. Depends on the sample of respondents chosen and on how questions are framed. It requires a large sample of respondents and each respondent may be required to express a large number of preferences. The estimation process is complex.

Additive aggregation methods

The simplest additive aggregation method entails the calculation of the ranking of each country according to each sub-indicator and summation of the resulting ranking, e.g., Information and Communication Technologies Index (Fagerberg, 2001). The method is based on ordinal information (the so called Borda rule). It is simple and independent of outliers. But, the absolute value of information is lost.

$$CI_c = \sum_{q=1}^{Q} Rank_{qc} \quad \text{for } c=1,\ldots,M. \tag{28}$$

The second method is based on the number of indicators that are above and below some benchmark. This method uses nominal scores for each indicator to calculate the difference between the number of indicators that are above and below an arbitrarily defined threshold around the mean, e.g., the Innovation Scoreboard (European Commission, 2001a).

$$CI_c = \sum_{q=1}^{Q} \cdot sgn\left[\frac{I_{qc}}{I_{EUq}} - (1+p)\right] \quad \text{for } c=1,\ldots,M. \tag{29}$$

The threshold value p can be arbitrarily chosen above or below the mean. As with the preceding method, it is simple and unaffected by outliers. However, the interval level information is lost. For example, assume that the value of indicator I for country a is 30% above the mean and the value for country b is 25% above the mean, with a threshold of 20% above the mean. Both country a and b are then counted equally as 'above average', in spite of a having a higher score than b.

By far, the most widespread linear aggregation is the summation of weighted and normalised sub-indicators:

$$CI_c = \sum_{q=1}^{Q} w_q I_{qc} \tag{30}$$

with $\sum_q w_q = 1$ and $0 \le w_q \le 1$, for all $q=1,..,Q$ and $c=1,...,M$.

Although widely used, this aggregation imposes restrictions on the nature of sub-indicators. In particular obtaining a meaningful composite indicator depends on the quality of the underlying sub-indicators and the unit of measurement of these sub-indexes. Furthermore, additive aggregations have important implications on the interpretation of weights.

When using a linear additive aggregation technique, a necessary and sufficient condition for the existence of a proper composite indicator is *preference independence*: given the sub-indicators $\{x_1, x_2, ..., x_Q\}$, an additive aggregation function exists if and only if these indicators are mutually preferentially independent[25] (Debreu, 1960; Keeney and Raiffa, 1976; Krantz et al., 1971).

Preferential independence is a very strong condition, as it implies that the trade-off ratio between two variables $S_{x,y}$ is independent of the values of the *Q-2* other variables, (Ting, 1971)[26]. From an operational point of view, this means that an additive aggregation function permits the assessment of the marginal contribution of each variable separately. These marginal contributions can then be added together to yield a total value. If, for example, environmental dimensions are involved, the use of a linear aggregation procedure implies that among the different aspects of an ecosystem, there are no synergies or conflict. This appears to be quite an unrealistic assumption (Funtowicz et al., 1990). For example, *"laboratory experiments made clear that the combined impact of the acidifying substances SO_2, NO_X, NH_3 and O_3 on plant growth is substantially more severe that the (linear) addition of the impacts of each of these substances alone would be."* (Dietz and van der Straaten, 1992). Additive aggregation thus could result in a biased composite indicator, i.e. it will not entirely reflect the information of its sub-indicators. The dimension and the direction of the error are not easily determined, and the composite can not be adjusted properly.

Non-compensatory multicriteria approach (MCA)

As a common practice, greater weight could be given to components which are considered to be more significant in the context of the particular composite indicator, (OECD, 2003). Yet, it can be shown that when using an additive or a multiplicative aggregation rule and sub-indicators are expressed as intensities (e.g. in pounds, litres or euro and not qualities – e.g. good, bad, medium – or in rankings) the substitution rates equal the weights of the variables up to a multiplicative coefficient[27] (Munda and Nardo 2003). As a consequence, weights in additive aggregations necessarily have the meaning of substitution rates (*trade-offs*) and do not indicate the importance of the indicator associated. This implies a compensatory logic, i.e. the possibility of offsetting a disadvantage on some variables by a sufficiently large advantage on other variables. For example, in the construction of the TAI index a compensatory logic (using equal weighting) would imply that one is willing to renounce, let's say, to 2% of *Patents granted to residents,* or to 2% of *University enrolment* in exchange of a 2% increase in *Electricity consumption*.

The implication is the existence of a theoretical inconsistency in the way weights are actually used and their real theoretical meaning. For the weights to be interpreted as "*importance coefficients*" (the greatest weight is placed beside the most important "dimension") non-compensatory aggregation procedures must be used to construct composite indicators (Podinovskii, 1994). This can be done using a non-compensatory multi-criteria approach.

When various variables are used to evaluate a set of countries, some of these variables may be in favour of one country while other variables may be in favour of another. As a consequence a conflict among the variables could arise. This conflict can be treated in the light of a non-compensatory logic by

taking into account the absence of preferential independence within a discrete non-compensatory multi-criteria approach (NCMC) (Munda, 1995; Roy, 1996; Vincke, 1992).

Given a set of sub-indicators $\boldsymbol{G}=\{x_q\}$, $q=1,...,Q$, and a finite set $\boldsymbol{M}=\{c\}$, $c=1,...,M$ of countries, assume that the evaluation of each country c with respect to an individual indicator x_q (i.e. the indicator score or variable) is based on an *interval or ratio* scale of measurement. For simplicity of exposition, it is also assumed that a higher value of an individual indicator is preferred to a lower one, i.e., the higher, the better. Further assume, the existence of a set of weights $w=\{w_q\}$, $q=1,2,...Q$, with $\sum_{q=1}^{Q} w_q = 1$, interpreted as *importance coefficients*. This information constitutes the impact matrix. For explanatory purposes, suppose only 5 of the countries included in the TAI dataset[28] and equal weights are given to all of the sub-indicators (**Table 25**).

Table 25. Impact matrix for TAI (five countries)

	Patents	Royalties	Internet	Tech exports	Telephones	Electricity	Schooling	University
Finland	187	125.6	200.2	50.7	3.080	4.150	10	27.4
USA	289	130	179.1	66.2	2.997	4.073	12	13.9
Sweden	271	156.6	125.8	59.7	3.096	4.145	11.4	15.3
Japan	994	64.6	49	80.8	3.003	3.865	9.5	10
Korea	779	9.8	4.8	66.7	2.972	3.653	10.8	23.2
weight	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

The mathematical problem is then how to use this information to rank all the countries from the best to the worst one in a complete pre-order (i.e. without any incomparability relation, see Roubens and Vincke, 1985). The following are important:

- Intensity of preference (how much country a is better than country b according to sub-indicator q);
- Number of indicators in favour of a given country;
- Weight attached to each indicator;
- Relationship of each country with respect to all the others.

The sources of uncertainty and imprecise assessment should be reduced as much as possible. Unfortunately Arrow's impossibility theorem (Arrow, 1963) clearly shows that no perfect aggregation convention can exist. Therefore, when aggregating, it is essential to check not only which properties are respected by a given ranking procedure, but also that the essential properties the specific problem are not lost.

The mathematical aggregation convention can be divided into two main steps:

- Pair-wise comparison of countries according to the whole set of sub-indicators used.
- Ranking of countries in a complete pre-order.

The first step results in a $M \times M$ matrix, E, called *outranking matrix* (Arrow and Raynaud, 1986, Roy, 1996). Any generic element of E: e_{jk} , $j \neq k$ is the result of the pair-wise comparison, according to all the Q sub-indicators, between countries j and k. Such a global pair-wise comparison is obtained by means of equation:

$$e_{jk} = \sum_{q=1}^{Q} (w_q(Pr_{jk}) + \frac{1}{2} w_q(In_{jk})) \quad (31)$$

where $w_q(Pr_{jk})$ and $w_q(In_{jk})$ are the weights of sub-indicators presenting a preference and an indifference relation respectively. In other words, the score of country *j* is the sum of the weights of sub-indicators, for which this country does better than country *I,* as well as – if any – half of the weights for the sub-indicators according to which the two countries do equally well (**31)**. $e_{jk} + e_{kj} = 1$ clearly holds.

The pair-wise comparisons are different from those in the AHP method – which belongs to the set of compensatory multicriteria methods, jointly with CA. In the latter, the question to be answered was whether I_q is more important than I_z, here, instead, the question is whether I_q is higher for country *a* or for country *b*. And if I_q is indeed higher for country *a* , it is the weight of sub-indicator *q*, which enters into the computation of the overall importance of country *a*, in a way consistent with the definition of weights as importance measures.

In the TAI example, the pair-wise comparison of such as Finland and the United States shows that Finland has better scores for the sub-indicators--*Internet* (weight 1/8), *Telephones* (weight 1/8), *Electricity* (weight 1/8) and *University* (weight 1/8). Thus the score for Finland is 4*1/8=0.5, while the complement to one is the score of the US. The resulting outranking matrix is (**Table 26**):

Table 26. Outranking impact matrix for TAI (five countries)

	Finland	USA	Sweden	Japan	Korea
Finland	0	0.5	0.375	0.75	0.625
USA	0.5	0	0.5	0.625	0.625
Sweden	0.625	0.5	0	0.75	0.625
Japan	0.25	0.375	0.25	0	0.75
Korea	0.375	0.375	0.375	0.25	0

The way the information are combined generates several possible ranking procedures (Young, 1988; Munda, 2004), each with pros and cons. One possible algorithm is the Condorcet-Kemeny-Young-Levenglick (CKYL) ranking procedure (Munda and Nardo 2003). According to CKYL, the ranking of countries with the highest likelihood is the one supported by the maximum number of sub-indicators for each pair-wise comparison, summed over all pairs of countries considered. More formally, all the *M(M–1)* pair-wise comparisons compose the outranking matrix *E*. Call *R* the set of all *M!* possible complete rankings of alternatives, $R=\{r_S\}$, $s=1,2,...,M!$. For each r_S, compute the corresponding score φ_s as the summation of e_{jk} over all the $\binom{M}{2}$ pairs *j,k* of alternatives.

That is, $\varphi_s = \sum e_{jk}$ *where* $j \neq k$, $s = 1,2,...M!$ *and* $e_{jk} \in r_s$ .The final ranking ($r*$) is the solution of:

$$r* \Leftrightarrow \varphi_* = \max \sum e_{jk} \quad \text{where } e_{jk} \in R \quad (32)$$

In the TAI example, the number of permutations obtained from 5 countries are 120, the first 5 are listed in **Table 27**. For example, the score of the first ranking (USA, Sweden, Finland, Japan and Korea) is obtained as follows: according to the impact matrix the comparison of US with the other countries yields

0.5 against Finland and Sweden, and 0.625 against Japan and Korea (overall 2.25). The comparison of Sweden yields 0.625 against Finland and Korea and 0.75 against Japan (overall 2). Finland obtains 0.625 against Korea and 0.75 against Japan (overall 1.375). Finally Japan obtains 0.75 against Korea. The final score of this ranking is then equal to 2.25+2+1.375+0.75=6.375.

Table 27. Permutations obtained from the outranking matrix for TAI and associated score

USA	Sweden	Finland	Japan	Korea	**6.375**
Sweden	Finland	USA	Japan	Korea	**6.375**
Sweden	USA	Finland	Japan	Korea	**6.375**
Finland	USA	Sweden	Japan	Korea	6.125
Finland	Sweden	USA	Japan	Korea	6.125
USA	Finland	Sweden	Japan	Korea	6.125

According to expression (32) the final ranking will be the permutation(s) with the highest score. In our example the first 3 permutations have the highest overall score, and thus all those can be considered as a winning ranking.

This aggregation method has the advantage to overcome some of the problems raised by additive or multiplicative aggregations, e.g., preference dependence, the use of different ratio or interval scale to express the same indicator and the meaning of trade-offs given to the weights. With this method, moreover, qualitative and quantitative information can be jointly treated. In addition, it does not need any manipulation or normalisation to assure the comparability of sub-indicators. The drawbacks on the other hand include the dependence of irrelevant alternatives, i.e. the possible presence of cycles/rank reversal in which in the final ranking, country *a* is preferred to *b*, *b* is preferred to *c* but *c* is preferred to *a* (the same problem highlighted for AHP with indicators). Furthermore, information on intensity of preference of variables is never used: if one indicator for country *a* is much less than the same indicator for country *b* produces the same ranking as the case in which this difference is very small[29]. Notice that with this method, the focal point is shifted to the determination of weights, which is crucial for the result[30].

