Designing Sustainable Governance Indicators: Criteria and Methodology

Martin Brusis

This chapter explains the concepts, assessment criteria and methodology behind the Bertelsmann Stiftung's Sustainable Governance Indicators (SGI), which evaluate and rank sustainable governance in all 30 member states of the OECD. Our aim in developing the SGI has been to measure political reform capacity, that is, the capacity of political actors to identify and implement "reforms," or changes needed to improve the status quo. Policymakers and scholars will naturally have different opinions about which reforms are necessary and how to ensure their success. But as long as there is no consensus about the best possible set of reforms, approaches that measure reform capacity in terms of the degree to which a given catalogue of reforms has been implemented will be criticized as oversimplified and mechanistic as well as biased in applying one standard for all. Clearly, executing checklists of externally devised reforms may not necessarily entail nor result in cogent reform capacity. Though less readily tangible, dynamic adaptation, institutional learning and innovation are more telling indicators of the capacity for reform in a given state. For these reasons, the SGI relies on an indirect measurement of reform capacity that compares the policy performance and executive governance of various states. A cross-national comparison of policy outcomes allows reform activities and their impact to be assessed *ex post*. Better policy outcomes or greater improvements – cross-nationally or cross-temporally – may result from a higher reform capacity in a given country. As the international consensus on "good" policy outcomes grows, identifying reform capacity *ex post* has become much more feasible.

At the same time, however, it must be said that an *ex-post* evaluation of policy outcomes of this type sheds little light on the so-called secrets of success or on whether outcomes can be attributed to deliberate strategy, charismatic leadership, favorable circumstances, pure chance or something else. Only a handful of these possible causal factors are relevant here, as our project focuses on those drivers of reform capacity that can be influenced by intentional, targeted policies. One of these factors, "executive governance", has yet to receive sufficient attention despite wide acknowledgment of its importance (Andeweg 2003). Executives have command over significant resources and translate popular preferences into policies. How they are governed may not guarantee the success of reforms, but it surely affects the chances of governments to succeed with their reform measures.

The concept of "executive governance" refers to the institutional arrangements of governing and comprises the mechanisms for and patterns of interaction between the core executive and its organizational environment, both within the executive itself and in the wider political system. The focus on governance also implies that the *institutional capacity* for reforms is compared, not the individual decisions of prime ministers or the leadership attributes of charismatic reformers.

This approach of institutional analysis raises the question how we can know whether a country's established model of executive governance enhances or reduces its reform capacity? While broadly agreeing on the desirability of democracy, scholars continue to argue over which institutional arrangements are superior: presidential or parliamentary systems of government, majoritarian or consensual democracies, and small or large public sectors. These macro-level categories refer to institutional features deeply embedded within a given country's institutional culture and tradition. As a result, they constitute the given framework conditions for capacity-building reformers that often lie beyond the scope of their discretion.

In contrast, the micro-level functions and processes required to "run a government" have increasingly been subjected to cross-national evaluation, transfer and learning. Driven by a growing interest in "good governance" and "performance management," political practitioners and scholars alike have been developing a body of best management practices. International organizations and agencies such as the Organization for Economic Cooperation and Development (OECD), the World Bank and the European Commission have used these best practices as reference points for benchmarking and peer review processes (Ben-Gera 2004; Nunberg 2000; OECD 2005). This evolving international public management know-how covers areas such as strategic planning, inter-ministerial coordination, the drafting of legal acts, monitoring, budgeting, auditing, task delegation, institutional learning devices, and public communication and consultation policies.

Drawing on this shared knowledge, the SGI assume that the extent to which a government has established best practices in performing its functions can be taken as an indication of better executive governance and one that is likely to enhance the country's institutional capacity for reform. It should be noted that greater institutional capacity does not necessarily generate improved socioeconomic performance or better performance in terms of policy outcomes or the quality of democracy, all of which can be influenced by the factors previously mentioned (e.g., favorable circumstances, charismatic leadership and so forth). Nevertheless, greater institutional capacity does improve the chances that political leaders will make decisions that

can fully harness a country's potential and maximize its performance. Furthermore, a micro-level evaluation of executive governance pinpoints observable deficiencies and may allow efforts to be guided in a more targeted way toward improving the quality of governance.

Concepts, questions and indicators

For the reasons explained above, the SGI incorporate a two-tiered system of measuring reform capacity that assesses both reform outcomes and the institutional potential for reform. These two distinct aspects are represented in two composite indicators: a Status Index and a Management Index. These two indices consist of 149 individual items–93 and 56, respectively. Seventy-four quantitative indicators are derived from information collected from public data sources. Experts for each country have provided 62 qualitative assessments as well as 13 quantitative indicators, such as the percentage of government-sponsored bills that were ultimately adopted in parliament. (For a detailed list of indicators, questions and sources, please see the appendix.)

Figure 1: Composition of the Status and Management Indices

	Status Index	**Management Index**
Dimensions	2	2
Categories	4	7
Criteria	18	15
Items	93	56
Of which: **Expert assessments**	26	36
Quantitative indicators	67	7
Quantitative expert indicators	-	13

The Status Index

The Status Index reflects the growing political and scholarly consensus on what good policy outcomes entail as well as the importance of a high-quality democracy as a framework for policy performance. The SGI's concept of democracy includes not only the rights of political participation and electoral competition but also the rule of law (Merkel 2004). Since all OECD member countries are democracies, the SGI's questions in this category focus on the quality rather than the presence of democracy. There are a series of questions designed to address whether citizens face discrimination in the electoral process, how citizens can access

public information, the degree to which the media are independent and diversified, how well states protect civil rights and whether the government and administration act predictably and in accordance with the law (criteria S1-S4).

To assess policy performance, the Status Index then examines four broad policy sectors that constitute political priorities for governments in responding to the key challenges that most advanced industrial states face: the global integration of markets and its effects for national economies and competitiveness; ageing societies and their effects for the sustainability of social security and pension systems; new security risks arising from terrorism, transnational crime, migration and its structural causes; and the depletion of natural resources in connection with global climate change. To address these challenges, governments need to devise and co-ordinate various interrelated policies that may be grouped in four broad sectors reflecting the challenges: economy and employment; social affairs; security; sustainability. These four policy sectors comprise the following policy areas:

(1) Economy and employment: labor market policy, enterprise policy, tax policy, budgetary policy (criteria S6-S9);

(2) Social affairs: health policy, social cohesion, family policy, pension policy (criteria S10-S13);

(3) Security: external/internal security policy; integration policy (criteria S14-S15);

(4) Sustainability: environmental policy, research and innovation policy, education policy (criteria S16-S18).

