Measuring good governance: A comparison of three datasets

- Working paper -

4/8/2008

Ingo Rohlfing, PhD

Research Associate

Department of Management, Economics and Social Sciences

University of Cologne

Herbert-Lewin-Str. 2

D-50931 Köln

rohlfing@wiso.uni-koeln.de

For comments I am grateful to Hans-Jürgen Andreß and Henning Lohmann. All errors are my own. Valuable research assistance was provided by Caroline Wehner. The paper is part of the project "*The quality of macrodata in the social sciences*" at the University of Cologne.

1. Introduction

Good governance is a paradigm gaining more and more popularity in the scientific and the political sphere. In the political realm, good governance is seen as the key for a positive performance of countries and the path for sustainable political development (Bouckaert and Van de Walle, 2003; Knack, Kugler and Manning, 2003). In the academia, an increasing number of empirical studies in the state of good governance all over the world (e.g., Kaufmann, Kraay and Mastruzzi, 2004), its determinants (e.g., Al-Marhubi, 2004), and effects (e.g., Kurtz and Schrank, 2007).

Any proper descriptive and causal analysis of governance requires its accurate measurement in the first place. In this paper, I aim to promote the analysis of governance by critically examining three widely perceived datasets measuring governance on the country-level.[1] The datasets are the *World Governance Indicators* (WGI), collected by the World Bank;[2] the *World Governance Survey* (WGS), which is hosted by the Overseas Development Institute (ODI) and was initiated with the assistance of the United Nations;[3] and the *Quality of Government* (QoG) data, generated by the Quality of Government Institute at the University of Göteborg.[4] Besides transparency, which is key to any informed assessment of a dataset (King, 1995; Winship, 2007), I will apply a comprehensive framework of analysis including five categories: *conceptualization*, *operationalization*, *measurement*, *aggregation*, and *representativeness* (cf., Rohlfing, 2008). For each of dataset, I will assess how it performs on these dimensions and detail specific strengths and shortcomings. The analysis of the three datasets shows that each marks a laudable effort toward the better measurement and understanding governance. Yet, each dataset displays some deficiencies on the five dimensions that need to be taken into account when using it.

In section two, I will introduce the criteria on the basis of which the datasets will be evaluated so as to lay the basis for the remainder of the paper. Section three covers a discussion of the conceptualization of governance underlying each dataset. In section four, I will discuss how governance is operationalized. The following section deals with the data collection and processing of sources and the raw data derived from them. Section six focuses

[1] See Knack et al. (2003) for a more policy-oriented discussion of specific indicators.
[2] http://info.worldbank.org/governance/wgi2007/home.htm. See Thomson (2007) for a study using parts of the WGI.
[3] http://www.odi.org.uk/wga_governance/Index.html. See Jacobs (2005) and Petrova (2007) for studies drawing on the WGS.
[4] http://www.qog.pol.gu.se/. See Treisman (2007) for an analysis using elements of the QoG and Ledet (2006) for an appraisal.

on how these scores are aggregated. The seventh section asks for the datasets' representativeness and suitability for generalization. The eighth section concludes.

2. What is high-quality data? A comprehensive framework

The evaluation of the datasets calls for a comprehensive framework specifying standards of high-quality data. The framework I use in this paper comprises five dimensions each of which contains a set of subcategories: conceptualization, operationalization, data collection and processing, aggregation, and representativeness. These dimensions represent the tasks a data producer performs in the course of compiling a dataset. At the same time, the guideline can be used by a data user for an evaluation of existing datasets. I cannot go in the details of all criteria in this paper inasmuch as they were subject to detailed discussions in varies bodies of research in the past (cf., Rohlfing, 2008). For this reason, I focus on the essentials of each criterion and the related subdimensions and refer the reader to the references for in-depth treatments of the issue in question.

On a basic level, conceptualization requires the specification of what one aims to analze, for example in the form of "governance is understood as…". The definition of a concept is essential because it lays the ground for the operationalization and measurement stage (Sartori, 1970). A sound conceptualization needs to take three criteria into account. The first standard is *concept specificity*, capturing the degree to which the definition of the subject matter is unambiguous. The quest for clarity extends to the constituent elements of a definition, also called attributes – good governance is when *A*, *B*, and *C* is present – and the way the elements are linked to each other. The most basic and common linkage of attributes is through "and" and "or" (Goertz, 2006, chap. 4) – governance is when A *and* B *and* C is given, or, alternatively, when A *or* B *or* C is present. The former understanding of governance is based on what I call an AND-concept, since the attributes are linked to each through "and".[5] Correspondingly, the latter example represents an OR-concept.[6] Being explicit about the type of concept is important for obvious substantive reasons. Moreover, the concept type is quintessential when it comes to the aggregation of indicator scores, which will be discussed below in more detail (cf., Goertz, 2006, chap. 4).

The second standard of conceptualization is *concept homogeneity*, meaning that the same definition should be applied to all units of analysis at any point in time (Przeworski and

[5] The "AND" is capitalized in order to denote that it refers to the logical AND, which becomes important in the aggregation stage (Goertz, 2006, chap. 4).

[6] What I call AND-concept is labeled *necessary/sufficient* concept by Goertz (2006), who also underscores the importance of distinguishing between AND- and OR-concepts (see also Collier and Mahon, 1993). OR-concepts are called *family resemblance* concepts. I prefer my labels because they are more intuitive and capture the essence of a concept in terms of how its elements are related to each other.

Teune, 1966). An example for a lack of concept homogeneity and, thus, comparability, is unemployment data before and after 1982. In this year, the International Labor Organization proposed a new definition of unemployment that was adopted by many countries. The change in the definition undermines the comparability of unemployment rates prior and after 1982 since it has direct implications of how governance is measured (Brandolini, Cipollone and Viviano, 2006). Finally, the *concept fit* of the definition needs to be maximal. Concept fit is given when irrelevant elements are excluded from the conceptualization and all relevant elements are included. Put otherwise, it can be understood as the degree to which a specific definition captures the essence of the subject of interest (Adcock and Collier, 2001). For example, there is scholarly consensus that social justice should not be part of a definition of democracy (Munck and Verkuilen, 2002). Depending on the subject matter, there may not be full agreement about what relevant and irrelevant attributes are. In many fields, however, there is consensus about (ir)relevance of some attributes, so an assessment of the concept fit is possible within limits. Regarding those concepts for which agreement is low, which is the case for governance (Kaufmann, Kraay and Mastruzzi, 2007b, 23), it becomes more important to determine the *internal consistency* of a definition.

The operationalization of a concept's attributes should meet two criteria. First, *attribute specificity* is mandatory, denoting that one must make explicit what indicator or indicators are used. In case one attribute is measured through multiple indicators, it is additionally necessary to specify the attribute type. Just as one can distinguish between AND- and OR-concepts, it is possible to differentiate AND- and OR-attributes. This means that one can measure an attribute through the indicators A *and* B *and* C or A *or* B *or* C (Goertz, 2006, chap. 4). Again, this aspect of attribute specificity will take center stage in the discussion of aggregation. The second criterion to consider on the operationalization dimension is an indicator's *content validity*, that is, the extent to which they are measures of what they are supposed to measure (Carmines and Zeller, 1979).

