19

CONCEPTUAL AND PRACTICAL ISSUES IN CONSTRUCTING COMPOSITE INDICES

Nadia Farrugia

Abstract. This chapter presents an analysis of the conceptual issues associated with the construction of composite indices. Composite indices, which are constructed by averaging a number of indicators or sub-indices, are multi-dimensional, in that they represent aggregate measures of a combination of factors. They are often used to simplify complex measurement constructs, and often have a strong political appeal due to the fact that they simplify complex matters into a single number. However, composite indices are often criticised for their subjectivity. Indeed the methodology used to construct an index generates considerable debate on various aspects, such as the weighting method used, possible correlation among the different sub-indices, missing variables, standardisation procedures and others. This paper will attempt to propose some desirable criteria for the construction of composite indices, including simplicity, ease of comprehension, and coverage issues and transparency. It will also discuss a number of methodological considerations including weighting.

1. Introduction

Composite indices are usually constructed by summing and then averaging a number of components (or sub-indices) to derive another variable. An observation on a composite index is usually constructed as a weighted (linear) aggregation of a number of component variables. Such indices are often used to measure multi-dimensional concepts which cannot be measured by a single indicator. They often have a strong political appeal due to the fact that they simplify complex matters into a single variable. In the context of policy making, composite indices are useful in drawing attention to a particular issue and its components, and in benchmarking or monitoring performance. However, such indices generate considerable debate on various aspects, such as the weighting method used, possible correlation among the different sub-indices, missing variables, standardisation procedures and others. In reality, the results of composite indices are sensitive to different methodological choices used in their computation.

This chapter proposes some desirable criteria for the construction of composite indices, including simplicity, ease of comprehension, coverage issues and transparency. It will also discuss a number of methodological considerations including weighting.

Composite indicators are increasingly being used to make cross-national comparisons of country performance in specified areas such as the economy, environment, globalisation, society and innovation/ technology/information. They are popular in benchmarking exercises where countries wish to measure their performance relative to other countries and are used to identify general trends, determine performance targets and set policy priorities.

Renowned composite indices include the Economic Competitiveness Index of the International Institute for Management Development, the Commonwealth Secretariat's Economic Vulnerability Index, the Economic Freedom of the World Index of Fraser Institute, the Environmental Performance Index of the Universities of Yale and Columbia, the Growth Competitiveness Index and the Current Competitiveness Index of the World Economic Forum, the Human Development Index of the United Nations, the Summary Innovation Index of the European Commission, and the Economic Resilience Index of the University of Malta.

The chapter is structured as follows. Section 2, which follows this introduction, discusses the main strengths and weaknesses of composite indices and gives an overview of the different frameworks of desirable attributes for developing composite indices. Section 3 discusses the main conceptual issues involved in constructing composite indices, focusing on the selection of variables, ways to deal with missing data, standardisation of variables, weighting and aggregation, and testing the composite index. Section 4 concludes the study.

2. Definition, Uses, Strengths and Weaknesses

Strengths and Weaknesses

There are many studies on the strengths and weaknesses of composite indices (see, for example, Saisana and Tarantola, 2002); Briguglio, 2003; Conway, 2005; Lievesley, 2005). This section summarises the views expressed in these works.

One of the major strengths of composite indices is that, since they summarise complex or multi-dimensional issues, often yielding a single-value measure of the issue under consideration, they facilitate the task of ranking subjects with regard to complex issues. They can also be used to assess progress over time on complex issues, since they are easier to interpret than trying to find a trend in many separate indicators. Composite indices can help to develop a common language for discussion and are an effective tool for communicating with policymakers and the public.

Given that quantification requires pre-definition of the issue or issues, composite indices help to set the direction for policymakers and to focus the discussion. They support decision making as they help justify certain priorities and can be used to set targets, establish standards and also promote accountability. They are thus essential for empirical work on the linkages between policy and performance. Indices may help disseminate information and can be used to make the public more aware of certain problems, for communication, and for alerting stakeholders about issues. Composite indices may also generate academic discussion and enhance awareness among scholars on the issues involved.

Such indices however have a number of weaknesses, principally associated with the subjectivity in their computation, especially in the choice of variables and the weighting procedure. Indeed, composite indices may send misleading policy messages if they are poorly constructed or misinterpreted, and may invite simplistic policy conclusions. They may also be misused, e.g., to support a desired policy, if the construction process is not transparent and if the methodology lacks sound statistical or conceptual principles. The selection of indicators and weights could be the target of political challenge.

Composite indices are averages of different sub-indices and the single value which they produce may conceal divergences between the individual components or sub-indices, possibly hiding useful information.

Furthermore, a composite index may require some form of trade-off between the sub-indices of the composite index and averaging would conceal, for example, situations where the effect of one variable cancels out the effect of another. It may thus disguise serious failings in some dimensions and increase the difficulty of identifying proper remedial action, and, may lead to inappropriate policies if dimensions of performance that are difficult to measure are ignored. Moreover, measurement problems may arise due to absence of data for certain variables or for certain countries, different methods of statistical compilation across countries and errors in measurements of the variables.

Quality Frameworks for Developing Statistics and Composite Indices

There have been several calls for a framework for classifying and evaluating composite indices. Drewnowki (1972: 77) claimed that one requires some "ordering principles for the selection of useful indicators and rejection of ill-conceived and inapplicable ones". Wish (1986: 97-98) similarly argued that "indicators require a systematic rationale for categorisation". Several organisations and individual researchers have defined desirable attributes for statistics and indicators (see IMF, 2003; OECD, 2003; Booysen, 2002; Briguglio, 2003; JRC-OECD, 2005). The next section discusses these desirable features.