Each policy area is evaluated by country experts and through indicators from public data sources. Country experts were asked to assess a set of questions and evaluate the extent to which a particular policy realizes specified objectives, such as the goal of fiscal sustainability in the case of budgetary policy. These objectives have been carefully selected and defined so as to avoid any ideological bias and to make sure that they would be broadly accepted and supported by citizens, policymakers and scholars alike across both political and value-based divisions.

For example, the objective of family policy (criteria S12) is stipulated as enabling women to combine parenting with participation in the labor market. In doing so, that particular question does not betray a preference for either a traditional single-wage-earner family model or so-called working mothers. Instead, the question presupposes that an optimal system of family

support would enable women to decide freely whether and when they would prefer to remain full-time mothers or to work full- or part-time.

Another example is the question regarding health care policy (criteria S10). This question avoids any biased leanings toward either a predominantly private or public health care system. Instead, the objective of public health care policies is defined as providing high-quality health care for the largest possible proportion of the population at the lowest possible costs. In an explanation appended to the question, experts are instructed to apportion less weight to the criterion of cost efficiency if the criteria of quality and inclusiveness can be considered fulfilled.

In selecting the performance indicators from public data, we have been careful to choose those indicators that are clear in meaning, do not invite ambiguous interpretations and are available for all OECD countries. We have also sought to avoid including model-specific indicators that might be seen as being biased in favor of particular types of economies. For example, the public sector's proportional contribution to GDP is not used as a performance indicator, because doing so would entail using a disputed interpretation as well as implying a bias in favor of either liberal, Anglo-American market economies with small public sectors or of social-democratic Scandinavian market economies with large public sectors. Moreover, the scholarly debate about the varieties of capitalism shows that there is no consensus about which institutional model is the most sustainable and that distinct models still persist in spite of all tendencies toward convergence (Hall and Soskice 2001; Howell 2003). Adopting such a distinction between two or three institutional models would not provide an appropriate basis for an evaluation because doing so would conceptually "freeze" certain features and ignore the dynamics of change, especially when it comes to continental European models.

Taken together, the SGI aim to provide a composite picture of a state's performance across various policy areas. As a result, some indicators refer to policy outputs rather than the impact of public policies on society (or "outcomes"). An example of an output indicator would be spending on pre-primary education as a percentage of GDP. Together with other outcome indicators such as the at-risk-of-poverty rate for children, this output indicator is considered sufficiently unambiguous as a descriptive measure of the success of family policy. Merging output and outcome indicators does not deny the existence of causal links between indicators, such as those between spending on pre-primary education and female employment rates. At the same time, however, from a conceptual standpoint, the SGI are less concerned with such links than with the overall nexus between policy performance and executive governance.

The Status Index also includes a few indicators that describe changes over time rather than levels, such as the inflation rate and the average annual growth rate of government spending on research and development. These and other indicators of change over time have been selected because they describe important policy effects that complement the performance profiles of states. That these two types of indicators (change and level) are combined should be viewed against the background of the SGI's generic design of assessing both aggregate policy performance and executive governance.

The SGI also compare countries in terms of basic socioeconomic parameters (criterion S5). These parameters assess each country's socioeconomic situation as revealed, for example, by real GDP per capita, the rate of potential GDP growth, the unemployment rate, growth of the labor force, the Gini coefficient, the inflation rate, the real interest rate and the share of foreign trade in GDP (criterion S5).[1]

In summary, the Status Index combines democratic, socioeconomic and policy performance items because it conceives of high democratic and socioeconomic standards as necessary scope conditions for policy-specific performance. Moreover, high-quality performance in terms of all of these three areas can be understood as outcomes that result from strategic and accountable executive governance.

The Management Index

The Management Index reflects the consensus practitioners and scholars have developed on what good governmental practices entail. The Index first examines the extent to which core executives act strategically and can rely on institutional capacities for strategic policy-making. This dimension, labeled "executive capacity," is based on a commonly accepted notion of governing that identifies the government or core executive as the key actor of governance (criteria M1 - M12) (Knack, Kugler and Manning 2003). The Management Index then analyzes the role of actors outside the executive and the extent to which these actors hold governments accountable, enhance the knowledge base of decisions and deliberate the normative appropriateness of policy decisions. This dimension of "executive accountability" reflects the degree of importance attained in governance by actors outside the executive (criteria M13 - M15) (Pierre and Peters 2005).

[1] The last-mentioned indicator is adjusted in order to control for the impact of the size of population.

In their theoretical account of governing, Pierre and Peters describe the state's dependence on these actors as follows: "states must be open to a wide range of information, including much that is uncomfortable and dissonant, if it is to be successful in governing. In other words, states must be in close contact with the society and utilize social information openly and accurately when governing. This further implies that the state is likely to be in close communication with societal actors who possess much of the information that would be required for effective governing and also generally that the state must be willing to engage in a formal or informal exchange of power over decisions for that information." (Pierre and Peters 2005, 46)

Both dimensions of executive capacity and executive accountability are further structured into categories and criteria. Four separate components of executive capacity are distinguished: policy preparation, policy implementation, the incorporation of external impulses and institutional learning. These components, in turn, refer to stages in the cycle of policy formation as well as to concepts deriving from the literature on Europeanization, globalization and policy learning (Common 2004; Dolowitz and Marsh 2000; Radaelli 2003; Wiesenthal 1995). In particular, the components address the following factors:

(1) Policy preparation: strategic planning and expert advice, inter-ministerial coordination, regulatory impact assessments (RIAs), consultation and communication policies (criteria M2-M6);

(2) Policy implementation: anticipation of veto actors in the legislative process, management of task delegation to ministers, agencies, subnational governments and private actors (criteria M7-M9);

(3) Incorporation of external reform impulses: governmental capacity to adapt to globalization, Europeanization or transnationalization as well as to import and export policies (criteria M10-M11);

(4) Institutional learning: governmental capacity to reform its own institutional arrangements and improve its strategic orientation (criterion M12).

Executive accountability is subdivided into three separate categories corresponding to actors or groups of actors that are considered to be key accountability providers in theories of democracy and governance (Pierre and Peters 2005, 46; Schedler 1999, 17; Schmitter 2004). The particular questions here ask: To what extent are citizens informed about government policies (criterion M13)? Is the parliament capable of evaluating and controlling the executive (criterion M14)? And are intermediary organizations (e.g., the media, political parties, interest associations) characterized by policy know-how and relevance (criterion M15)? As was the

case with the Status Index, country experts provide evaluations in response to the individual questions of the Management Index. In addition, the country experts also collect numerical data on, for example, the share of governmental bills adopted by parliament or the size of expert staff in parliament. Two further criteria require background data on the (changing) composition of cabinets (criterion M1) and parliaments (criterion M14.1) that are not aggregated.