The third important aspect of data quality concerns the collection and processing of raw data. With respect to data collection, one has to focus on the presence and extent of the *source coverage problem*, which involves three issues (cf., Azar, Cohen, Jukam and McCormick, 1973; Doran, Pendley and Antunes, 1973; Hazlewood and West, 1974; Jackman and Boyd, 1979; Lustick, 1996; Sommer and Scarritt, 1999). The first aspect is *source sampling*.[7] Source sampling is at hand when the sources that are used for the generation of data and the scoring of an indicator only captures a subset of all relevant information. In this

[7] This type of sampling is labeled as "source" sampling so as to distinguish it from sampling units of analysis, which are countries in the case of governance research.

context, “source” has a rather broad meaning and includes surveys as well as primary and secondary sources. A classic example of source sampling is survey research, which examines a subset of people belonging to the population one is interested in. Assume you want to know the overall number of unemployed people in a certain country. Evidently, counting the number of unemployed in the sample will underestimate the true extent of unemployment even if the sample is random. This is not a serious problem in survey research, however, since it is possible to estimate the number of unemployed people in the population within certain margins of error. This strategy is not feasible, on the other hand, when using newspaper reports as sources for the generation of democracy scores (cf., Bollen, 1993; Smith, 1969). The reason for this problem is that one does not know the overall number of violations of democratic principles in a country. This is different in survey research where the size of the population is known. In addition, there is uncertainty about the newspaper’s coverage of the democratic development in the country, i.e., how many violations are reported. In sum, the uncertainty about the utility of the sample of collected reports is rather large.

The second aspect of the source coverage problem is closely related to the first one and asks for the presence of a *source sampling bias*. The presence of such a bias is important to assess when the type of information matters in addition to its magnitude. For example, one can assume that international newspapers only report serious violations of democratic principles in semi- or non-democratic countries. When one uses this newspaper for the measurement of a country’s democratic quality, one confronts the problem of source sampling and, moreover, biased sampling because (the probably more numerous) infringements of democratic rules will not be covered of the newspaper (Bollen, 1993). Similarly, measures of central tendency like the mean or the variance are probably not generalizable to the population when derived from a biased sample.

The third and related subcategory asks for the degree of *intersource reliability* and comes into play when different sources are used for the scoring of one unit over time and/or several units at the same point in time. As research on international conflict has shown, different sources generally have divergent coverage of the same event or processes. For instance, a local Middle East newspaper includes more reports on cross-border incidents than any Western national newspaper (Azar et al., 1973). Thus, it is recommendable to check whether the data producer drew on different sources for the scoring of different units and, if so, what the degree of intersource reliability is. This can be done through simple correlations of the information derived from different sources or a factor analysis comparing the factor loadings of the sources (Hazlewood and West, 1974; Jackman and Boyd, 1979).

The fourth dimension on which data quality should be evaluated is aggregation. The standard the dataset should meet is the *attribute type congruence* and *concept type congruence* of aggregation. As I indicated above, the concept type and attribute type form the background for the aggregation of indicator scores. With respect to the scoring of attributes, the minimum-scoring indicator determines the score under an AND-type, while the maximum indicator matters for OR-attributes (Goertz, 2006, chap. 4). To give an example, assume that you measure the liberalism of a country as the degree of liberalism in several areas, one of which is the extent of trade protectionism. Protectionism is in turn measured through tariff levels and non-tariff barriers to trade (NTBs). On a hypothetical protectionism scale ranging from 1 (prohibitive protectionism) to 10 (free trade), a country scores 3 on tariffs and 6 on NTBs. When protectionism is an AND-attribute, that is, you measure protectionism through tariff levels *and* the pervasiveness of NTBs, the attribute receives a score of 3 because the lower score on the tariff-indicator is the only one that matters. Correspondingly, an OR-attribute would be scored with 6. A similar logic applies to the aggregation of the scores on multiple attributes to a single case score. The minimum-scoring attribute drives the score of the case under an AND-concept. On the other hand, the score on the largest indicator is assigned to a case when one is dealing with an OR-concept.

The final dimension moves beyond the level of individual cases and asks for the representativeness of the generated dataset. Representativeness concerns two issues. First, one needs to consider whether the units included in the dataset is a *sample*. If this is the case, one should further assess if the sample is *random* or *biased*. The second issue subsumed under the dimension of representativeness deals with the extent and nature of *missing data* for the included units (Hug, 2003). The literature on missing data distinguishes three types of missingness (Allison, 2001, 2-5; King, Honaker, Joseph and Scheve, 2001, 2-3): *missing completely at random (MCAR)*, *missing at random (MAR)*, and *missing not at random* (MNAR), which is also called *non-ignorable* (*NI*). Data is MCAR when the missingness is random. For example, the probability that one does not have information on birth rates is similar for developed and less-developed countries (and any other variable one can think of). MAR is in place if the missingness is systematically related to a third variable that is known.[8] Data on birth rates is probably better described as MAR because data is missing more often for less-developed than for developed states (Herrera and Kapur, 2007). In this instance, the Gross Domestic Product (GDP) is a good predictor for missing information on birth rates,

[8] The label MAR is somewhat unfortunate for the missingness is non-random. However, it is established in the literature, so I use it here.

which makes the data MAR.[9] Finally, data is MNAR when the missingness is systematic, but when the missingness cannot be predicted with a third variable. This would be the case when birth rate is more likely to be missing for less-developed countries and when you do not have GDP data or any other data with which the missingness can be predicted.[10] Distinguishing between these three types of missing data is important because the viability of different remedies hinges on the nature of the missingness. For example, the deletion of cases for which information is lacking (list-wise deletion) is viable when data is MCAR, but is rather likely to introduce a bias in the presence of MAR and MNAR. Since the literature on missingness is rather involved and extensive, I leave it with that and refer the reader to the references for details (cf., Allison, 2001; Hug, 2003; King et al., 2001; König, Finke and Daimer, 2005).

In the remainder of this paper, I will apply these criteria to the World Governance Indicators, the World Governance Survey, and the Quality of Government data. Each section will be concluded with a table comparing the three datasets on the dimensions that I just detailed.

3. Conceptualization

3.1. World Governance Indicators

The WGI fully meets the criterion of concept homogeneity because the same understanding of governance is applied to all units at any point in time. The specificity of the WGS's conceptualization of governance is more problematic because the reader is offered two somewhat distinct definitions. On the one hand, one finds a definition according to which governance "*includes the process by which governments are selected and replaced, the capacity of the government to formulate and implement sound policies, and the respect of citizens and the state for the institutions that govern economic and social interactions among them*" (Kaufmann et al., 2004, 254). On the other hand, governance is conceptualized through six so called clusters that are set in relation to one element of this definition (Kaufmann et al., 2004, 254-255). Table 1 details how the attributes and clusters are linked to each other.

[9] The predictor variable does not need to be causally linked to the variable for which data is missing (King et al., 2001, 3).
[10] Since an informed researcher knows what countries are less-developed even without GDP data, this example additionally presumes ignorance about the world.

Table 1: Attributes and clusters underlying the WGI

Attributes	Clusters
Selection of government	Voice and accountability
Replacement of government	Political stability and absence of violence
Formulation of policies	Government effectiveness
Implementation of policies	Regulatory quality
Citizen's respect for institutions	Rule of law
State's respect for institutions	Control of corruption

The table shows that the meaning of the attributes and clusters forming a pair is not fully the same. For example, "political stability and the absence of violence" has a broader meaning than "the replacement of government". A country can experience political instability in the form of frequent parliamentary elections, but may have perfectly stable governments that are orderly replaced in regular intervals. Likewise, stable governments may go along with violence in the streets or the country side. The other pairs of attributes and clusters display similar ambiguities. For instance, the disrespect for institutions by the state, that is, the politician's and official's disrespect, does not necessarily go along with corruption. Politicians and officials can also ignore institutions without being corrupt. Empirically, corruption may be the major consequence of disrespect. Conceptually, however, there is no perfect coincidence. All in all, the concept specificity of the WGI is low because it is unclear what the underlying conceptualization of governance is.