IMF (2003): Data Quality Assurance Framework

The IMF (2003) has developed the 'Data Quality Assurance Framework' (DQAF) to assess the overall quality of statistics produced by its member countries. 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 statistics activities. The IMF assesses the overall quality of statistics produced by its member countries based on the following five dimensions, namely (1) assurance of integrity: the features which support firm adherence to objectivity in the production of statistics; (2) methodological soundness: how current practices relate to internationally agreed methodological practices for specific statistical activities; (3) accuracy and reliability: the adequacy of the source data statistical techniques to portray the reality to be captured; (4) serviceability: the way in which users' needs are met in terms of timeliness of the statistical products, frequency, consistency and revision cycle; and (5) accessibility: whether effective data and metadata are easily available to data users and whether there is assistance to users.

OECD (2003): Quality Framework and Guidelines for OECD Statistics

The OECD's 'Quality Framework and Guidelines for OECD Statistics' (OECD, 2003) is built on seven dimensions, namely (1) relevance: a careful evaluation and selection of basic data has to be carried out to ensure that the right range of domains is covered in a balanced way, implying that relevance has to be evaluated considering the overall purpose of the indicator; (2) accuracy: the degree to which basic data correctly estimate or describe the quantities or characteristics that they are designed to measure; (3) credibility: refers to the fact that the data are perceived to be produced professionally in accordance with appropriate statistical standards and policies that are transparent, and implying that, other things being equal, data produced by official sources are preferred to other sources; (4) timeliness: reflects the length of time period between their availability and the event or phenomenon they describe; (5) accessibility: reflects how readily the data can be located and accessed from original sources; (6) interpretability: reflects the ease with which the user may understand and properly use and analyse the data; and (7) coherence: reflects the degree to which they are logically connected and mutually consistent.

Booysen (2002): Dimensions for Classifying and Evaluating Indicators

Booysen (2002) lists seven general dimensions for classifying and evaluating development indicators, which can be applied to other types of indicators. These are: (1) content - the aspects that the indicator measures; (2) technique and method - the method in which the indicator measures the concept, that is, quantitative (qualitative), objective (subjective), cardinal (ordinal), or uni-dimensional (multi-dimensional) manner; (3) comparative application - whether the indicator compares the concept (a) across space (cross-section) or time (time-series), and (b) in an absolute or relative manner; (4) focus - whether the indicator measures the concept in terms of input (means) or output (ends); (5) clarity and simplicity - how clear and simple the indicator is in its content, purpose, method, comparative application and focus; (6) availability - how readily available data are on the particular indicator across time and space; and (7) flexibility - how relatively flexible the indicator is in allowing for changes in content, purpose, method, comparative application and focus.

Briguglio: Desirable Characteristics for Developing Composite Indices

The desirable characteristics suggested by Briguglio (2003) refer to composite indices. In his work on vulnerability indices, the author suggested that a composite index should be constructed on the basis of a number of criteria including the following: simplicity and ease of comprehension; affordability; suitability for international and temporal comparisons; and transparency. He states that one of the advantages of simplicity and ease of comprehension is that these render the index easier to use by policymakers and other users of the index. Affordability, implying that data should be relatively easy to obtain and to process, is important because it permits replication by third parties for evaluation and verification. In turn, this is related to transparency.

In the case of cross-country analysis, Briguglio states that preferably, the data should be collected as a matter of routine in line with the information required for the management of a country. Suitability for international and temporal comparisons implies that the index is based on variables which are measured in a homogenous manner, internationally and temporally. Transparency requires that the methodology used should be clearly explained by those constructing the index. This is essential for validation, evaluation and quality control purposes.

JRC-OECD (2005): Handbook on Constructing Composite Indicators

The Eurostat framework (JRC-OECD, 2005) is based on seven dimensions, namely: (1) relevance: whether the data are what the user expects; (2) accuracy: whether the figures are reliable; (3) comparability: whether the data are in all necessary respects comparable across countries; (4) completeness: whether the domains for which statistics are available reflect the needs expressed by users; (5) coherence : whether the data are coherent with other data; (6) timeliness and punctuality: whether the user receives the data in time and according to pre-established dates; and (7) accessibility and clarity: whether the figure is accessible and understandable.

The *Handbook on Constructing Composite Indicators*, developed by JRC-OECD (2005), states that the construction of a composite index involves ten steps. These are:

1. *Theoretical framework:* Developed to provide the basis for the selection and combination of single indicators into a meaningful composite index, clearly defining the phenomenon to be measured and its sub-components.

2. *Data selection:* Indicators should be selected on the basis of their analytical soundness, measurability, coverage, relevance to the phenomenon being measured and relationship to each other.

3. *Multivariate analysis:* To investigate the overall structure of the indicators, assess the suitability of the data set, deciding whether the structure of the composite index is well-defined, if the set of the available indicators is sufficient or appropriate to describe the phenomenon, and explain the methodological choices, e.g., weighting and aggregation. Methods include: principle components analysis; factor analysis; and, the Cronbach coefficient alpha (c-alpha).

4. *Imputation of missing data:* The approaches for dealing with missing values include: data deletion; mean substitution; regression; multiple imputation; nearest neighbour; or ignoring the missing data, i.e., take the average index of the remaining indicators.

5. *Normalisation:* This is required to adjust the different variables on dimensions such as size, population or income and smoothened through

time against cyclical variability as well as to put the different variables on a common basis. Commonly used methods include: ranking of indicators across subjects (e.g., countries); standardisation (or z-scores); rescaling; distance to a reference subject; categorical scales; values above or below the mean; and percentage of differences over consecutive time points.

