Rankings and Ratings: Instructions for Use

Michaela Saisana & Andrea Saltelli*

Multidimensional measures (composite indicators, indices, ratings, league tables) can effectively underpin the development of data-driven narratives in support to policy. A controversy surrounds the use of these measures. We review some good and bad practices from the recent literature. We then discuss briefly a decalogue to develop a multidimensional measure. We argue in favor of a multi-modeling approach to represent different scenarios in the construction of an aggregate measure prior to drawing recommendations for policy making. Finally, we try to establish a link between the analytic use of well-designed aggregate measures and the development of a robust culture of evaluation of policies based on evidence. An application of these concepts and tools to the Rule of Law index developed by the World Justice Project is given.

Introduction

The media increasingly consumes statistic-based narratives. Media coverage is evidently on the rise of events such as the publishing of the World Economic Forum's World Competitiveness Index and the Environmental Performance Index, or the OECD-PISA study of verbal and numerical literacy of younger generations, not to mention more specialized measures such as the Transparency International's Corruption Index or the Center for Global Development's Commitment to Development Index, and many others. The economically literate press is particularly keen, with the *Economist*, a weekly journal, being a case in point. Its 'Market and Data' pages are rich with indices built from the *Economist*'s Intelligence Unit, and the journal's prose makes rich reference to measures developed by international organizations and NGO's, often interspersing these with 'classic', or official statistical data, such as demography of GDP.[1]

* European Commission – Joint Research Centre, Institute for the Protection and Security of the Citizen, Unit of Econometrics and Applied Statistics, michaela.saisana@jrc.ec.europa.eu; andrea.saltelli@jrc.ec.europa.eu.

[1] 'THE 1.3m people of Mauritius (…) enjoyed a GDP per person of only $200 (…) ranks first in the latest annual Ibrahim index, (…) 24th spot in the World Bank's global ranking for ease of doing business,' October 2009.

Hague Journal on the Rule of Law, 3 : 247–268, 2011

doi:10.1017/S1876404511200058

Multidimensional measures,[2] else termed composite indicators, are calculated as a function of variables and weights, ideally based on a conceptual or theoretical framework of the issue being tackled. The rising popularity of composite indicators (more than fivefold over the last five years[3]) maybe due to the temptation of simplification: 'the temptation of stakeholders and practitioners to summarize complex and sometime elusive process (e.g., sustainability or a single-market policy) into a single figure to benchmark country performance for policy consumption seems irresistible.'[4] Thus the construction of a composite indicator would be driven by the need for generation of narratives for advocacy in intellectual debates.

The use of aggregate measures is somewhat controversial, and it is beyond doubt that composite indicators are a value laden construct. The controversy on the use of aggregate measures may be seen to unfold along an *analytic* versus *pragmatic* axis. The core of the non-aggregators' argument is in the subjective nature of these measures – especially from what pertains to the selections of variables and their weighting, whilst the point of aggregators is on the practical use of composite indicators, their fitness for the purpose.

The recent 'beyond GDP' debate is also a witness of the new zeitgeist. Since the 1990s there has been a shift from considering just the *analytic* problems associated with the GDP[5] towards broadening the picture to alternative, including multidimensional, indicators of well-being.[6] Some authors[7] consider the 'lack of consensus' is a defining property of composite indicators. The use of weights, and their normative implications, will most likely remain controversial. Although it is possible to argue that weights are analytically derived, it is beyond doubt that composite indicators are a value laden construct. The best that can be achieved under these circumstances is that the normative elements are clearly spelled out and that the measure is technically accurate.

Several constituencies have come to accept an aggregate measure (and reach compromise on weighting) to be used for benchmarking best practices. In the European Commission, composite indicators are often built by a process of consensus involving member states authorities and expert groups (e.g., Summary

[2] In this paper the terms 'composite indicator', 'index', 'aggregate measure', and 'multidimensional measure' are used interchangeably. Also popular are terms such as rating and league table.

[3] The authors' query 'composite indicators' on Scholar Google resulted in 992 hits on October 2005 and are 5340 hits at the time of writing (December 2010).

[4] M. Saisana et al., 'Uncertainty and Sensitivity Analysis Techniques as Tools for the Analysis and Validation of Composite Indicators', in: 2 *Journal of the Royal Statistical Society* A 168 (2005), pp. 307-323

[5] J. Rifkin, *The European Dream* 2004, p. 70.

[6] J.E. Stiglitz et al., 'Report by the Commission on the Measurement of Economic Performance and Social Progress' 2009, <www.stiglitz-sen-fitoussi.fr>.

[7] L. Cherchye et al., 'One Market, One Number? A Composite Indicator of EU Internal Market Dynamics', in: 51 *European Economic Review* (2007), pp. 749-779.

Innovation Index, e-Readiness Business index). Many international organizations likewise audit their indices with large communities of experts and stakeholders. Subjectivity and fitness are both at play when constructing and adopting a multidimensional measure, where inter-subjectivity may be at the core of the exercise, such as when participative approaches (such as budget allocation or analytic hierarchy process) are used to assign weights. Thus, these only apparently conflicting properties underpin composite indicators' suitability for advocacy and we would add that, however good the scientific basis for a given good composite indicator, its acceptance relies on negotiation and peer acceptance.

Section 2 of the paper offers an overview of the various steps in the construction of a composite indicator and underlines the main issues and problems a practitioner would encounter. Section 3 presents an analysis of statistical and conceptual coherence in the WJP Rule of Law Index and an impact assessment of the modeling choices made in its development. Section 4 concludes with brief implications and preliminary findings on the World Justice Project Rule of Law Index and some general recommendations on how composite indicators should be seen.

Steps in the Construction of a Composite Indicator

Making proper conceptual and methodological choices to build a multidimensional measure is as much of an art as it is science. The critical literature review offered in the 2008 OECD *Handbook on Constructing Composite Indicators* discusses plurality of approaches, together with the advantages and limitations of each, and suggests an 'ideal sequence' of steps to construct a composite indicator. Each step is important, and the coherence of the whole process is equally important, since the choices made in one step have important implications for subsequent steps.

Table 1 presents a 'decalogue' to be followed in the construction of a composite indicator. These steps have been put in practice when auditing, upon request of their developers, multidimensional measures such as the UN Multidimensional Poverty Assessment Tool,[8] the Composite Learning Index,[9] the Environmental Performance Index,[10] the Alcohol Policy Index,[11] and the Index of African Governance.[12]

[8] M. Saisana and A. Saltelli, 'The Multidimensional Poverty Assessment Tool (MPAT): Robustness Issues and Critical Assessment', EUR 24310, European Commission, JRC-IPSC, Italy, 2010.

[9] A. Saltelli et al., *Global Sensitivity Analysis. The Primer* 2008.

[10] M. Saisana and A. Saltelli, 'Uncertainty and Sensitivity Analysis of the 2010 Environmental Performance Index', EUR 56990, European Commission, JRC-IPSC, Italy, 2010.

[11] D.A. Brand et al., 'Comparative Analysis of Alcohol Control Policies in 30 Countries', in: 4:4 *PLoS Medicine* (2007), pp. 752-759.