Geometric aggregation

An undesirable feature of additive aggregations is the implied full compensability, such that poor performance in some indicators can be compensated by sufficiently high values of other indicators. For example if an hypothetical composite were formed by inequality, environmental degradation, GDP per capita and unemployment, two countries, one with values 21, 1, 1, 1; and the other with 6,6,6,6 would have equal composite if the aggregation is additive and EW is applied. Obviously the two countries would represent very different social conditions that would not be reflected in the composite. If multi-criteria analysis entails full non-compensability, the use of a geometric aggregation (also called deprivational index) $CI_c = \prod_{q=1}^{Q} x_{q,c}^{w_q}$ is an in-between solution[31].

In the example above, the first country would have a much lower composite than the second, if the aggregation is geometric (2.14 for the first and 6 for the second). In a benchmarking exercise, countries with low scores in some sub-indicators thus would prefer a linear rather than a geometric aggregation. On the other hand, the marginal utility of an increase in the score would be much higher when the absolute value of the score is low: the first country increasing the second indicator by 1 unit would increase its composite from 2.14 to 2.54, while country 2 would go from 6 to 6.23. In other terms, the first country would increase its composite by 19% while the second only by 4%. Consequently, a country should be more interested in increasing those sectors/activities/alternatives with the lowest score in order to have a

higher chance to improve its position in the ranking, if the aggregation is geometric rather than linear (Zimmermann and Zysno, 1983).

Furthermore, the type of aggregation employed is strongly related with the method used to normalise raw data (*Step 5*). In particular, Ebert and Welsch (2004) has shown that the use of linear aggregations yields meaningful composite indicators, only if data are all expressed in partially comparable interval scale (i.e. temperature in Celsius of Fahrenheit) of type $f : x \to \alpha x + \beta_i \quad \alpha > 0$ (i.e. α fixed, but β_i varying across subindicators) or in a fully comparable interval scale (β constant); Non-comparable data measured in ratio scale (i.e. kilograms and pounds) $f : x \to \alpha_i x$ where $\alpha_i > 0$ (i.e. α_i varying across sub-indicators) can only be meaningfully aggregated by using geometric functions, provided that x is strictly positive. In other terms, except in the case of all indicators measured in different ratio scale, the measurement scale must be the same for all indicators when aggregating. Thus, care should be given when indicators measured in different scale coexist in the same composite. The normalisation method should be properly used to remove the scale effect.

Table 28 highlights the dependence of rankings to the aggregation methods used (in this case linear, geometric and based on the multi-criteria technique for the TAI dataset with 23 countries). Although in all cases, equal weighting is used, the resulting rankings are very different. For example, Finland ranks first according to the linear aggregation, second according to the geometric aggregation and third according to the multi-criteria. Note that Korea ranks sixteenth with GME, while its ranking is much higher according to the other two methods, while the reverse is true for Belgium.

Table 28. TAI country rankings by different aggregation methods

	Position in the ranking		
	LIN	**NCMC**	**GME**
Finland	1	3	2
United States	2	1	1
Sweden	3	2	3
Japan	4	4	4
Korea, Rep. of	5	9	16
Netherlands	6	8	5
United Kingdom	7	5	6
Singapore	8	12	18
Canada	9	11	13
Australia	10	9	14
Germany	11	7	8
Norway	12	6	11
Ireland	13	13	7
Belgium	14	17	9
New Zealand	15	15	17
Austria	16	15	12
France	17	14	10
Israel	18	18	15
Spain	19	20	19
Italy	20	19	21
Czech Republic	21	21	23
Hungary	22	23	22
Slovenia	23	22	20

Note: Dataset TAI, for 23 countries. Numbers refer to the position in ranking.

UNCERTAINTY AND SENSITIVITY ANALYSIS

Sensitivity analysis for modelers? Would you go to an orthopaedist who didn't use X-rays? This sentence by J.M. Furbringer is the essence of what sensitivity analysis is and why it is useful. The composite indicators development involves stages where subjective judgment has to be made: the selection of sub-indicators, the treatment of missing values, the choice of aggregation model, the weights of the indicators, etc. All these subjective choices are the *bones* of the composite indicator and shape, together with the information provided by the numbers themselves, the message communicated by the composite indicator. However, composite indicators can send misleading or non-robust policy messages if they are poorly constructed or misinterpreted.

Since the quality of a model also depends on the soundness of its assumptions, good modeling practice requires that the modeler provides an evaluation of the confidence in the model, assessing the uncertainties associated to the modeling process and the subjective choices undertaken. This is what sensitivity analysis does: it performs the 'X-rays' of the model by studying the relationship between information flowing in and out of the model.

More formally, sensitivity analysis is the study of how the variation in the output can be apportioned, qualitatively or quantitatively, to different sources of variation in the assumptions, and of how the given composite indicator depends upon the information fed into it. Sensitivity analysis is thus closely related to uncertainty analysis which aims to quantify the overall uncertainty in the countries' ranking as a result of the uncertainties in the model input. A combination of uncertainty and sensitivity analysis can help to gauge the robustness of the composite indicator ranking, to increase its transparency, to identify which countries are favoured or deteriorate under certain assumptions and to help framing a debate around the Index.

Next we describe how to apply uncertainty and sensitivity analysis to composite indicators. Our synergistic use of uncertainty and sensitivity analysis has recently been applied for the robustness assessment of composite indicators (Saisana *et al.* 2005a, Saltelli *et al.*, 2004) and has proven to be useful in dissipating some of the controversy surrounding composite indicators such as the Environmental Sustainability Index (Saisana *et al.* 2005b).

In the TAI case study we focus on 5 main uncertainties/assumptions: inclusion-exclusion of one indicator at-a-time, imputation of missing data, different normalisation methods, different weighting schemes and different aggregation schemes.

Let *CI* be the index value for country *c, c=1,…,M,*

$$CI_c = f_{rs}\left(I_{1,c}, I_{2,c}, ...I_{Q,c}, w_{s,1}, w_{s,2}, ...w_{s,Q}\right) \tag{33}$$

according the weighting model $f_{rs}, r = 1,2,3, \quad s = 1,2,3$, where the index *r* refers to the aggregation system (LIN, GME, NCMC) and index *s* refers to the weighting scheme (BAL, AHP, BOD). The index is based on Q normalised sub-indicators $I_{1,c}, I_{2,c}, ...I_{Q,c}$ for that country and scheme-dependent weights $w_{s,1}, w_{s,2}, ...w_{s,Q}$ for the sub-indicators. The most frequently used normalisation methods for the sub-indicators are based on the re-scaled (34a), standardised (34b), or on the raw indicator values (34c).

$$\begin{cases} I_{q,c} = \dfrac{x_{q,c} - \min(x_q)}{range(x_q)} & (34a) \\ I_{q,c} = \dfrac{x_{q,c} - mean(x_q)}{std(x_q)} & (34b) \\ I_{q,c} = x_{q,c} & (34c) \end{cases}$$

where $I_{q,c}$ is the normalised and $x_{q,c}$ is the raw value of the sub-indicator xq for country c.

Note that the re-scaled value (34a) can be used in conjunction with all the weighting schemes (BAL, AHP and BOD) for all aggregation systems (LIN, GME, NCMC). The standardised value (34b) can be used with weighting schemes (BAL, AHP) for aggregation systems (LIN, NCMC). And the raw indicator value (34c) can be used with weighting schemes (BAL, AHP) for aggregation systems (GME, NCMC).

The rank assigned by the composite indicator to a given country, i.e. $Rank(CI_c)$ is an output of the uncertainty/sensitivity analysis. The average shift in country rankings is also explored. This latter statistic captures the relative shift in the position of the entire system of countries in a single number. It can be calculated, as the average of the absolute differences in countries' rank with respect to a reference ranking over the *M* countries:

$$\overline{R}_S = \frac{1}{M}\sum_{c=1}^{M}\left|Rank_{ref}(CI_c) - Rank(CI_c)\right| \quad (35)$$

The reference ranking for the TAI analysis is the original rank given to the country by the original version of the index. The investigation of $Rank(CI_c)$ and $\overline{R}_S$ is the scope of the uncertainty and sensitivity analysis[32].

General framework

The analysis is conducted as a single Monte Carlo experiment, e.g. by eliminating all uncertainty sources simultaneously to capture all possible synergy effects among uncertain input factors. This involves the use of triggers, e.g. the use of uncertain input factors to decide which aggregation system and weighting scheme to adopt. A discrete uncertain factor, which can take integer values between 1 and 3 is used for the aggregation system and similarly for the weighting scheme. Other trigger factors are generated to select indicators to be omitted, the editing scheme, the normalisation scheme and so on, until a full set of input variables is available to compute $Rank(CI_c), \overline{R}_S$.

Uncertainty analysis (UA)

Various components of the CI construction process can introduce uncertainty in the output variables, $Rank(CI_c)$ and $\overline{R}_S$. The UA is essentially based on simulations that are carried on various equations that constitute the underlying *model*. The uncertainties are transferred into a set of scalar input factors, such that the resulting $Rank(CI_c)$ and $\overline{R}_S$ are non-linear functions of the uncertain input factors, and the estimated probability distribution (pdf) of $Rank(CI_c)$ and $\overline{R}_S$. Various methods are available for evaluating output

uncertainty. The following is the Monte Carlo approach, which is based on multiple evaluations of the model with k randomly selected model input factors. The procedure has six steps:

Step 1. Assign a pdf to each input factor X_i, $i = 1,2...k$. The first input factor, X_1 is used for the selection of the editing scheme (for the second TAI analysis only):

X_1	Estimation of missing data
1	Use bivariate correlation to impute missing data
2	Assign zero to missing datum

The second input factor X_2 is the trigger to select the normalisation method.

X_2	Normalisation
1	Rescaling (Equation 7.3a)
2	Standardisation (Equation 7.3b)
3	None (Equation 7.3c)

Both X_1 and X_2 are discrete random variables. In practice, they are generated by drawing a random number ζ, uniformly distributed between [0,1] and applying the so called *Russian roulette algorithm*, e.g. for X_1, select 1 if $\zeta \in [0,0.5)$ and 2 if $\zeta \in [0.5,1]$. Uncertain factor X_3 is generated to select which sub-indicator, if any, should be omitted.

ζ	X_3, excluded sub-indicator
$[0, \frac{1}{Q+1})$	None ($X_3 = 0$) all subindicators are used)
$[\frac{1}{Q+1}, \frac{2}{Q+1})$	$X_3 = 1$
…	…
$[\frac{Q}{Q+1}, 1]$	$X_3 = Q$

That is with probability $\frac{1}{Q+1}$ no sub-indicator will be excluded, while with probability $[1-\frac{1}{Q+1}]$ one of the Q sub-indicators will be excluded with equal probability. Clearly, one could have made the probability of $X_3 = 0$ larger or smaller than $\frac{1}{Q+1}$ and still sample the values $X_3 = 1,2,...Q$ with equal probability. A scatter-plot based sensitivity analysis would be used to track which indicator affects the output the most when excluded. Also recall that whenever a sub-indicator is excluded, the weights of the other factors are re-scaled to 1 to make the composite index comparable if either BAL or AHP. When BOD is selected the exclusion of sub-indicator leads to a re-execution of the optimisation algorithm.

Trigger X_4 is used to select the aggregation system

X_4	Aggregation Scheme
1	LIN
2	GME
3	NCMC

Note that when LIN is selected the composite indicators are computed as:

$$CI_c = \sum_{q=1}^{Q} w_{sq} I_{q,c} \tag{36}$$

while when GME is selected they are:

$$CI_c = \prod_{q=1}^{Q} \left(I_{q,c}\right)^{w_{4,q}} \tag{37}$$

When NCMC is selected the countries are ranked directly from the outscoring matrix.

X_5 is the trigger to select the weighting scheme;

X_5	Weighting Scheme
1	BAL
2	AHP
3	BOD

The last uncertain factor X_6 is used to select the expert. In this experiment, there are 20 experts. Once an expert is selected at runtime via the trigger X_6, the weights assigned by that expert (either for the BAL or AHP schemes) are assigned to the data. Clearly the selection of the expert has no bearing when BOD is selected ($X_5 = 3$). All the same this uncertain factor would be generated at each individual Monte Carlo simulation, given the row dimension of the Monte Carlo sample (*constructive dimension*) should be fixed in a Monte Carlo experiment, i.e. even if some of the sampled factors will not be active at a particular run, they will be all the same generated by the random sample generation algorithm. The constructive dimension of the Monte Carlo experiment, the number of random numbers to be generated for each trial, is hence $k = 6$ .Note that alternative arrangements of the analysis would have been possible.