6. *Weighting:* Although equal weighting is a commonly used method for, different weights may be assigned to component series in order to reflect their economic significance. Weights may be derived either from statistical models (principal components analysis, data envelopment analysis, regression analysis, unobserved components models), or based on public/expert opinion (budget allocation, public opinion, analytic hierarchy process, conjoint analysis). No matter which method is used, the assignment of weights involves essentially value judgements.

8. *Aggregation:* The most commonly used methods are additive techniques that range from summing up rankings in each indicator to aggregating weighted normalised indicators. However, additive aggregations imply requirements and properties, which are often not desirable and, at times, difficult to meet or burdensome to verify. Thus, the literature proposes other, and less widespread, aggregation methods such as multiplicative (e.g., geometric) aggregations or non-compensatory aggregations, such as the multi-criteria analysis.

8. *Sensitivity analysis:* Sensitivity analysis is the study of how variations in the outcomes of a model can be apportioned, qualitatively or quantitatively, to variations in the assumptions (Saltelli et al., 2004). Such a study can measure the extent to which scores of a given composite index depend on its components. Uncertainty analysis is used to quantify the overall variation in the scores of an index resulting from the uncertainties in the model input (Jamison and Sandbu, 2001).

9. *Link to other measures:* The relevance and interpretability of the results can be strongly reinforced by the comparison between the composite indicator and other well known and "classical" measures of relevant phenomena. In addition, the credibility of the indicator can benefit by its capacity to produce results which are highly correlated with the reference data.

10. *Visualisation:* Composite indicators can be visualised or presented in a number of different ways, which can influence their interpretation and be able to communicate a picture to decision-makers and other end-users quickly and accurately. This can be done using simple tabular tools or more complicated multi-dimensional graphics and interactive

software. Composite indicators should be transparent and capable of being decomposed back into their underlying indicators or values.

JRC-OECD (2005) argues that each phase of the composite index building process is important. The design of the theoretical framework can affect the relevance of the index; the multivariate analysis is important to increase its reliability; the imputation of missing data, as well as the normalisation and the aggregation procedure, can affect its accuracy.

Furthermore, JRC-OECD (2005) states that, while each step is extremely important, so is the coherence of the whole process. Choices made in one step can have important implications for other steps. Therefore, the composite indicator developer has not only to make the most appropriate methodological choices in each step, but also to identify if they fit well together.

Synthesis of Desirable Attributes of Composite Indices

As can be seen from the previous section, the quality frameworks for collecting statistics and constructing indices, including composite ones, are fairly similar. All the desirable attributes identified by the quality frameworks described above, when synthesised, can be reduced to seven, as listed below, in what the present author believes to be an order of importance:

1. *Accuracy.* This refers to the degree to which the composite index correctly estimates or describes the quantities or characteristics that it is designed to measure;

2. *Simplicity and ease of comprehension.* This refers to the ease with which the user understands and uses the composite index;

3. *Methodological soundness.* This requires that there is a logical connection between the different sub-indices based on sound conceptual principles;

4. *Consistency in measurement.* This requires that the index is based on variables which are measured in a homogenous manner;

5. *Transparency and Accessibility.* This requires that the methodology and the variables upon which the composite index was constructed are available for other researchers;

6. *Replicability.* This is associated with transparency and refers to the possibility that other researchers replicate the index to verify the results and evaluate the index;

7. Timeliness. This refers to the length of time period between the publication of the composite index and the event or phenomenon it purports to describe.

3. Conceptual Issues in Constructing Composite Indices

The construction of composite indices involves (1) selection of the components; (2) dealing with missing data; (3) normalisation of the observations; (4) weighting and aggregation of the components; and (5) testing and reviewing the results obtained.

Selection of Components

Defining the concept. The strengths and weaknesses of composite indices largely derive from the quality of the underlying variables, which summarise complex information of value to the observer. Before one starts to select the indicators to construct the composite index, one has to start by obtaining a precise definition of the concept to be measured. Then, on the basis of that precise definition, a researcher should search for suitable indicators to measure the defined concept.

Satisfying desirable attributes. Indicators should be selected according to the desirable attributes described above. Thus, indicators should be selected on the basis of their analytical soundness, measurability, coverage, relevance to the phenomenon being measured and relationship to one another.

Begging the question. It is important that when a composite index is designed to prove a hypothesis or some other relationship, it does not include among the indicators, those variables or relationships which it was designed to prove.

Reviewing data. Although the choice of indicators must be guided by the theoretical framework, the data selection process can be quite subjective as there may be no readily available indicators to measure the phenomenon in question. Prior to the search for indicators, it is useful to draft a tentative indicator set, i.e., an ideal set of indicators irrespective of their actual or potential availability. Every effort should be made to retain on this list indicators that are deemed important, even though the data may not be available and a researcher may have to rely on proxy variables.

Avoiding Redundancies. The number of variables making up a composite index should be as small as possible. This is due to various reasons, one of them being the fact that there is an element of trade-off between the richness of information and the ease of communication. Indeed, the more comprehensive a composite index is, the weaker it may be in adequately reflecting performance. Another reason is that combining too much information from diverse areas risks becoming meaningless. Furthermore, it can also be argued that there is a trade-off between the number of indicators and the cost of obtaining the information, with too many indicators rendering the composite index unaffordable. However, this does not imply that the composite index should have fewer indicators than necessary. Paraphrasing Albert Einstein, indicator sets should be as simple as possible, but not simpler (Bossel, 1999). The composite index must be made up of a comprehensive and compact set of variables, covering all relevant aspects, suggesting that a composite index has an optimal number of indicators.