In sum, the Management Index assumes that more accountability – in the form of public scrutiny, information channels and normative deliberation – improves a country's capacity to reform. Since the conventional approach of governing tends to view accountability and participation mechanisms as constraints on executive authority, this assumption might be challenged as counterintuitive. Reforming may be much easier for some governments as they can govern under much more conducive structural conditions and rely on enabling actor constellations. In order to avoid equating constellations of circumstances that are categorically different, assessing the reform capacity of a particular executive should also factor in the following contextual conditions:

(1) veto players (number and powers);

(2) the country's particular economic and social distress;

(3) attitudes among the population;

(4) path dependencies (historical and institutional development).

The SGI deal with these contextual conditions in the following ways:

Veto players

Veto player theory differentiates between institutional veto players (e.g., bicameral parliament, constitutional court, etc.) as defined in the constitution and the political veto players which are primarily the different parties that make up a governing coalition but also organized interest groups (Tsebelis 2002). While it is true that a greater number of veto players increases an adopted policy's degree of stability, veto players do not necessarily block improvements to the status quo. In fact, many authors have argued that veto players might even improve the quality of reforms by helping governments to better assess the potential impacts of a given reform. Doing so induces reformers to broaden their bases of support by accommodating veto-player critiques and thereby rendering reforms irreversible (c.f.,Benz 2003, 230).

For this reason, the Management Index does not attribute a veto function *a priori* to certain attributes of a given political system, such as a large number of (governing) parties, a bicameral parliament or a strong constitutional court. Instead, the Management Index uses empirical

data to examine whether veto players have a positive influence on reform policy or whether they just block it, driven by confrontational preferences. In addition, the Management Index switches perspectives and asks to what extent governments are able to anticipate veto players in the legislative process (Evans and Manning 2003).

One might still argue that each OECD country has a different number of veto points (i.e., potential blockade constellations) that their respective governments need to consider in legislative processes. For example, whereas British governments do not have to take into consideration either a constitutional court's concerns or those of a second parliamentary chamber with veto rights, German governments must take these two formidable veto players into account.

This comparison naturally raises the question as to whether a German government that must anticipate more veto points in the lawmaking process should have its success more positively evaluated than a British government that, in comparison, has fewer veto points requiring consideration.

For the purposes of the SGI, we have decided to treat states equally in this respect, regardless of whether their governments must anticipate two, three, four or five veto points. One could object to this methodology by pointing to the aforementioned example and argue that it is easier for the British government to attain the best assessment because it effectively has to anticipate only two veto points. At the same time, however, governments in systems with more veto points than the United Kingdom are aware of the additional veto points and therefore can—and indeed must—prepare for them accordingly. To permit *de facto* a bonus for states with veto-intensive systems could result in a situation in which such a state—despite having had several proposed laws fail approval by the constitutional court—is accorded a better score than a government in a system with less veto points that succeeded in having more laws adopted.

Furthermore, according to this methodology, a system based on the majority principle (e.g., the United Kingdom's) does not automatically receive a good rating for strategic capacity just for having comparatively low veto hurdles. Instead, a country's aggregate assessment also takes into account the assessments of the government's consultations with business and social actors and its communication with the public. For example, a British government that, thanks to its majority principle, is able to pass many laws while, at the same time, ignore societal interests, would receive a poor rating for the respective question. In comparison, systems with many veto points usually have various social interests represented among their veto players. As a result, governments in these systems that successfully anticipate veto points can general-

ly expect to receive positive ratings for the questions related to public communication and consultation.*Economic and social distress*

Some scholars have argued that economic and social crises endow newly elected governments with popular mandates for change that aid them in overcoming vested interests (Williamson and Haggard 1994). However, it may well be asked whether the reform pressure generated by economic and social crises facilitates reforms or whether the effects of such crises might also limit a government's scope for action and thereby actually impede reforms. Given this ambivalence, we tend to see national governments as the principle agents responsible for—and capable of—coping with the presence or absence of reform pressures arising from socioeconomic crises.

Attitudes among the population

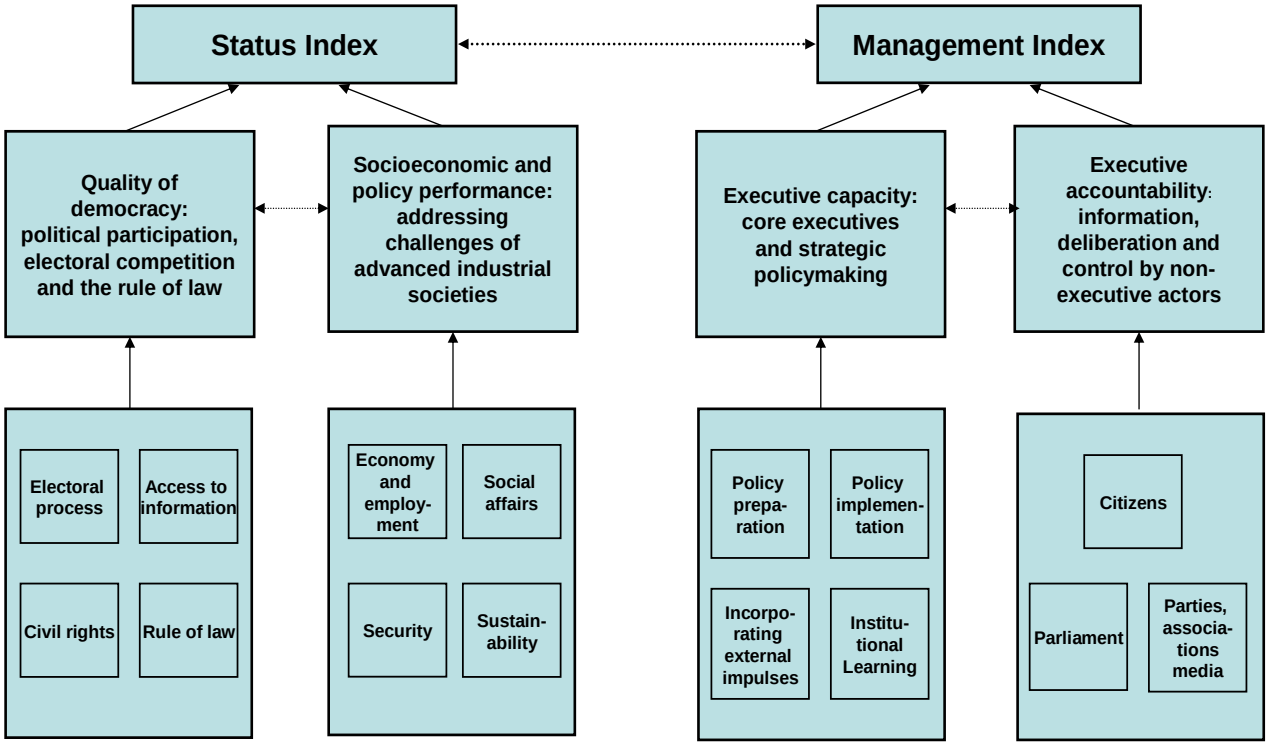
A higher popular willingness to accept costly reforms is likely to support the formulation and implementation of reforms. The assessment of executive accountability in the Management Index takes this contextual condition into account by asking what citizens in a given country know about the motives, objectives, effects and implication of governmental policies. A more profound popular knowledge of policies is assumed to increase the public acceptance of reforms.