An inspection of the selected indicators shows that they have higher content validity if they are supposed to be measures of the clusters. Moreover, the governance scores are always reported for the clusters and are not set in relation to the components of the definition (e.g., Kaufmann et al., 2004; Kaufmann, Kraay and Mastruzzi, 2007a). Thus, it seems safe to conclude that the clusters embody the WGI's actual understanding of governance. The concept type is not explicitly specified in the documentation. However, it clearly reads as if all clusters are considered essential elements of governance, which implies that the definition is of the AND-type.

With regard to concept fit, the WGI's conceptualization of governance is free of internal inconsistencies. Moreover, no undoubtedly superfluous attributes are included or relevant attributes excluded, so concept fit can be said to be high. However, two other things

need to be mentioned when talking about the substance of the definition. First, it contains procedural and outcome attributes. Voice and accountability, government effectiveness, rule of law, and control of corruption capture whether governance in a country complies with certain procedural standards. On the other hand, political stability and absence of violence and regulatory quality – understood as the market conformity of policies – are attributes focusing on the outcome of processes. Mixing procedural and substantive elements is not a problem per se. But it should be kept in mind when using the WGI because the outcome perspective on governance is disputed by those who prefer a purely procedural definition, which for example underlies the World Governance Survey.

Second, the WGI conceptualization bears much similarity to common definitions of democracy (cf., Munck and Verkuilen, 2002). This does not mean that the definitions are fully equivalent, which cannot be the case because there is widespread agreement that democracy should be exclusively conceptualized through procedural attributes (Munck and Verkuilen, 2002). However, rule of law, voice and accountability, the absence of corruption, and stability and the absence of violence are also integral characteristics of democracies. Again, this is not meant to be a problem, but it should be kept in mind that the WGI understanding of governance is close to common conceptualizations of democracy.

2.2. World Governance Survey

The concept specificity of the WGS is not optimal because there is some ambiguity about what the real definition of governance is. At one place in the accompanying material, governance is conceptualized as "*the formation and stewardship of the formal and informal rules that regulate the public realm, the arena in which state as well as economic and societal actors interact to make decisions*" (Hyden and Court, 2002, 13). On the other hand, the WGS's documentation contains a list of six basic principles that are seen as constitutive for good governance and that do not match closely with the first definition: *participation*, *decency*, *fairness*, *accountability*, *transparency*, and *efficiency* (Hyden and Court, 2002, 25). The picture becomes even more complicated because one also finds six *functional dimensions* and *institutional arenas* are introduced alongside with a specification of the role formal and informal rules perform in each arena (table 2) (Hyden and Court, 2002, 16-22). The functional dimensions do neither neatly correspond to the six principles of governance, nor do they conform to the definition presented above.[11] A closer analysis of the WGS's documentation reveals that the functional dimensions seem to be central because they are operationalized and

[11] It is possible to link participation to the socializing dimension and fairness to the adjudicatory dimension. Beyond that, it becomes difficult to set the dimensions in relation to attributes and basic principles.

taken for the measurement of governance (Hyden and Court, 2002, 31-34). Although the concept type is not clearly specified, it seems to argue that the six dimensions jointly represent governance, meaning that they form a concept of the AND-type.

Table 2: Functional dimensions and institutional areas of governance

Functional dimension	Institutional arena	Purpose of rules
Socializing	Civil society	Shape the way citizens raise and become aware of public issues
Aggregating	Political society	Shape the way issues are combined into policy by political institutions
Executive	Government	Shape the way policies are made by government institutions
Managerial	Bureaucracy	Shape the way policies are administered and implemented by public servants
Regulatory	Economic society	Shape the way state and market interact to promote development
Adjudicatory	Judicial system	Shape the setting for resolution of disputes and conflicts

Concept homogeneity is given because the same definition of governance is consistently applied. The WGS also meets the criterion of concept fit. There is no striking misconceptualization in the form of an included irrelevant or missing relevant attribute. Moreover, there are no internal inconsistencies, so concept fit is high. With respect to the substantive understanding of governance, table 2 shows that the WGS puts exclusive emphasis on rules and the input side of the political process. The exclusion of substantive elements like liberal economic policies is explained with the goal of developing a definition that does not draw on the ideal of market-oriented Western democracies (Hyden and Court, 2002, 13-16). In a similar vein, it is argued that the WGS's understanding of governance is not similar to the democratic quality of a country or the state of its democratic transition. A democratic deficit in a country should not automatically produce a low governance score because good governance is possible even if the state of democracy is poor (Hyden and Court, 2002, 27-28). It is not disputable per se to delineate the notion of governance from democratic transition or democratic quality. However, this argument is not entirely convincing because the six functional dimensions are also integral elements of a democracy (Munck and Verkuilen, 2002).

2.3. Quality of Government

According to the QoG, governance is in place "*when implementing laws and policies, government officials shall not take anything into consideration about the citizen/case that is not* beforehand *stipulated in the policy or the law*" [emphasis added] (Rothstein and Teorell, 2005). Thus, it is only the *impartiality* of policy implementation that is at the heart of good governance, meaning that the QoG conceptualization just comprises a single attribute: the impartiality of policy implementation. Homogeneity of the QoG conceptualization is given because this definition applies to all countries at all times. However, the QoG's internal consistency and concept fit are not optimal because a single-attribute definition is too minimalistic. In fact, the documentation of the QoG dataset spends much space on explaining why impartiality should be the sole attribute of good governance and why the latter does not require a high democratic quality of a country (Rothstein and Teorell, 2005). The argument goes that citizens and investors appreciate predictability. Citizens, for instance, are frustrated when they have to pay bribes in order to receive medical treatment. Investors are discouraged from investing in a country when the national officials are corrupt. For these reasons, predictability can be framed as an element of good governance (cf., Rothstein and Stolle, 2007).

Thus, making impartiality an attribute of a governance definition cannot be criticized as such. However, there are numerous studies also solid findings showing that participation and procedures matter, that is, the way laws come about (Scharpf, 1999). Assume that a dictator only produces laws that benefit the wealthy people, for example, by relieving them from paying any taxes, and that these policies are implemented impartially. Formally seen, this would be a perfect case of impartiality in the QoG sense. However, the lack of participation on the input side produces a policy that substantively discriminates against less wealthy people, discouraging them from working and paying taxes and thus undermining the country's development. The formulation of policies by dictators would pass under the narrow QoG conceptualization of governance, which, in my eyes, contradicts even a minimum understanding of good governance. Table 3 summarizes the discussion of conceptualization.

Table 3: Comparison of the conceptualization

Criterion	WGI	WGS	QoG
Number of attributes	Six	Six	One
Concept type	AND	AND	Not applicable
Concept homogeneity	Yes	Yes	Yes
Concept specificity	Low	Low	High
Concept fit/ Internal consistency	High	High	Low Definition too minimalist

4. Operationalization

4.1. World Governance Indicators

The operationalization of the six clusters of governance is detailed in table 4.