Reducing the number of variables. The number of variables used can be reduced by principal components analysis (PCA), a geometric method that reduces the number of variables by creating a new set of variables that are linear combinations of the existing variables. It transforms correlated variables into a new set of uncorrelated variables using a covariance matrix or a correlation matrix. The objectives of PCA are (1) dimensionality reduction; (2) the determining of linear combinations of variables; (3) feature selection: the choosing of the most useful variables; (4) visualisation of multi-dimensional data; (5) identification of underlying variables; and (6) identification of groups of objects or of outliers. 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. Factor analysis (FA) is also used as a tool in attempts to reduce a large set of variables into a smaller set. It is similar to PCA but it is based on a particular statistical model (Spearman, 1904). FA assumes that the data are 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. Similar to FA is correspondence analysis, a descriptive/exploratory technique designed to analyse simple two-way and multi-way tables containing some measure of correspondence between the rows and columns, which is, however, better suited for qualitative data. For a comprehensive description of this method, computational details and its applications, see Greenacre (1994). An extension of simple correspondence analysis to more than two variables is called multiple correspondence analysis.

Checking for correlation between the components or sub-indices. When one develops a composite index there is a risk of an element of overlap in what the different variables attempt to measure, especially if the different variables are made up of sub-indices. It is thus useful to carry out a rank correlation test to check for correlation between the different variables. If there is a high correlation between any two or more variables, it is suggested that one of the variables is discarded. This principle ties in with the principle of having a small number of variables, which helps in the operational function of the index as well as in the ease of comprehension of the index. If the composite index is made up of some highly correlated variables, this may cause the index to be biased in favour of these variables, as it implies that a higher weight is attached to these variables. Variables that are uncorrelated below a certain threshold should also be discarded as redundant. An alternative way to investigate the degree of correlation among a set of variables is the Cronbach coefficient alpha (c-alpha). If the correlation is high, then there is evidence that the indicators are measuring the same underlying construct. If the reliability coefficient increases after deleting a sub-indicator from the scale, one can assume that the sub-indicator is not highly correlated with other sub-indicators in the scale.

Cluster analysis. Cluster analysis and discriminant analysis can also be used to avoid redundancy. Cluster analysis is a multivariate procedure for detecting natural groupings in data and is sometimes used to aggregate the data in a composite index. It is based upon the placing of objects into more or less homogeneous groups, in a manner such that the relationship between groups is revealed. Cluster analysis lacks an underlying body of statistical theory and is more objective than subjective (Wulder, 2005). Homogeneous and distinct groups are delineated based upon assessment of distances or, in the case of Ward's method, an F-test (Davis, 1986).

Discriminant analysis. Another method is discriminant analysis, which can be used either to assess the adequacy of classification given group memberships of the objects under study, or used to assign objects to one of a number of (known) groups of objects. Although discriminant analysis is relatively robust to non-normality due to skewness, it is highly sensitive to outliers. Variables with significant outliers necessitate transformation prior to analysis. Linearity is also assumed for discriminant analysis (Wulder, 2005).

Dealing with Missing Data

When constructing a composite index comprising several different variables and a large number of countries, it is inevitable that some indicators will be unavailable for some countries. It is thus important to analyse what can be done in such a situation. Methods to deal with missing data can be split into two categories: single imputation, which substitutes a value for each missing value and multiple imputation, which replaces each missing value with a range of plausible values. There is an extensive literature on the analysis of missing data, including Little (1997), Yuan (2000), Rubin (1987), Lavori et al. (1995) and Schafer (1997).

Excluding subjects from the analysis. One way to deal with missing data is to exclude the subject (e.g., country) from the composite index construction if it includes an unavailable observation. The argument in favour of such a rule is that a missing observation may cause the results to be biased in favour of the other available indicators and may render the composite indicator values incomparable with other countries. Another argument in favour of excluding subjects with missing indicator values is that aggregating only the available indicators can also negatively affect the credibility of the composite index, as an analyst would have to check which indicators are available and which are not, and it would be difficult to compare the results across time and space. The disadvantage with such an approach is that the researcher ends up with a smaller sample. For this reason, imputation methods are sometimes applied.

Single imputation methods. As mentioned earlier, single imputation methods substitute a value for each missing value. A brief description of the main single imputation methods, namely case deletion, mean/median/mode estimation, cold and hot deck imputation and regression imputation, raw maximum likelihood and expectation maximisation imputation, follows. This list is by no means exhaustive. Case deletion simply omits the missing records from the analysis. However, this approach ignores possible systematic differences between complete and incomplete samples and may produce biased estimates. Cold deck imputation recovers an observation by checking whether the observation is available for a previous year. This option is useful in composite indices components with scores that are not expected to change much over a few units of time. This method is a very popular missing data imputation procedure. Mean/median/mode estimation replaces missing data with the mean of non-missing values. The disadvantage is that the standard deviation and standard errors are underestimated. Hot deck imputation involves stratifying and sorting the data by key covariates, and then replacing missing data from another record in the same strata. More simply, it involves analysing the dataset and checking whether there is a similar subject (e.g., country) with the similar characteristics, and then replacing the missing indicator with the indicator available in the similar country. However, here again, underestimation of standard errors can be a problem. Regression imputation imputes each independent variable on the basis of other independent variables in the model, but may produce biased estimates. It is also likely to over fit the data and result in correlations to be unrealistically high. Also, for every country the missing observation is conditioned by the observations in other countries. In general, single imputation results in the sample size being over estimated with the variance and standard errors being underestimated. Graham et al. (2003) referred to single imputation methods as "unacceptable methods". Multiple imputation methods were developed in order to overcome these problems.