Path dependencies

A country's particular path of historical and institutional development is likely to determine the scope and choice of available reform policies (North 1990; Thelen and Steinmo 1992). The SGI's design encompasses different conceptual and methodological strategies to reflect the formative role of such paths. One strategy, of course, is to use an approach of indirect measurement. While the Status Index focuses on policy outcomes and avoids evaluating policies on the basis of chosen instruments alone, the Management Index focuses on functions and processes without judging executive governance on the basis of institutional settings that are considered to be superior. Since the conditions under which path dependence constrains or enhances reform capacity are not well-known, the SGI's approach of indirect measurement seeks to avoid implying such conditions in both its concepts and questions.

The Management Index also uses another strategy that implicitly addresses the extent to which countries are locked into extant and constraining paths of institutional development. Therefore, there are questions under the category of "institutional learning" that aim to evaluate whether government actors monitor their own institutional arrangements of governing and improve strategic capacity by means of targeted institutional reforms.

Figure 2: Components of the Status and Management Index



Measurement

To operationalize and measure the concepts used in constructing the SGI, we decided to rely on a combination of statistical data drawn from official sources as well as the qualitative assessments of country experts. Statistical data are generally more reliable than expert opinions, particularly when they are collected by official institutions and by using methods that conform to cross-national standards. At the same time, however, such data often do not adequately cover the full meaning of a concept. We therefore believe that complex concepts can be measured best through the use of expert assessments that take the country-specific context into account and provide "thick" descriptions capturing the nuances of phenomena. Nevertheless, one must always remember that the responses of experts are prone to bias by subjective perceptions and thereby pose problems of intercoder reliability (Munck and Verkuilen 2002).

The SGI's expert survey questionnaire is designed to improve the validity of expert assessments through the use of six tools and procedural steps. First, many assessment questions are formulated so as to elicit detailed factual evidence rather than broad—and consequently more

subjective—assessments. In fact, many questions ask for responses that may be cross-checked by responses to other questions, statistical data or data from opinion surveys.

Second, the questionnaire provides detailed explanations of and four tailored response options for each question. This information is intended to illustrate the purpose of a question, to structure the way the expert words his or her assessment, and to provide a standardized framework for the production of the country reports. The experts were instructed to adapt the standardized response options to the individual context of the particular country they were evaluating and to substantiate their ratings (numerical assessment) with evidence in their country report (in the following: "expert report"). The rating scale for each question ranges from one to 10, with one being the worst and 10 being the best. The scale is differentiated by four response options provided for each question. Although the written assessments do not allow for a direct reconstruction of the numerical ratings, they do provide an explanatory background for them.

A pretest was conducted in order to evaluate how the SGI's conceptual framework would be understood and interpreted by experts without detailed prior knowledge of the project. We used the results of this pretest to revise the questionnaire and add more explanatory information. Moreover, as part of the assessment and review process, experts entered their textual and numerical assessments into an online database that contains all information on the questions and allowed them and others to keep track of the survey process.

Third, each OECD member state was examined by three leading scholars with established expertise in their respective countries. To identify subjective bias and reduce any distortion it might cause, the experts were selected so as to represent both domestic and external views as well as the viewpoints of political scientists and economists. Each expert was tasked with writing assessments for "their" country, which resulted in the production of three individual and yet parallel country assessments ("expert reports") for each country. The experts were instructed to assess the situation in their countries as of March 2007 and to take into account the period between January 2005 and March 2007 when explaining their ratings. Although many experts knew each other personally and cooperated to some extent in compiling the so-called expert indicators, we checked and ensured that the ratings and written assessments were made independently.

In completing the questionnaire, each expert had to provide ratings for 62 questions, which means that the evaluations for all 30 countries entailed a total of 1,860 scores, or ratings. For 65 percent of the ratings, the expert scores deviated by two levels or fewer, which results in a standard deviation of equal to or less than 0.94. The following table shows the median stand-

ard deviations of the expert ratings for all SGI questions, indicating the degree of congruence among the experts' opinions. Since the formulation of a question, explanation and response options led experts with very different backgrounds to supply very similar ratings, the low standard deviation could be interpreted as a measure of a rating's precision.

At the same time, caution should be exercised here with this interpretation, as the motivation behind selecting three experts per country was indeed to benefit from the input of a variety of political orientations as well as professional experience. In other words, a high standard deviation is a side-effect of the survey's design—not necessarily an indication of a validity problem. Moreover, the review process eliminated all measurement errors resulting from cases in which obvious misunderstandings produced high standard deviations. Organized as a discursive process between and among the experts and reviewers, the review process allowed the participants to clarify concepts, define the exact meaning of questions and agree on conventions of interpretation that would ensure reliable evaluations.