[12] M. Saisana et al., 'A Robust Model to Measure Governance in African Countries', Report 23773, European Commission, JRC-IPSC, Italy, 2009.

We offer some considerations and hints on Steps 1 to 7 in the following section.

Table 1. A decalogue for composite indicator construction

Step 1. Theoretical/conceptual framework
Step 2. Data selection
Step 3. Data treatment
Step 4. Multivariate analysis
Step 5. Normalization
Step 6. Weighting and aggregation
Step 7. Uncertainty and sensitivity analysis
Step 8. Relation to other indicators
Step 9. Decomposition into the underlying indicators
Step 10. Visualization of the results

Source: OECD (2008) Handbook on composite indicators

Theoretical framework

The controversy surrounding multidimensional measures can perhaps be put into context if one considers these measures as models, in the mathematical sense of the term. Models are inspired from systems (natural, biological, social) that one wishes to understand. While a causality entailment structure defines the natural system, and a formal causality system entails the formal system, no rule of encoding the formal system given the real system, i.e., to move from perceived reality to model, was ever agreed.[13]

The formalization of the system generates an image, the theoretical framework, that is valid only within a given information space. As a result, the model of the system will reflect only some of the characteristics of the real system, together with the choices made by the scientists on how to observe the reality. When building a model to describe a real-world phenomenon, formal coherence is a necessary property, yet not sufficient. The model in fact should fit the objectives and intentions of the user, i.e., it must be the most appropriate tool for expressing the set of objectives that motivates the whole exercise. No matter how subjective and imprecise the theoretical framework is, it implies the recognition of the multidimensional nature of the phenomenon to be measured, and the effort of specifying the single aspects and their interrelation.

[13] R. Rosen, *Life Itself* 1991.

Most of the issues described with a composite indicator are complex problems, (for example, welfare, quality of education, or the rule of law). The complexity of the issue is reflected by the multi-dimensionality and multi-scale representation of the issue in the theoretical framework. If we accept a definition of the theoretical framework that requires integrating a broad set of (probably conflicting) points of view, as well as requires and the use of non-equivalent representative tools, then the problem becomes to reduce the complexity in a measurable form. In other words, non-measurable issues, such as the rule of law, need to be replaced by intermediate objectives whose achievement can be observed and measured.

Reducing an entire system into parts has limits when crucial properties of the entire system are lost: often the individual pieces of a puzzle hide the whole picture. As suggested by Box et al., 'all models are wrong, some are useful.'[14] The quality of a composite indicator is thus in its fitness or function to purpose. In practice, a framework should clearly define the phenomenon to be measured and its subcomponents and guide the selection of indicators. Ideally, this process would be based on what is desirable to measure and not on data availability. Transparency throughout the entire process is a necessary (but not a sufficient) condition for the credibility of a composite indicator.

Data selection

Strengths and limitations of composite indicators are strongly related to data quality. Ideally, variables should be selected on the basis of their relevance, analytical soundness, timeliness, accessibility, and other features. While the choice of indicators must be guided by the theoretical framework, 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. Use of soft data may entail the risk of introducing measurement error in the overall scores. 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, is often required. Furthermore, one has to make sure that the type of the selected variables – input, output or process indicators – match the definition of the composite indicator.

Data treatment

Data sets are rarely ever complete. Values for some countries/years may not be available. Imputation of missing data is the art of filling empty spaces in a data

[14] G. Box et al., *Statistics for Experimenters* 1978.

matrix.[15] In general there are three methods for dealing with missing data: a) case deletion; b) single imputation; or c) multiple imputation.[16] The first method, also called complete case analysis, 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. Furthermore, standard errors will in general be larger in a reduced sample given that less information is used. 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-maximization imputation, or multiple imputation, e.g., Markov Chain Monte Carlo algorithm. Data imputation could lead to the minimization of bias and the use of 'expensive to collect' data that would otherwise be discarded by case deletion. 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. 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, or whether extreme values influence the whole exercise.

Besides estimating missing data, preliminary treatment of indicators consists of treating eventual outliers, so as to avoid them becoming unintended benchmarks during the normalization step. Furthermore, outliers can have a strong impact on the correlation structure, and hence, can potentially introduce bias in the interpretation of the results. There are many methods suitable for outlier detection, but in the context of composite indicator building, the combined use of skewness and kurtosis could be particularly apt. A skewness value greater than 1 together with a kurtosis value greater than 3.5 (both in absolute terms) could flag problematic indicators that need to be treated before the final index construction.[17]

Multivariate analysis

The interrelationships between selected indicators is an important element to be taken into account, since they can lead to composite indicators that confuse and mislead both decision-makers and the general public. This step is helpful in assessing the statistical and conceptual coherence of the framework. The analysis can be carried out across individual indicators or countries.

[15] A.P. Dempster and D.B. Rubin, 'Introduction', in W.G. Madow et al., (eds.) *Incomplete Data in Sample Surveys (Vol. 2): Theory and Bibliography* 1983, pp. 3-10.

[16] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data* 2002.

[17] R.A. Groeneveld and G. Meeden, 'Measuring Skewness and Kurtosis', in: 33 *The Statistician* (1984), at pp. 391-399.

Grouping information on individual indicators. The analyst must first decide whether the nested structure of the composite indicator is well-defined and if the selected indicators are sufficient and appropriate to describe the phenomenon. This decision must couple expert opinion with the statistical structure of the data set. Principal component analysis (or Factor analysis, correspondence analysis)[18] or Cronbach Alpha[19] can be used to explore whether the conceptual dimensions of the phenomenon are supported statistically by the selected data. If not, a re-consideration of the indicators might be needed.

Grouping information on countries. In this type of analysis the most similar countries are grouped and studied separately. Cluster analysis[20] is a useful 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 and psychiatry to archaeology. Cluster analysis can also be used to impute missing data (Step 3 above) by using the available information on the countries that belong to the same cluster as the country in question.

Normalization

The individual variables underlying a composite indicator are often expressed in different measurement units. For example, unemployment can be measured in number of persons, health in number of diseases, survey data in high, medium, low, or important, not important. Before aggregation, variables need to be rendered comparable. There are several normalization methods[21] and practitioners should take into account data properties, as well as the objectives of the composite indicator, when choosing a suitable method. The most popular normalization methods are:

a. *Ranking* – this is the simplest normalization technique, which not affected by outliers, however does not preserve the information on the distance between countries.
b. *Standardization* (or z-scores) – converts indicators with a mean of zero and standard deviation of one.
c. *Min-Max* – normalizes indicators within [0, 1] range by subtracting the minimum value and dividing by the range of the indicator values.
d. *Distance to target* – normalizes indicators by dividing the country value with a reference/target value.

[18] B. Manly, *Multivariate Statistical Methods* 1994.

[19] L.J. Cronbach, 'Coefficient Alpha and the Internal Structure of Tests', in: 16 *Psychometrika* (1951), pp. 297-334.

[20] M.R. Anderberg, *Cluster Analysis for Applications* 1973.