Multiple imputation methods. Contrary to single imputation methods, which substitute a value for each missing value, multiple imputation replaces each missing value with a range of plausible values. The main types of multiple imputation methods are the regression method, the propensity score method and the Markov Chain Monte Carlo algorithm. A brief description of each follows. In the parametric regression method, a regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then fitted and used to impute the missing values for each variable (Rubin, 1987). This method is useful for monotone missing data patterns. The propensity score is the conditional probability of assigning to a particular treatment given a vector of observed covariants (Rosenbaum and Rubin, 1983). In the propensity score method, a propensity score is generated for each variable with missing values to indicate the probability of that observation being missing. The observations are then grouped based on these propensity scores, and an approximate Bayesian bootstrap imputation (Rubin, 1987) is applied to each group (Lavori et al., 1995). In the Markov Chain Monte Carlo algorithm, one constructs a Markov Chain—a sequence of random variables in which the distribution of each element depends on the value of the previous one—long enough for the distribution of the elements to stabilise to a common distribution (Schafer, 1997). By repeatedly simulating steps of the chain, it simulates draws from the distribution of interest. This method is useful for an arbitrary missing data pattern.

Overview of imputation methods. The imputation of missing data affects the accuracy of the composite index and its credibility (see Allison, 2000). Furthermore, even if timeliness can be improved, extensive use of imputation techniques can undermine the overall quality of the indicator and its relevance. Regression coefficients for predictors with large fractions of imputed data will show substantial biases towards zero (Landerman et al., 1997). Roth (1994), Little and Rubin (2002) and Wothke (1998) reviewed different imputation methods and concluded that case deletion and mean substituting missing data handling methods are inferior when compared with multiple imputation methods. Regression methods are somewhat better, but not as good as hot deck imputation.

It should be observed that multiple imputation theories are still relatively new and are still being developed. Although at present there is still some scepticism about this methodology, it is important to state that the superiority of multiple imputation to traditional methods is based on mathematical fact, not belief or opinion (Wayman, 2003). One can calculate the efficiency of multiple imputation using a ratio developed by Rubin (1987), which analyses the relative increase in variance due to non-response.

Quantifying qualitative data. Sometimes quantitative data may not be available for some indicators or else may be restricted to limited coverage and only qualitative information may be obtained. For this reason, a researcher may have to transform qualitative data into a quantitative format. One way in which this can be carried out is by categorising an occurrence (in terms of intensity or frequency) along a multi point mapping scale, such as the Likert scale. The points on the scale can be for example, from 1 to 7, with 1 being the lowest possible occurrence and 7 the highest possible. The wider the spread of the scale, the more possible will it be to derive meaningful standard deviations of the averages obtained, but there is a limit to how many meaningful categories one can work with (Briguglio, 2003). This approach also permits non-linearity such as, for example, in cases where the occurrence grows or declines exponentially or when it takes a U-shaped or S-shaped pattern. It should be noted that, however, linear mapping is the most common procedure. The main defect of this method relates to the subjectivity of the category groupings and the choice between linear and non-linear relationships.

Normalisation of the Variables

Since the indicators which make up a composite index very rarely have the same units, indicators should be standardised, i.e., converted to a similar unit, in order to render them comparable. Freudenberg (2003) and Jacobs et al. (2004) list a number of normalisation methods. It should be noted that the selection of a suitable method 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 and normalisation may reduce the difference between results if there are large outliers. Robustness tests might be needed to asses their impact on the outcomes. If a composite index is made up of a number of sub-indices and these sub-indices are normalised, it may be useful to re-standardise the composite index after aggregation has been carried out. The two most common methods of normalisation are rescaling and standardisation.

Rescaling is a commonly used normalisation method, where observations are given values of between 0 and 1 by means of the following formula:

$$XSij = (Xij - MinXj) - (Maxj - MinXj)$$

where:
- XSij is the value of the normalised observation for subject i (e.g., a country) of component j;
- Xij is the actual value of the same observation; and
- MinXj and MaxXj are the minimum and maximum values of all observations of component j for all subjects (e.g., for all countries) considered.

Standardisation (or z-scores) is very similar to the above method and converts indicators to a common scale with a mean of zero, as in the following equation:

$$XSij = (Xij - MeanXj) / SDXj$$

where:

- XSij is the value of the standardised observation for subject i (e.g., a country) of component j;
- Xij is the actual value of the same observation;
- Mean Xj is the mean value of all observations of component j for all subjects (e.g., for all countries) considered; and
- SDXj is the mean value of all observations of component j for all subjects (e.g., for all countries) considered.

It should be noted that these normalisation procedures can lead to outlier observations being highly influential in determining outcomes (Atkins et al., 2000). It is therefore useful for the researcher to examine carefully the effect of outliers on relative scores when normalising data.

Another rescaling or normalisation method used is expressing an observation in terms of a relative position of a given indicator vis-à-vis a reference point and transforming observations such that values around the mean receive 0, whereas the ones above/or below a certain threshold receive 1 and -1 respectively. This method has been used by the Summary Innovation Index (European Commission, 2001). Other methods include expressing the observations as a ratio to a given value or in terms of percentage differences from a given value.

The normalisation phase is crucial both for the accuracy and the coherence of final results. An 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.

