WIDER Working Paper No. 2013/068

Evaluating governance indexes

Critical and less critical questions

Rachel M. Gisselquist*

July 2013

Abstract

Recent years have seen a proliferation of 'composite indicators' or 'indexes' of governance. Such measures can be useful tools for analysing governance, making public policy, building scientific knowledge, and even influencing ruling elites, but some are better tools than others and some are better suited to certain purposes than others. This paper provides a framework of ten questions to help users and producers of governance indexes to evaluate them and consider key components of index design. In reviewing these ten questions—only six of which, it argues, are critical—the paper offers examples from some of the best known measures of governance and related topics. It advances two broad arguments: First, more attention should be paid to the fundamentals of social science methodology, i.e., questions about concept formation, content validity, reliability, replicability, robustness, and the relevance of particular measures to underlying research questions. Second, less attention should be paid to some other issues commonly highlighted in the literature on governance …/

Keywords: governance, indexes, composite indicators, research methods, development
JEL classification: B40, C43, H11, H83

*UNU-WIDER, Helsinki, email: gisselquist@wider.unu.edu

This study has been prepared within the UNU-WIDER project 'ReCom–Foreign Aid: Research and Communication', directed by Tony Addison and Finn Tarp.

UNU-WIDER gratefully acknowledges specific programme contributions from the governments of Denmark (Ministry of Foreign Affairs, Danida) and Sweden (Swedish International Development Cooperation Agency—Sida) for ReCom. UNU-WIDER also gratefully acknowledges core financial support to its work programme from the governments of Denmark, Finland, Sweden, and the United Kingdom.

ISSN 1798-7237 ISBN 978-92-9230-645-8

… measurement, i.e., questions about descriptive complexity, theoretical fit, the precision of estimates, and correct weighting. The paper builds upon a thorough review of the literature and the author's three years of research in practice as co-author of a well-known governance index.

Acknowledgements

I am grateful to Robert I. Rotberg and Aniket Bhushan for their comments on this paper and our discussions about governance indexes, as well as to the Centre for International Governance Innovation (CIGI) and the North-South Institute (NSI) for hosting a June 2013 workshop on 'Measuring Governance Effectiveness: National and International Dimensions' and to its participants for thoughtful advice.

UNU World Institute for Development Economics Research (UNU-WIDER)
Katajanokanlaituri 6 B, 00160 Helsinki, Finland

Typescript prepared by Janis Vehmaan-Kreula at UNU-WIDER.

1 Introduction

Recent years have seen a proliferation of measures of governance. Some of the best known include the World Bank's Worldwide Governance Indicators (WGI) and Country Policy and Institutional Assessment (CPIA), Transparency International's Corruption Perceptions Index, and Freedom House's Freedom in the World. We might also add to this list measures designed to capture closely related topics, such the UNDP's Human Development Index, Vision of Humanity's Global Peace Index, and the Legatum Prosperity Index.

Most of these measures combine, in some way, several metrics to get a single score, grade, or rating. Such 'composite indicators' or 'indexes' thus capture multiple facts or dimensions of a complex, multi-dimensional concept in a way that facilitates straightforward evaluation and comparison. Such measures can be useful tools for analysing governance, making public policy, building scientific knowledge, and even influencing ruling elites, but some are better tools than others and some are better suited to certain purposes than others. This paper provides a framework of ten questions to help users of governance indexes evaluate them and decipher which is which, as well as to help producers of governance indexes consider key components of index design.

In reviewing these ten questions—only six of which, it is argued here, are critical—this paper offers examples from some of the best known measures of governance and related topics. It advances two broad arguments through this discussion: First, more attention should be paid in governance measurement to the fundamentals of social science methodology, i.e., questions about concept formation, content validity, reliability, replicability, robustness, and the relevance of particular measures to underlying research questions. Second, less attention should be paid to some other issues commonly highlighted in the literature on governance measurement, i.e., questions about descriptive complexity, theoretical fit, the precision of estimates, and correct weighting.

Most of the questions reviewed here are not brand new to the literature on governance measurement. Some are also identical to the core questions social scientists are taught to ask in designing any measurement project. However, to the best of the author's knowledge, this is the first time these ten questions have been addressed together in this way and applied to the evaluation of diverse governance indexes. In identifying four of these questions as less critical this paper also challenges commonly accepted arguments in the literature on governance measurement. The point of this exercise is not to single out particular indexes for criticism or praise. All of the indexes discussed here are because they are in widespread use and have some major strengths.

This paper focuses on composite indicators of governance but the framework is broadly applicable to the evaluation of 'single' governance indicators as well, although some questions (in particular those on weighting and aggregation) will not be relevant. More broadly, it may also be used to consider composite indicators on topics other than governance. However, individual topics warrant individual attention too. The four questions identified as not necessarily critical in governance measurement, for instance, may be more important in measurement of other complex concepts for which causal theories are more developed, underlying data stronger, and various components easier to measure.

The framework presented here builds both on a review of the literature, and on three years of research in practice, specifically the author's experience in working with Robert I. Rotberg to design and co-author the first two editions of the Ibrahim Index of African Governance (IIAG), one of the best-known measures of governance for Africa, described by Mitra (2013: 489) as 'a well-established index that has become a reference point for governments and NGOs'.[1] Rotberg and Gisselquist, then based at Harvard University, developed the first pilot edition of the IIAG in 2007 and the second edition in 2008, both with financial support from the Mo Ibrahim Foundation.[2] Subsequent editions of the IIAG have been compiled by Ibrahim Foundation staff in house in London.

This paper draws on several examples from that research, and the argument presented here is (hopefully) also reflected in the author's work on the IIAG. However, this paper is not intended as a review or defense of that particular governance measure. The framework presented here can equally be used to consider strengths and weaknesses in the IIAG, several of which are discussed explicitly below.

The paper begins with a brief discussion of governance measurement and its relevance to research and policy analysis. It then introduces broad critiques that emerge from the existing literature on indexes. The rest of the paper focuses on the framework for evaluation, drawing on the literature and examples from various measurement projects to discuss each of the ten questions in turn. The conclusion briefly notes additional practical questions that will be important for some users and producers of governance indexes.

2 Governance measures and their uses

Governance is a contested concept (see Gisselquist 2012). This paper is intentionally agnostic about the many specific and competing definitions of governance. Broadly, the act of governance is understood here as the exercise of political power to manage a community's affairs (see Weiss 2000). The quality of governance, sometimes called government performance, is understood to refer then to the quality of this exercise of power and the quality of this management, including (but not limited to) its outcomes in terms of the quality of public goods and services received by citizens (see, e.g., Bratton 2011, 2013; Rotberg 2013; Rotberg and Gisselquist 2009; Stiglitz et al. 2009).

---

[1] In developing the IIAG, for instance, detailed 'indicator evaluation sheets' were prepared by the authors' team of research assistants for dozens of variables considered for inclusion in the pilot IIAG, a rich source of information about the strengths and weaknesses of available data. The author also worked closely with research assistants and affiliates at Harvard to develop more detailed research notes on various difficult-to-measure topics, such crime, corruption, gender equity, and inequality, and with research affiliates throughout Africa to compile available data from national government and other sources. Over 70 individuals contributed to the project as research assistants or affiliates and provided valuable insights. Please see Rotberg and Gisselquist (2009: 2-4) for a full listing of names. In developing the model used in the IIAG, the strengths and weaknesses of multiple weighting and aggregation options were considered, various methods were applied and tested, and numerous experts consulted. Finally, responses to formal and informal presentations, along with published critiques and other written feedback, provided invaluable information not only on the technicalities of government measurement, but also on the uses and understandings of governance measures by diverse audiences, including scholars, practitioners, and members of the public, in both developing and developing countries.

[2] The authors also published the 'Index of African Governance' in 2009 using the same model as their 2007 and 2008 IIAGs.

Measurement is a major topic in the large literature on governance (see Andrews et al. 2010; Arndt 2008; Davis et al. 2012; Hallerberg and Kayser 2013; Oman and Arndt 2010; Rothstein and Teorell 2012). Governance can be assessed at multiple levels, but the discussion here focuses on measures at the country level. Rotberg, Bhushan, and Gisselquist (2013) identifies over 100 such index projects and databases that seek to measure either governance broadly defined or some core component of it.

Governance measures have been used and produced by both scholars and practitioners. However, international organizations like the World Bank have largely taken the lead (Arndt 2008; Davis et al. 2012; Thomas 2007). With several important exceptions, scholars have not played a leading role in governance measurement (see Holmberg and Rothstein 2012; Levi 2006). This arguably stems in part from a desire to remain independent from what are often heated ideological and normative debates, for instance over the good governance agenda in development policy. As a result, although 'who governs?' and 'how well?' are central questions for political science in particular, many academic political scientists have simply remained apart from discussions about measuring governance (see Putnam 1993).