[21] Saltelli et al., *Global Sensitivity Analysis*.

Weighting and aggregation

Combining the information on the underlying indicators necessitates deciding a suitable weighting method and aggregation rule to employ. Weighting implies a 'subjective' evaluation, which is particularly sensitive in case of complex, interrelated, and multidimensional phenomena. The menu of weighting methods is rather large, and continues increasing with practitioners' creativity. Ideally, weights should reflect the importance of each indicator to the overall composite. Most composite indicators rely on equal weighting, where all normalized variables are given the same weight.

Statistical models, such as *principal components analysis* or *factor analysis,* could be used to account for the highest variation in the data set, using the smallest possible number of factors.[22] Weighting only intervenes to take into account the overlapping information of two or more correlated indicators, and it is not a measure of the theoretical importance of the indicators. Another statistical method, called *data envelopment analysis* (DEA) is extremely parsimonious about weighting assumptions as it lets the data decide on the weights and is sensitive to national priorities.[23] DEA employs linear programming tools to estimate an efficiency frontier to be used as a benchmark to measure the relative performance of countries. However, given that the weights are derived from the data, they are also subject to eventual data measurement errors.

Multiple regression models can handle a large number of indicators (see Hair et al., 2006).[24] This approach can be applied in cases where the model input are indicators related to various policy actions and the model output is the target. The regression model, thereafter, could quantify the relative effect of each policy action on the output. However, this implies the existence of a 'dependent variable' (not in the form of a composite indicator) that accurately and satisfactorily measures the target in question. Measuring the influence of a number of independent variables on this policy target is a reasonable question. Alternatively, such an approach could be used for forecasting purposes. In the general case of multiple output indicators, canonical correlation analysis – a generalization of multiple regression – could be applied. In any case, there is always the uncertainty that the relations

[22] G. Nicoletti et al., 'Summary Indicators of Product Market Regulation with an Extension to Employment Protection Legislation', OECD, Economics department working papers No. 226, ECO/WKP(99)18, 2000.

[23] W. Melyn and W.W. Moesen, 'Towards a Synthetic Indicator of Macroeconomic Performance: Unequal Weighting when Limited Information Is Available', Public Economic research paper 17, CES, KU Leuven, 1991; L. Cherchye et al., 'Creating Composite Indicators with DEA and Robustness Analysis: The Case of the Technology Achievement Index', in: 59 *Journal of Operational Research Society* (2008), pp. 239-251.

[24] J.F. Hair et al., *Multivariate Data Analysis,* 6th edn., 2006.

captured by the regression model for a given range of inputs and output, may not be valid for different ranges

Alternatively, *participatory methods* that incorporate various stakeholders, from experts to citizens and politicians, can be used to assign weights. In the *budget allocation* approach, experts are given a budget of N points, to be distributed over a number of individual indicators, paying more for those indicators whose importance they want to stress.[25] The budget allocation is optimal for a maximum of ten to twelve indicators, since a large number of indicators 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.[26] The *analytic hierarchy process* (AHP)[27] is also used for multi-attribute decision making, since it assesses the importance of an indicator based on a number of pair-wise comparisons. The resulting weights are less sensitive to errors of judgment. However, since the AHP is based on comparisons of indicator pairs, it is applicable only to few indicators.

Whatever method is used to derive weights, no consensus is likely to exist. This should not preclude the development of a composite indicator, but highlight instead the danger of presenting any composite indicator as 'objective.' At best, weights reflect priorities that have been informed by popular or expert judgments (including the analyst). Assumptions ad implication of the used weighting system should be always made clear and tested for robustness. Soundness and transparency should guide the entire exercise.

The impact of the weighting method used, though significant in most cases, it is not the single most important source of uncertainty. In our experience, other factors have had the same level of impact, if not higher, on final scores and rankings. Examples of such factors include the hierarchical structure chosen to represent the framework, or even the aggregation rule.

Different aggregation rules are possible. Individual indicators could be summed up, multiplied, or aggregated using non-linear techniques. Each technique implies different assumptions, and has specific consequences. *Linear aggregation* is useful when the underlying indicators are correlated and full compensability between indicators is allowed, whilst *geometric aggregation* (multiplication between indicators, where weights appear as exponents) are appropriate when less compensability among indicators is envisaged. Assume that we were to calculate scores in a

[25] B. Moldan et al., 'Sustainability Indicators: Report of the Project on Indicators of Sustainable Development', SCOPE 58, 1997.

[26] J. Parker, *Environmental Reporting and Environmental Indices*, PhD Dissertation, Cambridge, UK, 1991.

[27] R.W. Saaty, 'The Analytic Hierarchy Process: What It Is and How It Is Used', in: 9 *Mathematical Modelling* (1987), pp. 161-176.

dimension of the rule of law, formed by three sub-components for two countries: Country A with values 5, 5, 6; and Country B with values 5, 9, 2. These two countries would have equal scores in the dimension if the arithmetic average is used (assuming equal weights just to make the case). Obviously the two countries represent very different conditions which would not be reflected in the dimension's score. Here, a proper aggregation rule would be one that places country B in a lower position than country A because of the very low score in one of the sub-components. The geometric average fits this purpose:

$$Dimension = \left(I_1^{1/3} \cdot I_2^{1/3} \cdot I_3^{1/3}\right)$$

The expression above embodies imperfect compensability across the three sub-components. It thus addresses one of the most serious criticisms of linear aggregation formula, which allows for perfect compensability across dimensions. Some compensability is inherent in the definition of any index that increases with the value of its components. Adopting the geometric average within a component, as opposed to the arithmetic average, produces lower component values, with the largest changes occurring in households with uneven performance across subcomponents.

When different goals are equally legitimate and important, and in addition trade-offs exist between the dimensions of a composite indicator (namely negative correlations between dimensions) then a non-compensatory logic may be necessary. If the analyst decides that absence of corruption cannot compensate an inefficient criminal justice system, just to remain in a rule of law measurement context, then neither the linear nor the geometric aggregation are suitable. Instead, a non-compensatory *multi-criteria approach* (MCA) will formalize the idea of finding a compromise between two or more legitimate goals.[28] Non-compensatory MCA provides an overall ranking that is only based on the weights and on the sign of the difference between country values for a given indicator. The magnitude of the difference between country scores for a given indicator is ignored and no aggregation formula is employed. Consequently, no compensation occurs.

Like any other method, multi-criteria analysis has pros and cons. At least in its basic form this approach does not reward outliers, i.e., those countries having large advantages (disadvantages) in individual indicators since it keeps only the ordinal information. Another disadvantage is the computational expensiveness when the number of countries is high (the number of permutations to calculate grows exponentially).

[28] G. Munda, *Social Multi-criteria Evaluation for a Sustainable Economy* 2008.