Figure 3: Standard deviations for expert ratings, median values by question

Democracy		Policy performance		Executive capacity				Executive accountability	
S1.1	0.471	S6.1	0.644	M2.1	0.943	M9.1	0.943	M13.1	1.247
S1.2	0.816	S7.1	0.943	M2.3	0.943	M9.2a	0.471	M14.8	0.644
S1.3	0.471	S8.1	0.471	M3.1	1.247	M9.2b	0.943	M14.9	0.471
S2.1	0.471	S9.1	0.816	M3.2	1.124	M9.2c	0.816	M14.10	0.471
S2.2	0.816	S10.1	0.658	M3.3	0.943	M9.3a	0.943	M14.11	0.943
S2.3	0.816	S11.1	0.943	M3.4	1.124	M9.3b	0.816	M14.12	0.486
S3.1	0.816	S12.1	0.943	M3.5	0.486	M9.3c	1.095	M14.13	0.471
S3.2	0.880	S13.1	1.095	M3.6	0.943	M10.1	0.943	M15.1a	0.943
S4.1	0.816	S14.1	0.816	M4.1	0.816	M11.1a	0.943	M15.2c	0.943
S4.2	0.644	S14.2	0.816	M4.2	1.095	M11.2	1.247	M15.3a	0.943
S4.3	0.644	S14.3	0.943	M4.3	1.000	M12.1	0.816	M15.3b	0.943
		S15.1	0.816	M5.1	0.88	M12.2	1.247		
		S16.1	0.943	M6.1	0.943				
		S17.1	0.816						
		S18.1	0.880						

Fourth, the countries examined by the SGI were subdivided among seven regional coordinators according to geographical location. These regional coordinators, who are political scientists with both comparativist and area expertise, were each responsible for three to five of the 30 OECD countries. Selecting information from each expert report according to the criteria of validity and objectivity, the regional coordinators integrated this information into a synthesized "country report" and gave numerical ratings based on those provided in the three expert reports.

Fifth, the regional coordinators reviewed their ratings collectively so as to make it possible to draw comparisons across the entire OECD world. As part of the discussions making up the

review process, each regional coordinator was required to explain, defend and eventually re-calibrate his ratings and assessments. To make any changes agreed to during the review process more transparent, the coordinators also agreed to choose the median of the three country expert ratings as the default score and, if a deviation from the median score was deemed necessary, to keep the score within the range of ratings provided in the experts' reports. During the review process, the regional coordinators deviated from the respective median values in 31 percent of the total number of 1860 scores provided in the reports. In turn, two percent of these scores exceeded the range defined by the expert ratings and each of these deviations was justified in the body of the country reports.

Sixth, as part of a second round of reviews, an advisory body composed of renowned scholars and practitioners and charged with making strategic decisions discussed and approved the ratings. This second review resulted in changes to 6 percent of the total scores and entailed a slight reduction in the proportion of scores deviating from the country expert median in the expert reports (down to 29 percent from 31 percent) and exceeding the range of the ratings in the experts' reports (down to 1 percent from 2 percent).

The SGI's other main sources of data are quantitative indicators collected from publicly available statistics. Giving country experts and coordinators access to these indicators through the online database allowed them to rely on an equal basis of standardized information. Values missing from public statistics were supplemented with those from previous years or other sources. If the latter was not possible, then the missing value was imputed using the median of the available values. In the cases of at-risk-of-poverty rates (items S11.2, S12.4 and S13.3), the missing values for non-EU member states were imputed by regressing at-risk-of-poverty rates on relative poverty rates obtained from the Luxembourg Income Study (LIS).[i] These imputation techniques were applied to ensure the comparability of data drawn from different sources.

Furthermore, for some countries, values are missing because the structures of their political systems make it impossible to derive certain indicators. Such "system-missing" values occur, for example, for countries that do not have second chambers of parliament (item M8.1), presidents with legislative veto rights (item M8.2), constitutional courts capable of vetoing legislation (item M8.3), and expert staff assigned either to individual parliamentary deputies (item M14.5) or to parliamentary groups of parties (item M14.6). In such cases, these specific indicators were not included in the aggregation process for such countries.

In contrast, countries in which governments did not assess the potential socioeconomic impact of draft laws (item M4.1, RIAs) were given the worst possible score (i.e., one) because the concept behind the SGI assumes that the application of RIAs enhances executive capacity. In the cases of these countries, the related questions for items M4.2 and M4.3 have also been given the lowest score.

Weighting and aggregation

While the expert ratings are based on a unified scale ranging from one to 10, the quantitative indicators are provided using different scales and units of measurement. In order to aggregate the latter into composite indices, the indicators had to first be standardized. This was accomplished by calculating the relative distance from the best performing state and assigning a value to this distance using a scale ranging from one to 10. In cases where lower values of indicators denoted better performance, the scores were inverted so as to guarantee that higher scores always represented better performance. This technique of standardization through a linear transformation was chosen for the SGI because it is both intuitively plausible and easier to understand than, for example, a z-transformation or a transformation based on a logistic function (Matthes and Schröder 2004).

In addition, the chosen method of standardization has desirable effects insofar as it generates scales with identical ranges and fixed end points, limits the influence of outlier values and increases the distance between values lying within a narrow interval, so as to emphasize the relative position of states vis-à-vis other states (Giovannini et al. 2005). To check the robustness of our standardization approach, we also calculated the Status Index and Management Index using z-transformed and logistic-function-transformed values. These standardization methods produced very similar rankings, with the scope of rank shifts being limited to a maximum of three ranks (see appendix).

This standardization procedure was modified for the following expert-provided quantitative indicators:

M14.2/ M14.3 – Number of parliamentary committees/ average number of (sub-) committee members: As there is no strictly monotonous, linear relationship between the number and size of committees and their ability to monitor the executive, value ranges were determined according to accepted “best practices” in committee organization (Schnapp and Harfst 2005). For example, countries with 12-18 committees and an average number of 13-25 deputies per

(sub-) committee were given a score of 10. Scores for countries with more or fewer and larger or smaller committees were depreciated as a result of applying the above-mentioned linear transformation;

M14.6/ M14.7 – Expert support staff per faction/expert support staff per deputy: As the distributions of the values for these indicators were skewed by extremely high values for the U.S. Congress, these values were logtransformed prior to standardization;

M15.2a – Fragmentation of the party system: This indicator is based on the effective number of parliamentary parties, a standard measure of party system fragmentation (Laakso and Taagepera 1979). The underlying assumption is that more fragmented party systems with many small parties are less capable of generating program-based competition (as opposed to factionalism), which is believed to improve executive accountability. However, the causal relationship is also not strictly monotonous, and assuming a monotonous relationship would introduce a bias in favor of systems with either dominant parties or two parties. For this reason, we set a cutoff value of five (Sartori 1976) and depreciated the scores only for countries with an effective number of parties exceeding this threshold.