Another issue is the role of measurement in contemporary political science, where one leading school of thought sees the principal focus of the field instead in the testing of causal theories (see King et al. 1994). This paper suggests, on the contrary, that concept formation on core topics such as governance is central to the building of social scientific knowledge and further that established methods of concept formation and measurement can and should be applied to governance indexes (see Collier and Gerring 2009; Gerring 1999; Sartori 1984). As Laitin (1995: 455-56) argues, conceptual formation has been and should remain central to the field of political science: 'It is hard to think about the political world without [concepts such as 'charisma' and 'the division of labor'], even if their causal role in any political processes is obscure. And many other such concepts guide our thinking and theorizing today …. Such concepts are theoretical in the sense that they combine discrete facts common to our daily life into a category, helping us to see the confusing universe in which we live in a more patterned way'.

Concept formation and measurement are also necessary for the testing of causal theories. Indeed, scholars may have shied away from governance measurement themselves, but that has not stopped them from using existing governance measures developed at and by practitioner organizations in their analyses, all too often with little critical examination of these measures (see Apaza 2009).

The best known composite indicator of governance, the Worldwide Governance Indicators (WGI), underscores such links between measurement, theory development, and policy. The WGI comes out of a long-standing research programme at the World Bank led by Daniel Kaufmann and Aart Kraay, with the assistance of Pablo Zoido-Lobatón and Massimo Mastruzzi. It has been used both in World Bank publications to identify and describe governance trends around the world (e.g., World Bank 2007) and in scholarly journal articles to test major theoretical propositions such as the relationship between governance and growth (e.g., Kaufmann and Kraay 2002; Kaufmann et al. 2007a; Kurtz and Schrank 2007a, 2007b). It is also regularly cited in policy discussions and debates, especially with reference to foreign aid. The Millennium Challenge Corporation, for instance, employs one of the most explicit frameworks for identifying countries that qualify for its assistance and directly incorporates four of the six composite indicators of governance released by the WGI: Control of Corruption, Government Effectiveness, Regulatory Quality, and the Rule of Law. The

WGI is especially important to its assessment of 'ruling justly', one of its core criteria, for which WGI composite indicators comprise three of six data points (Millennium Challenge Corporation 2012). Not surprisingly, there is also a healthy body of work analysing and critiquing the WGI (e.g., Apaza 2009; Arndt and Oman 2006; Knoll and Zloczysti 2011; Kurtz and Schrank 2007a and 2007b; Langbein and Knack 2009; Thomas 2009).

Another measure of governance broadly defined that highlights links between research and policy is the Ibrahim Index of African Governance (IIAG), introduced above. The annual results of the IIAG are covered in many African and international newspapers and have received attention by policy makers and donors. Scholarly journal articles and research reports have also analysed and used our work (e.g., Farrington 2009, 2010, 2011; Frankel forthcoming; McFerson 2009; Mitra 2013; Paruolo et al. 2013; Saisana, Annoni, and Nardo 2009).[3]

Some measures of governance and its components have achieved especially widespread public attention. The annual results of Transparency International's Corruption Perceptions Index (CPI), for instance, are regularly reported and debated in the mainstream press. Indeed, many organizations like Transparency International use such measurement projects as advocacy tools to raise their public profile and advance their policy agenda (see Sampson 2010). Whether or not part of an explicit strategy, Freedom House's Freedom in the World similarly draws attention to its work on civil and political liberties around the world; the Index of Economic Freedom produced by the Heritage Foundation, in collaboration with the *Wall Street Journal*, helps to frame public debate on economic issues in (self-described) 'conservative' terms, highlighting free enterprise, limited government, and individual freedom; Vision of Humanity's Global Peace Index raises awareness of the importance of peace for human progress; and so on.

The vast majority of country-level governance index projects produce one or more composite indicators that combine multiple indicators to get single scores, grades, or ratings. The WGI, for instance, provides six composite indicators, each designed to capture a different aspect of governance. Although some users of the WGI aggregate these into a single comprehensive score, the WGI itself does not do so.

Some governance measurement projects also spotlight individual indicators. The 2007 and 2008 reports on the IIAG, for instance, highlighted the importance of disaggregation and made available all of the underlying figures used in the construction of the index, along with detailed source notes, and comparative scores at the indicator, sub-category, and category levels, as well as overall. Development data tools such as the Gapminder allow for comparative analysis of indicators using interesting visualizations.[4] The 'Dashboard of Sustainability' is another software package that further allows users to construct new composites, selecting the indicators to be included and their weights (Consultative Group on Sustainable Development Indices 2003; Jesinghaus 2012).[5] Other governance measurement projects such as the Afrobarometer and the World Governance Assessment focus on the compilation of disaggregated data on governance.

---

[3] Rotberg and I have not been involved in the IIAG project since 2009. Although published after that date, most of these publications are based on the 2008 IIAG that we authored. Frankel (forthcoming) uses our 2009 IAG. Editions of the IIAG since 2009 employ a somewhat different model.

[4] See http://www.gapminder.org/.

[5] See http://esl.jrc.ec.europa.eu/envind/dashbrds.htm. Developed by the International Institute for Sustainable Development (IISD) and the European Commission's Joint Research Centre (JRC).

Indexes and composite indicators are not new. Gross domestic product (GDP) is a composite indicator (of the total value of goods and services produced in a country) although it is rarely described as such (OECD and Joint Research Centre (JRC) of the European Commission 2008, 22). Composite indicators cover a variety of topics, from economic progress to environmental sustainability to educational quality (see Bandura 2008, 2011; Saltelli 2007). The OECD/JRC *Handbook on Constructing Composite Indicators* is an excellent text that describes the steps in producing a composite indicator and techniques for imputing missing data, normalizing or scaling data for comparability, weighting and aggregation of indicators and components, and robustness and sensitivity checks.

Several issues may be unique to governance measurement as opposed to measurement of many other topics. As discussed further below, one is the complexity and contested nature of the concept of governance itself, and a second is the difficulty of measuring many of its dimensions and underlying components, as well as the extremely poor quality of relevant statistics and the difficulties inherent in collecting information on topics like corruption.

Largely for this latter reason, governance measures have employed perhaps a wider range of types of data than measures of other topics. These include not only standard national statistics such as those compiled in the census and by nationally-representative surveys, but also ratings by experts based on their knowledge of particular cases (e.g., Freedom House's Freedom in the World), surveys of internal or external observers (e.g., World Economic Forum's Executive Opinion Survey), aggregation of multiple such elite surveys or of these and other information sources (e.g., Corruption Perceptions Index, Worldwide Governance Indicators), figures or ratings based on questionnaires that take into account both expert background knowledge and systematic review of relevant legislation or other documentary sources (e.g., Doing Business, World Justice Project Rule of Law Index), systematic coding of reports or other documents based on a detailed protocol to produce simple scores or ratings (e.g., Cingranelli-Richards Human Rights Dataset), and estimated data or imputed figures for key indicators (e.g., Save the Children's Mothers' Index, Human Development Index). Given the diversity of ways in which measurement projects define and operationalize governance, it is not surprising that there can be significant variation across assessments.

3 A brief review of major critiques

Critiques and reviews of existing governance indicators and indexes comprise a large literature. As discussed above, many well-cited critiques of governance indicators, such as Arndt and Oman (2006), have focused on the most widely used governance index, the WGI. Other sourcebooks and research notes also provide more general discussions (Sudders and Nahem 2007; Williams 2011). Several collections of essays have brought together key thinkers to analyse various aspects of governance indicators in depth (e.g., Davis et al. 2012; Hallerberg and Kayser 2013; Hertie School of Governance 2013). Not surprisingly, the authors of governance measures themselves have been particularly active in these debates. Hyden, Court, and Mease (2003), Kaufmann and Kraay (2008), and Rotberg (2004), for instance, identify major weaknesses in the literature that their own projects explicitly seek to address.

Other scholars have also weighed in on the use of composite indicators more generally, what Ravallion (2012) labels 'mashup indices'. In 2008 and 2009, the Commission on the

Measurement of Economic Performance and Social Progress, chaired by Joseph Stiglitz, explored the limits of GDP and the use of alternative measures of well-being, including composite indexes (Stiglitz et al. 2009). Høyland et al. (2012), among others, have bemoaned 'the tyranny' of international index rankings.