Table 4: Attributes, clusters, and indicators of the WGI

Clusters	Indicators
Voice and accountability	Political process, Civil liberties, Political rights, Independence of the media
Political stability and absence of violence	Likelihood that the government in power will be destabilized or overthrown by unconstitutional or violent means
Government effectiveness	Quality of public service provision, Quality of the bureaucracy, Competence and independence of the civil service, Credibility of government's commitment to policies
Regulatory quality	Incidence of market-unfriendly policies, Burdens imposed by excessive regulation
Rule of law	Incidence of crime effectiveness and predictability of the judiciary, Enforceability of contracts,
Control of corruption	Extent of corruption

The indicators that are subsumed under the first cluster show that the WGI does not aim to measure the selection of the government in a narrow sense, but the democratic quality of a country more generally. The second, third, and fourth indicator indeed are measures of

democracy (cf., Munck and Verkuilen, 2002). The first indicator, political process, is somewhat unspecific because it is not clear what this indicator measures in particular. Is it important that the process is rule-driven, that specific types of actors are involved in the process (parliament, parties, direct democracy, etc.? The operationalization of the second cluster is problematic too. Political stability and the absence of violence is a rather broad attribute, whereas the indicator is narrower by merely measuring the constitutional and non-violent removal of the government from power. Thus, this indicator lacks content validity to some degree. This problem could be solved by adjusting the attribute label to what the indicator measures or by using indicators that measure instability and violence more generally.

The operationalization of the cluster formulation of policies exhibits also exhibits some problems. First, the indicators "quality of public service provision" and "quality of the bureaucracy" are redundant because public services are provided by the bureaucracy. Thus, it is unclear what is gained by using both indicators. Second, the WGI focuses on the effectiveness of the government at the neglect of the parliament. This problem reappears at the operationalization stage because one only finds an indicator measuring the government's commitment to policies. However, if one is interested in the stability of policies, one should additionally consider the parliament's commitment because it is also involved in policy-making in many countries. Because of this, the indicator measuring the policy commitment is too narrow and lacks content validity.

A problem concerning the operationalization of "regulatory quality" is that it is measured through the incidence of market-unfriendly policies and the burdens imposed by excessive regulation. These indicators have a strong normative loading because of their emphasis of market-conform policies and deregulation. This normative bias is questionable inasmuch as it implicitly assumes that these are always the best economic policies to pursue. There is no such unambiguous link in practice because some interventions of the state in market processes may do more good than bad (Scharpf, 1999). For this reason, the indicators measuring regulatory quality lack content validity too.

The operationalization of the cluster "rule of law" is justifiable inasmuch as the two indicators measure the extent to which the law is obeyed. On the other hand, the measurement of "control of corruption" is questionable. This attribute asks for the instruments the state applies so as to diminish and prevent corruption. The indicator, however, measures the actual extent of corruption. These two issues are not identical, since corruption can be pervasive even if the state tries to control it and non-existent in the absence of any anti-corruption

measures. As a matter of fact, the WGI is interested in the extent of corruption, since it is said in the documentation that "control of corruption" is defined "as the use of public office for private gain" (Kaufmann et al., 2007b, 24). In this view, the attribute simply seems to be mislabeled and is more accurately described as "extent of corruption". Finally, I find that the attribute specificity of the WGI is not optimal. It is left implicit how the indicators are linked to each other, given that the attribute involves multiple indicators. This creates some ambiguity about how indicator scores should be aggregated to an attribute score.

4.2. World Governance Survey

The six functional dimensions governance has according to the WGS are measured through five indicators each. An inspection of the indicators shows that they mostly exhibit content validity (table 5). One may wonder whether "ensuring freedom from fear" and "ensuring freedom from want" should be subsumed under the executive dimension. The reason for this question is that the indicators refer to the outcome of policy making and not the *way* policies are produced, which is what the WGS aims to measure. Beyond that minor point, however, the selection and assignment of indicators to the dimensions appear justified. In contrast, it is to criticize that the attribute specificity is not optimal because the links between the indicators are not specified. On the basis of the WGS's documentation, it is most reasonable to assume that they are integral measures of the attributes, meaning that the five attributes of the AND-type.

The selection of indicators is heavily inspired by the *Universal Declaration of Human Rights* (UDHR). The UDHR was taken because most countries in the world have signed it and because of a consensus among people that human rights are important (Hyden and Court, 2002, 23). Deriving indicators from the UDHR is not disputable in principle. However, it casts doubt on the claim that the WGS's understanding of governance is different from the Western conception of democracy in which human rights play a central role. In fact, a comparison of the indicators underlying common democracy indices and the WGS shows a substantial degree of overlap (cf., Munck and Verkuilen, 2002, 10). Thus, governance as measured by the WGS is closer to existing measures of democracy than the documentation makes one believe.

Table 5: WGS's operationalization of governance[12]

Functional dimension	Purpose of rules	Indicators
Socializing	Shape the way citizens raise and become aware of public issues	Freedom of expression Freedom of peaceful action Freedom from discrimination Opportunity for consultation Public duties
Aggregating	Shape the way issues are combined into policy by political institutions	Representativeness of legislature Political competition Aggregation of public preferences Role of legislative function Accountability of elected officials
Executive	Shape the way policies are made by government institutions	Ensuring freedom from fear Ensuring freedom from want Willingness to make tough decisions Political-military relations Attitude to peace
Managerial	Shape the way policies are administered and implemented by public servants	Scope for policy advice Meritocracy in bureaucracy Accountability of appointed officials Transparency Equal access to public service
Regulatory	Shape the way state and market interact to promote development	Security of property Equal treatment Obstacles to business Consultation with private sector International economic considerations
Adjudicatory	Shape the setting for resolution of disputes and conflicts	Equal access to justice Due process Accountability of judges Incorporation of International Human Rights Norms Predisposition to conflict resolution

4.3. Quality of Government

The QoG dataset is a meta-dataset that brings together existing data on variables that are related to governance. The variables are assigned one of the three following categories: *what it is* (WII), *how to get it* (HTG), and *what you get* (WYG). WII variables capture the "core areas of the QoG compound" (Teorell, Holmberg and Rothstein, 2007, 13). HTG variables comprise factors that are expected to promote good governance and WYG indicators refer to the consequences of Quality of Government. The inclusion of these three types of variables in one dataset seems to be beneficial from a user perspective because it makes it easy to empirically examine the state, causes, and effects of good governance.

[12] A more detailed description of the indicators can be found in Hyden and Court (2002, 31-34).

However, the category of WII variables, which matters for the operationalization of governance, is problematic. The shortcoming is that not all indicators in the WII category are supposed to measure the extent of impartiality, which is good governance according to the QoG. This becomes apparent in the description of this category as containing "variables pertaining to the core areas of the QoG compound" (Teorell et al., 2007, 13). A case in point is the inclusion of indicators measuring the democratic quality of a country because it is stressed that good governance does not require democracy (Rothstein and Teorell, 2005). The problem for the data user is that it remains unclear what variables of the WII category are meant to be valid measures of good governance. This leaves the data user uncertain about which indicators to use and not to use. Taking into account that the WII category comprises 113 indicators, there is a large menu for choice and room for much confusion about what indicators to select. The failure to single out specific indicators of governance makes the QoG's attribute specificity minimal. As a consequence of this, it is not possible to make more detailed statements of how the QoG operationalizes governance. Table 7 summarizes the discussion of operationalization.