Uncertainty and sensitivity

The construction of composite indicators involves stages where judgments have to be made, such as selection of data, data quality, data treatment (e.g., imputation), data normalization, weighting method, weights, and aggregation rule. Hence, the message brought by the composite indicator deserves analysis and corroboration.[29] *Uncertainty analysis* (UA) and *sensitivity analysis* (SA) can be instrumental in this respect.[30] UA focuses on how the sources of uncertainty propagate through the structure of the composite indicator, and how the sources affect the composite scores. SA studies how much each individual source of uncertainty contributes to the variance of a country's composite indicator score or rank. The synergistic use of UA and SA has proven to be more powerful[31] than the application of UA alone. When assessing the impact of the various sources of uncertainty (data errors), the composite indicator is no longer a magic number corresponding to crisp data treatment and subjective choices, but reflects uncertainty and ambiguity in a more transparent and defensible fashion.

Case Study: The WJP Rule of Law Index

The definition of the term 'rule of law' is the subject of wide debate in a post-natural law world.[32] A meta-analysis of the many current definitions concludes that they refer to five constituent elements: a) capacity of legal rules/principles to guide people in the conduct of their affairs; b) efficacy; c) stability; d) supremacy of legal authority for citizens and governmental actors; and e) availability of impartial institutions of enforcement.[33]

These five constituent elements of the various definitions of rule of law are included in the definition of the WJP Rule of Law Index that is built on a rules-based system. Four universal principles are upheld: a) the government and its officials and agents are accountable under the law; b) the laws are clear, publicized, stable, and fair, and protect fundamental rights, including the security of persons and property; c) the process by which the laws are enacted, administered, and

[29] M. Saisana et al., 'A Robust Model to Measure Governance in African Countries', Report 23773, European Commission, JRC-IPSC, Italy, 2009.

[30] Saltelli et al., *Global Sensitivity Analysis*.

[31] Saisana et al., 'Uncertainty and Sensitivity Analysis Techniques as Tools for the Analysis and Validation of Composite Indicators'.

[32] M.A. Thomas, 'What Do the Worldwide Governance Indicators Measure?', in: 22:1 *European Journal of Development Research* (2010), pp. 31-54; S. Skaaning, 'Measuring the Rule of Law', in: 63:2 *Political Research Quarterly* (2010), pp. 449-460.

[33] J. Fallon and H. Richard, '"The Rule of Law" as a Concept in Constitutional Discourse', in: 97:1 *Columbia Law Review* (1997) pp. 1-56.

enforced is accessible, fair, and efficient; and d) access to justice is provided by competent, independent, and ethical adjudicators, attorneys or representatives, and judicial officers who are of sufficient number, have adequate resources, and reflect the makeup of the communities they serve. These four principles rigorously define what the researchers aimed to measure by building the WJP Rule of Law before they set out to measure it.

The WJP Rule of Law Index is not the first attempt to capture governance or rule of law worldwide. Several measures of public sector capacity exist, for example the *World Bank Governance Indicators*, *Index of State Credibility*, *Participatory Development and Good Governance*, *Corruption Perception Index*, *International Country Risk Guides*, *Civil Liberties and Political Rights*, *Index of Economic Freedom*, and *Index of African Governance*. In all these attempts, quantifying the complex concepts underlying the rule of law with single index numbers raises several practical challenges, such as the selection of indicators, the quality of data, and the statistical combination of these into a model, as described earlier. Yet, if done properly, the exercise could yield a useful tool capable of assessing nations' efforts in delivering the rule of law to their citizens. The tool could be used for benchmarking purposes across space and time, monitoring changes, identifying problems, and contributing to priority setting and policy formulation.

The assessment of conceptual and statistical coherence of the WJP Rule of Law framework and the estimation of the impact of the modeling assumptions on a country's performance are necessary steps to ensure the transparency and reliability of the WJP Rule of Law Index and enable policymakers to derive more accurate and meaningful conclusions. A careful assessment of the WJP Rule of Law Index was guided by two key questions.

a. Is the Index conceptually and statistically coherent?
b. What is the impact of key modeling assumptions on the Rule of Law Index results?

Conceptual and statistical coherence of the Rule of Law framework

The *WJP Rule of Law Index 2010* framework (version 3.0) is shown in Table 2.

Currently, the WJP Rule of Law Index Framework is populated with roughly 500 variables grouped in nine factors:[34] 1) limited Government Powers; 2) absence of corruption; 3) clear, publicized and stable Laws; 4) order and security; 5) fundamental rights; 6) open government; 7) regulatory enforcement; 8) access to

[34] The conceptual framework for the *WJP Rule of Law Index* 2010 comprises a tenth factor on Informal justice. These ten factors are further disaggregated into forty nine sub-factors. The scores of these sub-factors are built from over five hundred variables (survey items) drawn from assessments of the general public (thousand respondents per country) and local legal experts.

Table 2. WJP Rule of Law Index – Conceptual framework[35]

The *WJP Rule of Law Index*™, version 3.0

Version 3.0 of the *Index* is composed of 10 factors derived from the WJP's universal principles. These factors are divided into 49 sub-factors which incorporate essential elements of the rule of law.

Factor 1: Limited Government Powers

1.1	Government powers are effectively limited by the fundamental law
1.2	Government powers are effectively limited by the legislature
1.3	Government powers are effectively limited by the judiciary
1.4	Government powers are effectively limited by independent auditing and review
1.5	Government officials are sanctioned for misconduct
1.6	Freedom of opinion and expression
1.7	The State complies with international law
1.8	Transition of power occurs in accordance with the law

Factor 2: Absence of Corruption

2.1	Government officials do not request or receive bribes
2.2	Government officials exercise their functions without improper influence
2.3	Government officials do not misappropriate public funds or other resources

Factor 3: Clear, Publicized and Stable Laws

3.1	The laws are comprehensible to the public
3.2	The laws are publicized and widely accessible
3.3	The laws are stable

Factor 4: Order and Security

4.1	Crime is effectively controlled
4.2	Civil conflict is effectively limited
4.3	People do not resort to violence to redress personal grievances

Factor 5: Fundamental Rights

5.1	Equal treatment and non-discrimination are effectively guaranteed
5.2	The right to life and security of the person is effectively guaranteed
5.3	Due process of law and rights of the accused are effectively guaranteed
5.4	Freedom of opinion and expression is effectively guaranteed
5.5	Freedom of belief and religion is effectively guaranteed
5.6	Freedom from arbitrary interference with privacy is effectively guaranteed
5.7	Freedom of assembly and association is effectively guaranteed
5.8	Fundamental labor rights are effectively guaranteed

Factor 6: Open Government

6.1	Administrative proceedings are open to public participation
6.2	Official drafts of laws and regulations are available to the public
6.3	Official information is reasonably available

Factor 7: Regulatory Enforcement

7.1	Government regulations are effectively enforced
7.2	Government regulations are applied and enforced without improper influence
7.3	Due process is respected in administrative proceedings
7.4	The Government does not expropriate private property without adequate compensation

Factor 8: Access to Civil Justice

8.1	People are aware of available remedies
8.2	People can access and afford legal counsel in civil disputes
8.3	People can access and afford civil courts
8.4	Civil justice is impartial
8.5	Civil justice is free of improper influence
8.6	Civil justice is free of unreasonable delays
8.7	Civil justice is effectively enforced
8.8	ADR systems are accessible, impartial, and effective

Factor 9: Effective Criminal Justice

9.1	The criminal investigation system is effective
9.2	The criminal adjudication system is timely and effective
9.3	The correctional system is effective in reducing criminal behavior
9.4	The criminal justice system is impartial
9.5	The criminal justice system is free of improper influence
9.6	Due process of law and rights of the accused are effectively protected

Factor 10: Informal Justice

10.1	Informal justice systems are timely and effective
10.2	Informal justice systems are impartial and free of improper influence
10.3	Informal justice systems respect and protect fundamental rights

[35] Mark D. Agrast et al., *The World Justice Project Rule of Law Index 2010* 2010.

civil justice; and 9) effective criminal justice. A tenth factor on informal justice will be included in 2012.