Weighting

One of the key issues in the construction of composite indices is the choice of the weighting for summing the components. Almost all quality dimensions are affected by this choice, especially accuracy, coherence and interpretability (JRC-OECD, 2005). This is also one of the most criticised characteristics of composite indices, and the one which generates most debate. An important issue relates to whether equal weights or differential weights are to be used, and if the latter are chosen, how to derive the differential weights. Aggregation/weighting questions have been extensively studied in the literature on productivity indices (Balk, 2002). This section will provide an analysis of equal and differential weights, as well as of the main theories that can be used to derive differential weights.

Equal weighting. Most composite indices rely on equal weighting, i.e., all variables are given the same weight. This could be a result of the fact that all variables making up the composite index are deemed to be of equal importance to the concept to be measured, but it could also be a result of lack of consensus on an alternative, or insufficient knowledge. When using equal weights, it may happen that, by combining variables with a high degree of correlation, one may introduce an element of double counting into the index. It is thus useful to test the indicators for statistical correlation, for example with the Pearson correlation coefficient (Manly, 1994), and choosing only indicators exhibiting a low degree of correlation or giving less weight to correlated indicators. Furthermore, minimising the number of variables in the index using the methods described earlier may be desirable on other grounds, such as transparency and parsimony.

Differential weighting. It is often argued that equal weights render the concept too simplistic and that instead, indicators should be weighted and aggregated according to the underlying theoretical framework of the concept being measured. The OECD (2003) states that "greater weight should be given to components which are considered to be more significant in the context of the particular composite indicator". It should be noted that when equal weights are applied, if the variables are grouped into components and those are further aggregated into the composite, then applying equal weights to the variables may imply an unequal weighting of the component (JRC-OECD, 2005).

Subject-specific or indicator-specific weights. Problems arise in the determination of differential weights and whether they should be country-specific or indicator-specific. In the case of the former one can argue that country-specific weights render the composite index incomparable between different countries. On the other hand, indicator-specific weights may imply that although an indicator may have less socio-economic and/or political implications for one country compared to another, it will have to be given the same importance in the composite index according to the weight applied.

Changing weights over time. 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. If instead, the objective of that analysis is that of defining best practices or that of setting priorities, then weights should necessarily change over time. In the construction of price indices, a Laspeyres index is used for constant weights, while a Paasche index is used for changing weights.

Weights reflecting the statistical quality of the data. 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.

Regression method. In deriving weights using the regression method, one uses as a dependent variable a proxy of the composite index and this is regressed on a number of explanatory variables which represent the components of the composite index. The coefficients on the explanatory variables of the estimated equation are taken as weights for averaging the components of the index. Since this approach lets the data produce the weights, it does not require normalisation of the observations. The procedure has a number of methodological defects, which limit the operationality and reliability of a composite index aggregated using this method. The most important methodological defect is that if the dependent variable is considered to be a proxy for the variable to be indexed, one need not go through a cumbersome regression procedure to compute the index (Briguglio, 2003). Other defects are that the regression may produce negative coefficients, which would imply the use of negative weights; the regression may produce weights which are very small, almost significant and comparatively relatively large coefficients, which would also imply the use of weights which are very small and weights which are comparatively large in the same index. Another defect is that, since the coefficients pertain to data with different units and varying distributions, it is not possible to estimate the weight of each variable in the composite index.

Stochastic weights. This technique was developed by Anders Hoffmann, ex-OECD and a co-author of the Handbook on Constructing Composite Indicators (JRC-OECD, 2005), and it generates sets of random weights each of which sums to 1. The main defect of this procedure is that the weights are generated in a procedure which assigns too much value to chance. Also, this procedure does not assign a greater weight to components which are considered to be more significant in the context of the particular composite index.

Participatory methods. Participatory methods can be used to assign weights, either those that incorporate various stakeholders (Moldan and Billharz, 1997) or those that make use of public opinion polls (Parker, 1991). 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, 1997).

Weights based on the precautionary principle. Closely related to the above, that is, based on expert opinion, is the determination of weights based on the precautionary principle. In this procedure, experts assign differential weights to the various components, and a large weight is assigned to that component which is expected to be crucial to attaining the phenomenon the composite indicator is attempting to measure.

"Benefit-of-the-Doubt" weighting system. The "benefit-of-the-doubt" weighting system, proposed by Melyn and Moesen (1991), chooses the weights such that the evaluated country has a maximal composite index value. Also referred to as endogenously-weighted composite indicators, the method is based on the Data Envelopment Analysis (DEA) method (Farrell, 1957; Charnes et al., 1978). The core idea is that a country's relatively good performance in some dimensions is indicative of the fact that this country considers the concerned policy dimensions as relatively more important (Van Puyenbroeck, 2005). This method has a high political acceptance as no other weighting scheme yields a higher composite index value. The principle is also easy to communicate: if another country, say Country B, gets a higher overall score using Country A's assigned weighting scheme, this implies that Country B is outperforming Country A. The "benefit-of-the-doubt" approach is useful when individual expert opinion is available, but when experts disagree about the right set of weights. A possible criticism of the "benefit-of-the-doubt" approach is that it makes performance look better than what it really is, since the selected weights can deviate from the true but

(unknown) priorities. The method also does not exclude extreme scenarios where all the relative weight is assigned to a single indicator, which would then completely determine the overall index value. Some restrictions can be imposed as in Cherchye et al. (2004) where they did not allow the sum of weights in each category to exceed the sum of weights in another category by more than 20 percent.