One major critique of national governance indexes is that they are inherently flawed because they are too simple. The developers of the World Governance Assessment (WGA), for instance, argue that projects that rank countries in a single index are 'not very helpful … in getting a better understanding of what really happens to governance on the ground in these countries' and may also 'stigmatize' countries based on the perceptions of external experts (Hyden et al. 2003: 4). The WGA proposes instead a broader and more descriptive exercise that relies on interviews in each country with a cross-section of experts to compile information on six policy arenas (civil society, political society, government, bureaucracy, economic society, and judiciary) and six principles (participation, fairness, decency, accountability, transparency, and efficiency) (Hyden, Court, and Mease 2004). Other critics point specifically to the inability of a country-level index to fully capture the significant variations in governance that exist at sub-national levels (e.g., Harttgen and Klasen (2012), UNDP's Global Program on Democratic Governance Assessments). Still other critics highlight the related question of 'actionability'—i.e., they ask, how does diagnosing poor public management, weak service provision, or low levels of accountability, for instance, provide policy-relevant insight into how to fix these deficiencies? (e.g., Williams 2011).

This broad critique raises important questions about the value of indexes relative to other sorts of governance assessment, but overall it seems to be a straw man: Indexes are useful tools for engagement with a non-expert audience and may be their only source of information because they lack the time or interest to engage more broadly. Experts, on the other hand, rarely use national governance indexes as the single means of governance assessment for particular countries. Rather indexes are designed and used to facilitate cross-national comparisons, to explore trends over time, and to identify relationships for further study. Used in this way, governance indexes are entirely compatible with other sorts of assessments. The only question then might be whether the cost and time of compiling them add value, i.e., is it better to simply start with specific analyses or case studies without also doing a blunt comparative assessment using an index?

In terms of actionability, it is true that many indexes are not designed to point to key actionable steps. However, the comparisons they highlight may suggest those steps. For instance, an index that identifies improvements in the rule of law in Mozambique relative to other African countries, helps to identify Mozambique as a model for such reform. More broadly, at least a few governance measurement projects have been designed precisely to address the actionability question: two examples include the World Bank's Actionable Governance Indicators and Service Delivery Indicators.

Other major critiques highlighted in the extant literature accept that indexes may be useful but raise particular concerns about particular measures. One set of criticisms focuses on the validity of particular governance measures, arguing that they do not match what (at least in the critic's view) is a common understanding of governance. Kaufmann and Kraay (2008: 10), for instance, draw a strict distinction between governance outcomes and development outcomes and criticize the IIAG as an 'extreme example' of mixing both together, which they consider problematic because it 'risks making the links from governance to development tautological'. Debates have also focused on whether governance should be conceptualized

and measured in terms of inputs, process, or outcomes (e.g., McFerson 2009) and objective data or perceptions (e.g., Kaufmann and Kraay 2008; Rotberg and West 2004). Other critics relatedly highlight the failure of existing measures of governance to separate out the effect of government action from that of other contextual factors, such as poverty (e.g., Andrews et al. 2010). A number of studies argue that particular indexes should be made broader by including more components, or should be better designed in terms of weighting and methods of aggregation (e.g., Høyland et al. 2012; Mitra 2013; Noorbakhsh 1998).

A second set of critiques highlights the weak theoretical base of much work on governance measurement. Andrews (2008) argues that the production of 'indicators without theory' promotes a model of governance that 'resembles a set of well meaning but problematic proverbs', borrowing loosely from a variety of theoretical traditions (see also Thomas 2009). The result then has limited use in building scientific knowledge or informing development policy. Indeed, Grindle (2004) levels a closely related critique of the ad hoc nature of the entire good governance agenda (see also Grindle 2007, 2010).

Another major set of critiques focuses on the quality of data used in governance measurement and particular data-generating efforts (see, e.g., Delapalme 2011; Rotberg and Gisselquist 2007, 2008, 2009; Round 2012). Many critics argue that some types of data, such as subjective assessments by external experts, are biased and inappropriate measures of governance. A closely related critique notes the lack of transparency in the presentation of indicator data in some measurement projects (e.g., Arndt and Oman 2006: 11; Thomas 2009).

Still another significant group of work highlights lack of precision in the presentation of aggregate scores, rankings, or estimates (e.g., Høyland et al. 2012; Kaufmann and Kraay 2008). These latter five broad critiques of governance indexes are discussed in more depth in the next section of this paper.

4 Critical and less critical questions: a framework for evaluation

Basic social science methodology provides the core framework for evaluating governance indexes and the literature on governance measurement highlights a handful of other issues, for a total of ten core questions that users and producers of governance indexes should ask. The first six of these questions are critical. These deal with concept formation, content validity, reliability, replicability, robustness, and the relevance of particular measures to underlying research questions. Through the discussion below, I argue that the final four are less critical, even though each occupies significant attention in the extant literature on governance measurement. These deal with descriptive complexity, theoretical fit, the precision of estimates, and correct weighting.

Question 1. What exactly does it measure?

Producers of governance indexes often fail to clearly answer this question, perhaps because the answer seems so obvious ('governance'). Concept specification is a simpler step for concepts with more broadly agreed definitions or common theoretical frameworks; because the concept of governance is contested, common understandings of the term cannot be assumed. At a minimum, this conceptualization should resonate with common understandings of the term, popular or scholarly, and the differences between what is considered 'governance' and related concepts like democracy and development should be made clear.

More broadly, Gerring (1999: 367)'s eight criteria provide a useful checklist for evaluating conceptual 'goodness':

1. Familiarity – how familiar is the concept (to a lay or academic audience)?
2. Resonance – does the chosen term ring (resonate)?
3. Parsimony – how short is (a) the term and (b) its list of defining attributes (the intension)?
4. Coherence – how internally consistent (logically related) are the instances and attributes?
5. Differentiation – how differentiated are the instances and the attributes (from other most-similar concepts)? How bounded, how operationalizable, is the concept?
6. Depth – how many accompanying properties are shared by the instances under definition?
7. Theoretical utility – how useful is the concept within a wider field of inferences?
8. Field utility – how useful is the concept within a field of related instances and attributes?

Many governance indexes basically skip the concept specification step and define governance by how it is operationalized. This is arguably the main weakness of the WGI in particular; the literature on the project is impressive but it has not included a detailed discussion of its central concept or of the six components of governance that it identifies. Project literature instead offers brief one to two sentence definitions of the concepts of governance and each of its six components. As Thomas (2009) argues, these brief definitions are largely divorced from other discussions of these concepts in the literature, from which they appear to borrow haphazardly, and no justification is offered. For instance, the dimension of 'Voice and Accountability' apparently takes the notion of 'voice' from Hirschman's classic *Exit, Voice and Loyalty*, but it does not clarify the conceptual linkages or how exactly 'accountability' fits in.

Indeed, many if not most governance indexes can be criticized on similar grounds. The IIAG, for instance, has included a longer discussion of what governance is drawing on Rotberg (2004), but more could be said (see also Rotberg 2007, 2009; Rotberg and West 2004). As the work on the IIAG project has expanded under the leadership of the Ibrahim Foundation, furthermore, little attention has been paid to concept specification. The 2012 IIAG offers the following definition of governance:

> Governance, as defined by the Foundation, is considered from the viewpoint of the citizen. The definition is intentionally broad so as to capture all of the political, social and economic goods and services that any citizen has the right to expect from his or her state, and that any state has the responsibility to deliver to its citizens. The IIAG is unique in that it assesses governance by measuring outputs and outcomes. This definition of governance does not focus on de jure measurements, but rather aims to capture attainments or results, reflecting the actual status of governance performance in a given context—be it national, regional or continental. … (Mo Ibrahim Foundation 2012, 1)

While definitions like these tend to be clear and specific in the sense that they tend to refer to detailed lists of categories, sub-categories, and indicators, they confuse the concept with its measurement and stretch common understandings of the term: Governance is a contested concept, but it is certainly not commonly understood to refer specifically to, for instance, 88 indicators grouped into four categories in the IIAG. In Gerring's terms, such definitions are not 'familiar', they lack resonance, they are not parsimonious, and they are not clearly differentiated from other concepts. Because they do not include discussion of how the various components relate, it is also unclear if they are 'coherent'.

These weaknesses do not mean that such measures should never be used, but they do highlight some major limits to their theoretical and field utility. Studies that use these measures will need to specify concepts post-hoc and to the extent that measures were not developed with these concepts in mind, they will likely be only second best operationalizations of them (e.g., Kurtz and Schrank 2007a and 2007b).