Step 2. Generate randomly *N* combinations of independent input factors $\mathbf{X}^l$, $l = 1,2,...N$ (a set $\mathbf{X}^l = X_1^l, X_2^l, \ldots, X_k^l$ of input factors is called a sample). For each trial sample $\mathbf{X}^l$ the computational model can be evaluated, generating values for the scalar output variable Y^l, where Y^l is either $Rank(CI_c)$, the value of the rank assigned by the composite indicator to each country, or $\overline{R}_S$, the averaged shift in countries' rank.

Step 3. Close the loop over *l*, and analyse the resulting output vector $\boldsymbol{Y}^l$, with $l = 1, ..., N$.

The generation of samples can be performed using various procedures, such as simple random sampling, stratified sampling, quasi-random sampling or others (Saltelli *et. al*, 2000a). The sequence of $\boldsymbol{Y}^l$ gives the pdf of the output *Y*. The characteristics of this pdf, such as the variance and higher order moments, can be estimated with an arbitrary level of precision that is related to the size of the simulation N .

Sensitivity analysis using variance-based techniques

A necessary step when designing a sensitivity analysis is to identify the output variables of interest. Ideally these should be relevant to the issue tackled by the model (Saltelli *et al.*, 2000b, 2004). It has been noted earlier that composite indicators can be considered as models. When several layers of uncertainty are simultaneously present, composite indicators could become a non-linear, possibly non-additive model. As argued by practitioners (Saltelli et al., 2000a, EPA, 2004), for non-linear models, robust, "model-free" techniques should be used for sensitivity analysis. Sensitivity analysis using variance-based techniques are model free and display additional properties convenient for the present analysis, such as:

- 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, e.g. in fractional factorial design (Box et al. 1978);
- they are quantitative, and can distinguish main effects (first order) from interaction effects (higher order);
- they are easy to interpret and to explain;
- they allow for a sensitivity analysis whereby uncertain input factors are treated in groups instead of individually;
- they can be justified in terms of rigorous settings for sensitivity analysis.

To compute a variance based sensitivity measure for a given input factor X_i, start from the fractional contribution to the model output variance, i.e. the variance of Y where Y is either $Rank(CI_c)$, and $\overline{R}_S$) due to the uncertainty in X_i :

$$V_i = V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i)) \tag{38}$$

Fix factor X_i, e.g. to a specific value x_i^* in its range, and compute the mean of the output Y averaging over all factors but factor X_i: $E_{\mathbf{X}_{-i}}(Y|X_i = x_i^*)$. Then, take the variance of the resulting function of x_i^* over all possible x_i^* values. The result is given by Equation (38), where the dependence from x_i^* has been dropped. V_i is a number between 0 (when X_i does not gives a contribution to Y at the first order), and $V(Y)$, the unconditional variance of Y , when all factors other than X_i are non influential at any order. Note that the following is always true:

$$V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i)) + E_{X_i}(V_{\mathbf{X}_{-i}}(Y|X_i)) = V(Y) \tag{39}$$

where the first term (39) is called a main effect, and the second one the residual. An important factor should have a small residual, e.g. a small value of $E_{X_i}(V_{\mathbf{X}_{-i}}(Y|X_i))$. This is intuitive as the residual

measures the expected reduced variance that one would achieve if one could fix X_i. Rewrite this as $V_{\mathbf{X}_{-i}}(Y|X_i = x_i^*)$, a variance conditional on x_i^*. Then the residual $E_{X_i}(V_{\mathbf{X}_{-i}}(Y|X_i))$ is the expected value of such conditional variance, averaged over all possible values of x_i^*. This would be small if X_i is influential. A first order sensitivity index is obtained through normalising the first-order term by the unconditional variance:

$$S_i = \frac{V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i))}{V(Y)} = \frac{V_i}{V(Y)} \tag{40}$$

One can compute conditional variances corresponding to more than one factor, e.g. for two factors X_i and X_j, the conditional variance would be $V_{X_iX_j}(E_{\mathbf{X}_{-ij}}(Y|X_i, X_j))$, and the second-order term variance contribution would become:

$$V_{ij} = V_{X_iX_j}(E_{\mathbf{X}_{-ij}}(Y|X_i, X_j)) - V_{X_i}(E_{\mathbf{X}_{-i}}(Y|X_i)) - V_{X_j}(E_{\mathbf{X}_{-j}}(Y|X_j)) \tag{41}$$

where clearly V_{ij} is different from zero only if $V_{X_iX_j}(E_{\mathbf{X}_{-ij}}(Y|X_i, X_j))$ is larger than the sum of the first-order term relative to factors X_i and X_j.

When all *k* factors are independent from one another, the sensitivity indices can be computed using the following decomposition formula for the total output variance *V(Y)* :

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \sum_i \sum_{j>i} \sum_{\substack{l>j \\ j>i}} V_{ijl} + \ldots + V_{12\ldots k} \tag{42}$$

Terms above the first order (42) are known as interactions. A model without interactions among its input factors is said to be *additive*. In this case, $\sum_{i=1}^{k} V_i = V(Y)$, $\sum_{i=1}^{k} S_i = 1$ and the first order conditional variances of equation (38) are all needed to decompose the model output variance. For a non-additive model, higher order sensitivity indices, responsible for interaction effects among sets of input factors, have to be computed. However, higher order sensitivity indices are usually not estimated, as in a model with *k* factors, the total number of indices (including the S_i 's) that needs to be estimated would be as high as 2^k-1. Instead a more compact sensitivity measure is used. The total effect sensitivity index concentrates on a single term for all the interactions, involving a given factor X_i. To give an example, for a model of *k*=3 independent factors, the three total sensitivity indices would be:

$$S_{T1} = \frac{V(Y) - V_{X_2X_3}(E_{X_1}(Y|X_2, X_3))}{V(Y)} = S_1 + S_{12} + S_{13} + S_{123} \tag{43}$$

And analogously:

$$S_{T2} = S_2 + S_{12} + S_{23} + S_{123} \tag{44}$$
$$S_{T3} = S_3 + S_{13} + S_{23} + S_{123}$$

The conditional variance $V_{X_2X_3}(E_{X_1}(Y|X_2,X_3))$ in equation (43) can be written in general terms as $V_{\mathbf{X}_{-i}}(E_{X_i}(Y|\mathbf{X}_{-i}))$ (Homma and Saltelli, 1996). This is the total contribution to the variance of *Y* due to non-X_i , i.e., to the *k*-1 remaining factors, such that $V(Y)-V_{\mathbf{X}_{-i}}(E_{X_i}(Y|\mathbf{X}_{-i}))$ includes all terms. In general, $\sum_{i=1}^{k} S_{Ti} \geq 1$.

Given (39), the total effect sensitivity index can also be written as:

$$S_{Ti} = \frac{V(Y)-V_{\mathbf{X}_{-i}}(E_{X_i}(Y|\mathbf{X}_{-i}))}{V(Y)} = \frac{E_{\mathbf{X}_{-i}}(V_{X_i}(Y|\mathbf{X}_{-i}))}{V(Y)} \qquad (45)$$

For a given factor, X_i ,a significant difference between S_{Ti} and S_i signals an important role of interaction for that factor in *Y*. Highlighting interactions among input factors helps to improve our understanding of the model structure. Estimators for both (S_i, S_{Ti}) are provided by a variety of methods reviewed in Chan *et al.* (2000). Here the method of Sobol' (1993), in its improved version due to Saltelli (2002) is used. The method of Sobol' uses quasi-random sampling of the input factors. The pair (S_i, S_{Ti}) gives a fairly good description of the model sensitivities, which for the improved Sobol' method is of $2n(k+1)$ model evaluations, where *n* represents the sample size, required to approximate the multidimensional integration implicit in the E and V operators above to a plain sum. *n* can vary in the hundred-to-thousand range.

When the uncertain input factors X_i are dependent, the output variance cannot be decomposed, as in equation (42). The S_i, S_{Ti} indices, defined by (38) and (45) are still valid sensitivity measures for X_i, though their interpretation has changed, e.g. S_i could carry over the effects of other factors that can be positively or negatively correlated to X_i (see Saltelli and Tarantola, 2002), while S_{Ti} can no longer be decomposed meaningfully into main effect and interaction effects. The S_i, S_{Ti}, for the case of non-independent input factors, could also be interpreted in as "settings" for sensitivity analysis.

A description of two settings linked to S_i, S_{Ti} is discussed below[33].

Factors' Prioritisation (FP) Setting. Suppose a factor that, once "discovered" in its true value and fixed, would reduce the most V(Y). The true values for the factors however are unknown. The best choice one can make would be the factor with the highest S_i, whether the model is additive, and the factors are independent or not.

Factors' Fixing (FF) Setting. Can one fix a factor [or a subset of input factors] at any given value over their range of uncertainty without reducing significantly the variance of the output? One can only fix those (sets of) factors whose S_{Ti} is zero.

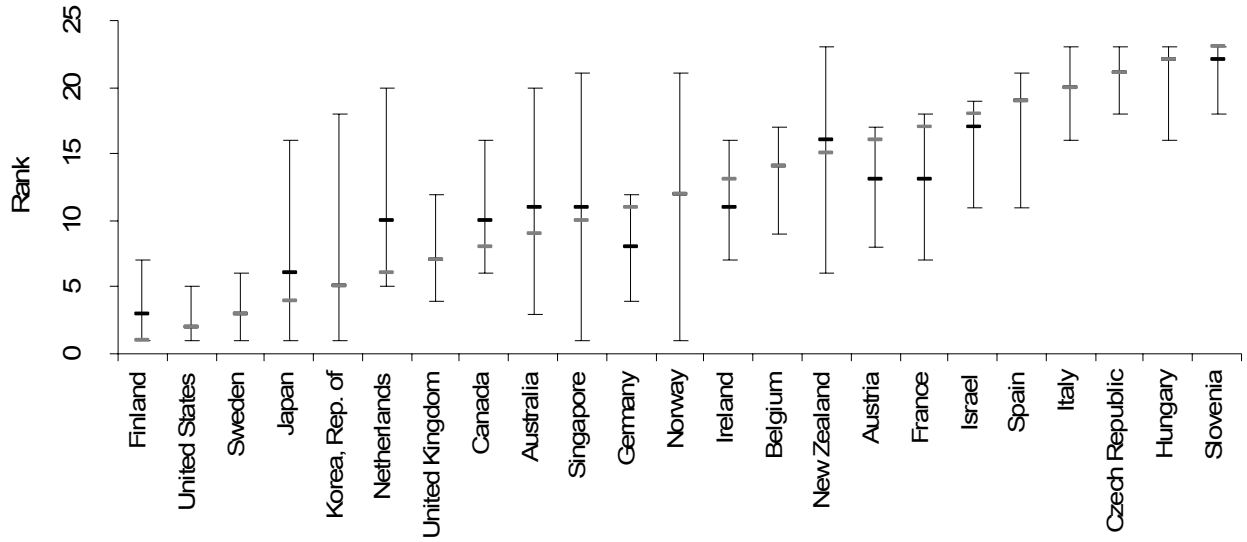
The extended variance-based methods, including the improved version of Sobol', for both dependent and independent input factors, are implemented in the freely distributed software SIMLAB (Saltelli *et al.*, 2004).

Analysis 1

The first analysis is run without imputation, i.e. by censoring all countries with missing data. As a result, only 34 countries in theory could be analysed. Other countries from rank (original TAI) 24 are also dropped, e.g., Hong Kong, as this is the first country with missing data. The analysis is restricted to the set of countries whose rank is not altered by the omission of missing records. The uncertainty analysis for the remaining 23 countries is given in **Figure 18** for the ranks, with countries ordered by their original TAI position, ranging from Finland, rank=1, to Slovenia, rank=23. Note that the choice of ranks, instead of composite indicator values, is dictated by the use of the NCMC aggregation system.

The width of the 5^{th} – 95^{th} percentile bounds and the ordering of the medians (black hyphen) often are at odds with the ordering of the original TAI (grey hyphen). Although one could still see the difference between the group of leader and that of laggards, there are considerable differences between the new and the original TAI. If the uncertainty within the system were a true reflection of the status of knowledge and the (lack of) consensus among experts on how TAI should be built, we would have to conclude that TAI is not a robust measure of country technology achievement.