Aggregation

Linear and geometric aggregation. 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. Furthermore, linear aggregations reward base-indicators proportionally to the weights, while geometric aggregations reward those countries with higher scores (JRC-OECD, 2005).

Aggregation methods and weighting systems. 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. The assumption of preference independence is essential for the existence of a linear aggregation rule (Munda and Nardo, 2003). Thus, from a mathematical point of view, given the variables, X1, X2, ..., Xn, an additive aggregation function exists if and only if these variables are mutually preferentially independent. In terms of policy, when geometric aggregation is used, a country with low scores on one indicator will need a much higher score on the others to improve its situation, implying that in benchmarking exercises, countries with low scores prefer a linear rather than a geometric aggregation. Also, 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.

Non-compensatory multi-criteria aggregation. The averaging procedure can hide important information in the individual components. For example, three bad scores can possibly yield the same average as two good scores and one very bad score. A method proposed to take consideration of this problem is the non-compensatory multi-criteria approach (see Munda, 2005). This method involves assigning various weights for each component, leading to various permutations. This method is, however, very computationally costly when the number of subjects is large, as the number of permutations to be calculated increases exponentially.

Subjective choices. 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 objectives must be clearly stated at the outset, and the chosen model must be checked to see to what extent it fulfils the goals. No matter which technique is used, weights are effectively value judgements. In many instances, choosing differential component weights on conceptual grounds may be a fruitless exercise. McGillivray et al. (2008) show that, with regard to the Human Development Index, if the components of an index are correlated, variable weights produce results which are indistinguishable from those produced by equal weights.

Testing and Reviewing the Results Obtained

Uncertainty and sensitivity analysis. When one constructs a composite index, it is useful to test the robustness of that index, which depends on a number of factors including missing data, the choice of the imputation algorithm and the choice of weights. This is usually done by means of uncertainty and sensitivity analysis, the iterative use of which during the development of a composite index could improve its structure (Saisana et al., 2005; Tarantola et al., 2000). The method includes the inclusion and exclusion of certain components, modelling data errors based on the available information on variance estimation, using alternative and normalisation schemes and using different weighting and aggregation schemes. Again here such analysis will never yield totally objective results, however, it may enhance the transparency of the exercise.

Outliers. It is also useful to check the index results for any outliers through, for example, a visual inspection of the data or by plotting the data in scatter diagrams. An outlier can be due to an error in inputting the data. When it is derived from correct data, one may consider leaving it out as this can bias the results, especially when carrying out the normalisation procedure. However, the exclusion of outliers is somewhat subjective and may also lead to biased results. Again here, in the interest of transparency, results with and without the outliers should be reported.

Analysing the results obtained. When one analyses the scores and/or rankings derived from a composite index, it is important to assess not just the final results obtained but also the results of the sub-components of the index, in order to see whether the average score of the component is meaningful. It may be useful, for example, to discuss whether a given score derived from averaging three components with very different values can be interpreted differently from a similar score derived from three components with a similar values.

4. Conclusion

This chapter has shown that composite indices have their advantages and disadvantages and, as Saisana et al. (2002) argue, it is not likely that the debate on the practical uses of composite indicators will ever be settled. All things considered, composite indicators should be accepted for what they are: simplified representations of complex realities, often based on subjective choices of component sub-indices.

However, the importance of composite indices should not be understated. If an index is based on sound methodological criteria and constructed according to the desirable attributes described above, in particular transparency, then it could be a valuable instrument. The Human Development Index is a case in point. Although this index is by no means perfect, it has served to highlight the importance of education and health in human development and has generated considerable debate on the need to augment simple GDP per capita indicators.

References

Allison, P. (2000). "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research*, Vol. 28(3): 301-309.

Atkins, J., Mazzi, S., and Easter, C. (2000). "Commonwealth Vulnerability Index for Developing Countries: The Position of Small States" *Economic Paper, No 40*. London: Commonwealth Secretariat.

Balk. B. (2002). *The Residual. On Monitoring and Benchmarking Firms, Industries and Economies with Respect to Productivity*. Rotterdam: Erasmus Research Institute of Management.

Booysen, F. (2002). "An Overview and Evaluation of Composite Indices of Development," *Social Indicators Research*, Vol. 59(2): 115–51.

Bossel, H. (1999). *Indicators for Sustainable Development: Theory, Method, Applications—A Report to the Balaton Group*. Canada: International Institute for Sustainable Development.

Briguglio, L. (2003). "The Vulnerability Index and Small Island Developing States: A Review of Conceptual and Methodological Issues," Paper prepared for the AIMS Regional Preparatory Meeting on the BPoA+10 Review, 1-5 September 2003, Praia, Cape Verde.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978). "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, Vol. 2: 429-444.

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*, Vol. 42: 919-955.

Conway, P. (2005). "OECD Composite Indicators: Pros and Cons, Weighting, Sensitivity, Updating Issues." Paper presented at the Workshop on European Indices and Scoreboards, 24-25, October 2005, Brussels.

Davis, J. (1986). *Statistics and Data Analysis in Geology*. Toronto: John Wiley and Sons.

Drewnowski, J. (1972). "Social Indicators and Welfare Measurement: Remarks on Methodology," *Journal of Development Studies*, Vol. 8: 77–90.

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

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

Farrell, M.J. (1957). "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society*, Vol. 120(3): 253-290.

Freudenberg, M. (2003). "Composite Indicators of Country Performance - A Critical Assessment." *Science, Technology and Industry Working Paper, 2003/16*. Paris: OECD.

Graham, J.W., Cumsille, P.E., and Elek-Fisk, E. (2003). Methods for Handling Missing Data," In Schinka, J.A. and Velicer, W.F. (eds) *Research Methods in Psychology*. New York: John Wiley and Sons: 87-114.