These 10 factors are further disaggregated into 49 sub-factors. The scores of these sub-factors are built from variables drawn from assessments of the general public (a thousand respondents per country) and local legal experts. The major rationale for aggregating this wealth of roughly five hundred variables (in 2010) on rule of law dimensions was to reduce the measurement errors: errors may be correlated among sources, but so long as there is some idiosyncratic or source-specific error, the resulting index may be more accurate than any randomly selected single source.

Statistical quality features of the Index have been assessed through univariate and multivariate analyses, and global sensitivity analysis. Univariate analysis has been carried out at the variable level and focused on the presence of missing data, outliers, and potentially problematic variables due to highly asymmetric distributions (skewness). The raw data used in this paper were provided by the developers in [0, 1] scale and they represented average scores of public or expert opinion on 473 variables. Most of these variables are not affected by outliers or skewed distributions, except for fifteen variables spread across six dimensions of the rule of law. Given the high number of variables combined in building each of the factors, the skewed distributions of those fifteen variables do not bias the results.

Other data quality tests focused on missing data. The 2010 dataset is characterized by excellent data coverage (99.96 percent, 473 variables × 35 countries). Data coverage per factor is very good or excellent for most countries, except for four countries that miss more than 25 percent of the values on some factors or sub-factors:

- Indonesia, Liberia, Singapore, and South Korea on *Fundamental labor rights* (sub 5.8); *Regulatory Enforcement* (F.7); and *Access to Civil Justice* (F.8);
- Liberia and Indonesia on *Equal treatment and absence of discrimination* (sub factor 5.1).

Hence, those factor/sub-factor scores for the aforementioned countries should be interpreted with caution.

A further data quality issue relates to the treatment of missing values. The WJP Rule of Law Index team opted not to impute missing data, but instead to calculate country scores per sub-factor and factor by a weighted average of available variable scores for a given country. Although this approach can be a good starting point, it has notable shortcomings, as, in essence, it implies replacing missing variable scores per country with the weighted average of the available variable scores for the given country. We tested the implications of 'no imputation' versus the hot-

deck imputation method and discuss this below in the second part of the assessment, together with the other modeling assumptions.

Principal component analysis (PCA) was used to assess to which extent the conceptual framework is confirmed by statistical approaches and to identify eventual pitfalls. PCA was applied at the sub-factor level. Overall, the analysis confirms the WJP Rule of Law Index structure, as within each of the nine dimensions a single latent factor is identified, which captures more than 65 percent of the variance (best result for Limited government powers, where the single latent factor summarizes 83 percent of the data variance). A more detailed analysis of the correlation structure within and across the nine WJP dimensions confirms the expectation that the sub-factors are more correlated to their own dimension than to any other dimension and all correlations are strong and positive. Hence, no-reallocation of sub-factors is needed. An eventual refinement of the framework concerns three pairs of sub-factors that represent strong collinearity ($r > .90$): sub-factor 1.2 with 1.3, sub-factor 7.1 with 7.2, and sub-factor 8.5 with 8.7. It is recommended that these pairs sub-factors are combined together (this implies assigning them 0.5 weight each when all other sub-factors underlying a factor receive a weight of 1 each).

Had the WJP Rule of Law Index team attempted to further aggregate the nine factors into an overall Index by using a weighted arithmetic average of the factors, this choice would be supported by the data: PCA shows that the nine factors share a single latent factor that captures more than 80 percent of the total variance and all nine factors correlate with the single latent factor with loadings over 0.78. Hence, the nine WJP factors are not distinct, but partially overlapping aspects of rule of law. When deciding on equal or non-equal weighting for the nine dimensions, one should bear in mind two points: (a) that most of the factors are strongly correlated to each; and (b) that two factors – *Order and Security* and *Open Government* – appear to describe slightly different aspects of rule of law than the remaining (and highly correlated) factors. These remarks suggest that an equal weighting scheme would not guarantee equal contribution of those two Factors with respect to the remaining factors on the overall Index classification. A further consideration is that the *Absence of Corruption* (F.2) is so highly correlated with *Regulatory Enforcement* (F.7) or *Access to Civil Justice* (F.8) or *Effective Criminal Justice* (F.9) (at 0.92 or more), which does not justify presenting it as a standalone aspect of the rule of law. All other WJP factors, though correlated to each other, they are communicating partially overlapping but not tautological aspects of the rule of law.

Global sensitivity analysis has been employed in order to evaluate a sub-factor's contribution to the variance of the factor scores. The assumption made by the WJP Rule of Law Index team was that all sub-factors receive equal weights in

building the respective factor (calculated as a simple average of the underlying sub-factors). Our tests focused herein on identifying whether a factor is statistically well-balanced in its sub-factors. There are several approaches to test this, such as eliminating one sub-factor at a time and comparing the resulting ranking with the original factor ranking, or using a simple (e.g., Pearson or Spearman rank) correlation coefficient. A more appropriate measure aptly named 'importance measure' (henceforth S_i) has been applied here, also known as correlation ratio or first order sensitivity measure.[36] The S_i describes 'the expected reduction in the variance of factor scores that would be obtained if a given sub-factor could be fixed.' Estimating the S_i's for the sub-factors within each factor, the results are rather reassuring: all sub-factors are important in classifying countries across the concept represented by the relevant factor, though some sub-factors are slightly more important than others. Three exceptions are shown in Table 3. For the *Regulatory Enforcement*, one can question the contribution of sub-factor 7.4 on the basis of its low S_i (=.472) compared to that of the other sub-factors (>.872). Similar for the *Access to Civil Justice*, where the contribution of sub-factor 8.2 is just 0.346 when for sub-factors 8.5 and 8.7 the contribution is greater than 0.85. Finally, on *Effective Criminal Justice*, the contribution of sub-factor 9.1 is low compared to the contribution of the other sub-factors.

In the case that the WJP Rule of Law Index team decided to summarize the nine factors with an overall Index by simply averaging them, the *Si* values would have been comparable to each other, ranging between 0.61 and 0.93 (Table 4). The most influential factors would have been *absence of corruption*, and *regulatory enforcement*. The least influential factors would have been *order and security*, and *open government* (as already anticipated in the previous paragraphs given the lower correlation of those factors with the remaining).