For most indicators, there are no broadly agreed-upon, absolute benchmarks that denote top- or bottom-level performance. This is the case either because performance is assumed to increase or decrease continually or because established benchmarks (e.g., the threshold of a general government deficit of 3 percent of GDP, which the European Union uses as an eligibility criterion for membership in the Economic and Monetary Union) remain contested among scholars and policymakers. For this reason, we decided to define empirical, relative benchmarks by assigning scores of one and 10 to the worst and best performing state, respectively, within the given set of countries.

This benchmarking technique made full use of the range given by the scale. At the same time, however, it also caused a certain degree of divergence between indicators and expert ratings, as the latter rarely chose the lowest possible scores. This means that the choice of empirical, relative benchmarks led to a situation in which, for the quantitative indicators, the worst performers were assigned a score of one even if they performed only slightly worse than other countries. In contrast, without a method of standardization for the expert ratings, the worst performers here would have “suffered” less from a small gap in relation to better-performing countries. Since we did not want to treat quantitative indicators and expert ratings in different ways, we rescaled all expert ratings so as to generate distributions with identical ranges. To

avoid the emergence of any discrepancies between expert ratings and written assessments, we also include the original, non-standardized expert ratings in the country reports.

In order to integrate individual items into a composite index, weights have to be assigned to all individual items. The SGI's method of weighting these items has been guided by three considerations:

In the first place, we decided that weights should reflect the conceptual status of items, criteria, categories and dimensions that are components of the key SGI concepts of democracy, policy performance and executive governance. Once these concepts were disaggregated into their components, theoretical reasoning was used to identify, define and juxtapose these components. For example, the idea of distinct stages in the policy cycle inspired the disaggregation of the "executive capacity" dimension into categories such as policy preparation, implementation and learning. In contrast, our prior empirical knowledge about, for example, the impact of effective inter-ministerial coordination on the preparation of policies was mainly based on the experiences of practitioners, case-based evidence, intuition and common sense.

Our knowledge has been particularly limited when it comes to the interaction of individual components with each other, for example, on how inter-ministerial coordination, regulatory impact assessments and strategic planning jointly affect policy preparation. This uncertainty about effects and interrelations suggests that components might best be considered hypotheses about the presence or fulfillment of a concept (Goertz 2006, 53-58). For example, by defining inter-ministerial coordination as a component of policy preparation, one must assume that effective inter-ministerial coordination improves policy preparation. On the more aggregate level of SGI categories, it is contended that effective mechanisms of policy preparation in combination with effective implementation and institutional learning increase the strategic capacity of executives. However, we do not know precisely how much individual components contribute to the aggregate concept and whether certain components reinforce or hamper the contributions of other components.

Given these uncertainties, the safest strategy for building indices is to assume, on the one hand, that all components possess equal status as hypotheses about the presence and fulfillment of aggregate concepts and, on the other hand, that each component may partially, but not fully substitute for the effect of other components. The corollary for the construction of the index at this point is to assign equal weights to all components and choose an additive method of aggregation.

Second, the SGI has been operationalized as a combination of an expert survey and a compilation of so-called hard statistical data. This methodological choice is motivated by taking two facts into consideration: On the one hand, OECD member states are well-charted by numerous datasets, and there are official, cross-national datasets that provide information that is more reliable than subjective assessments of experts. On the other hand, statistical data cover only very specific aspects of more complex realities and ignore a context that can allow for a fuller understanding of an indicator's particular meaning. In order to take this complexity more fully into account, experts were asked to provide contextualized assessments that were then subjected to a review process.

In this way, the combination of expert assessments and statistical indicators assumes that both types of observations have specific strengths and weaknesses, that they cannot fully substitute for each other, and that neither of them is epistemologically superior to the other (Collier, Brady and Seawright 2004, 252-258). For this reason, we decided to assign equal weight to the expert assessments and the sets of indicators within the policy areas constituting the performance assessment of the Status Index.

Third, the Status Index represents an integrated measurement of the quality of democracy and the policy performance in OECD member states. Since most OECD member states are stable, functioning democracies, one might be tempted to infer that the assessment of democracy should be given less weight than the performance assessment. However, the results of the SGI expert surveys and other democracy assessments indicate that the quality of democracy varies even in consolidated democracies. Moreover, a high-quality democracy may be viewed as being a necessary framework for strategic policymaking, reform capacity and meaningful policy performance. For this reason, we have assigned equal weight in the Status Index to the dimensions of policy performance and quality of democracy.

The Status Index and the Management Index scores are thus derived by calculating the arithmetic means of the scores for their respective two dimensions (i.e., the Status Index's quality of democracy and policy performance and the Management Index's executive capacity and executive accountability). The individual dimension scores in the Status Index are derived by calculating the arithmetic means of the criteria scores , which are also derived by calculating the arithmetic means of their respective components. The categories (i.e., the four broad policy sectors) of the Status Index do not imply a theoretical status within the SGI conceptual framework. They are descriptive and used only to group policy areas. There are therefore no calculations at the category level in the Status Index.

The Management Index's dimension scores likewise represent the arithmetic means of their equally weighted component scores, but the Management Index contains two additional levels of disaggregation so as to reflect the greater diversity of governing practices and mechanisms addressed by the individual questions. The "executive capacity" dimension is disaggregated into stages of the policy process (e.g., preparation, implementation etc.) that constitute four categories, each of which consists of between one and five criteria that are used to group activities such as regulatory impact assessment or the anticipation of veto players. In addition, a distinction is drawn between single items (e.g., with M9.1, which asks about how the government achieves its own policy objectives) and sets of items that are closely related to each other by using lettered annotations (e.g., a, b, c, etc.). For example, intra-executive monitoring mechanisms are viewed as forming such a set, consisting of: organizational incentives limiting ministerial self-interest (M9.2a), the monitoring of line ministries (M9.2b), executive agencies (M9.2c) and internal auditing arrangements (M9.2d). These items are weighted equally, and together are assigned the same weight as M9.1.

The two composite indicators—that is, the Status Index and the Management Index— provide scores and ranks for each of the 30 states. The ranking is based on the score that is precise to the second decimal place. If two or more states have the same score at this level of precision, they are ranked equally.

In order to confirm the robustness of our chosen approach of weighting, we tested various other weighting models. For the Status Index, for example, we first increased the weight given to the policy performance dimension. Doing so could theoretically be justified by arguing that most OECD member states are stable democracies where policy delivery matters more than rights and procedures. Second, we explored a model with equally weighted components, thereby weighting the expert ratings according to the total number of items forming a policy area/ criterion. Such a model would reflect the assumption that policy areas with more quantitative performance indicators are better charted by hard statistical data and are therefore less dependent on subjective expert opinions. Third (and fourth), we calculated aggregate values based solely upon expert ratings and, conversely, on quantitative indicators. These models would correspond to qualitative or quantitative research designs, respectively. As figure 4 shows, the aggregate correlations between these models and our default model are fairly high, which confirms the robustness of our weighting.