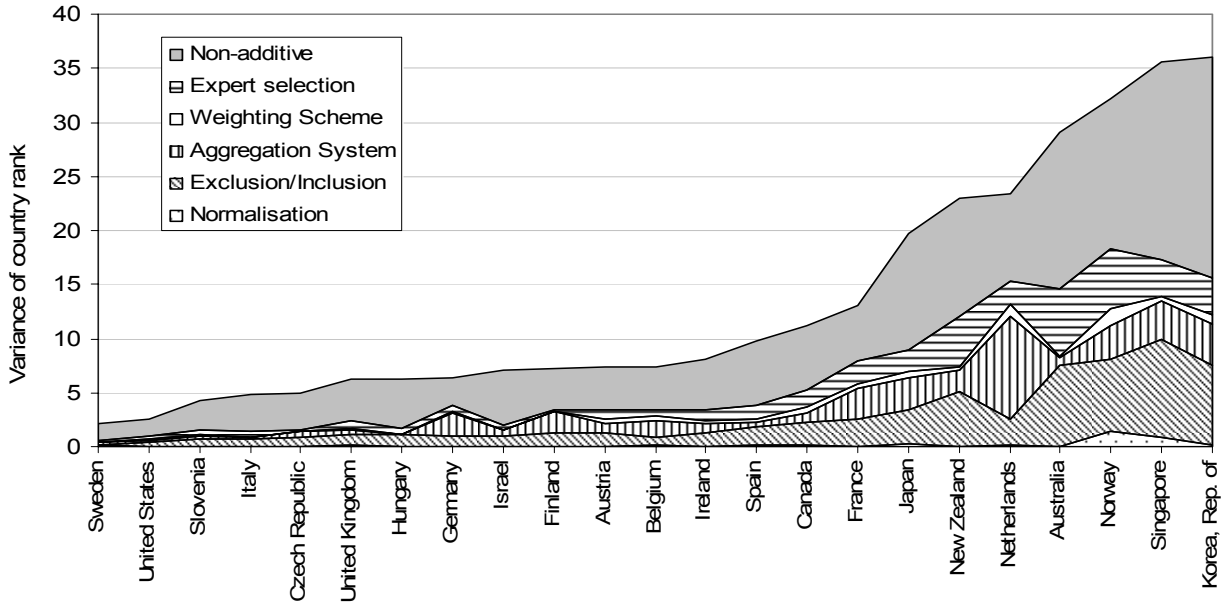
Figure 18. Uncertainty analysis of TAI country rankings



Note: Results show the country rankings according to the original TAI 2001 (light grey marks), and the median (black mark) and the corresponding 5^{th} and 95^{th} percentiles (bounds) of the distribution of the MC-TAI for 23 countries. Uncertain input factors: normalisation method, inclusion-exclusion of a sub-indicator, aggregation system, weighting scheme, expert selection. Countries are ordered according to the original TAI values.

Figure 19 shows the sensitivity analysis based on the first order indices calculated by the Sobol' (1993) method and the improved version due to Saltelli (2002). The total variance for each country's rank is presented along with the part that can be decomposed according to the first order conditional variances. The aggregation system, followed by the inclusion-exclusion of sub-indicators and expert selection are the most influential input factors. The countries with the highest total variance in ranks are the middle-performing countries, while the leaders and laggards in technology achievement have low total variance. The non-additive, non-linear part of the variance that is not explained by the first order sensitivity indices ranges from 35% for the Netherlands to 73% for the United Kingdom, whilst for most countries it exceeds 50%. This underlines the necessity of computing higher order sensitivity indices that capture the interaction effect among the input factors.

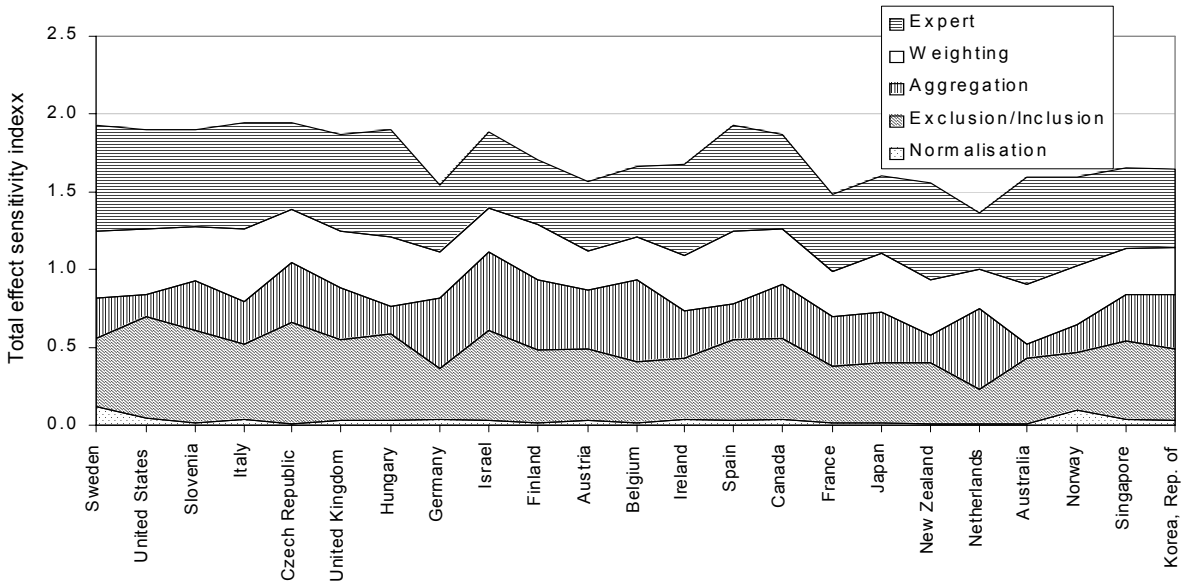
Figure 19. Sobol' sensitivity measures of first orderTAI results



Note: Results are based on the first order indices. Decomposition of country's variance according to the first order conditional variances. Aggregation system, followed by the inclusion-exclusion of sub-indicator and expert selection are the most influential input factors. The part of the variance that is not explained by the first order indices is noted as non-additive. Countries are ordered in ascending order of total variance.

Figure 20 shows the total effect sensitivity indices for the variance of each country's rank. The total effect sensitivity indices concentrate on one single term for all the interactions involving each input factor. The indices add up to a number greater than one due to existing interactions, which seem to exist among the identified influential factors.

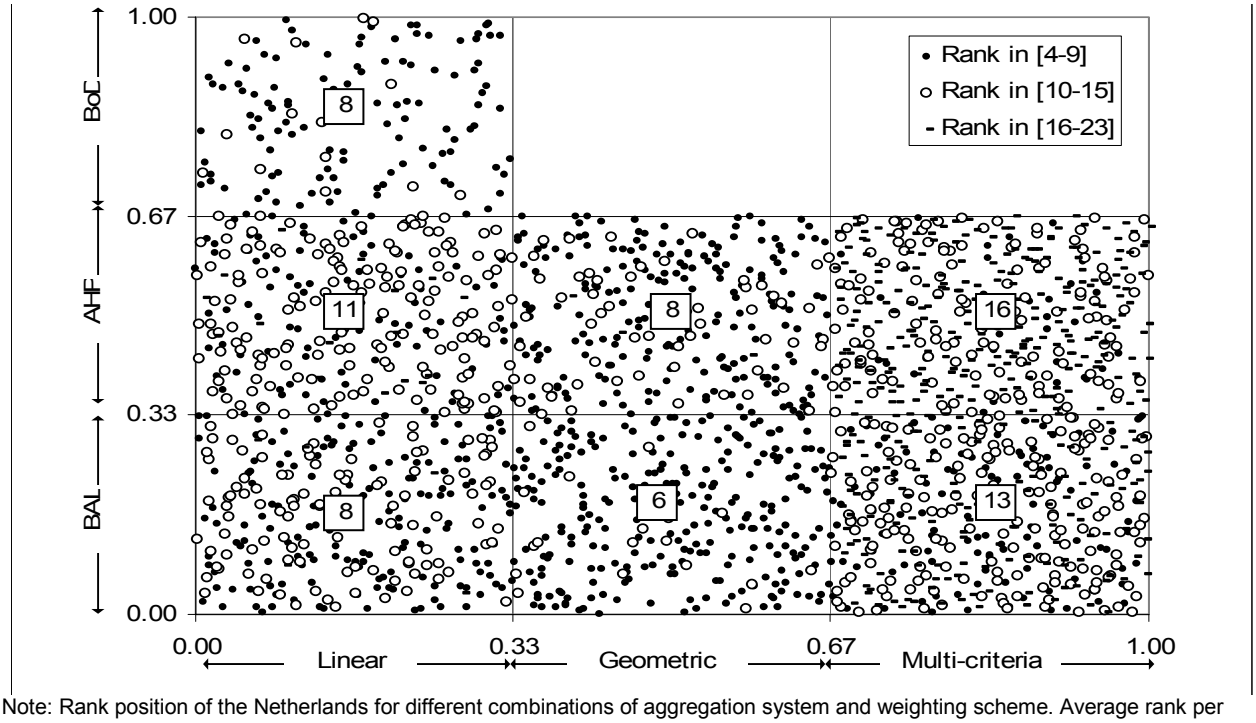
Figure 20. Sobol' sensitivity measures of TAI total effect indices



Note: Results are based on the total effect indices. Aggregation system inclusion-exclusion of sub-indicator and expert selection present most of the interaction effects. Countries are ordered in ascending order of total variance.

If the TAI model were additive with no interactions between the input factors, the non-additive part of the variance in Figure 21 would have been zero. In other words, the first order sensitivity indices would have summed to 1, and the sum of the total effect sensitivity indices in Figure 22 would have been 1. Yet, the sensitivity indices show the high degree of non-linearity and additivity for the TAI model and the importance of the interactions. The high effect of interactions for the Netherlands, which also has a large percentile bound, is further explored. **Figure 21** shows that the Netherlands is favoured by the combination of "geometric mean system" with "BAL weighing", and not favoured by the combination of "Multi criteria system" with "AHP weighting". This is a clear interaction effect. In depth analysis of the output data reveals that as far as inclusion – exclusion is concerned, it is the exclusion of the sub-indicator "Royalties" leading to worsening of the Netherlands' ranking under any aggregation system.

Figure 21. Netherlands ranking by aggregation and weighting systems



Note: Rank position of the Netherlands for different combinations of aggregation system and weighting scheme. Average rank per case is indicated in the box. The interaction effect between aggregation system and weighting scheme is clear.

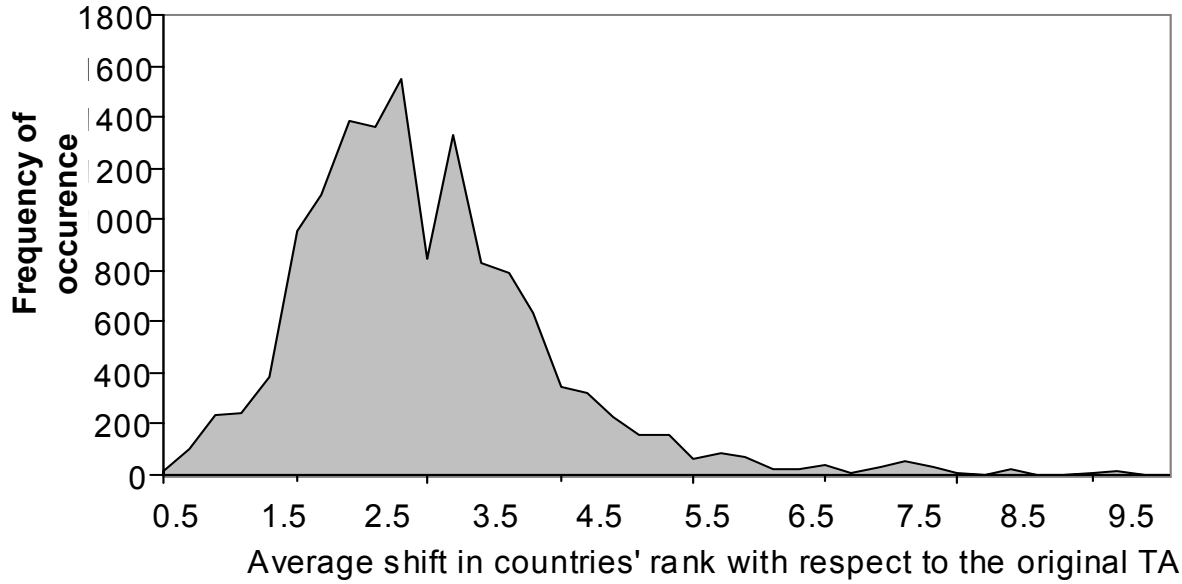
Figure 22 shows the histogram of values for the average shift in the rank of the output variable (Equation 7.2) with respect to the original TAI rank. The mean value is almost 3 positions, with a standard deviation slightly above 1 position. The input factors -- the aggregation system plus inclusion/exclusion at the first order -- affect this variable the most (**Table 29**). When the interactions are considered, both weighting scheme and expert choice become important. This effect can be seen in **Figure 23**. In some cases the average shift in country's rank when using NCMC can be as high as 9 places.