Table 3. Importance measures for the three WJP Rule of Law Index sub-factors

WJP Rule of Law Index factors and sub-factors	Importance measure (S_i)
Regulatory enforcement	
Government regulations are effectively enforced (#7.1)	0.920
Government regulations are applied and enforced without improper influence (#7.2)	0.872
Government does not expropriate without adequate compensation (#7.4)	0.472 (*)
Access to civil justice	
People can access and afford legal advice and representation (#8.2)	0.346 (*)

[36] Saltelli et al., *Global Sensitivity Analysis*.

Table 3. Continued

WJP Rule of Law Index factors and sub-factors	Importance measure (S_i)
People can access and afford civil courts (#8.3)	0.665
Civil justice is impartial (#8.4)	0.548
Civil justice is free of improper influence (#8.5)	0.889
Civil justice is not subject to unreasonable delays (#8.6)	0.522
Civil justice is effectively enforced (#8.7)	0.852
ADRs are accessible, impartial, and effective (#8.8)	0.689
Effective criminal justice	
Criminal investigation system is effective (#9.1)	0.438 (*)
Criminal adjudication system is timely and effective (#9.2)	0.849
Criminal system is impartial (#9.4)	0.629
Criminal system is free of improper influence (#9.5)	0.842
Due process of law and rights of the accused (#9.6)	0.615

Note: (*) sub-factors with much lower contribution to the variance of the relevant factor than the equal weighting expectation.

Table 4. Importance measures for the nine WJP Rule of Law Index factors

WJP Rule of Law Index Factors	Importance measure (S_i)
Limited government powers	0.880
Absence of corruption	0.934
Clear, publicized and stable laws	0.845
Order and security	0.734
Fundamental rights	0.852
Open government	0.610
Regulatory enforcement	0.929
Access to civil justice	0.859
Effective criminal justice	0.865

Impact of modeling assumptions on the WJP Rule of Law Index results

The aim of the robustness analysis is to assess to what extent the modeling choices in building each factor in the Rule of Law Index might affect country classification. We have dealt with these uncertainties in order to check their simultaneous

and joint influence on the results, with a view to better understand their implications. In the present exercise the data are assumed to be error-free and already normalized. The assessment was based on a combination of a Monte Carlo experiment and a multi-modeling approach. This type of assessment respects the fact that the country scores or ranks associated with composite indicators are generally not calculated under conditions of certainty, even if they are frequently presented as such.[37] The Monte Carlo experiment was based on some hundreds of 'complete' datasets built upon estimation of missing data with hot-deck imputation (single imputation) or multiple imputation. The original dataset (without any imputation) was also included.

The multi-modeling approach involved exploring plausible combinations of the two key assumptions needed to build the index: the weighting issue and the aggregation formula. We simulated a total of nine models that could have been used to build the factors in the WJP Rule of Law.

Assumption on the weighting scheme: the factors are built assuming equally weighted sub-factors. We tested two alternative and legitimate weighting schemes: factor analysis derived weights (upon factor rotation and squared factor loadings),[38] or cross-efficiency data envelopment analysis.[39] Practitioners use this approach to counter stakeholder objections that a given weighting scheme is not fair to a country because it does not reflect certain stakeholders' priorities.[40]

Assumption on the aggregation rule: The factors are built using an arithmetic average (a linear aggregation rule) of the sub-factors. Decision-theory practitioners have challenged aggregations based on additive models because of inherent theoretical inconsistencies and because of the fully compensatory nature of linear aggregation, in which a comparative high advantage on few indicators can compensate a comparative disadvantage on many indicators.

Besides the arithmetic average, we considered three different approaches to aggregate the sub-factors: a geometric average, a Borda rule, and a Copeland rule.[41] In the geometric average, sub-factor scores are multiplied as opposed to summed in the arithmetic average. In the models where geometric averaging was used, we re-scaled the normalized data onto a 1-100 range for technical reasons. The Borda

[37] Saisana et al., 'Uncertainty and Sensitivity Analysis Techniques as Tools for the Analysis and Validation of Composite Indicators'; Saisana et al., 'Rickety Numbers: Volatility of University Rankings and Policy Implications'.

[38] Nicoletti et al., 'Summary Indicators of Product Market Regulation with an Extension to Employment Protection Legislation'.

[39] T.R. Sexton et al., 'Data Envelopment Analysis: Critique and Extensions', in R.H. Silkman (ed.), *Measuring Efficiency: An Assessment of Data Envelopment Analysis, Vol. 32* 1986.

[40] Cherchye, 'Creating Composite Indicators with DEA and Robustness Analysis: The Case of the Technology Achievement Index'.

[41] Munda, *Social Multi-criteria Evaluation for a Sustainable Economy.*

rule is the following: given *N* countries, if a country is ranked last, it receives no points; it receives 1 point if it is ranked next to the last. The scoring process continues like this up to *N-1* points awarded to the country ranked first. The Copeland rule is a non-compensatory multi-criteria method and is summarized as follows: compare country A with every other country B. Score +1 if a majority of the sub-factors prefers A to B, -1 if a majority prefers B to A, and 0 if it is a tie. Summing up those scores over all countries B (B≠A), yields the Copeland score of country A.

The Monte Carlo simulation comprises 1,500 runs (combining assumptions on missing data estimation, weighting and aggregation approach). Table 5 reports the original country ranks and the 95 percent confidence interval for the simulated median rank for all nine factors. Overall, all country ranks on all nine factors lay within the simulated intervals. Few exceptions are found for factor 4 (Ghana ranks 26, slightly better than expected [28, 30]); factor 5 (Bulgaria ranks 16, slightly better than expected [18, 19]); factor 6 (El Salvador ranks 27, slightly better than expected [29, 35]); and factor 7 (Dominican Republic ranks 17, slightly better than expected [19, 22]). Confidence intervals for the median rank are narrow enough for all countries (less than 3 positions) to allow for meaningful inferences to be drawn. Exceptionally, few countries have slightly wider intervals: El Salvador (4-6 positions on factor 1 and factor 6); Croatia (4 positions on factor 2); Ghana (4 positions on factor 3); Thailand (4 positions on factor 3 and factor 8); Colombia (5 positions on factor 5); Nigeria, Indonesia, and Kenya (4-5 positions on factor 6); and India (4 positions on factor 7). Results are extremely robust for factor 1 and factor 2, where 16-19 of the 35 countries have an exact simulated median rank (zero interval) that coincides with the relevant WJP factor rank. All things considered, the majority of the countries just sees ± 1 positions shift due to the methodological assumptions.