Table 6: Comparison of operationalization

Criterion	WGI	WGS	QoG
Indicators per attribute	Depends on attribute	Five	Cannot be assessed
Attribute specificity	Moderate	High	Low
Attribute type	AND	AND	Cannot be assessed
Content validity	Moderate	High	Cannot be assessed

5. Data collection and processing

5.1. World Governance Indicators

The WGI draws on information that is provided by a variety of international, national, public, private, and non-governmental organizations. All information is taken either from expert surveys or surveys of citizens and companies. The exclusive use of subjective data is justified with the argument that objective data is not broadly available for many countries over time (Kaufmann et al., 2004, 271). I discuss the use of sources and the aggregation procedure together since they are closely intertwined in practice. The process is as follows. First, the answers to the selected survey items are standardized and assigned to one of the six clusters. The average score is calculated when more than one question from a survey can be subsumed under a specific cluster. The beneficial aspect of using multiple sources for the scoring of

countries is that one can calculate margins of error, which are the smaller, the higher the correlation between the sources. Afterwards, a single score is compiled for each dimension by estimating an unobserved components model. In a nutshell, the measurement procedure assigns higher weight to sources that correlate strongly with each other. This approach is based on the argument that a high degree of correlation between sources can be interpreted as systematic and reliable information about governance. On average, the assignment of weights to sources increases somewhat the importance of business sources at the expense of household surveys and public agencies (Kaufmann et al., 2007a, 11-13).[13] The weights are then applied to the scores generated from each source and an aggregated attribute score is created for each country and dimension. These scores are not further aggregated, meaning that one obtains six dimension-specific governance scores for each country and year.

One should appreciate that the WGI makes transparent what sources and survey questions are used in the scoring process (Kaufmann et al., 2007a, 39-69). Moreover, the website of the WGI contains links to the websites of almost all providers of the raw data. However, there are some serious issues that can be subsumed under the source coverage problem. First of all, the choice of surveys from which information is extracted is not entirely transparent. The criterion seems to be that a source yields information for more than one or two points of observation (Kaufmann et al., 2007a, 8-9). This is a reasonable guideline, but it remains unclear whether the sources the WGI is based on all available sources satisfying this criterion or whether there are additional sources that are not used. For this reason, it cannot be determined whether the underlying surveys represent a sample or the population of all suitable sources and whether the sample is biased or randomly drawn.[14]

An additional problem concerning the use of sources is the distinction between representative and non-representative sources. The latter are defined as those which do not cover a sufficiently large number of countries and/or do not cover a reasonable mix of poor and wealthy states (Kaufmann et al., 2004, 256-258). The assignment of sources as representative and non-representative is important because the latter receive less weight in the estimation of the unobserved components model (Kaufmann et al., 2007a, 70-75). Unfortunately, it is neither made transparent how it was decided what a representative and a

[13] The results one obtains are insensitive to the applied weighting scheme (Kaufmann et al., 2007a, 13).
[14] A related point that has been raised by others already is the strong reliance on business surveys (Kaufmann et al., 2007b, 12-14). This could be interpreted as using a biased set of sources because business surveys may favor the view of multi-national corporations and business actors more generally, thus producing unbalanced governance pictures of countries. While the share of business sources is larger than the share of non-business surveys, it has been shown that the information extracted from the two types of sources is not systematically different. For this reason, the somewhat unbalanced set of sources does not seem to produce biased empirical results.

non-representative source is, nor does it have seem to be checked in how far the governance scores are sensitive to different assignment rules.

Another issue that can be subsumed under source sampling refers to the extraction of items from one source. First, it is difficult to determine what items were *not* used for the scoring of clusters and on what basis this was decided. This is a shortcoming because a data user cannot replicate the data collection process. Moreover, it would be interesting to see whether the governance scores are sensitive to different extraction rules. Second, there are no explicit guidelines about how the items were assigned to one of the six dimensions. Again, this makes it impossible to replicate the scoring process and to check whether different assignment procedures result in different governance scores.[15]

Concerning the extent of *intersource reliability* and the quality of the utilized sources in more general, it is stated that each source uses comparable methodologies from year to year (Kaufmann et al., 2007a, 6). Consistency in the data collection is beneficial because as a procedural change undermines the comparability of scores over time (Hazlewood, 1973, 174). However, it is also noted that the sources have different country coverage, implying that different sources are used for the scoring of divergent countries (Kaufmann et al., 2007a, 6). This is an important point because the cross-section comparability of data is as important as its diachronic comparability (Azar et al., 1973; van Deth, 1998). An analysis of the former requires an analysis of how the surveys were performed. This has not been done, so there is some uncertainty about the data quality and comparability on the cross-section dimension.

The WGI has been criticized already on the ground that different sources are used for the scoring of countries (Arndt and Oman, 2006; Knack, 2006). The founders of the WGI respond that the alternative would be to limit the number of country-years to those for which one has a sufficiently high overlap of sources (Kaufmann et al., 2007b, 6). They discard this option because of their intention to compile a dataset with extensive coverage. One may make this argument, but it does not relieve one from a careful assessment of the extent of intersource reliability. Another counterargument is that the WGI contains margins of error, which are the larger, the smaller the number of available sources for a country is in a given year. However, the margins of error are of little help if the raw data is of low quality because then it is misleading anyway. Therefore, it remains is indispensable to evaluate the quality of each source by assessing how the survey was conducted.

[15] The fact that most of the sources are publicly available, what is frequently emphasized in the documentation, does not ensure replicability because one needs to know the details of the measurement procedure, which are not known.

Furthermore, it is not examined in how far the expert surveys suffer from *method factors*.[16] Method factors refer to systematic measurement error in the data that derives from biased expert judgments (cf., Bollen and Paxton, 1998). It has been found, for example, that the personal familiarity of an expert with countries tends to bias the scoring of these countries. One means to determine the extent and nature of method factors is structural equation estimation (Bollen, 1993; Bollen and Paxton, 2000). Since no such instrument has been applied, the quality of the expert surveys and the governance scores derived from them suffers from some degree of uncertainty that is not reflected in the margins of error included in the WGI.

An additional problem related to the expert surveys is that they may not be independent of the other sources the WGI draws on (called "correlation of errors" in the WGI material (Kaufmann et al., 2007b). For example, the WGI uses the country assessments of risk rating agencies. These assessments may in turn be based on other sources used for the construction of the WGI, which would mean that the information that is gathered from these sources is not independent. If there is indeed interdependence between different sources, it would not be surprising that some of them correlate strongly with each other. Ultimately, this would inflate the weights of the interdependent sources in the estimation of the unobserved components model and thus undermine the interpretability of the results. The same problem results when sources draw on the same raw data like newspapers and news agencies that almost necessarily provide partial and, most often, biased reports about certain events (e.g., Smith, 1969; Sommer and Scarritt, 1999). As a matter of fact, methodological research in conflict analysis shows that a source coverage problem exists, since newspapers only provide incomplete pictures about events like cross-border incidents (Azar et al., 1973) or oppressive measures in non-democratic countries (Bollen, 1993).