Table 29. Sobol' sensitivity measures of first order and total effects on TAI results

Input Factors	*First order (S_i)*	*Total effect (S_{Ti})*	S_{Ti} - S_i
Normalisation	0.000	0.008	0.008
Exclusion/Inclusion of sub-indicator	0.148	<u>0.435</u>	<u>0.286</u>
Aggregation system	<u>0.245</u>	0.425	0.180
Weighting Scheme	0.038	<u>0.327</u>	<u>0.288</u>
Expert selection	0.068	<u>0.402</u>	<u>0.334</u>
Sum	0.499	1.597	

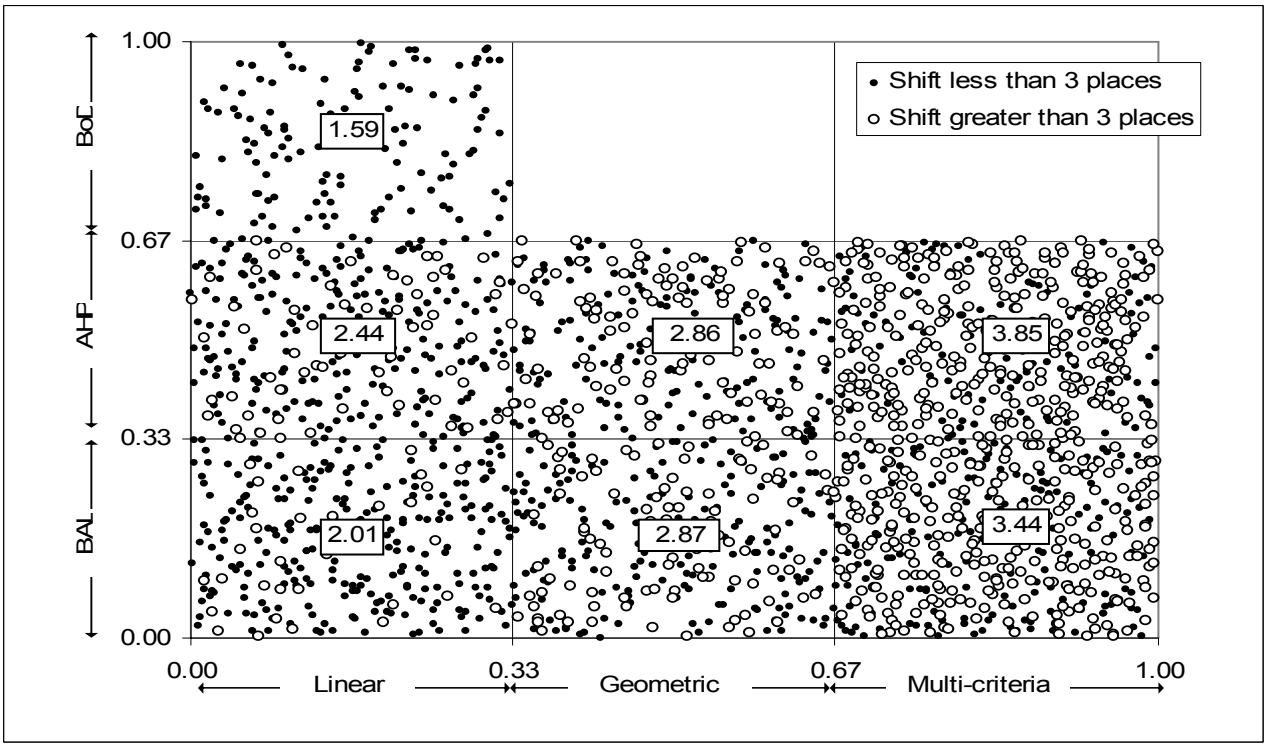
Note: Average shift in countries' rank with respect to the original TAI . Significant values are underlined.

Figure 22. Uncertainty analysis for TAI output variable



Note: Average shift in countries' rank with respect to the original TAI. Uncertain input factors: normalisation method, inclusion-exclusion of a sub-indicator, aggregation system, weighting scheme, expert selection.

Figure 23. Average shift in TAI country rankings by aggregation and weighting combinations



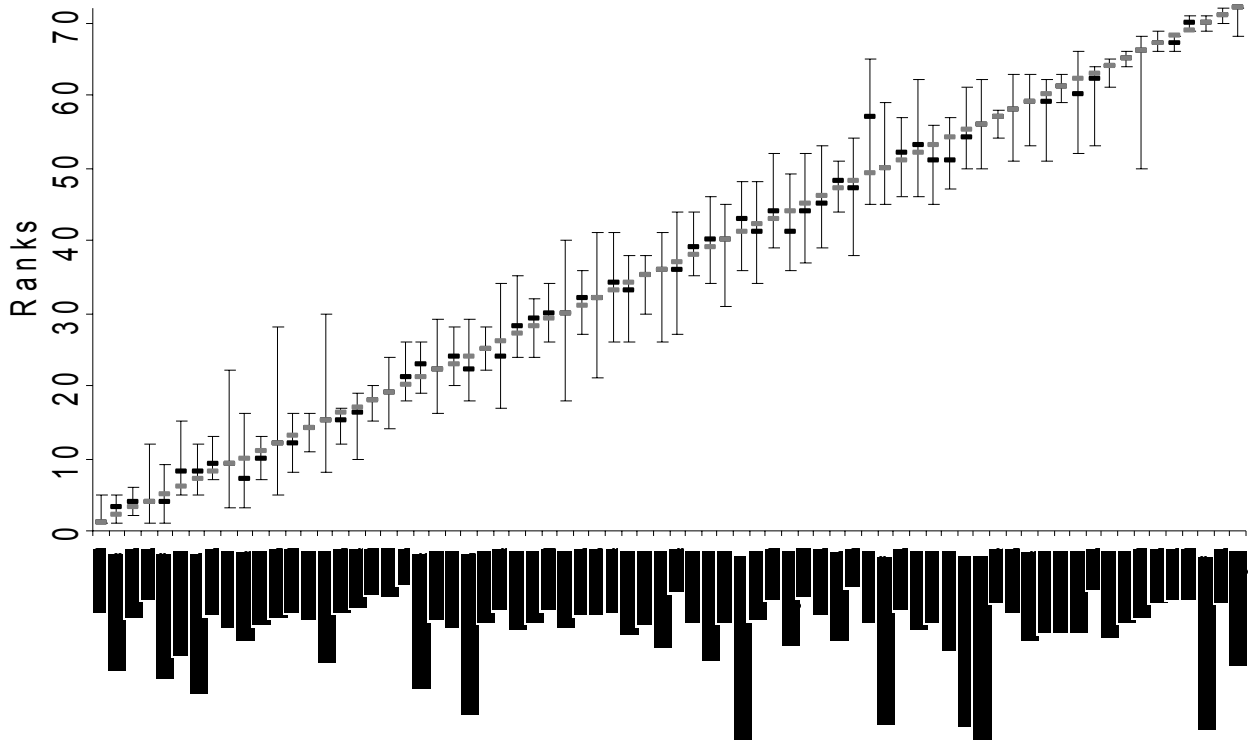
Note: Average shift in countries' rank with respect to the original TAI for different combinations of aggregation system and weighting scheme. Average value per case is indicated in the box.

Analysis 2

In this analysis, it is assumed that the TAI stakeholders have agreed on a linear aggregation system. In fact, one might argue that the choice of the aggregation system is to some extent dictated by the use of the index, and by the expectation of its stakeholders. For instance, if stakeholders believe that the system should be non-compensatory, NCMC would be adopted. Eventually, this would lead to an on average medium–good performance, being worth more for a country than a performance, which is very good on some sub-indicators and bad in others. A GME approach would follow the progress of the index overtime in a scale-independent fashion.

Given these considerations, the second analysis is based on the LIN system, as in the original TAI. The uncertainty analysis plot (**Figure 24**) shows a much more robust behaviour of the index, with fewer inversions of rankings, when median-TAI and original TAI are compared. As far as for the sensitivity, the uncertainty arising from imputation does not seem to make a significant contribution to the output uncertainties, which are also dominated by weighing, inclusion-exclusion, expert selection. Even when, as in the case of Malaysia, imputation by bivariate approach ends into an unrealistic number of patents being imputed for this country (234 patents granted to residents per million people), its rank's uncertainty is still insensitive to imputation. The sensitivity analysis results for the average shift in ranking output variable (Equation 7.2) is shown in **Table 30**. Interactions are now between expert selection and weighing, and considerably less with interaction with inclusion-exclusion.

Figure 24. Uncertainty analysis of TAI country rankings



Note: Uncertainty analysis results showing the countries' rank according to the original TAI 2001 (light grey marks), and the median (black mark) and the corresponding 5th and 95th percentiles (bounds) of the distribution of the MC-TAI for 72 countries. Uncertain input factors: imputation, normalisation method, inclusion-exclusion of a sub-indicator, weighting scheme, expert selection. A linear aggregation system is used. Countries are ordered according to the original TAI values.

Table 30. Sobol' sensitivity measures and average shift in TAI rankings

Input Factors	*First order (S_i)*	*Total effect (S_{Ti})*	$S_{Ti} - S_i$
Imputation	0.001	0.005	0.004
Normalisation	0.000	0.021	0.021
Exclusion/Inclusion of sub-indicator	0.135	0.214	0.078
Weighting Scheme	0.212	0.623	0.410
Expert selection	0.202	0.592	0.390
Sum	0.550	1.453	

Note: Significant values are underlined.

The use of one strategy versus another in indicator building might lead to a biased picture of the country performance, depending on the severity of the uncertainties. As shown by the preceding analyses, if the constructors of the index disagree on the aggregation system, it is highly unlikely for a robust index to emerge. If uncertainties exist in the context of a well-established theoretical framework, e.g. a participatory approach within a linear aggregation scheme is favoured, the resulting country rankings could be fairly robust in spite of the uncertainties.

Both imputation and normalisation do not affect significantly countries ranking when uncertainties of higher order are present. In the current set-up, the uncertainties of higher order are expert selection and weighing scheme (second analysis). *A fortiori* normalisation does not affect output, when the very aggregation system is uncertain (first analysis). In other words, when the weights are uncertain, it is unlikely that normalisation and editing will affect the country ranks.

The aggregation system is of paramount importance. It is recommended that indicator constructors agree on a common approach. Once the system is fixed, it is the choice of the aggregation methods and of the experts that – together with indicator inclusion – exclusion, dominates the uncertainty in the country ranks. However, note that even in the second analysis, when the aggregation system is fixed, the composite indicator model is strongly non additive, which reinforces the case for the use of quantitative, Monte Carlo based approach to robustness analysis.

REFERENCES

Adriaanse A. (1993), Environmental policy performance. A study on the development of indicators for environmental policy in the Netherlands. SDV Publishers, The Hague.

Anderberg, M.R. (1973), Cluster Analysis for Applications, New York: Academic Press, Inc.

Arrow K.J. (1963), Social choice and individual values, 2d edition, Wiley, New York.

Arrow K.J., and Raynaud H. (1986), Social choice and multicriterion decision making, M.I.T. Press, Cambridge.

Arundel A. and Bordoy C. (2002, Methodological evaluation of DG Research's composite indicators for the knowledge based economy. Document presented by DG RTD at the Inter-service consultation meeting on Structural Indicators on July 11th 2002.

Binder, D.A. (1978), "Bayesian Cluster Analysis," Biometrika, 65, 31 -38.