Complementary to the uncertainty analysis, sensitivity analysis has been used to identify which of the modeling assumptions have the highest impact on country classification. Almost all combinations of modeling assumptions lead to similar country classifications (90 percent of the countries shift up to ± 1 position). The choice of factor analysis derived weights versus equal weights for the sub-factors underlying a factor is non-influential, and neither is the choice of arithmetic versus geometric average. Allowing for country-specific weights (cross-efficiency DEA) also does not significantly influence the results. The highest impact is due the assumption of a non-compensatory aggregation (Copeland rule). Assuming no other change compared to the WJP methodology, but for the use of Copeland rule, Indonesia would lose 16 positions (moving from 16 to 32) on factor 3 (see Figure 1). Currently, Indonesia is ranked 16 because it offsets low scores on sub-factors 3.2 and 3.3 (ranked 29 and 28, respectively) with an excellent

Table 5. WJP factor rank and simulated 95 percent confidence interval for median rank

	F.1	F.2	F.3	F.4	F.5	F.6	F.7	F.8	F.9
Albania	28 [26 ,29]	31 [31 ,31]	25 [23 ,25]	14 [14 ,15]	22 [23 ,24]	34 [33 ,35]	32 [32 ,32]	31 [31 ,31]	22 [21 ,24]
Argentina	33 [32 ,33]	20 [20 ,20]	31 [31 ,32]	25 [24 ,25]	21 [20 ,23]	29 [28 ,29]	28 [28 ,29]	20 [20 ,23]	28 [27 ,29]
Australia	3 [3 ,4]	6 [6 ,8]	5 [5 ,6]	6 [6 ,9]	6 [5 ,6]	7 [7 ,7]	5 [3 ,5]	6 [6 ,7]	8 [8 ,9]
Austria	4 [4 ,4]	3 [4 ,4]	6 [5 ,6]	3 [3 ,4]	1 [1 ,1]	11 [11 ,12]	3 [3 ,5]	4 [3 ,4]	1 [1 ,1]
Bolivia	32 [32 ,33]	30 [30 ,32]	33 [33 ,33]	30 [28 ,30]	30 [30 ,32]	26 [25 ,26]	30 [29 ,30]	29 [29 ,29]	35 [34 ,35]
Bulgaria	29 [29 ,30]	29 [29 ,29]	20 [20 ,21]	17 [17 ,17]	16 [18 ,19]	23 [22 ,23]	25 [24 ,25]	24 [21 ,24]	26 [25 ,27]
Canada	6 [6 ,6]	5 [5 ,5]	4 [3 ,4]	5 [5 ,6]	4 [4 ,4]	4 [3 ,4]	6 [6 ,6]	8 [6 ,8]	9 [9 ,10]
Colombia	20 [20 ,20]	22 [22 ,24]	21 [22 ,25]	32 [32 ,32]	29 [27 ,31]	10 [9 ,10]	15 [15 ,15]	15 [13 ,16]	31 [29 ,31]
Croatia	27 [27 ,28]	23 [21 ,25]	30 [27 ,30]	10 [7 ,10]	19 [19 ,21]	32 [30 ,32]	29 [29 ,30]	22 [22 ,24]	21 [20 ,21]
Dominican Rep	26 [26 ,27]	28 [28 ,28]	14 [13 ,14]	31 [31 ,31]	28 [28 ,30]	21 [22 ,22]	17 [19 ,22]	19 [18 ,19]	24 [23 ,24]
El Salvador	23 [20 ,23]	19 [18 ,19]	22 [22 ,24]	21 [20 ,21]	17 [16 ,17]	27 [29 ,35]	16 [16 ,17]	23 [22 ,23]	30 [30 ,30]
France	8 [8 ,8]	7 [7 ,8]	8 [7 ,8]	8 [8 ,11]	9 [9 ,10]	6 [5 ,6]	9 [8 ,9]	9 [9 ,10]	6 [6 ,6]
Ghana	12 [12 ,12]	18 [18 ,19]	23 [21 ,25]	26 [28 ,30]	14 [14 ,15]	18 [19 ,21]	23 [23 ,23]	21 [20 ,22]	16 [15 ,16]
India	14 [14 ,14]	25 [25 ,27]	13 [12 ,12]	23 [23 ,25]	20 [19 ,21]	9 [10 ,11]	24 [24 ,28]	27 [27 ,28]	23 [22 ,23]
Indonesia	18 [18 ,18]	27 [27 ,27]	16 [17 ,18]	19 [19 ,21]	25 [23 ,25]	17 [17 ,21]	21 [20 ,22]	32 [32 ,32]	19 [18 ,19]
Japan	5 [5 ,5]	8 [8 ,8]	3 [3 ,5]	2 [2 ,2]	8 [8 ,9]	8 [8 ,8]	4 [4 ,4]	10 [9 ,10]	2 [2 ,2]
Jordan	22 [22 ,23]	12 [12 ,12]	15 [15 ,15]	15 [12 ,15]	31 [30 ,31]	35 [33 ,35]	12 [12 ,14]	17 [15 ,17]	15 [15 ,17]
Kenya	35 [35 ,35]	34 [34 ,34]	35 [34 ,35]	29 [29 ,30]	34 [34 ,34]	30 [27 ,31]	34 [33 ,34]	33 [33 ,34]	25 [23 ,25]
Liberia	24 [24 ,24]	33 [33 ,34]	26 [26 ,27]	35 [35 ,35]	27 [26 ,28]	16 [16 ,16]	35 [35 ,35]	34 [33 ,34]	33 [32 ,33]
Mexico	21 [21 ,22]	32 [32 ,32]	17 [16 ,16]	27 [26 ,28]	24 [23 ,24]	13 [11 ,13]	31 [31 ,32]	30 [30 ,31]	34 [32 ,34]
Morocco	25 [24 ,25]	21 [21 ,24]	27 [26 ,27]	22 [22 ,22]	23 [21 ,22]	33 [31 ,33]	27 [26 ,27]	25 [23 ,26]	17 [17 ,18]
Netherlands	2 [2 ,2]	2 [2 ,4]	2 [2 ,2]	9 [9 ,9]	3 [3 ,3]	2 [2 ,2]	2 [2 ,2]	3 [2 ,3]	4 [4 ,4]
Nigeria	30 [28 ,30]	24 [23 ,25]	29 [29 ,31]	33 [33 ,35]	32 [31 ,33]	28 [29 ,34]	22 [18 ,21]	18 [16 ,19]	29 [28 ,31]
Pakistan	34 [34 ,34]	35 [35 ,35]	34 [34 ,34]	24 [23 ,24]	35 [35 ,35]	31 [31 ,31]	33 [33 ,34]	35 [34 ,35]	32 [32 ,33]
Peru	19 [19 ,20]	17 [17 ,17]	19 [18 ,19]	28 [26 ,27]	15 [14 ,15]	25 [23 ,24]	18 [18 ,20]	26 [25 ,26]	27 [26 ,29]
Philippines	17 [17 ,17]	26 [24 ,26]	24 [21 ,24]	20 [18 ,20]	26 [27 ,27]	19 [18 ,20]	20 [18 ,20]	28 [28 ,30]	20 [20 ,21]
Poland	10 [10 ,10]	13 [13 ,13]	18 [18 ,19]	7 [6 ,7]	10 [10 ,10]	14 [13 ,14]	14 [13 ,14]	13 [13 ,16]	12 [12 ,12]
Singapore	11 [11 ,13]	4 [3 ,4]	7 [7 ,8]	1 [1 ,2]	12 [12 ,12]	20 [17 ,19]	7 [7 ,10]	1 [2 ,4]	5 [5 ,5]
South Africa	13 [13 ,13]	15 [15 ,16]	10 [10 ,10]	34 [34 ,34]	18 [16 ,18]	12 [12 ,13]	13 [12 ,13]	12 [13 ,16]	18 [16 ,19]
South Korea	15 [15 ,15]	11 [11 ,11]	11 [11 ,11]	13 [13 ,15]	7 [7 ,7]	5 [5 ,6]	10 [10 ,11]	5 [5 ,7]	11 [10 ,11]
Spain	7 [7 ,8]	9 [9 ,11]	12 [13 ,14]	12 [12 ,13]	5 [5 ,6]	15 [15 ,16]	11 [10 ,12]	7 [6 ,8]	10 [10 ,10]
Sweden	1 [1 ,1]	1 [1 ,1]	1 [1 ,1]	4 [3 ,4]	2 [2 ,2]	1 [1 ,1]	1 [1 ,1]	2 [1 ,2]	3 [3 ,3]
Thailand	16 [16 ,16]	14 [14 ,14]	28 [28 ,32]	16 [16 ,16]	13 [13 ,13]	24 [25 ,26]	19 [17 ,19]	16 [13 ,17]	13 [13 ,14]
Turkey	31 [31 ,31]	16 [16 ,16]	32 [29 ,32]	18 [18 ,19]	33 [33 ,33]	22 [20 ,22]	26 [26 ,27]	14 [14 ,16]	14 [14 ,14]
USA	9 [9 ,9]	10 [10 ,11]	9 [8 ,9]	11 [11 ,12]	11 [11 ,11]	3 [3 ,4]	8 [7 ,9]	11 [11 ,11]	7 [7 ,7]