As a response to a similar critique (Arndt and Oman, 2006; Knack, 2006), Kaufman, Kraay, and Mastruzzi (2007b, 16-18) acknowledge that this may be a problem. However, they claim that the sources they use never are perfect copies of other sources. In this case, the aggregation of information from different sources is always superior to the use of a single source. While this is correct, however, a non-perfect overlap between sources still drives up their correlation and may result in the spread of inaccurate information about governance. Kaufman, Kraay, and Mastruzzi (2007b, 18-20) further argue on the basis of a statistical analysis that their sources are not interdependent. If the sources are interdependent, the correlations between expert surveys should be high. In contrast, the correlations between

[16] The problem of method factors is recognized and discussed, but discarded on the basis of arguments whose validity can be questioned .

expert surveys and larger surveys of individuals and firms should be low because these are expected to be independent. Since the correlations are high in both cases, they conclude that there is no interdependence among expert surveys. This conclusion is premature, however, because an alternative explanation is equally compatible with the evidence. The expert surveys may be based on the large surveys (and perhaps other expert surveys as well), which would almost automatically produce high correlations. In my eyes, the question of interdependence can hardly be answered on the ground of statistical calculations. Instead, it is necessary to find out what the sources are that the expert use to form their opinion. The list of sources could be compared with the sources underlying the WGI so as to find out whether there is interdependence or not. Notwithstanding the broad base of sources and the statistically sophisticated aggregation technique, I conclude that the data collection stage involves serious problems that cast doubt on the validity of the WGI governance scores.

5.2. World Governance Survey

The WGS is based on surveys of local elites in the countries of interest. The use of objective data is discarded because the WGS is interested in processes and not in output and performance measures. Surveys of foreign experts that are integral to the WGI are considered inappropriate because they are believed to be too far removed from the developments in other countries. Surveys of the citizens in a country are rejected for three reasons: they are too costly; it is difficult to draw random samples of participants; and pre-tests revealed that citizens have a lack of "detailed knowledge and understanding of specific governance" (Court, Hyden and Mease, 2002a, 5). Members of the local elite have this knowledge and thus are most suitable for the collection of governance data. The participants were asked to locate their country for each item on a scale from 1 (lowest score) to 5 (highest score). The answers are averaged for each item in order to account for measurement error and added up to a single governance score for each country. Since there are 30 items each of which can take a score from 1 to 5, the range of governance scores goes from 30 to 150 (Court et al., 2002a, 11-12).

In the case of the WGS, *source sampling* captures the selection of survey participants and the information obtained from them. The experts were selected from seven different groups so as to achieve a cross-section of local elites and avoid evaluation biases (Court et al., 2002a, 7-8).[17] The WGS shows, for example, that participants from the political and administrative realm tend to give better governance scores than NGO experts (Court et al.,

[17] These fields are: high-ranking civil servants, long-standing parliamentarians, business persons, senior judges and lawyers, respected academics, consultants or policy-advisors, heads or senior officials in local NGOs, editors or senior reporters in the media, and any other relevant category (Court et al., 2002a, 7).

2002a, 15-16).[18] Because of this, selecting experts from different spheres is a viable strategy to improve measurement validity. It can be questioned, however, whether the members of each group are equally suited to answer each item in the questionnaire. For instance, it does not seem given that the members of NGOs or local business persons know the extent to which there is a merit-based system for recruitment in the civil service. It may be the case that the local experts have sufficiently broad knowledge, but this should not be taken for granted. Another point that has not been considered is that the answers of all participants may be upward-biased in non-democratic countries. If the elites of non-democratic countries are allowed to participate at all in the WGS, then they may overrepresent the quality of governance in their country because of fear of oppression. In this view, the exclusive use of local experts may produce an *source sampling bias*.

Another potentially problematic manifestation of source sampling concerns the response rates that are reported for each country (Court et al., 2002a, 9). It is to be welcomed that the national response rates are presented. However, the response rates per group in each country are missing.[19] This information would be interesting to know because the WGS's founders themselves acknowledge that the answers obtained from different groups tend to be systematically different (see above). Thus, imbalances in the responses per group may somewhat bias the results. Moreover, the documentation mentions that data is missing for some countries because the response rates of different groups were too diverse (Court et al., 2002a, 8). Unfortunately, the criteria for the exclusion of countries are not made explicit. There is no indication about how skewed the distribution of responses should be in order to be considered useless. A similar critic holds true for the removal of countries because of an overall response rate that is too low because it is not explicated what how low "too low" is.[20] For this reason, one can only speculate about the precise reasons for why six countries were excluded from the analysis.[21]

A third point of criticism going under the rubric of source sampling and potential sampling bias refers to the non-use of the national response rates for the choice of countries. For example, Chile is included in the study with a response rate of just .31. 113 questionnaires were sent out, which implies that the number of completed surveys is 35 and that there are

[18] A special problem that is not accounted for is that the participants were asked in 2000 to score their country for the years 2000 and 1995. As is well-known in survey research, answers concerning the past should be dealt with cautiously because of fading memories.

[19] One only finds information about the median for each group across all countries and the differences between the means of specific groups (Court et al., 2002a, 15-16).

20 At one point, it is written that the country coordinators should deliver 35 completed surveys (Court et al., 2002a, 7). However, this figure simply seems to be the number of surveys to collect in order to get paid by the founders of the WGS.

[21] These countries are Barbados, Korea, Nepal, Nigeria, Papua-New Guinea, and Samoa (Court et al., 2002a, 8).

five to six respondents from each group of elites on average. It seems reasonable to ask whether 35 returned questionnaires are a sufficiently high number. If one assumes that the country coordinator contacted 113 elite members because of his belief that all their opinions matter, a response rate of .31 is rather small and casts doubt on the validity of the data. This is even more the case when one considers that the respondents are probably not a random sample of all potential respondents. The WGS is based on the filled-in questionnaires that were returned by local experts. There is no indication for a second or third wave of questionnaire releases, which is customary practice in survey research so as to drive up the response rate and avoid sampling bias. Thus, there is a certain risk that the returned questionnaires represent a source sampling bias.

With respect to the processing of the raw data, it is positive that the quality of the governance scores is checked through reliability and validity tests. The reliability score is high and indicates the viability of the survey method. Measurement validity of the WGS seems to be given as well because the governance scores correlate reasonably high with comparable indicators of the WGI and other datasets (Court et al., 2002a, 17-22).

5.3. Quality of Government

As mentioned before, the QoG subsumes variables under three categories: *what it is* (WII), *how to get it* (HTG), and *what you get* (WYG). The broad coverage of the QoG is beneficial because it does not only allow for descriptive inference (keeping in mind that the set of purportedly valid indicators is not clearly specified), but also includes variables that drive the quality of good governance and its consequences. However, the breadth of the QoG dataset comes at the expense of depth because the quality of the datasets that were drawn together is not examined. The failure to provide quality assessments introduces an unknown degree of uncertainty into any analysis drawing on sources in the QoG dataset. Furthermore, the variety of sources that supposedly measure similar constructs like unemployment makes it necessary to provide information that goes beyond the time-series cross-section coverage of the included datasets. The reason for this is that datasets measuring the same concept comprise different data are likely to produce divergent results in an empirical analysis (cf., Azar et al., 1973; Deken and Kittel, 2006; Doran et al., 1973). Thus, tests for convergent validity or factor analyses comparing different datasets are important so as to find out whether they are substitutable (cf., Hazlewood and West, 1974; Jackman and Boyd, 1979). Without such tests, it may happen that researchers analyzing exactly the same empirical relationship come to contradictory conclusions simply because they selected different datasets that are comprised by the QoG. A summary of the discussion of measurement is presented in table 7.