Greenacre, M. J. (1994). "Multiple and Joint Correspondence Analysis." In Greenacre, M.J. and Blasius, J. (eds) *Correspondence Analysis in the Social Sciences - Recent Developments and Applications*. London: Academic Press: 141-161.

IMF (2003). "Data Quality Assessment Framework and Data Quality Programme, Fifth Review of the Fund's Data Standards and Initiatives." Prepared by the Statistics Department of the IMF in consultation with other departments.

Jacobs, R., Smith, P. and Goddard, M. (2004). "Measuring Performance: an Examination of Composite Performance Indicators," *Technical Paper Series No. 29*. UK: Centre for Health Economics, University of York.

Jamison, D.T. and Sandbu, M.E. (2001). "WHO Ranking of Health System Performance," *Science*, Vol. 293: 1595-6.

Jesinghaus, J. (1997). "Current Approaches to Valuation." In Molden, B. and Bilharz, S. (eds) *Indicators of Sustainable Development*. Chichester: John Wiley and Sons: 84-91.

JRC-OECD (2005). "Handbook on Constructing Composite Indices: Methodology and User Guide." *OECD Statistics Working Paper*. Paris: OECD.

Landerman, L.R., Kenneth C.L. and Pieper, C.F. (1997). "An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values," *Sociological Methods and Research*, Vol. 26: 3-33.

Lavori, P.W., Dawson, R. and Shera, D. (1995): "A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data," *Statistics in Medicine*, Vol. 14: 1913-1925.

Lievesley, D. (2005). "The Politics of Performance Indicators." Paper presented at the Workshop on European Indices and Scoreboards, 24-25, October 2005, Brussels.

Little, R.J.A. (1997). "Biostatistical Analysis with Missing Data." In Armitage, P. and Colton, T. (eds) *Encyclopedia of Biostatistics*, London: Wiley and Sons.

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

Manly, B. (1994). *Multivariate Statistical Methods*. UK: Chapman and Hall.

McGillivray, M., Noorbakhsh, F. and Gwynn, B. (2008). "Methodological Issues in the Construction of Composite Indices: A Case Study of the Human Development Index." This Volume, Chapter 20.

Melyn, W. and Moesen, W. (1991). Towards a Synthetic Indicator of Macroeconomic Performance: Unequal Weighting when Limited Information is Available." *Public Economic Research Raper No.17*. Leuven: CES, KU.

Moldan, B. and Billharz, S. (eds) (1997). *Indicators of Sustainable Development*. Chichester: John Wiley and Sons.

Munda, G. and Nardo, M. (2003). "On the Methodological Foundations of Composite Indicators Used for Ranking Countries." Universitat Autonoma de Barcelona, Dept. of Economics and Economic History.

Munda, G. (2005). "Aggregation Issues: Non-Compensatory Multi Criteria Analysis." Paper prepared for the Workshop on European Composite Indicators and Scoreboards, Brussels, 24-25, October 2005.

OECD (2004). *Quality Framework and Guidelines for OECD Statistical Activities*. Available at: http://www.oecd.org/statistics/.

Parker J. (1991). "Environmental Reporting and Environmental Indices." PhD Dissertation, UK: Cambridge University.

Rosenbaum, P. R., and Rubin, D. B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70: 41-55.

Roth, P.L. (1994). "Missing Data: A Conceptual Review for Applied Psychologists," *Personal Psychology*, Vol. 47: 537-560.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Saisana, M. and Tarantola, S. (2002). *State-of-the-Art Report on Current Methodologies and Practices for Composite Indicator Development.* Italy: Joint Research Centre of the European Commission.

Saltelli, A., Tarantola, S., Campolongo, F., Ratto, M. (2004). *Sensitivity Analysis in Practice.* New York: Wiley and Sons.

Saisana M., Tarantola S. and Saltelli A. (2005). "Uncertainty and Sensitivity Techniques as Tools for the Analysis and Validation of Composite Indicators," *Journal of the Royal Statistical Society A*, Vol. 168(2): 1-17.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Spearman, C. (1904). "General Intelligence, Objectively Determined and Measured," *American Journal of Psychology*, Vol. 15: 201-293.

Tarantola S., Jesinghaus J. and Puolamaa M. (2000). "Global Sensitivity Analysis: a Quality Assurance Tool in Environmental Policy Modelling." In Saltelli A., Chan K., Scott M. (eds) *Sensitivity Analysis*, New York: John Wiley and Sons: 385-397.

Van Puyenbroeck, T., (2005). "Composite Indicators and Data Envelopment Analysis." Paper presented at the Workshop on European Indices and Scoreboards, Brussels 24-25, October 2005.

Wayman, J.C., (2003). "Multiple Imputation for Missing Data: What is it and How can I use it?" Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.

Wish, N.B. (1986). "Are We Really Measuring the Quality of Life?" *American Journal of Economics and Sociology*, Vol. 45: 93–99.

Wothke, W. (1998). "Longitudinal and Multi-group Modeling with Missing Data." In Little, T.D., Schnabel, K.U. and Baumert, J. (eds) *Modeling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples.* Mahwah, NJ: Lawrence Erlbaum Publishers: 219-240.

Wulder, M., (2005). *A Practical Guide to the Use of Selected Multivariate Statistics.* Canada: Canadian Forest Services.

Yuan, Y., (2000). *Multiple Imputation for Missing Data: Concepts and New Development.* Rockville, MD: SAS Institute Inc.