F.1. Limited Government Powers; F.2. Absence of Corruption; F.3. Clear, Publicized and Stable Laws; F.4. Order and Security; F.5. Fundamental Rights; F.6. Open Government; F.7. Regulatory Enforcement; F.8. Access to Civil Justice; F.9. Effective Criminal Justice.

performance on sub-factor 3.1 (ranked 5). Similarly, Ghana would move from rank 18 to 25 on factor 6, if compensation had not been allowed (currently Ghana compensates for low performance on sub-factors 6.2 (ranked 28) and 6.3 (ranked 22) with a very good performance on sub-factor 6.1 (ranked 9) (see Figure 1). Interestingly, when combining the Copeland rule with the use of hot-deck imputation, the impact of the non-compensatory aggregation rule is less pronounced. Under this assumption, Indonesia, for example, would lose only 9 positions (moving from 16 to 25) on factor 3 because after imputation its rank on sub-factor 3.3 improves significantly (from 28 to 20), although its rank on sub-factor 3.2 slightly deteriorates (from 29 to 34).

This analysis, by assessing the impact of the modeling choices, gives more transparency in the entire process, and can help to appreciate the *WJP Rule of Law Index* results with respect to the assumptions made during the development phase.

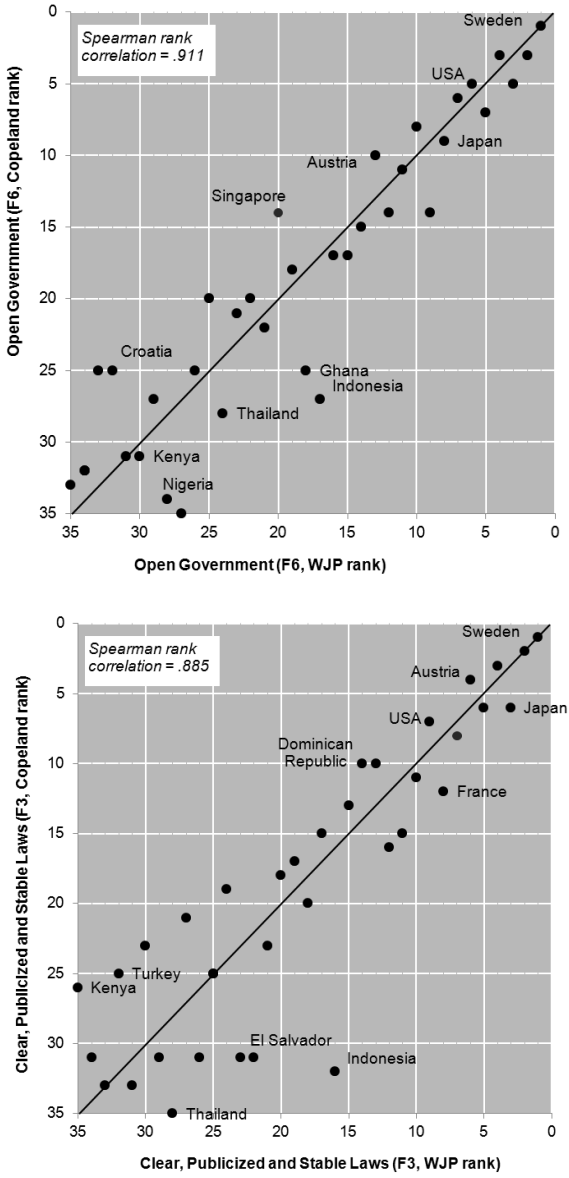

Figure 1. Compensability: WJP ranks v. ranks obtained by a non-compensatory approach (Copeland rule)

Conclusions

Our society is changing so fast that we need to know as soon as possible when things go wrong. Without rapid alert signals, appropriate corrective action is impossible. This is where composite indicators could be used as yardstick. Whether or not one accepts composite indicators for the purpose of benchmarking performance, one might find itself, even unwillingly, exposed to a composite indicator published in the news.

In this paper, we presented composite indicators as multi-dimensional and multi-scale representations of complex phenomena. We briefly explored the main steps necessary for their sound construction; emphasized the need for a coherent and viable theoretical framework; and underlined the main issues related to weighting and aggregation, in relation to compensability.

We offered an assessment of the WJP Rule of Law Index, showing that it is statistically and conceptually coherent, and that almost all factors are well balanced in their underlying sub-factors, as conceptualized. A slight mismatch between the weights and the actual importance of the underlying sub-factors was found for three factors – *Clear, Publicized and Stable Laws*; *Access to Civil Justice*; and *Effective Criminal Justice*. Country classifications across the nine factors were also fairly robust to methodological changes related to the estimation of missing data, weighting, or aggregation rule (90 percent of the countries shift less than ± 1 position). Finally, in the case that the WJP Rule of Law Index team decided to build an overall Index by simply averaging the nine factors, this choice would have been statistically supported with two reservations: (a) the contribution of *Order and Security*, and *Open Government*, whose weights should be slightly greater than the weights of the remaining factors, in order to guarantee equal contribution to the overall Index country classification; and (b) the presentation of *Absence of Corruption* as one of the nine factors of the WJP Rule of Law Index, given the very high correlation with three other factors on *Regulatory Enforcement* or *Access to Civil Justice* or *Effective Criminal Justice* (at 0.92 or more).

A bottle-neck conclusion is that composite indicators should never be seen as a goal, *per se*, regardless of their quality. They should be seen, instead, as a starting point for initiating discussion and attracting public interest and concern.