Users of governance indexes should be especially wary of projects that purport to measure 'new' concepts. Creative linguistics can make a lot of sense from a marketing and advocacy point of view, but are problematic from a conceptual standpoint. Two well-respected examples include the Bertlesmann Stiftung's Transformation Index (BTI), a composite index of 'transformation' built from two other indexes on 'status' and 'management' and the Legatum Prosperity Index, an explicit effort to redefine 'prosperity' beyond material wealth. Each of these concepts can be seen as problematic according to Gerring's criteria: They are used in an unfamiliar way and do not (arguably) resonate; their relationship with more commonly used concepts is not fully specified (e.g., is 'transformation' exactly the same as democratization and economic liberalization?); it is not clear if their components form a coherent whole (e.g., social capital, effective governance, human rights, health, security, quality of life, etc. are all 'good things', but why these good things and not others?); and because they are not defined with reference to a theoretical framework, their theoretical utility is not obvious.

This is not to say that conceptual innovation is not possible. One of the best examples of an index measuring a new concept is the UNDP's Human Development Index. The key difference here, however, is the underlying rigor of the concept, with its explicit grounding in Sen's work on human capabilities (Stanton 2007). Freedom House's *Freedom in the World* offers another good example: It defines 'freedom' specifically as 'the opportunity to act spontaneously in a variety of fields outside the control of the government and other centers of potential domination—according to two broad categories: political rights and civil liberties'. These two categories are in turn linked to commonly accepted definitions grounded in the Universal Declaration of Human Rights:

> Political rights enable people to participate freely in the political process, including the right to vote freely for distinct alternatives in legitimate elections, compete for public office, join political parties and organizations, and elect representatives who have a decisive impact on public policies and are accountable to the electorate. Civil liberties allow for the freedoms of expression and belief, associational and organizational rights, rule of law, and personal autonomy without interference from the state. (Freedom House 2012)

Thus, although 'freedom' is in a sense redefined (more narrowly than its everyday usage), Freedom House effectively links the concept as defined to existing frameworks, which helps to differentiate it from other concepts and to make its component parts coherent.

Question 2. Does the operational definition capture the concept?

Once a concept has been properly specified, the next step logically and chronologically, is to operationalize it. An operational definition should identify the component(s) to be included in the measurement and specify how these components are put together in a manner that is consistent with the core concept. In the case of governance, a multi-dimensional concept, this generally involves the aggregation of various categories, (sometimes) sub-categories, and indicators. Index producers have a number of options in terms of how to normalize data and weight and aggregate components.

In the case of single indicators, lack of conceptual validity can often be spotted quite straightforwardly. For instance, a proposal to measure the quality of education in secondary schools based solely on an assessment of the quality of textbooks would have obvious problems of validity. However, for many if not most composite indicators of governance, assessing conceptual validity is far less clear. One approach that is commonly used in assessing the validity of measures of governance and other topics is to assess the measure against other measures of the same concept or against other measures of different concepts that theory suggests should be related to it in a particular way. Asking whether such relationships hold is worthwhile, but as I argue under question 8, it is less critical in assessing the validity of governance measures than in some other areas.

The Cingranelli-Richards (CIRI) Human Rights Dataset provides almost a textbook example of what the specification of an operational definition involves. CIRI includes two composite indexes, the Physical Integrity Rights Index and the Empowerment Rights Index. As summarized in Figures 1a and 1b, each is an additive index of four and seven indicators respectively. Each indicator has a score of between 0 and 2 (which means that each indicator is weighted equally in the construction of each index). Scores for indicators are assigned based on the US State Department Country Reports on Human Rights Practices and Amnesty International's Annual Reports according to a detailed coding protocol (Cingranelli and Richards 2008, 2010). Cingranelli and Richards (1999) and Richards et al. (2001) discuss these indexes and their methodologies in greater detail.

Figure 1a. CIRI physical integrity rights

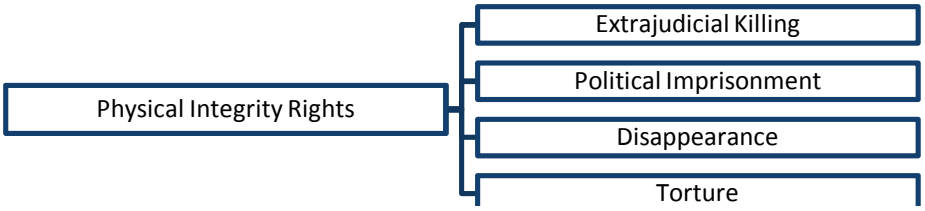
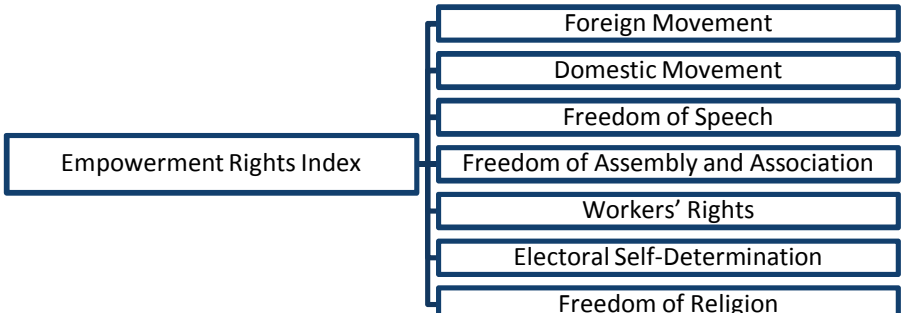


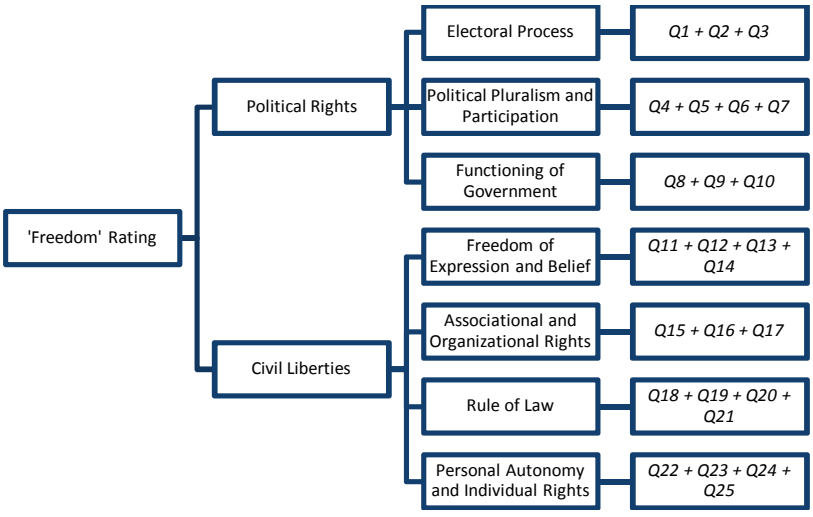
Figure 1b. CIRI empowerment rights[6]



Source: Author's work.

CIRI's operational definition is by no means the only way to capture these concepts. For instance, Freedom House's *Freedom in the World* offers an alternative approach focusing on the closely related concepts of 'political rights' (PR) and 'civil liberties' (CL). As summarized in Figure 2, PR and CL scores are derived from scores on three and four sub-categories respectively (Freedom House 2012). Each of these sub-categories aggregates the values of three to four questions, each with a value of between 0 to 4. There are also two discretionary questions included under PR. The values of these questions are added together for a possible score of 40 for PR and 60 for CL. The total score of each category is then used to assign a value of between 1 and 7 for PR and CL, which is also benchmarked against the previous year. PR and CL scores are averaged (i.e., weighted equally) to derive the overall score upon which the freedom rating is assigned.

Figure 2. Freedom House's Freedom in the World



Source: Author's work.

Overall, this operational definition follows quite straightforwardly from the core concept of 'freedom' as defined above, but several components of it call for a bit more explanation and justification. The sub-categories of 'functioning of government' under political rights and 'rule of law' under civil rights in particular are somewhat unusual additions in that they do not follow clearly from the referenced Universal Declaration of Human Rights, nor are they

6 This is based on the 'new' empowerment rights index.

included in other similar listings of such rights (e.g., 'rule of law' is not included in CIRI's empowerment rights index). Weighting and some of the details of the aggregation method are also not explicitly justified in the discussion: Although the methodology is relatively simple, several debatable claims about the relationships among components and their relative importance are implied by the chosen method:

- PR and CL are equally important to 'freedom' because they receive equal weight in deriving the final value.

- PR and CL are compensatory; better scores in PR can make up for poor scores in CL and vice versa. (CIRI's indexes are also compensatory.)