Table 7: Comparison of measurement

Criterion	WGI	WGS	QoG
Type of sources	Expert polls & surveys	Surveys of elites in a country	Surveys, subjective data, objective data
Selection rule for sources	Sufficient time-series coverage Exhaustiveness of sources unclear	Elites from different areas (politics, business, NGOs, etc.) Sample probably biased	Relevance for good governance
Number of sources per country	Varies from 1 to more than 10	One survey (33-42 responses, response rates .31 to .84)	Varies across countries
Assessment of quality of sources	Weak	Reliability and validity tests Choice of respondents probably biased	None

6. Aggregation

6.1. World Governance Indicators

Both the WGS's concept and attributes are of the AND-type. Thus, achieving attribute type congruence of aggregation requires taking the minimum-scoring indicator as the attribute score. In a second step, the concept type congruence of aggregation calls for assigning the minimum score on all attributes as the case score. With respect to the latter task, I have mentioned above that the WGI does only deliver cluster-specific governance figures and refrains from calculating a single governance score for each country-year. The non-aggregation of the cluster scores is explained with the argument that these are more informative than a country-year governance score, which is an argument that cannot be disputed.[22]

However, the aggregation of indicator scores to cluster scores is rather problematic. The basic point is that the WGI completely ignores the indicators in the scoring process. As I mentioned above, the WGI details indicators for each attribute. Yet, they do not figure explicitly in the scoring stage because the survey items are directly assigned to an attribute and not the indicators belonging to it. This is an inappropriate procedure because the WGI

[22] Aggregating the attribute scores can neither be criticized, since it is a matter of taste and of the research question whether aggregated or non-aggregated attribute scores are more appropriate.

concept of governance involves multi-indicator attributes, which makes it necessary to score attributes by aggregating the indicator scores. As a consequence of the failure to proceed that way, the WGI lacks attribute type congruence and contains governance scores that are of dubious validity.

In a first step, the WGI averages the scores of all items from one survey that can be subsumed under the same cluster. Within-survey aggregation presumes that these items are substantively equivalent and that they can be averaged so as to diminish random measurement error. The assumption of the substitutability of items can be questioned (Knack and Manning, 2000). In order to exemplify this point, assume there are two countries: one is marked by a highly corrupt judiciary (low score) and a public administration with high integrity (high score). The second country has modest corruption scores for both the bureaucracy and the judicial system. Both countries would appear moderately corrupt in the WGI because of medium average corruption scores. This impression would be misleading, however. Averaging corruption scores means that a country can compensate for high corruption in one branch of the state (e.g., the judiciary) by being rather integer in another branch (e.g., the administration). This line of argument is not convincing because a strongly corrupt judicial system should be a sufficient condition for bad governance. It discourages people and investors because of a lack of predictability and the failure to enforce the rule of law. It is not obvious how an integer administration should compensate for a corrupt judiciary in practice. In this view, averaging the scores on items is unwarranted and the minimum-scoring indicator should be taken as the attribute score instead (Goertz, 2006, chap. 4).

A similar critic applies to the estimation of the unobserved components model in the second step of the aggregation procedure. Simply spoken, the unobserved components technique derives a single cluster score from the indicator scores. Again, this procedure allows a high score on one indicator to compensate for a low figure on another indicator belonging to the same cluster. Imagine that two sources yield low scores for “civil liberties” and high figures for “independence of the media”, which are two indicators that belong to the voice and accountability cluster. The cluster score would lay in between the high and the low score because of the way unobserved components estimation works. This procedure does not achieve attribute type congruence because the attribute is of the AND-type, which requires taking the minimum indicator score as the attribute score. Substantively, the minimum-rule is sensible because it seems counterintuitive to argue that a country is moderately democratic when there is no free media at all. A country must have a free and independent media in order to qualify as a democracy independently of how it scores on other democracy indicators. This

conceptual aspect is not reflected in the WGI's aggregation technique because high figures on one indicator may compensate for low figures on another indicator.

For these reasons, I conclude that the WGI lacks attribute type congruence and contains governance scores that do not match the underlying logic of operationalization. As a response to a related critic of the WGI's scoring procedure, the documentation says at one point that the WGI draws on an "implicit definition of corruption" that is obtained from the aggregation of data from different sources (Kaufmann et al., 2007b, 6-8). This argument turns the standard measurement procedure upside down because the careful conceptualization and operationalization of a concept should precede the scoring and aggregation process and not vice versa (Sartori, 1970).

6.2. World Governance Survey

The WGS's aggregation of indicator scores lacks soundness too. The WGS data is based on an AND-concept of governance, that is, it involves six elements that are all considered integral of governance. Moreover, each of the six elements is measured through five indicators each of which captures an essential aspect of this dimension, meaning that the attributes are of the AND-type as well. Altogether, it follows that the minimum indicator should be taken for the scoring of a country-year (Rohlfing, 2008).

Table 8: Rank of countries in WGS under different aggregation procedures for 1995 and 2000

Country	1995		2000	
	Rank (sum)	Rank (minimum)	Rank (sum)	Rank (minimum)
Argentina	8	5	9	4
Bulgaria	7	7	10	10
Chile	15	14	15	14
China	3	8	7	7
India	16	12	14	13
Indonesia	1	1	6	9
Jordan	14	11	13	11
Kyrgyzstan	9	10	4	8
Mongolia	10	13	11	12
Pakistan	4	3	2	1
Peru	2	2	8	6
Philippines	13	9	5	2
Russia	6	6	3	5
Tanzania	11	16	12	15
Thailand	12	5	16	16
Togo	5	4	1	3
Kendall's tau	.70		.65	

This aggregation procedure is not applied because the WGS adds the scores on all indicators, thus compensating low scores on some indicators through high scores on others. A comparison of the ranking of countries according to the sum of scores and minimum-indicator scores shows that the inaccurate aggregation technique matters (table 8). Kendall's rank correlation coefficient *k-tau* is just .70 for 2000 and .65 for 1995.

6.3. Quality of Government

The QoG does not specify what indicators of the *what it is* category are supposed to be valid measures of good governance, not to speak of the calculation of specific governance scores. Since no indicator scores are aggregated, it is not possible to evaluate the quality of the QoG on the aggregation dimension. Table 9 summarizes the discussion of aggregation.

Table 9: Comparison of aggregation

Criterion	WGI	WGS	QoG
Aggregation technique	Averaging within sources Unobserved components model for each dimension	Adding up all indicator scores	Cannot be assessed
Attribute type congruence	No	No	Cannot be assessed
Concept type congruence	No	No	Cannot be assessed

7. Representativeness

7.1. World Governance Indicators

The WGI offers an impressive amount of governance scores for the years 1996, 1998, 2000, 2002, 2004, 2005, and 2006. The WGI has global coverage, implying that it represents the full population of countries. With respect to the standards of representative data, it follows that the missing data problem is the only criterion applying to the WGI. The country coverage reaches from a minimum of 154 states on the control-of-corruption dimension in 1996 to a maximum of 212 countries on the government-effectiveness dimension in 2006. In general, the country coverage increases on all dimensions over time and in its current version the WGI now offers data for almost all countries in the world. While this is to be welcomed, it gives rise to the question whether data is systematically missing for the years and dimensions with a relatively smaller coverage. This question is based on the general observation that data is missing more often for less-developed countries that perform poorly on governance indicators than for

developed states (Herrera and Kapur, 2007). Thus, the data for 1996 and 1998 may suffer from missingness that is missing at random (MAR) or missing not at random (MNAR). These two types of missingness do not allow for simple listwise deletion, since this would produce methodological problems when the data is used for a regression. Instead, it would be necessary to apply the more demanding techniques of imputation and statistical modeling taking the missing data problem into account.