Boscarino J.A., Figley C.R., and Adams R.E. (2004), "Compassion Fatigue following the September 11 Terrorist Attacks: A Study of Secondary Trauma among New York City Social Workers", International Journal of Emergency Mental Health, Vol. 6, No. 2 • 2004, 1-10.

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

Bryant F.B., and Yarnold P.R. (1995), Principal components analysis and exploratory and confirmatory factor analysis. In Grimm and Yarnold, Reading and understanding multivariate analysis. American Psychological Association Books.

Chan, K., Tarantola, S., Saltelli, A. and Sobol', I. M. (2000), Variance based methods. In Sensitivity Analysis (eds A. Saltelli, K. Chan, M. Scott) pp. 167-197. New York: John Wiley & Sons.

Charnes A., Cooper W.W., Lewin A.Y., and Seiford L.M. (1995), Data Envelopment Analysis: Theory, Methodology and Applications. Boston:Kluwer.

Cherchye L. (2001), "Using data envelopment analysis to assess macroeconomic policy performance", Applied Economics, 33, 407-416

Cherchye L., and Kuosmanen T. (2002), "Benchmarking sustainable development: a synthetic meta-index approach", EconWPA Working Papers.

Cherchye, L., Moesen W.and Van Puyenbroeck T. (2004), "Legitimately Diverse, Yet Comparable: on Synthesising Social Inclusion Performance in the EU", Journal of Common Market Studies, 42, 919-955.

Commission of the European Communities (1984), The regions of Europe: Second periodic report on the social and economic situation of the regions of the Community, together with a statement of the regional policy committee, OPOCE, Luxembourg.

Cortina, J.M. (1993), " What is coefficient alpha? An examination of theory and applications", Journal of Applied Psychology, 78, 1, 98-104

Cox, D., Fitzpatrick, R., Fletcher, A., Gore, S., Spiegelhalter, D. and Jones, D. (1992), "Quality-of-life assessment: can we keep it simple?", J.R. Statist. Soc. 155 (3), 353-393.

Cronbach, L. J. (1951),"Coefficient alpha and the internal structure of tests.", Psychometrika, 16, 297-334.

Davis, J. (1986); Statistics and Data Analysis in Geology , John Wiley & Sons, Toronto, 646p.

Pan American Health Organization (1996), Annual report of the Director. Healthy People, Healthy Spaces 1996, Official Document No. 283, Washington, D.C. 20037, U.S.A. http://165.158.1.110/english/sha/ops96arx.htm

Debreu G. (1960), Topological methods in cardinal utility theory, in Arrow K.J., Karlin S. and Suppes P. (eds.) Mathematical methods in social sciences, Stanford University Press, Stanford.

Dempster A.P. and Rubin D.B. (1983), Introduction pp.3-10, in Incomplete Data in Sample Surveys (vol. 2): Theory and Bibliography (W.G. Madow, I. Olkin and D.B. Rubin eds.) New York: Academic Press.

Dietz F.J.and van der Straaten J. (1992), "Rethinking environmental economics: missing links between economic theory and environmental policy", Journal of Economic Issues, Vol. XXVI No. 1, pp. 27-51.

Dunteman, G.H. (1989), Principal components analysis, Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 69.

European Commission , DG ENTR (2001), European Innovation Scoreboard, Brussels.

Ebert U. and Welsch H. (2004), "Meaningful environmental indices: a social choice approach", Journal of Environmental Economics and Management, vol. 47, pp. 270-283.

Emam K., Goldenson D., McCurley J. and Herbsleb J., (1998), "Success or failure? Modeling the likelihood of software process improvement", International software engineering research network, Technical Report ISERN-98-15.

Environmental Protection Agency (EPA), Council for Regulatory Environmental Modeling (CREM), "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 (2000), Business Climate Indicator, DG ECFIN, European Commission, Brussels.

European Commission (2001a), Summary Innovation Index, DG Enterprise, European Commission, Brussels.

European Commission (2001b), Internal Market Scoreboard, DG MARKT, European Commission, Brussels.

European Commission (2004a), Economic Sentiment Indicator, DG ECFIN, Brussels, http://europa.eu.int/comm/economy_finance/index_en.htm

European Commission (2004 b), Composite Indicator on e-business readiness, DG JRC, European Commission, Brussels.

Everitt, B.S. (1979), "Unresolved Problems in Cluster Analysis," Biometrics, 35, 169 -181.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999), "Evaluating the use of exploratory factor analysis in psychological research", Psychological Methods, 4: 272-299.

Fagerberg J. (2001), Europe at the crossroads: The challenge from innovation-based growth in the Globalising Learning Economy, B. Lundvall and D. Archibugi eds., Oxford Press.

Feldt, L.S., Woodruffe, D.J. and Salih, F.A. (1987), "Statistical Inference for Coefficient Alpha", Applied Psychological Measurement, 11,1, 93-103.

Forman E.H. (1983), "The analytic hierarchy process as a decision support system", Proceedings of the IEEE Computer society.

Freudenberg, M. (2003), "Composite indicators of country performance: a critical assessment", OECD, Paris.

Funtowicz S.O., Munda G. and Paruccini M. (1990), "The aggregation of environmental data using multicriteria methods", Environmetrics, Vol. 1(4), pp. 353-36.

Funtowicz S.O.and Ravetz J.R. (1990), Uncertainty and quality in science for policy, Kluwer Academic Publishers, Dordrecht.

Gentle, J.E., Härdle, W. and Mori, Y. (Eds.) (2004): *Handbook of Computational Statistics: Concepts and Methods*, Springer

Girardin P., Bockstaller C. and Van der Werf H., (2000), "Assessment of potential impacts of agricultural practives on the environment: the AGRO*ECO method", Environmental impact Assessment Review, vol.20, pp. 227-239.

Gorsuch, R. L. (1983), Factor Analysis. Hillsdale, NJ: Lawrence Erlbaum. Orig. ed. 1974.

Gough C., Castells, N. and Funtowicz S., (1998), "Integrated Assessment: an emerging methodology for complex issues", Environmental Modelling and Assessment, n.3, 19-29.

Green, S.B., Lissitz, R.W. and Mulaik, S.A.(1977), "Limitations of coefficient alpha as an index of test unidimensionality", Educational and Psychological Measurement, 37, 827-838.

Green P.E., and Srinivasan V., (1978), "Conjoint analysis in consumer research: issues and outlook", Journal of Consumer Research 5, 103-123.

Golub, Gene H. & van der Vorst, Henk A. (2000): Eigenvalue computation in the 20th century, Journal of Computational and Applied Mathematics, Vol. 123, Iss. 1-2.

Grubb D., and Wells W. (1993), "Employment regulation and patterns of work in EC countries", OECD Economic Studies, n. 21 Winter, 7-58, Paris.

Hair J.F., Anderson R.E., Tatham R.L. and Black W.C. (1995), Multivariate data analysis with readings, fourth ed. Prentice Hall, Englewood Cliffs, NJ.

Hartigan, J.A. (1975), Clustering Algorithms, New York: John Wiley & Sons, Inc.

Harvey A. (1989), Forecasting, structural time series models and the Kalman filter, Cambridge University Press, Cambridge UK.

Hatcher, L. (1994), A step-by-step approach to using the SAS system for factor analysis and structural equation modeling. Cary, NC: SAS Institute. Focus on the CALIS procedure.

Hattie, J. (1985), "Methodology Review: Assessing unidimensionality of tests and items", Applied Psychological Measurement, 9, 2, 139-164.

Hollenstein, H. (1996), "A Composite Indicator of a Firm's Innovativeness. An Empirical Analysis Based on Survey Data for Swiss Manufacturing", Research Policy, 25, 633-45.

Hollenstein, H. (2003), "Innovation Modes in the Swiss Service Sector. A Cluster Analysis Based on Firm-level Data", Research Policy, 32(5), 845-863.

Homma, T. and Saltelli, A. (1996), "Importance measures in global sensitivity analysis of model output", Reliability Engineering and System Safety, 52(1), 1-17.

Hutcheson, G., and Sofroniou N.(1999), The multivariate social scientist: Introductory statistics using generalized linear models, Thousand Oaks, CA: Sage Publications.

Jacobs, R. P. Smith and M. Goddard (2004), "Measuring performance: an examination of composite performance indicators, Centre for Health Economics", Technical Paper Series 29.

Jae-On K., and Mueller C.W. (1978a), Introduction to factor analysis: What it is and how to do it. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 13.

Jae-On K., and Mueller C.W. (1978b), Factor Analysis: Statistical methods and practical issues. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 14.

Jamison, D. and Sandbu, M. (2001), "WHO ranking of health system performance", Science, 293, 1595-1596.

Jencks, S.F., Huff, E.D. and Cuerdon, T. (2003), "Change in the quality of care delivered to Medicare beneficiaries, 1998-1999 to 2000-2001", Journal of the American Medical Association, 289(3): 305-12.

Kahma N. (2000), "Measuring environmental quality: an index of pollution", Ecological Economics, vol. 35 pp. 191-202.

Kahn J.R. and Maynard P. (1995), "Conjoint Analysis as a Method of Measuring Use and Non-Use Values of Environmental Goods", paper presented at the American Economic Association.

Kahn J.R. (1998), "Methods for aggregating performance indicators", mimeo, University of Tennessee.

Kahnna N. (2000), "Measuring Environmental quality: an index of pollution", Ecological Economics, 32, 191-202.

Karlsson J. (1998), A systematic approach for prioritizing software requirements, PhD. Dissertation n. 526, Linkoping, Sverige.

Kaufmann D., Kraay A. and Zoido-Lobatón P. (1999), Aggregating Governance Indicators, Policy Research Working Papers, World Bank, http://www.worldbank.org/wbi/governance/working_papers.html

Kaufmann D., Kraay A. and Zoido-Lobatón P. (2003), "Governance matters III: governance Indicators for 1996-2002", mimeo, World Bank.

Keeney R. and Raiffa H. (1976), Decision with multiple objectives: preferences and value trade-offs, Wiley, New York.

Keynes, J. M. (1891), The Scope and Method of Political Economy. London: Macmillan.

King's Fund (2001), The sick list 2000, the NHS from best to worst, http://www.fulcrumtv.com/sick%20list.htm

Kline, R.B. (1998), Principles and practice of structural equation modelling, NY: Guilford Press. Covers confirmatory factor analysis using SEM techniques. See esp. Ch. 7.

Koedijk K. and Kremers J. (1996), "Market opening, regulation and growth in Europe", Economic Policy (0)23, October.

Korhonen P., Tainio R. and Wallenius J., (2001), "Value efficiency analysis of academic research", European Journal of Operational Research, 130, 121-132.

Krantz D.H., Luce R.D., Suppes P. and Tversky A. (1971), Foundations of measurement, vol. 1, Additive and polynomial representations, Academic Press, New York.

Lawley, D. N. and Maxwell A. E. (1971), Factor analysis as a statistical method, London: Butterworth and Co.

Levine, M.S. (1977), Canonical analysis and factor comparison, Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 6.

Little R.J.A. and Schenker N. (1994), Missing Data, in Handbook for Statistical Modeling in the Social and Behavioral Sciences (G. Arminger, C.C Clogg, and M.E. Sobel eds.) pp.39-75, New York: Plenum.

Little R.J.A (1997), Biostatistical Analysis with Missing Data, in Encyclopedia of Biostatistics (p. Armitage and T. Colton eds.) London: Wiley.

Little R.J.A. and Rubin D.B. (2002), Statistical Analysis with Missing Data, Wiley Interscience, J. Wiley &Sons, Hoboken, New Jersey.

Mahlberg B. and Obersteiner M. (2001), Remeasuring the HDI by data Envelopment analysis, Interim report IR-01-069, International Institute for Applied System Analysis, Laxenburg, Austria.

Manly B. (1994), Multivariate statistical methods, Chapman & Hall, UK.

Massam B.H. (2002), "Quality of life: public planning and private living", Progress in Planning, vol. 58(3), pp. 141-227.

Massart, D.L. and Kaufman, L. (1983), The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, New York: John Wiley & Sons, Inc.

McDaniel, C. and Gates R. (1998), Contemporary Marketing Research. West Publishing, Cincinnati, OH.

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

Miller, M.B. (1995), "Coefficient Alpha: a basic introduction from the perspectives of classical test theory and structural equation modelling", Structural Equation Modelling, 2, 3, 255-273.

Milligan, G.W. and Cooper, M.C. (1985), "An Examination of Procedures for Determining the Number of Clusters in a Data Set," Psychometrika, 50, 159 -179.