- Electoral Process, Political Pluralism and Participation, and Functioning of Government are each slightly more important to freedom than are Civil Liberties, Freedom of Expression and Belief, Associational and Organizational Rights, Rule of Law, and Personal Autonomy and Individual Rights. Each of the first three sub-categories is weighted roughly 0.17 (=0.5*1/3) in the overall index, while each of the second four sub-categories is weighted 0.125 (=0.5*1/4) in the overall index. (A similar point can be made about the weighting implicit in most other governance indexes as well.)

Another key issue in evaluating an operational definition has to do with the type of indicators included, e.g., indicators of outcomes versus inputs and of perceptions versus objective facts. As discussed above, what type of indicators are the best has been the subject of some debate in work on governance measurement, but overall the choice should be consistent with the concept as defined. If the index is based on a model of the inputs that create good governance, for instance, an operational definition that relies on inputs is appropriate. It would not be appropriate for a concept of governance focused on the delivery of goods to citizens. Similarly, if citizen satisfaction is key to the concept of good governance as defined, use of data on the perceptions of citizens makes sense. Use of data measuring the perceptions of external observers, however, would not fit and would require explanation. In practice also, given the difficulty of measuring some components of governance such as corruption, composite indicators may need to draw on indicator data that do not perfectly fit the underlying concept. This should at least be clearly explained.

In practice, improving the fit between a concept and its operational definition is often an iterative process, involving adjustments over time as new techniques are proposed to fix identified weaknesses that only become clear as the measure is used. Ongoing discussions over the best way to calculate the Human Development Index offer one such example (see Alkire 2010; United Nations Development Program 2011).

Question 3. How good (reliable, valid, and complete) are the data used to construct the measure?

Methods of index construction can sometimes mask the basic principle of 'garbage in, garbage out'. A measure can only be as good and reflective of the evidence as its base components, no matter how technically sophisticated. Assessing whether the operationalization of a concept is valid involves not only assessing the design features highlighted under Question 2, but also assessing implementation, in particular the quality (reliability, validity, completeness) of the data used in the measure. Validity, as above, refers

to whether the measure accurately captures what it purports to capture. Reliability refers to the consistency of the measure. To take a relatively simple example, a measure of pre-trial detention in a country would be *reliable* if multiple sources (from prison officials to observers from the International Committee of the Red Cross) all provided the same estimate. But, it would not be *valid* if all of these estimates actually pertained to pre-trial detention in the capital city only.

These sorts of data problems are more common than most users of governance measures might expect. Several studies suggest related inconsistencies between measures taken from various data sources, stemming both from validity and reliability challenges (see e.g., Razafindrakoto and Roubaud 2010; Shyaka, Murangwa, and Alibata 2009). One simple example is offered by recent literacy rates reported for Zimbabwe—which at 91.2 per cent (2007), up from 89.5 per cent (2004), have been the highest in Africa after the Seychelles' (Rotberg and Gisselquist 2009: 276-77; UNESCO Institute of Statistics 2009). These rates rely on national reporting and Zimbabwe adopts a relatively minimal definition based on 'the population aged 15 years and above who have completed at least grade three' (UNESCO 2000).

Whether or not we agree that the Zimbabwean definition of literacy is a valid one for use in Zimbabwe, the use of varying definitions across countries complicates the validity of cross-national comparison using these data. Similar challenges with the measurement of some indicators such as maternity mortality, for instance, have been addressed by using available measurements to estimate new figures, although again these estimates are only as good as the model from which they are derived (Say et al. 2007).

Survey designers pay a lot of attention to reliability, making sure, for instance, that sampling is done correctly so that estimates that should be nationally representative, are. Reliability is also a challenge for those collecting other types of data used in governance measures. Governance measurement projects that rely on coding based on detailed protocols, such as the Polity IV and CIRI projects, for instance, adopt various procedures to insure and monitor inter-coder reliability. Validity can also pose major challenges, even with expert coding. For instance, a rating of institutional quality for a country for a given year would be reliable if it was consistent with multiple sources, but it would not be valid if it was in fact based on assessment of institutional quality in the previous year (due to lack of new information or the assumption of stability). For countries and topics for which information is hard to come by, multiple experts may assess institutional quality for this country in the same way, giving reliable, but not valid, ratings.

A final issue is the completeness of the data used. Missing data means that the measure will provide a less evidence-based assessment than it would if information were complete. Given the extremely poor quality of statistics in many developing countries and particularly for sub-Saharan Africa, the amount of missing data in many compilations is notable.

Data on crime and rule of law issues pose especially difficult challenges stemming not only from the different definitions in use across countries but also from inherent challenges of data collection (Rotberg and Gisselquist 2007: 79-90; Stone 2012). For instance, one country may compile national data on murder but not on manslaughter, while some countries compile information on reported crimes and others on convictions only. The compilation of such statistics relies on the functioning of the criminal justice system; very weak rule of law institutions and cultural norms may deter reporting of crimes to official authorities.

Measuring issues like corruption poses particular challenges because of incentives to conceal this crime.

Question 4. Is the measure (including all of its sub-components) transparent and replicable?

In order to answer questions 2 and 3, the users of governance indexes need to be able to dig deep into data and methods. All too often this is not possible because this information is not made available by index producers. Far too many index projects—on governance and other topics—only make public final scores or rankings. This lack of transparency means that far too many index projects disregard a basic principle of good science, replicability. It is now the norm in top journals in economics and political science that the datasets used in quantitative analyses be made available when work is published, and published governance indexes should be held to the same standard.

Being fully transparent about data and methods poses clear risks and challenges for index producers. The first risk is that it makes finding mistakes much easier, giving index producers an incentive to obscure information to avoid such embarrassment. This is the principal reason that users of governance measures should be concerned if index producers avoid making their data and methods available.

A second risk for index authors has to do with concerns about the use of their data and methods. Once these are in the public domain, index authors have little control over how the information is used. Authors who have invested a lot of resources into compiling an index have natural incentives to use these data to the fullest in their own publications, for instance, before releasing them for use by others.

A third challenge is simply lack of time or care in presentation. Making detailed information available takes work. The producers of policy relevant indexes often do not enjoy the same amount of control and time in publishing their results as authors of articles in scholarly journals and it may take several editions of an index for its authors to catch up with all of the writing and reporting needed for full transparency.

A fourth challenge has to do with public interest and the audience for the index. Many indexes are ultimately produced as advocacy tools, a way to engage with policy makers and the wider public. To fulfill this function an index must be presented in a way that is engaging to this audience and detailed notes on sources and methodology may be of little interest. Websites offer a very simple solution here: Publish the simple index results for the wider audience and make the details available on the web.

A fifth and especially difficult issue is what to do about underlying data that is proprietary or should not be in the public domain. For instance, it is generally not considered ethical to release individual level survey data that can be used to identify individual respondents. The World Bank's Country Policy and Institutional Assessment (CPIA) relatedly cannot release its raw data because it would reveal assessment of country politics by Bank staff. Solutions in such cases are complicated, but the basic principle should be to make as much of the information available as possible and, even if it is not made available to the broader public, to make it available to other experts who might evaluate it.

Beyond good science methods, practitioners should also care about transparency and replicability because they carry political weight. Indeed, the literature on good governance

itself emphasizes the value of transparency and of strengthening processes of accountability and engagement, principles that are increasingly emphasized in calls for more open data (see Round 2012; United Nations 2013; Williams 2009).

Question 5. How sensitive and robust is the measure to different data and design choices?

As the discussion of questions 1 to 4 suggests, data and design choices in governance measurement are not an exact science. There are no hard and fast rules about which data sources are best, which indicators should be included, what methods of normalization should be used, and how data should be aggregated and weighted. The best indexes describe and justify each of these choices, including examination of the impact of various specific choices on the robustness of their results.

As the Dashboard of Sustainability can help to illustrate, index producers in the extreme can make indexes say almost anything they want by carefully selecting which indicators to include and adjusting the weighting of components. Overall scores obviously change dramatically if one component is weighted 100 per cent, but even smaller adjustments can have major significance.

Sensitivity and robustness assessment can be done in multiple ways and can be designed to assess multiple aspects of index design. One good example of how to assess the effect of data sources is provided by the CIRI project, which conducted analyses using different underlying sources to code its indicators (Cingranelli and Richards 1999; Richards, Gelleny, and Sacko 2001). Another example is offered by the work of the JRC, which has conducted sensitivity analyses for multiple indexes (including our 2008 IIAG) focused on the impact of different weighting choices (Paruolo et al. 2013; Saisana, Annoni, and Nardo 2009; Saisana, Saltelli, and Tarantola 2005).