Figure 4: Effects of different weighting models, compared with the chosen model of aggregation

	Correlation with default index values	+/-3 ranks	+/-2 ranks
Democracy 30% performance 70% ratings = indicators	.996	Korea /-	Japan / Portugal, Spain
Democracy 50% performance 50% components equally weighted	.999	-/-	-/-
Democracy 30% performance 70% components equally weighted	.995	Korea /-	Japan / Finland, Poland, Portugal, Spain
Democracy 50% performance 50% only ratings weighted	.998	-/-	-/ Iceland
Democracy 30% performance 70% only ratings weighted	.992	-/ Portugal	Japan, Sweden /-
Democracy 50% performance 50% only indicators weighted	.995	-/-	-/ Czech Republic
Democracy 30% performance 70% only indicators weighted	.988	Japan, Korea, Luxembourg / Netherlands	Greece, Iceland / Czech Republic, Poland, Portugal, Spain, United Kingdom

In sum, the Sustainable Governance Indicators have been designed to assess reform capacity indirectly by measuring both policy outcomes and executive governance. This indirect approach also entails a degree of methodological, theoretical and political self-restraint, in that it assumes that there is no single recipe for reform to be written by social scientists. Moreover, any attempt to communicate the "right" reform will always arouse suspicions of ideological motives behind the proposed reforms..

Our measurement of executive governance combines state-centric and societal notions of governance, monitors the impact of veto players, and focuses on micro-level functions and processes of governance that reflect a growing consensus on best practices beyond traditional macro-categories of political systems. In contradistinction to the existing composite indicators and comparative assessments, the SGI explore policy outcomes and governmental practices in greater detail, and they do so for a larger sample of states while using more recent data. Moreover, the SGI integrate expert assessments and statistical data in order to combine the advantages of both types of information. As all disaggregate data are published together with the aggregate indices, the data may be customized by recombining them in different ways or with other datasets.

Acknowledgement

The author is indebted to Margaret Kraus (Calculus Consult) who provided important advice and feedback on statistical and technical issues, imputed missing values, explored different standardization procedures and constructed excel sheets for the aggregation of scores.

References

Andeweg, Rudy B. "On Studying Governments". *Governing Europe*, Ed. Jack Hayward and Anand Menon, Oxford 2003. 39-60.

Ben-Gera, Michal. *Co-Ordination at the Centre of Government: The Functions and Organisation of the Government Office*. OECD, SIGMA paper 35, Paris 2004.

Benz, Arthur. "Konstruktive Vetospieler in Mehrebenensystemen". *Die Reformierbarkeit der Demokratie. Innovationen und Blockaden*, Ed. Renate Mayntz and Wolfgang Streeck, Frankfurt/New York 2003. 205-238.

Collier, David, Henry E. Brady, and Jason Seawright. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology". *Rethinking Social Inquiry. Diverse Tools, Shared Standards*, Ed. Henry E. Brady and David Collier, Lanham 2004. 229-265.

Common, Richard. "Organisational Learning in a Political Environment. Improving Policy-Making in UK Government". *Policy Studies* (25) 1 2004. 35-49.

Dolowitz, David, and David Marsh. "Learning from Abroad: The Role of Policy Transfer in Contemporary Policy-Making". *Governance* (13) 2000. 5-24.

Evans, Gord, and Nick Manning. *Helping Governments Keep Their Promises. Making Ministers and Governments More Reliable Through Improved Policy Management*. World Bank, Internal Discussion Paper IDP-187, Washington 2003.

Giovannini, Enrico, Michela Nardo, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, and Anders Hoffman. *Handbook on Constructing Composite Indicators. Methodology and User Guide*. OECD, OECD Statistics Working Paper STD/DOC(2005)3, Paris 2005, doi:10.1787/533411815016.

Goertz, Gary. *Social Science Concepts: a User's Guide*. Princeton 2006.

Hall, Peter, and David Soskice (ed.). *Varieties of Capitalism. The Institutional Foundation of Comparative Advantage*. Oxford 2001.

Howell, Chris. "Varieties of Capitalism. And Then There Was One?" *Comparative Politics* (36) 1 2003. 103-124.

Knack, Steve, Mark Kugler, and Nick Manning. "Second Generation Governance Indicators". *International Review of Administrative Sciences* (69) 3 2003. 345-364.

Laakso, Markku, and Rein Taagepera. "Effective Number of Parties. A Measure with Application to Western Europe". *Comparative Political Studies* (12) 1 1979. 3-27.

Matthes, Jürgen, and Christoph Schröder. "Rahmenbedingungen für Unternehmen - Zur Aggregation von Weltbankdaten". *IW-Trends* 4 2004. 1-20.

Merkel, Wolfgang. "Embedded and Defective Democracies". *Democratization* (11) 5 2004. 33-58.

Munck, Gerardo L., and Jay Verkuilen. "Conceptualizing and Measuring Democracy. Evaluating Alternative Indices". *Comparative Political Studies* (35) 1 2002. 5-34.

North, Douglass C. *Institutions, Institutional Change and Economic Performance*. Cambridge 1990.

Nunberg, Barbara. *Ready for Europe: Public Administration Reform and European Union Accession in Central and Eastern Europe*. Washington D.C. 2000.

OECD. *Modernising Government: The Way Forward*. Organisation for Economic Cooperation and Development, Paris 2005, http://www.oecd.org/document/15/0,2340,en_2649_33735_35405455_1_1_1_1,00.html.

Pierre, Jon, and B. Guy Peters. *Governing Complex Societies. Trajectories and Scenarios*. New York 2005.

Radaelli, Claudio. "The Europeanization of Public Policy". *The Politics of Europeanization*, Ed. Kevin Featherstone and Claudio Radaelli, Oxford 2003. 27-56.

Sartori, Giovanni. *Parties and Party Systems. A Framework for Analysis*. Cambridge 1976.

Schedler, Andreas. "Conceptualizing Accountability". *The Self-Restraining State: Power and Accountability in New Democracies*, Ed. Andreas Schedler, Larry Diamond, and Marc F. Plattner, Boulder 1999. 13-52.

Schmitter, Philippe C. "The Ambiguous Virtues of Accountability". *Journal of Democracy* (15) 4 2004. 47-60.