Moldan, B., Billharz, S. and Matravers, R. (1997), Sustainability Indicators: Report of the Project on Indicators of Sustainable Development, SCOPE 58. Chichester and New York: John Wiley & Sons.

Moldan B., Hak T., Kovanda J., Havranek M. and Kuskova P. (2004), "Composite Indicators of Environmental Sustainability", OECD World Forum on Key Indicators, Palermo, 10-13 November 2004, proceedings, see http://www.oecd.org/dataoecd/43/48/33829383.doc

Muldur U. (2001), Technical annex on structural indicators. Two composite indicators to assess the progress of member States in their transition towards a knowledge based economy, DG RTD, Brussels.

Munda G. (1993), "Fuzzy information in multicriteria environmental evaluation models", PhD dissertation Vrije Universiteit te Amsterdam.

Munda G. (1995), Multicriteria evaluation in a fuzzy environment, Physica-Verlag, Contributions to Economics Series, Heidelberg.

Munda G. (2004), MCDA and Sustainability Decisions, forthcoming in J. Figueira, S. Greco and M. Ehrgott (eds.) – State of the art of multiple-criteria decision analysis, Kluwer, Dordrecht.

Munda, G. and Nardo, M. (2003), "On the methodological foundations of composite indicators used for ranking countries", OECD/JRC Workshop on Composite Indicators of Country Performance, Ispra, Italy, May 12.

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

Nanduri M., Nyboer J. and Jaccard M. (2002), "Aggregating physical intensity indicators: results of applying the composite indicator approach to the Canadian industrial sector", Energy Policy, 30, 151-162.

Nardo, M., Tarantola S., Saltelli A., Andropoulos C., Buescher R., Karageorgos G., Latvala A., and Noel F. (2004), "The e-business readiness composite indicator for 2003: a pilot study", EUR 21294.

Nicoletti G., S. Scarpetta and O. Boylaud, (2000), "Summary indicators of product market regulation with an extension to employment protection legislation", OECD, Economics department working papers No. 226, ECO/WKP(99)18. http://www.oecd.org/eco/eco.

Nilsson R. (2000), "Confidence Indicators and Composite Indicators", CIRET conference, Paris, 10-14 October 2000

NISTEP (National Institute of Science and Technology Policy), (1995), Science and Technology Indicators, NISTEP Report No. 37, Japan.

Norman G. R. and Streiner D. L. (1994), Biostatistics: The bare essentials. St. Louis, MO: Mosby.

Nunnaly J. (1978), Psychometric theory, New York: McGraw-Hill.

OECD (1999), Employment Outlook, Paris.

OECD (2003), Quality Framework and Guidelines for OECD Statistical Activities, www.oecd.org/statistics.

OECD (2004), Learning for Tomorrow's World - First Results from PISA 2003, Programme for International Student Assessment, http://www.pisa.oecd.org/dataoecd/1/60/34002216.pdf

Parker J. (1991), "Environmental reporting and environmental indices", PhD Dissertation, Cambridge, UK.

Pett M.A., Lackey N.R. and Sullivan J.J. (2003), Making sense of factor analysis: The use of factor analysis for instrument development in health care research. Thousand Oaks, CA: Sage Publications.

Pré Consultants (2000), The Eco-indicator 99. A damage oriented method for life cycle impact assessment. http://www.pre.nl/eco-indicator99/ei99-reports.htm

Podinovskii V.V. (1994), "Criteria importance theory", Mathematical Social Sciences, 27, pp. 237 - 252.

Porter M. and Stern S. (1999), The new challenge to America's prosperity: finding from the Innovation Index, Council on Competitiveness, Washington D.C.

Puolamaa M., Kaplas M., and Reinikainen T. (1996), Index of Environmental Friendliness. A methodological study, Eurostat.

Raykov T. (1998b), "Cronbach's Alpha and Reliability of Composite with Interrelated Non-homogenous Items", Applied Psychological Measurement, 22, 375-385.

Rosen R. (1991), Life Itself: A Comprehensive Inquiry into Nature, Origin, and Fabrication of Life. Columbia University Press 1991.

Roubens M. and Vincke P. (1985), Preference Modelling, Springer Verlag, Heidelberg.

Roy B. (1996), Multicriteria methodology for decision analysis, Kluwer, Dordrecht.

Saaty T. L. (1980), The Analytic Hierarchy Process, New York: McGraw-Hill.

Saaty R.W. (1987), "The analytic hierarchy process: what it is and how it is used", Mathematical Modelling, 9, 161-176.

Saisana M. and Tarantola S. (2002), State-of-the-art report on current methodologies and practices for composite indicator development, EUR 20408 EN, European Commission-JRC: Italy.

Saisana M., Tarantola S. and Saltelli A. (2005a), "Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators", Journal of the Royal Statistical Society A, 168(2), 1-17.

Saisana M., Nardo M. and Saltelli A. (2005b) Uncertainty and Sensitivity Analysis of the 2005 Environmental Sustainability Index, in Esty D. Levy M., Srebotnjak T. and de Sherbinin A. 2005 Environmental Sustainability Index: Benchmarking National Environmental Stewardship. New Have: Yale Center for Environmental Law and Policy, p.75-78.

Saltelli A. Nardo M., Saisana M. and Tarantola S. (2004) Composite Indicators - The Controversy and the Way Forward, OECD World Forum on Key Indicators, Palermo, 10-13 November 2004, http://www.oecd.org/dataoecd/40/50/33841312.doc

Saltelli A. (2002)," Making best use of model valuations to compute sensitivity indices", Computer Physics Communications, 145, 280-297.

Saltelli A., Chan, K. and Scott M. (2000a), Sensitivity analysis, Probability and Statistics series, New York: John Wiley & Sons.

Saltelli A. and Tarantola S. (2002),"On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal", Journal of American Statistical Association, 97 (459), 702-709.

Saltelli A., Tarantola S. and Campolongo F. (2000b), "Sensitivity analysis as an ingredient of modelling", Statistical Science, 15, 377-395.

Saltelli A., Tarantola S., Campolongo F. and Ratto M. (2004), Sensitivity Analysis in practice, a guide to assessing scientific models, New York: John Wiley & Sons. A software for sensitivity analysis is available at http://www.jrc.cec.eu.int/uasa/prj-sa-soft.asp.

Sharpe, A. (2004), Literature Review of Frameworks for Macro-indicators, Centre for the Study of Living Standards, Ottawa, CAN.

Schumpeter J.A. (1933), "The common sense of econometrics", Econometrica 1: 5-12.

Sobol' I. M. (1993)," Sensitivity analysis for non-linear mathematical models", Mathematical Modelling & Computational Experiment 1, 407-414.

Sobol' I. M. (1967), "On the distribution of points in a cube and the approximate evaluation of integrals", USSR Computational Mathematics and Physics, 7, 86–112.

Spath H. (1980), Cluster Analysis Algorithms, Chichester, England: Ellis Horwood.

SPRG (2001), Report of the Scientific Peer Review Group on Health Systems Performance Assessment, Scientific Peer Review Group (SPRG), WHO: Geneva. http://www.who.int/health-systemsperformance/sprg/report_of_sprg_on_hspa.htm

Storrie D. and Bjurek H. (1999), "Benchmarking European labour market performance with efficiency frontier technique", Discussion Paper FS I 00-2011.

Storrie D. and Bjurek H. (2000), Benchmarking the basic performance indicators using efficiency frontier techniques, Report presented to the European commission, DG employment and social affairs.

Tarantola S., Jesinghaus J. and Puolamaa M. (2000), Global sensitivity analysis: a quality assurance tool in environmental policy modelling. In Sensitivity Analysis (eds Saltelli A., Chan K., Scott M.) pp. 385-397. New York: John Wiley & Sons.

Tarantola S., Saisana M., Saltelli A., Schmiedel F. and Leapman, N. (2002), Statistical techniques and participatory approaches for the composition of the European Internal Market Index 1992-2001, EUR 20547 EN, European Commission: JRC-Italy.

Tarantola S., Liska R., Saltelli A., Leapman N. and Grant C. (2004), The Internal Market Index 2004, EUR 21274EN, European Commission: JRC-Italy

Ting H.M. (1971), Aggregation of attributes for multiattributed utility assessment, Technical report n. 66, Operations Research Center, MIT Cambridge Mass.

Ülengin B., Ülengin F. and Güvenç Ü., (2001), "A multidimensional approach to urban quality of life: the case of Istanbul", European Journal of Operational Research, 130, 361-374.

United Nations (1992, 1999, 2000, 2001), Human Development Report. United Kingdom: Oxford University Press. http://www.undp.org

U.K. Government (2004), “Sustainable Development Indicators in Your Pocket 2004”.

U.S. Department of Energy and Energy Information Administration (1995), Measuring energy efficiency un the United States’ economy: a beginning. U.S. Department of Energy, Washington, DC.

Vichi M. and Kiers H. (2001), "Factorial k-means analysis for two-way data", Computational Statistics and Data Analysis, 37(1), 49-64.

Vincke Ph. (1992), Multicriteria decision aid, Wiley, New York.

Widaman K. F. (1993), "Common factor analysis versus principal components analysis: Differential bias in representing model parameters?", Multivariate Behavioral Research 28: 263-311. Cited with regard to preference for PFA over PCA in confirmatory factor analysis in SEM.

World Economic Forum (2002), Environmental Sustainability Index. http://www.ciesin_org/indicators/ESI/index.html.

World Economic Forum (2004), Global Competitiveness Report.

WHO (2000), Overall Health System attainment. http://www.who.int/whr2001/2001/archives/2000/en/contents.htm

Young H.P. and Levenglick A. (1978), "A consistent extension of Condorcet’s election principle", SIAM Journal of Applied Mathematics, 35, pp. 285-300.

Young H.P. (1988), "Condorcet’s theory of voting", American Political Science Review, Vol. 82, No. 4, pp. 1231-1244.

Zimmermann H.J. and Zysno P. (1983), "Decisions and evaluations by hierarchical aggregation of information", Fuzzy Sets and Systems, 10, pp.243-260.

APPENDIX: TECHNOLOGY ACHIEVEMENT INDEX

Table A.1. List of sub-indicators of the Technology Achievement Index

Indicator	Unit	Definition
CREATION OF TECHNOLOGY		
PATENTS	Patents granted per 1,000,000 people	Number of patents granted to residents, to reflect the current level of invention activities (1998)
ROYALTIES	US $ per 1,000 people	Receipts of royalty and license fees from abroad per capita, so as to reflect the stock of successful innovations of the past that are still useful and hence have market value (1999)
DIFFUSION OF RECENT INNOVATIONS		
INTERNET	Internet hosts per 1,000 people	Diffusion of the Internet, which is indispensable to participation in the network age (2000)
EXPORTS	%	Exports of high and medium technology products as a share of total goods exports (1999)
DIFFUSION OF OLD INNOVATIONS		
TELEPHONES	Telephone lines per 1,000 people (log)	Number of telephone lines (mainline and cellular), which represents old innovation needed to use newer technologies and is also pervasive input to a multitude of human activities (1999)
ELECTRICITY	kWh per capita (log)	Electricity consumption, which represents old innovation needed to use newer technologies and is also pervasive input to a multitude of human activities (1998)
HUMAN SKILLS		
SCHOOLING	Years	Mean years of schooling (age 15 and above), which represents the basic education needed to develop cognitive skills (2000)
ENROLMENT	%	Gross enrolment ratio of tertiary students enrolled in science, mathematics and engineering, which reflects the human skills needed to create and absorb innovations (1995-1997)