The final values of an index will obviously change when different data and methods are used. The purpose of such analysis is to make clear the impact of specific data and design choices that may not otherwise be obvious, pointing either to the need for further justification of these choices, or for changes.

Question 6. Does the measure allow the analyst to address key questions of interest?

Addressing key questions of interest has to do both with whether the concept and operational definition are relevant to the question at hand, including the theoretical framework within which it is specified, and the ability of the measure to capture empirically what needs to be explained, including country coverage, time coverage, and the level of analysis at which measurement is taken (see Gingerich 2013). Whether the index can be used to predict trends is also important for many index users. Sometimes the answer to this question is obvious—the IIAG is not relevant to addressing questions about governance trends in Asia, nor can the WGI be used to answer questions about governance trends prior to 1996, the first year for which it was compiled.

Often, however, it is not. Even assessing seemingly straightforward issues such as comparability over time can be open to debate. The CPI, for instance, has explicitly noted that because it relies on different sources in each year, it should not be used to study

corruption over time.[7] The WGI similarly uses different sources in each year but its authors argue that it can be used to study variation over time (Kaufmann et al. 2007b). In any case, both the CPI and WGI are regularly used to analyse governance over time.

Question 7. Does the measure fully capture governance in all its complexity?

A common critique of national governance indicators is that they do not fully capture governance because they should include more components (e.g., Farrington 2010) or relatedly because they rely on national aggregates that do not capture sub-national variation (e.g., Gingerich 2013). The number of indicators included is also a common selling point of various indexes.

Conceptual development is important as highlighted in question 1, but focusing on this question is generally ill-advised because it tends to lead us to miss the forest for the trees. Descriptive complexity in itself should not be the objective of governance indexes for two reasons. First, as Gerring's criteria highlight, there is value in parsimony. A challenge for most measurement projects is to develop an operational definition that is just complex enough to capture key concepts or processes and to facilitate comparisons but not to fully describe every aspect. Identifying particular indicators left out of governance indexes thus is generally not much of a challenge. Simplicity is also only problematic when it involves excluding core dimensions of governance or means that the index does not then capture what it purports to.

Indeed, one school of thought suggests that indicators should be essentially stripped down to include only the components that are most important to the measure (see Langbein and Knack 2009). Statistical tools such as principal components analysis can be used to do this. This approach also should not be given too much weight in evaluating governance measures as it makes sense for some indexes but not all. Particularly for indexes in which part of the objective is conceptual development and for multidimensional topics like governance in which it is not clear how components fit together, strong arguments can also be made for more rather than less descriptive complexity.

A second reason that a focus on descriptive complexity can be misplaced has to do with how governance indexes are used. As discussed above, governance indexes are useful tools in comparing complex sets of information and exploring trends and relationships. They are essentially useful first cuts that can be used to develop and test broad hypotheses at the aggregate level. However, an analyst trying to understand the mechanisms and processes behind the sort of broad relationships described in indexes should be expected to look beyond such overall aggregates, for instance through careful qualitative case studies or quantitative analysis of more disaggregated data. Sole focus on developing an increasingly complex single aggregate measure often means that this sort of (equally important) disaggregated analysis is ignored.

[7] See, e.g., http://www.transparency.org/cpi2011/in_detail#myAnchor6. The 2012 CPI, however, employs a slightly different method for greater comparability.

Question 8. Does the measure behave as theory predicts?

Theory is important to rigorous governance measurement in the sense that measures should be developed with reference to and within theoretical frameworks so that they can speak to and be used to test theoretical propositions. However, several theory-based approaches to assessing the validity of indicators that have received significant attention in the literature may not be as damning to particular governance measures as this work suggests.

The first approach has to do with assessing the validity of a measure by assessing it against other measures of the same concept with the implication that valid measures of the same concept should be consistent (see Thomas 2009). One challenge of using this standard with regard to governance measurement goes back to question 1 above: Given the diversity of ways in which governance is defined, we need to be careful about comparing measures of 'governance' that are actually designed to capture different things. A more fundamental challenge given the current state of governance measurement is that, even if we can be sure that measures should be equivalent, this approach is only as good as existing measures. Even the most well-cited measures of governance have weaknesses, as explored in this paper. Given these weaknesses, it is not at all clear that a new indicator of governance that does not correlate well with existing governance measures should be rejected.

The second approach has to do with assessing the validity of a measure by exploring how it is related to other concepts in a manner consistent with theory. In particular, the authors of the WGI have paid significant attention to using the relationship of governance with growth to assess governance measures (Kaufmann and Kraay 2002; Kaufmann and Kraay 2008). The problem with this approach is that we simply do not know if the hypothesized relationship between governance and growth is correct (Khan 2009; Resnick and Birner 2006). Thus evaluating the goodness of constructs on whether they fit theoretical hypotheses limits our ability to actually use these constructs to test these hypotheses, as well as to build conceptual knowledge and theory more generally.

Question 9. How precise are index values and are confidence intervals specified?

Governance indexes produce overall scores that are imprecise and uncertain. Given this, the numerical differences between some scores or ranks may in fact not be very meaningful. One way to capture this sort of imprecision and uncertainty in survey research is to report confidence intervals or margins of error in addition to point estimates, an approach often preferred by statisticians and common, for instance, even in popular reporting on the results of pre-election polling. In general, the smaller the sample size, the larger the possibility of measurement error and thus the larger the margin of error or confidence interval. For this reason, the WGI and the CPI report 90 per cent 'confidence intervals' and Kaufmann and Kraay (2008: 18), among others, have argued that such reporting should be done for other governance indicators as well.

The underlying desire here for clear, accurate reporting of results is important, but in practice the reporting of confidence intervals in governance indexes can be misleading and does not necessarily reflect best practice. Given the nature of the data used in governance indexes, what are reported as confidence intervals are not necessarily the same thing as what we normally think of as confidence intervals. For instance, the WGI is not a survey but a compilation of data from multiple sources, some of which are surveys of public or elite opinion (e.g., Afrobarometer, Latin American Public Opinion Project, Gallup International

Millennium Survey) and some of which are based on coding by experts (e.g., International Country Risk Guide, EIU Country Risk Service, Freedom in the World, Index of Economic Freedom). The WGI's 'confidence intervals' are calculated as the point estimate plus and minus 1.64 times its estimated standard error, which is estimated based on the number of sources and the authors' estimation of the accuracy of these sources (Arndt and Oman 2006: 64). The estimation of the standard error is not based on the survey sample size, but rather on the number of assessments for each country and the degree to which their scores are consistent with each other.

This calculation further involves the key assumption that sources are independent of each other and thus that correlation of their scores is due to better measurement of the 'real' value of governance. However, even the WGI's authors recognize that this is a major assumption (Kaufmann et al. 1999). It is likely for instance that some of its sources in fact draw on each other in their assessments (Arndt and Oman 2006).

The uncertainty and imprecision surrounding scores are worth noting. Indexes that rank countries in particular may effectively obscure this imprecision because assigning ranks may suggest to users that small differences in scores are more meaningful than they are. Thus, one approach is simply not to rank. Another is to assign grouped scores or ranks, rather than reporting single figures for each country. Freedom House's *Freedom in the World,* for instance, assigns only one of three overall values (free, partially free, and not free). Both of these approaches, however, also mean than some information about overall results is obscured.

Question 10. Is the weighting 'correct'?

Weighting in governance indexes is generally derived with one of two principles in mind, the degree of confidence in each component's accuracy and the relative importance of each component to governance. The WGI, for instance, weights components based on the first principle, while the IIAG has generally relied on the second.

What both of these principles imply for weighting, however, is not clear in governance indexes. In practice, for instance, the first tends to rely on the problematic assumption that sources' errors are uncorrelated, which is touched on above (Arndt and Oman 2006). The second is problematic because most concepts of governance are simply not very specific about the relative importance of different governance dimensions. Rotberg and West (2004), for instance, proposes a hierarchy of political goods in which security is especially important, but this does not tell us precisely how much security should be weighed in quantitative terms relative to other dimensions of governance. Security may also be considered so important that the value of governance should be zero if it does not exist, regardless of values on all other dimensions.

In practice, many governance indexes appear to sidestep the weighting question by equally weighting core components, which avoids making a statement about their relative importance. As question 5 highlights, attention to the impact of weighting on overall measures is a useful way to consider the impact of these decisions and to justify design choices.