Schnapp, Kai-Uwe, and Phillip Harfst. "Parlamentarische Informations- und Kontrollressourcen in 22 westlichen Demokratien". *Zeitschrift für Parlamentsfragen* (36) 2 2005. 348-370.

Thelen, Kathleen, and Sven Steinmo. "Historical Institutionalism in Comparative Politics". *Structuring Politics. Historical Institutionalism in Comparative Politics*, Ed. Sven Steinmo, Kathleen Thelen, and Frank Longstreth, Cambridge 1992. 1-32.

Tsebelis, George. *Veto Players. How Political Institutions Work*. New York/Princeton 2002.

Wiesenthal, Helmut. "Konventionelles und unkonventionelles Organisationslernen. Literaturreport und Ergänzungsvorschlag". *Zeitschrift für Soziologie* (24) 2 1995. 137-155.

Williamson, John, and Stephen Haggard. "The Political Conditions for Economic Reform". *The Political Economy of Policy Reform*, Ed. John Williamson, Washington D.C. 1994. 527-596.

Appendix:

Status Index: Different standardization methods compared

	[0,1]-standardization		Log-standardization		z-standardization		Rank differences	
Country	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**[0,1] - log**	**[0,1] - z**
Austria	6.91	14	5.71	14	6.06	12	0	2
Belgium	6.61	17	5.59	17	5.69	16	0	1
Canada	7.89	8	6.04	8	6.69	8	0	0
Czech Republic	6.13	18	5.28	19	5.21	18	-1	0
Denmark	8.21	6	6.28	6	6.95	6	0	0
Finland	8.64	2	6.46	1	7.29	2	1	0
France	6.06	19	5.29	18	5.13	19	1	0
Germany	7.47	10	5.84	11	6.41	11	-1	-1
Greece	4.16	28	4.42	28	3.73	28	0	0
Hungary	5.22	25	4.97	25	4.56	25	0	0
Iceland	7.51	9	5.99	9	6.44	10	0	-1
Ireland	7.32	11	5.94	10	6.47	9	1	2
Italy	4.61	26	4.61	26	3.99	26	0	0
Japan	5.7	22	5.21	20	4.78	21	2	1
Luxembourg	6.64	16	5.67	15	5.83	15	1	1
Mexico	3.72	29	4.36	29	3.62	29	0	0
Netherlands	8.23	5	6.29	5	7.13	4	0	1
New Zealand	8.41	4	6.30	4	7.12	5	0	-1
Norway	8.71	1	6.45	3	7.23	3	-2	-2
Poland	4.22	27	4.47	27	3.90	27	0	0
Portugal	5.89	20	5.10	23	4.74	23	-3	-3
Slovakia	5.53	23	5.12	22	4.79	20	1	3
South Korea	5.42	24	5.08	24	4.60	24	0	0
Spain	5.73	21	5.15	21	4.77	22	0	-1
Sweden	8.58	3	6.45	2	7.29	1	1	2
Switzerland	7.98	7	6.09	7	6.77	7	0	0
Turkey	2.41	30	3.87	30	3.11	30	0	0
United Kingdom	7.02	13	5.74	12	6.02	13	1	0
United States	6.73	15	5.59	16	5.67	17	-1	-2

Z-standardization: Quantitative indicators and expert ratings (x) are standardized by subtracting the item-specific mean ($\bar{x}$) and dividing by the standard deviation.

$$y = \frac{x - \bar{x}}{SDEV_x}$$

Standardization based upon a logistic function: This procedure reduces the influence of extreme values while preserving the distance information of indicators. In a first step, quantitative indicators and expert ratings are z-transformed. The standardized values (x) are included into the following logistic function:

$$F(x) = \frac{1+9}{1+e^{-c \cdot x}}$$

The higher c, the steeper F(x) and the more small differences from the mean are expanded. Following Matthes and Schröder (2004), we calculate c as the square root of the coefficient of

variation of the respective indicator. This method ensures that equal absolute differences are not dependent on the value of the mean and have a larger effect if the variation is low.

Management Index: Different standardization methods compared

	[0,1]-standardization		Log-standardization		z-standardization		Rank differences	
Country	**Score**	**Rank**	**Score**	**Rank**	**Score**	**Rank**	**[0,1] - log**	**[0,1] - z**
Australia	6.53	11	5.89	10	6.09	11	1	0
Austria	6.54	10	5.79	13	6.10	9	-3	1
Belgium	4.92	25	4.89	28	4.69	25	-3	0
Canada	6.97	8	6.13	8	6.60	7	0	1
Czech Republic	4.75	27	4.98	26	4.49	27	1	0
Denmark	8.07	2	6.68	3	7.48	2	-1	0
Finland	7.94	3	6.69	2	7.36	3	1	0
France	5.06	24	5.18	23	5.06	22	1	2
Germany	6.31	15	5.81	12	6.04	12	3	3
Greece	3.33	30	4.37	30	3.70	30	0	0
Hungary	5.55	19	5.35	19	5.16	19	0	0
Iceland	7.44	5	6.44	6	6.88	6	-1	-1
Ireland	7.01	7	6.14	7	6.42	8	0	-1
Italy	4.89	26	5.02	25	4.50	26	1	0
Japan	5.5	21	5.40	18	5.25	18	3	3
Luxembourg	6.33	14	5.68	15	5.91	15	-1	-1
Mexico	5.36	22	5.28	22	5.01	23	0	-1
Netherlands	6.67	9	5.88	11	6.10	10	-2	-1
New Zealand	7.4	6	6.46	5	7.05	5	1	1
Norway	8.48	1	6.90	1	7.54	1	0	0
Poland	4.06	29	4.70	29	4.09	29	0	0
Portugal	5.55	19	5.34	20	5.13	20	-1	-1
Slovakia	5.6	18	5.31	21	5.09	21	-3	-3
South Korea	5.85	17	5.56	17	5.43	17	0	0
Spain	5.07	23	5.08	24	4.80	24	-1	-1
Sweden	7.85	4	6.63	4	7.31	4	0	0
Switzerland	6.46	13	5.75	14	5.96	13	-1	0
Turkey	4.7	28	4.92	27	4.45	28	1	0
United Kingdom	6.11	16	5.57	16	5.50	16	0	0
United States	6.52	12	5.98	9	5.94	14	3	-2

[i] As no LIS values were available for Iceland (and partially for Korea), the mean of the poverty rates for those four countries with Gini coefficients resembling Iceland most closely was used in the regression.