Table A.2. Raw data for the sub-indicators of the Technology Achievement Index

		PATENTS	ROYALTIES	INTERNET	EXPORTS	TELEPHONES (log)	ELECTRICITY (log)	SCHOOLING	ENROLMENT
1	Finland	187	125.6	200.2	50.7	3.08	4.15	10	27.4
2	United States	289	130	179.1	66.2	3.00	4.07	12	13.9
3	Sweden	271	156.6	125.8	59.7	3.10	4.14	11.4	15.3
4	Japan	994	64.6	49	80.8	3.00	3.86	9.5	10
5	Korea, Rep. of	779	9.8	4.8	66.7	2.97	3.65	10.8	23.2
6	Netherlands	189	151.2	136	50.9	3.02	3.77	9.4	9.5
7	United Kingdom	82	134	57.4	61.9	3.02	3.73	9.4	14.9
8	Canada	31	38.6	108	48.7	2.94	4.18	11.6	14.2
9	Australia	75	18.2	125.9	16.2	2.94	3.94	10.9	25.3
10	Singapore	8	25.5	72.3	74.9	2.95	3.83	7.1	24.2
11	Germany	235	36.8	41.2	64.2	2.94	3.75	10.2	14.4
12	Norway	103	20.2	193.6	19	3.12	4.39	11.9	11.2
13	Ireland	106	110.3	48.6	53.6	2.97	3.68	9.4	12.3
14	Belgium	72	73.9	58.9	47.6	2.91	3.86	9.3	13.6
15	New Zealand	103	13	146.7	15.4	2.86	3.91	11.7	13.1
16	Austria	165	14.8	84.2	50.3	2.99	3.79	8.4	13.6
17	France	205	33.6	36.4	58.9	2.97	3.80	7.9	12.6
18	Israel	74	43.6	43.2	45	2.96	3.74	9.6	11
19	Spain	42	8.6	21	53.4	2.86	3.62	7.3	15.6
20	Italy	13	9.8	30.4	51	3.00	3.65	7.2	13
21	Czech Republic	28	4.2	25	51.7	2.75	3.68	9.5	8.2
22	Hungary	26	6.2	21.6	63.5	2.73	3.46	9.1	7.7
23	Slovenia	105	4	20.3	49.5	2.84	3.71	7.1	10.6
24	Hong Kong, China (SAR)	6		33.6	33.6	3.08	3.72	9.4	9.8
25	Slovakia	24	2.7	10.2	48.7	2.68	3.59	9.3	9.5
26	Greece			16.4	17.9	2.92	3.57	8.7	17.2
27	Portugal	6	2.7	17.7	40.7	2.95	3.53	5.9	12
28	Bulgaria	23		3.7	30	2.60	3.50	9.5	10.3
29	Poland	30	0.6	11.4	36.2	2.56	3.39	9.8	6.6
30	Malaysia			2.4	67.4	2.53	3.41	6.8	3.3
31	Croatia	9		6.7	41.7	2.63	3.39	6.3	10.6
32	Mexico	1	0.4	9.2	66.3	2.28	3.18	7.2	5
33	Cyprus			16.9	23	2.87	3.54	9.2	4
34	Argentina	8	0.5	8.7	19	2.51	3.28	8.8	12
35	Romania	71	0.2	2.7	25.3	2.36	3.21	9.5	7.2
36	Costa Rica		0.3	4.1	52.6	2.38	3.16	6.1	5.7
37	Chile		6.6	6.2	6.1	2.55	3.32	7.6	13.2
38	Uruguay	2		19.6	13.3	2.56	3.25	7.6	7.3
39	South Africa		1.7	8.4	30.2	2.43	3.58	6.1	3.4
40	Thailand	1	0.3	1.6	48.9	2.09	3.13	6.5	4.6
41	Trinidad and Tobago			7.7	14.2	2.39	3.54	7.8	3.3
42	Panama			1.9	5.1	2.40	3.08	8.6	8.5

43	Brazil	2	0.8	7.2	32.9	2.38	3.25	4.9	3.4
44	Philippines		0.1	0.4	32.8	1.89	2.65	8.2	5.2
45	China	1	0.1	0.1	39	2.08	2.87	6.4	3.2
46	Bolivia	1	0.2	0.3	26	2.05	2.61	5.6	7.7
47	Colombia	1	0.2	1.9	13.7	2.37	2.94	5.3	5.2
48	Peru		0.2	0.7	2.9	2.03	2.81	7.6	7.5
49	Jamaica		2.4	0.4	1.5	2.41	3.35	5.3	1.6
50	Iran, Islamic Rep. of	1			2	2.12	3.13	5.3	6.5
51	Tunisia		1.1		19.7	1.98	2.92	5	3.8
52	Paraguay		35.3	0.5	2	2.14	2.88	6.2	2.2
53	Ecuador			0.3	3.2	2.09	2.80	6.4	6
54	El Salvador		0.2	0.3	19.2	2.14	2.75	5.2	3.6
55	Dominican Republic			1.7	5.7	2.17	2.80	4.9	5.7
56	Syrian Arab Republic				1.2	2.01	2.92	5.8	4.6
57	Egypt		0.7	0.1	8.8	1.89	2.94	5.5	2.9
58	Algeria				1	1.73	2.75	5.4	6
59	Zimbabwe			0.5	12	1.56	2.95	5.4	1.6
60	Indonesia			0.2	17.9	1.60	2.51	5	3.1
61	Honduras				8.2	1.76	2.65	4.8	3
62	Sri Lanka			0.2	5.2	1.69	2.39	6.9	1.4
63	India	1		0.1	16.6	1.45	2.58	5.1	1.7
64	Nicaragua			0.4	3.6	1.59	2.45	4.6	3.8
65	Pakistan			0.1	7.9	1.38	2.53	3.9	1.4
66	Senegal			0.2	28.5	1.43	2.05	2.6	0.5
67	Ghana				4.1	1.08	2.46	3.9	0.4
68	Kenya			0.2	7.2	1.04	2.11	4.2	0.3
69	Nepal			0.1	1.9	1.08	1.67	2.4	0.7
70	Tanzania, U. Rep. of				6.7	0.78	1.73	2.7	0.2
71	Sudan				0.4	0.95	1.67	2.1	0.7
72	Mozambique				12.2	0.70	1.73	1.1	0.2

Note: The first 23 countries are used as case study in the Handbook. Units are given in Table A.1.

ENDNOTES

[1] The technique of PCA was first described by Karl Pearson in 1901. A description of practical computing methods came much later from Hotelling in 1933. For a detailed discussion on the PCA, *see* Jolliffe, (1986), Jackson (1991) and Manly (1994). Social scientists may also find the shorter monograph by Dunteman, (1989) to be helpful.

[2] For reasons of clarity in this section we substitute the indexing $q=1,...Q$ with the indexing $i=1,...,Q$ and $j=1,...,Q$.

[3] Golub, Gene H. & van der Vorst, Henk A. (2000): *Eigenvalue computation in the 20th century*, Journal of Computational and Applied Mathematics, Vol. 123, Iss. 1-2. And Gentle, James E.; Härdle, Wolfgang; Mori, Yuichi (Eds.) (2004): *Handbook of Computational Statistics: Concepts and Methods*, Springer

[4] Euclidean distances can be greatly influenced by variables that have the largest values. One way around this problem is to standardise the variables.

[5] The name is based on the route that follows the grid of roads, as in most American cities it is not possible to go directly between two points.

[6] The value that divides in two equal parts the distribution of the random variable

[7] The value with the highest frequency

[8] A variant of unconditional mean imputation is the fill-in via conditional mean. The regression approach is one possible method. Another common method (called imputing means within adjustment cells) is to classify the data for the sub-indicator with some missing values in classes and impute provisionally the missing values of that class with the sample mean of the class. Then sample mean (across al classes) is then calculated and substituted as final imputation value.

[9] If the observed variables are dummies for a categorical variable then the prediction (7*) are respondent means within classes defined by the variable and the method reduces to that of imputing means with adjustment cells.

[10] Define $SSE = \sum_i (x_{ih} - \hat{x}_{ih})^2$, $SST = \sum_i (x_{ih} - \bar{x}_h)^2$, then $R^2 = 1 - (SSE / SST)$, $MSE = SSE/(M - r - k)$, where k is the number of coefficients in the regression and $(M-r)$ the number of observations. $RMS = \sum_i (\hat{x}_{ih} - \bar{x}_h)^2$ and $C_k = (SSE_k / MSE) - (M - r) + 2k$ where the SSE_k is computed from a model with only k coefficients and MSE is computed using all available regressors.

[11] Other iterative methods include the Newton-Raphson algorithm and the scoring method. Both involve a calculation of the matrix of second derivatives of the likelihood, which, for complex pattern of incomplete data, can be a very complicate function of θ. As a result these algorithms often require algebraic manipulations and complex programming. Numerical estimation of this matrix is also possible but careful computation is needed.

[12] For NMAR mechanisms one needs to make assumption on the missing-data mechanism and include them into the model, see Little and Rubin, (2002), Ch. 15.

[13] In a sample of *n* observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is a fault. Such values are called outliers (F.H.C. Marriott, 1990, A dictionary of statistical terms, Longman Scientific & Technical, Fifth edition, p.223). Eurostat adopts this definition of outlier.

[14] http://europa.eu.int/comm/economy_finance/publications/european_economy/2001/b2001_0809_en.pdf)

[15] Other methods are available, e.g. the Maximum Likelihood or the principal Factor centroids. Notice that these methods usually supply very different weights especially when the sample size of FA is small.

[16] Weights are normalized squared factor loading, e.g. 0.24 = (0.79^2)/2.64 which is the portion of the variance of the first factor explained by the variable Internet.

[17] To preserve comparability final weights could be rescaled to sum up to one.

[18] DEA has also been used in production theory, for a review see Charnes et al., (1995).

[19] We present the method as it has been used in Cherchye et al., (2004), and Cherchye and Kuosmanen, (2002).

[20] Additional constraints could be imposed. Country-specific restrictions to reflect prior information can also be added. Notice that constraints should not be given to the absolute value of weights but placed on the relative weights. The result of this approach is therefore relative weights or trade offs to be legitimately used in linear and geometric aggregations.

[21] In our example we imposed the requirement for each sub-indicator to weight at least 10% and no more than 15% of the total.

[22] However, precisely since BOD weights have the meaning of relative weights, the weights as originally produced by the algorithm could be normalised afterwards so as to sum up to one facilitating the comparison with the results of other methods.

[23] In 1991, 400 German experts were asked to allocate a budget to several environmental indicators related to an air pollution problem. The results were consistent, although the experts came from opposing social spheres like the industrial and the environmental sectors (Jesinghaus in Moldan and Billharz, 1997).

[24] The exercise was carried out at JRC interviewing experts in the field.

[25] A subset of indicators Y is *preferentially independent* of Y^C (the complement of Y) only if any conditional preference among elements of Y, holding all elements of Y^C fixed, remain the same, regardless of the levels at which Y^C are held. The variables $x_1, x_2, ..., x_Q$ are *mutually preferentially independent* if every subset Y of these variables is preferentially independent of its complementary set of evaluators.

[26] $\frac{\partial S_{x,y}}{\partial z} = 0$, $\forall x, y \in Y, \forall z \in Y^C$, see the previous note above.

[27] Suppose that country a is evaluated according to some criteria/sub-indicators $(x_1(a),...,x_Q(a))$, then the *substitution rate at a*, of sub-indicator j with respect to sub-indicator r (taken as a reference) is the amount $S_{jr}(a)$ such that, country b whose evaluations are: $x_l(a) = x_l(b), \forall l \neq j, r$; $x_j(b) = x_j(a) - 1$;

and $x_r(b) = x_r(a) + S_{jr}(a)$ is indifferent to country *a*. Therefore, $S_{jr}(a)$ is the amount which must be added to the reference sub-indicator in order to compensate the loss of one unit on sub-indicator *j* keeping constant the others. While for additive aggregations the substitution rate is constant, in the multiplicative aggregation it is proportional to the relative score of the indicator with respect to the others.

[28] Data are not normalized. Normalization does not change the result of the multicriteria method whenever it does not change the ordinal information of the data matrix.

[29] To prevent this problem it is possible to set thresholds of this type: if the difference between two countries in the indicator I is more than x%, then give to the country with the highest score a much higher weight. If the difference is less than x% give nearly the same weight. However, more precision comes at the expenses of ad hoc threshold and weighting values.

[30] Examples of MCA include the Environmental Sustainability Indicator 2005, agro-ecological indicators (Girardin et al., 2000) and an indicator of quality of life of three towns near to Puerto Vallarta, Mexico (Massam, 2002) within a project on the effects of tourism on the quality of life of small communities near international tourist resorts.

[31] Compensability of aggregations is widely studied in fuzzy sets theory, for example Zimmermann and Zysno, (1983) use the geometric operator $(\prod_q I_q)^{(1-\gamma)}(1-\prod_q(1-I_q))^{\gamma}$ where γ is a parameter of compensation: the larger is γ the higher is the degree of compensation between operators (in our case sub-indicators).

[32] The multi-criteria approach MCA produces ranks for countries. The focus of the analysis thus is $Rank(CI_c)$ rather than the raw value of the index CI_c.

[33] For proof, see Saltelli et al., (2004).