5 Conclusion

This paper has argued that the users and producers of governance measures should, first, pay more attention to the fundamentals of social science methodology, i.e., questions about concept formation, content validity, reliability, replicability, robustness, and the relevance of particular measures to underlying research questions. They should also, second, consider other issues commonly highlighted in the literature on governance measurement, i.e., questions about descriptive complexity, theoretical fit, the precision of estimates, and correct weighting, but these questions are less critical than the first set. Considered within this framework several governance indexes stand out as great measures, some better on some aspects than others.

In practice, the users and producers of governance measures also need to take more practical concerns into account. The first of these is cost. For instance, some data gaps might be filled if nationally-representative surveys are undertaken in every country, but this may not be financially feasible or even advisable given other pressing needs. Second, in deciding whether to produce a new governance index, analysts should consider whether it will add value. Given the dozens of existing governance indexes, it is worth considering whether the 'new' measure will really assess something new or do so in a new and better way? Third, another key issue, particularly for practitioners, is legitimacy. Will the governance assessment be considered legitimate by those being scored and/or by the users of the scoring? Overall, this seems more likely when the producers of such assessments are impartial and the methods and data are transparent. In some cases, who does the assessment (e.g., local or foreign experts) may also be worth considering. This is one reason that duplication of efforts may, in some instances, be justified.

Evaluating governance indexes and measures is closely tied to concept specification and measurement in the social sciences more generally. The best governance indexes, as judged by the framework proposed here, thus should essentially 'combine discrete facts … into a category, helping us to see the confusing universe in which we live a more patterned way' (Laitin 1995: 455-56).

References

Alkire, S. (2010). *Human Development: Definitions, Critiques, and Related Concepts.* New York: United Nations Development Programme. 2010/01.

Andrews, M. (2008). 'The Good Governance Agenda: Beyond Indicators without Theory'. *Oxford Development Studies*, 36(4): 379-407.

Andrews, M., R. Hay, and J. Myers (2010). 'Can Governance Indicators Make Sense? Towards a New Approach to Sector-Specific Measures of Governance'. *Oxford Development Studies*, 38(4): 391-410.

Apaza, C.R. (2009). 'Measuring Governance and Corruption through the Worldwide Governance Indicators: Critiques, Responses, and Ongoing Scholarly Discussion'. *Political Science & Politics*, 42(1): 139-43.

Arndt, C. (2008). 'The Politics of Governance Ratings'. *International Public Management Journal*, 11(3): 275-97.

Arndt, C. and C. Oman (2006). *Uses and Abuses of Governance Indicators*. Paris: OECD Development Centre.

Bandura, R. (2008). *A Survey of Composite Indices Measuring Country Performance: 2008 Update* New York: United Nations Development Programme, Office of Development Studies.

Bandura, R. (2011). 'Composite Indicators and Rankings: Inventory 2011'. Unpublished.

Bratton, M. (2011). 'The Democracy-Governance Connection'. In E. Lust and S. Ndegwa (eds), *Governing Africa's Changing Societies: Dynamics of Reform*. Boulder: Lynne Rienner Publishers.

Bratton, M. (2013). 'Governance: Disaggrating Concept and Measurement'. *APSA-Comparative Politics Newsletter*, 23(1): 12-13.

Cingranelli, D.L. and D.L. Richards (1999). 'Measuring the Level, Pattern, and Sequence of Government Respect for Physical Integrity Rights'. *International Studies Quarterly*, 43(2): 407-18.

Cingranelli, D.L. and D.L. Richards (2008). *The Cingranelli-Richards (CIRI) Human Rights Data Project Coding Manual Version 7.30.08*

Cingranelli, D.L. and D.L. Richards (2010). *Short Variable Descriptions for Indicators in the Cingranelli-Richards (CIRI) Human Rights Dataset. Document Version 11.22.10*

Collier, D. and J. Gerring (2009). *Concepts & Method in Social Science: The Tradition of Giovanni Sartori*. Oxford and New York: Routledge.

Consultative Group on Sustainable Development Indices (2003). *The Dashboard Manual*. International Institute for Sustainable Development (IISD) and the European Commission's Joint Research Centre (JRC).

Davis, K.E., A. Fisher, B. Kingsbury, and S.E. Merry (eds) (2012). *Governance by Indicators: Global Power through Quantification and Rankings, Law and Global Governance*. Oxford: Oxford University Press.

Delapalme, N. (2011). 'African Governance: The Importance of More and Better Data'. *Governance*, 24(1): 1-3.

Farrington, C. (2009). 'Putting Good Governance into Practice (I): The Ibrahim Index of African Governance'. *Progress in Development Studies*, 9(3): 249-55.

Farrington, C. (2010). 'Putting Good Governance into Practice (II): Revising and Extending the Ibrahim Index of African Governance'. *Progress in Development Studies*, 10(1): 81-86.

Farrington, C. (2011). 'Putting Good Governance into Practice (III): Instrumental and Intrinsic Aspects of Empowerment in Local Governmental Contexts'. *Progress in Development Studies*, 11(2): 151-61.

Frankel, J. (forthcoming). 'Mauritius: African Success Story'. In S. Edwards, S. Johnson and D. Weil (eds), *NBER Volume on African Economic Successes*. Chicago: University of Chicago Press.

Freedom House (2012). 'Methodology'. http://www.freedomhouse.org/report/freedom-world-2012/methodology (11 June 2013).

Gerring, J. (1999). 'What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences'. *Polity*, 31(3): 357-93.

Gingerich, D.W. (2013). 'Governance Indicators and the Level of Analysis Problem: Empirical Findings from South America'. *British Journal of Political Science,* 43(03): 505-40.

Gisselquist, R.M. (2012). 'Good Governance as a Concept, and Why this Matters for Development Policy'. Working Paper 2012/30. Helsinki: UNU-WIDER.

Grindle, M.S. (2004). 'Good Enough Governance: Poverty Reduction and Reform in Developing Countries'. *Governance*, 17(4): 525-48.

Grindle, M.S. (2007). 'Good Enough Governance Revisited'. *Development Policy Review,* 25(5): 533-74.

Grindle, M.S. (2010). 'Good Governance: The Inflation of an Idea'. Faculty Research Working Paper Series, RWP 10-023, Harvard Kennedy School.

Hallerberg, M. and M. Kayser (2013). 'Measuring Governance'. *APSA-Comparative Politics Newsletter,* 23(1): 1-2.

Harttgen, K. and S. Klasen (2012). 'A Household-Based Human Development Index'. *World Development*, 40(5): 878-99.

Hertie School of Governance (2013). *The Governance Report 2013*. Oxford: Oxford University Press.

Holmberg, S. and B. Rothstein (eds) (2012). *Good Government: The Relevance of Political Science*. Cheltenham, UK, and Northampton, Massachusetts: Edward Elgar.

Høyland, B., K. Moene, and F. Willumsen (2012). 'The Tyranny of International Index Rankings'. *Journal of Development Economics*, 97(1): 1-14.

Hyden, G., J. Court, and K. Mease (2003). *Making Sense of Governance: The Need for Involving Local Stakeholders*. London: Overseas Development Institute.

Hyden, G., J. Court, and K. Mease (2004). *Making Sense of Governance: Empirical Evidence from Sixteen Developing Countries*. Boulder, Colorado, and London: Lynne Rienner.

Jesinghaus, J. (2012). 'Welcome to the Millennium Development Goals Dashboard!', 12 September. http://esl.jrc.ec.europa.eu/ (10 June 2013).

Kaufmann, D. and A. Kraay (2002). 'Governance without Growth'. *Economia*, 3(1): 169-229.

Kaufmann, D., A. Kraay, and M. Mastruzzi (2007a). 'Growth and Governance: A Reply'. *Journal of Politics,* 69(2): 555-62.

Kaufmann, D., A. Kraay, and M. Mastruzzi (2007b). *The Worldwide Governance Indicators Project: Answering the Critics*. Washington, DC: World Bank. 4149.

Kaufmann, D., A. Kraay, and P. Zoido-Lobatón (1999). *Aggregating Governance Indicators*. Washington, DC: World Bank.

Kaufmann, D. and A. Kraay (2008). 'Governance Indicators: Where Are We, Where Should We Be Going?'. *The World Bank Research Observer,* 23(1): 1-30.

Khan, M.H. (2009). 'Governance, Growth and Poverty Reduction'. Department of Economic and Social Affairs, United Nations.

King, G., R.O. Keohane, and S. Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Knoll, M. and P. Zloczysti (2011). *The Good Governance Indicators of the Millennium Challenge Account: How Many Dimensions Are Really Being Measured?* Berlin: Deutsches Institut für Wirtschaftsforschung.

Kurtz, M.J. and A. Schrank (2007a). 'Growth and Governance: A Defense'. *Journal of Politics*, 69(2): 563-69.

Kurtz, M.J. and A. Schrank (2007b). 'Growth and Governance: Models, Measures, and Mechanisms'. *Journal of Politics*, 69(2): 538-54.

Laitin, D. (1995). 'Disciplining Political Science'. *American Political Science Review*, 89(2): 454-56.

Langbein, L. and S. Knack (2009). 'The Worldwide Governance Indicators: Six, One, or None?'. *The Journal of Development Studies*, 46(2): 350-70.

Levi, M. (2006). 'Why We Need A New Theory of Government'. *Perspectives on Politics*, 4(1): 5-19.

McFerson, H.M. (2009). 'Measuring African Governance: by Attributes or by Results?'. *Journal of Developing Societies*, 25(2): 253-74.

Millennium Challenge Corporation (2012). *Guide to the MCC Indicators and the Selection Process for Fiscal Year 2013*. Washington, DC: Millennium Challenge Corporation.

Mitra, S. (2013). 'Towards a Multidimensional Measure of Governance'. *Social Indicators Research*, 112(2): 477-96.

Mo Ibrahim Foundation (2012). *The 2012 Ibrahim Index of African Governance (IIAG) - Methodology*. London: Mo Ibrahim Foundation.

Noorbakhsh, F. (1998). 'The Human Development Index: Some Technical Issues and Alternative Indices. *Journal of International Development*, 10(5): 589-605.

OECD and Joint Research Centre (JRC) of the European Commission (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide.

Oman, C. and C. Arndt (2010). *Measuring Governance*. Paris: OECD.

Paruolo, P., M. Saisana, and A. Saltelli (2013). 'Ratings and Rankings: Voodoo or Science?'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(3): 609-34.

Putnam, R. (1993). *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Ravallion, M. (2012). 'Mashup Indices of Development'. *The World Bank Research Observer*, 27(1): 1-32.

Razafindrakoto, M. and F. Roubaud (2010). 'Are International Databases on Corruption Reliable? A Comparison of Expert Opinion Surveys and Household Surveys in Sub-Saharan Africa'. *World Development*, 38(8): 1057-69.

Resnick, D. and R. Birner (2006). 'Does Good Governance Contribute to Pro-Poor Growth? A Review of the Evidence from Cross-Country Studies'. Development Strategy and Governance Division, International Food Policy Research Institute.

Richards, D.L., R. Gelleny, and D. Sacko (2001). 'Money With A Mean Streak? Foreign Economic Penetration and Government Respect for Human Rights in Developing Countries'. *International Studies Quarterly*, 45(2): 219-39.

Rotberg, R.I. (2007). 'On Improving Nation-State Governance'. *Daedalus*, 136(1): 152-55.

Rotberg, R.I. (2009). 'Governance and Leadership in Africa: Measures, Methods, and Results'. *Journal of International Affairs*, 62(2): 113-26.

Rotberg, R.I. (2013). 'On Governance and Global Governance: What and How to Measure'. Presented at the Measuring Governance Effectiveness: National and International Dimensions, a conference sponsored by the Centre for International Governance Innovation and the North-South Institute Waterloo, Canada.

Rotberg, R.I. (2004). 'Strengthening African Governance: Ranking Countries Would Help'. *The Washington Quarterly*, 28(1): 71-81.

Rotberg, R.I., A. Bhushan, and R.M. Gisselquist (2013). 'The Indexes of Governance'. *Measuring Governance Effectiveness: National and International Dimensions, a conference sponsored by the Centre for International Governance Innovation and the North-South Institute*.

Rotberg, R.I. and R.M. Gisselquist (2007). *Strengthening African Governance – Ibrahim Index of African Governance: Results and Rankings 2007*. Cambridge, MA: Mo Ibrahim Foundation; Kennedy School of Government, Harvard University; and World Peace Foundation.

Rotberg, R.I. and R.M. Gisselquist (2008). *Strengthening African Governance – Ibrahim Index of African Governance: Results and Rankings 2008*. Cambridge, MA: Mo Ibrahim Foundation; Kennedy School of Government, Harvard University; and World Peace Foundation.

Rotberg, R.I. and R.M. Gisselquist (2009). *Strengthening African Governance – Index of African Governance: Results and Rankings 2009*. Cambridge, MA: Kennedy School of Government, Harvard University, and World Peace Foundation.

Rotberg, R.I. and D.L. West (2004). *The Good Governance Problem: Doing Something About It*. Cambridge, MA: World Peace Foundation and Harvard Kennedy School. 39.

Rothstein, B. and J. Teorell (2012). 'Defining and Measuring Quality of Government'. In S. Holmberg and B. Rothstein (eds). *Good Government: The Relevance of Political Science*. Cheltenham, UK & Northhampton, Massachusetts, USA: Edward Elgar.

Round, J.I. (2012). *Aid and Investment in Statistics for Africa*. Working Paper No. 2012/93. Helsinki: UNU-WIDER.

Saisana, M., A. Saltelli, and S. Tarantola (2005). 'Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators'. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2): 307-23.

Saisana, M., P. Annoni, and M. Nardo (2009). *A Robust Model to Measure Governance in African Countries*. European Commission Joint Research Centre, Institute for the Protection and Security of the Citizen.

Saltelli, A. (2007). 'Composite Indicators between Analysis and Advocacy'. *Social Indicators Research*, 81(1): 65-77.

Sampson, S. (2010). 'Diagnostics: Indicators and Transparency in the Anti-corruption Industry'. In S. Jansen, E. Schröter and N. Stehr (eds). Wiesbaden *Transparenz: Multidiszplinäre Durchsichten durch Phönomene und theorien des Undurchsichtigen*: VS Verlag für Sozialwissenschaften.

Sartori, G. (1984). *Social Science Concepts: A Systematic Analysis*. Beverly Hills, California: Sage Publications.

Say, L., M. Inoue, S. Mills, and E. Suzuki (2007). *Maternal Mortality in 2005: Estimates Developed by WHO, UNICEF, UNFPA, and The World Bank*. Geneva: WHO.

Shyaka, A., Y. Murangwa, and M. Alibata (2009). 'The Index of African Governance: Rwanda Response'. In R.I. Rotberg and R.M. Gisselquist (eds), *Strengthening African Governance – Index of African Governance: Results and Rankings 2009*. Cambridge, MA: Kennedy School of Government, Harvard University, and World Peace Foundation.

Stanton, E.A. (2007). *The Human Development Index: A History*. Amherst, Massachusetts: Political Economy Research Institute (PERI), University of Massachusetts-Amherst.

Stiglitz, J.E., A. Sen, and J.-P. Fitoussi (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress.*

Stone, C. (2012). 'Problems of Power in the Design of Indicators of Safety and Justice'. In K.E. Davis, A. Fisher, B. Kingsbury and S.E. Merry (eds), *Governance by Indicators: Global Power through Quantification and Rankings*. Oxford: Institute for International Law and Justice, New York University School of Law, and Oxford University Press.

Sudders, M. and J. Nahem (2007). *Governance Indicators: A Users' Guide*. Oslo: Oslo Governance Centre, United Nations Development Programme.

Thomas, M.A. (2007). 'The Governance Bank'. *International Affairs*, 83(4): 729-45.

Thomas, M.A. (2009). 'What Do the Worldwide Governance Indicators Measure?'. *European Journal of Development Research*, 22(1): 31-54.

UNESCO (2000). 'The EFA 2000 Assessment: Country Reports'. *Education for All the 2000 Assessment: Republic of Zimbabwe*, New York: UNESCO. section 6.3.2.

UNESCO Institute of Statistics (2009). 'National Literacy Rates for Youths (15-24) and Adults (15+)'. http://stats.uis.unesco.org (3 June 2009).

United Nations (2013). *A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development. The Report of the High-Level Panel of Eminent Persons on the Post-2015 Development Agenda*. New York: United Nations.

United Nations Development Program (2011). *Human Development Report 2011. Sustainability and Equity: A Better Future for All*. New York: UNDP.

Weiss, T.G. (2000). 'Governance, Good Governance and Global Governance: Conceptual and Actual Challenges'. *Third World Quarterly*, 21(5): 795-814.

Williams, A. (2009). 'On the Release of Information by Governments: Causes and Consequences'. *Journal of Development Economics*, 89: 124-38.

Williams, G. (2011). *What Makes a Good Governance Indicator?* Brighton: The Policy Practice. 6.

World Bank (2007). *A Decade of Measuring the Quality of Governance. Governance Matters 2007: Worldwide Governance Indicators, 1996–2006*. Washington, DC: World Bank.