7.2. World Governance Survey

The currently available data has been compiled in the first phase of the WGS project, the major aim of which was to test the viability of measuring governance through expert surveys.[23] Because of this, the country coverage is relatively small with a number of 16 states. On the time-series dimension, the WGS offers data for 1995 and 2000. The number of observations could be increased by disaggregating the WGS country scores to dimension-specific scores. While the viability of disaggregation depends on the research interest, this strategy would increase the number of observations from 32 to 192, which would allow for large-n analyses.

The WGS does not involve missing data for the included countries because all of the first-phase countries for which sufficient data was lacking were eliminated from the study. This means that the technique of listwise deletion was applied to the raw data and the available dataset is a processed version.[24] As explained in section two, listwise deletion should be applied when data is missing completely at random (MCAR), denoting that missingness is not systematically related to the scores of the variable for which data is missing or any other variable (Allison, 2001, 2-5). It has not been assessed whether the six excluded countries are systematically different from the remaining 16 states, implying that it has neither been examined whether listwise deletion is a viable tool. This implies, furthermore, that the interpretation of the empirical results for the 16 WGS countries may be undermined by the elimination of the six countries lacking sufficient data. They may be different from the remaining states in a way that is systematically related to governance, which impedes the generation of inferences derived from the 16 included countries. In addition, it is not explained why the 22 countries were selected in the first place, whether they are conceived of as a sample and, if yes, what the population is from which this sample was drawn. This is not a point of concern within the WGS project inasmuch as the insights of the first phase are not

[23] The second phase covering a larger number of countries is currently running, but no comprehensive data has been made available by now see for a first discussion .
[24] The excluded countries are Barbados, Korea, Nepal, Nigeria, Papua-New Guinea, and Samoa (Court et al., 2002a, 8)

generalized beyond the countries under analysis (Court, Hyden and Mease, 2002b).[25] Yet, researchers who intend to use the WGS data and aim to generalize should carefully specify what the population of interest is and then assess if the WGS countries are a random sample or not.

7.3. Quality of Government

The QoG data includes one cross-section and a time-series cross-section dataset that are both global in coverage. The cross-section data has been compiled for 2002, the time-series cross-section dataset covers the period from 1946 to 2005. The notion of global coverage should not conceal, however, that there is a lot of missing data for the included units. The extent and nature of the missingness varies from dataset to dataset that is included in the *what it is* category of the QoG. Moreover, it should be recalled that not all variables in this category are supposed to be measures of good governance. Without a better specification of how impartiality is operationalized, it is impossible to make more detailed assessments of the nature and extent of missing data. If a data user picks a certain variable from the QoG on the basis of her own theoretical and empirical reasoning, it is indispensable to assess the missing data problem oneself before feeding the variable into the empirical analysis.

Table 10: Comparison of representativeness

Criterion	WGI	WGS	QoG
Cross-section coverage	Global	16 countries	Global
Time-series coverage	1996, 1998, 2000, 2002, 2004, 2005, 2006	1995, 2000	1946-2005
Random sample	Does not apply	Cannot be assessed	Does not apply
Missing data	Potentially for less-developed and non-democratic countries for 1996	Yes Countries excluded through listwise deletion	Cannot be assessed

8. Conclusion

Good governance has been subject to an increasing number of empirical studies in the past. The accuracy of any empirical analysis of governance hinges on the quality of the data that is used. In order to promote the understanding of governance, I have examined three datasets

[25] The exclusion of six countries that I just discussed is a problem nonetheless, since the missing data problem does also apply when generalization is not intended.

measuring the state of governance in a country. The generation of the datasets represents a laudable effort toward a comprehensive understanding of good governance. However, my discussion has shown that all three data collections have their deficiencies that make it necessary to interpret any results derived from them with caution.

The WGI have broad cross-section coverage and draws on a large number of sources. The theoretical and conceptual foundation of the data generation procedure is weak, however. This casts doubt on the quality of the governance scores inasmuch as careful conceptual thinking should precede the scoring process. Additional uncertainty about the governance scores derives from the neglect to assess the quality of the underlying sources. The WGS is theoretically and conceptually more solid and conducts some quality checks of the generated scores. The major problem is the aggregation procedure, which does not correspond with the way governance is conceptualized. In addition, the surveys on which the WGS is based do not meet the standards of good survey research. The QoG combines a large stock of existing datasets that enable one to perform comprehensive studies in which governance can be used as an independent and a dependent variable. However, the understanding of good governance as the impartiality of policy implementation is too minimalistic. Moreover, it is not discussed how impartiality should be operationalized. As a consequence of this, the data user is left alone with the large amount of assembled data in an empirical analysis. In sum, the three datasets are only suitable within certain limits for the analysis of governance and much work for the improvement of data is still ahead of governance research.

References

Adcock, Robert, and David Collier (2001): Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95(3):529-546.

Allison, Paul David (2001): *Missing Data*. Thousand Oaks: Sage.

Al-Marhubi, Fahim (2004): The Determinants of Governance: A Cross-Country Analysis. *Contemporary Economic Policy* 22(3):394-406.

Arndt, Christiane, and Charles Oman (2006): *Uses and Abuses of Governance Indicators*. Paris: OECD.

Azar, Edward E., Stanley H. Cohen, Thomas O. Jukam, and James M. McCormick (1973): The Problem of Source Coverage in the Use of International Events Data. *International Studies Quarterly* 16(3):373-388.

Bollen, Kenneth A. (1993): Liberal Democracy - Validity and Method Factors in Cross-National Measures. *American Journal of Political Science* 37(4):1207-1230.

Bollen, Kenneth A., and Pamela Paxton (1998): Detection and Determinants of Bias in Subjective Measures. *American Sociological Review* 63(3):465-478.

Bollen, Kenneth A., and Pamela Paxton (2000): Subjective Measures of Liberal Democracy. *Comparative Political Studies* 33(1):58-86.

Bouckaert, G., and S. Van de Walle (2003): Comparing Measures of Citizen Trust and User Satisfaction as Indicators of 'Good Governance': Difficulties in Linking Trust and Satisfaction Indicators. *International Review of Administrative Sciences* 69(3):329-343.
Brandolini, Andrea, Piero Cipollone, and Eliana Viviano (2006): Does the ILO Definition Capture All Unemployment? *Journal of the European Economic Association* 4(1):153-179.
Carmines, Edward G., and Richard A. Zeller (1979): *Reliability and Validity Assessment*. Beverly Hills, Calif.: Sage Publications.
Collier, David, and James E. Mahon (1993): Conceptual Stretching Revisited - Adapting Categories in Comparative-Analysis. *American Political Science Review* 87(4):845-855.
Court, Julius, Goran Hyden, and Ken Mease (2002a): Assessing Governance: Methodological Challenges. *World Governance Survey Discussion Paper 2*.
Court, Julius, Goran Hyden, and Ken Mease (2002b): Governance Performance: The Aggregate Picture. *World Governance Survey Discussion Paper 3*.
Deken, Johan De, and Bernhard Kittel (2006): "Putting the Chain Saw into Social Expenditures. Retrenchment and the Problems of Using Aggregate Data." In *Welfare Reform in Advanced Societies: Exploring the Dynamics of Reform*, edited by Nico Siegel, and Jochen Clasen, Cheltenham: Edward Elgar.
Doran, Charles F., Robert E. Pendley, and George E. Antunes (1973): Test of Cross-National Event Reliability: Global Versus Regional Data Sources. *International Studies Quarterly* 17(2):175-203.
Goertz, Gary (2006): *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.
Hazlewood, Leo A. (1973): Concept and Measurement Stability in the Study of Conflict Behavior within Nations. *Comparative Political Studies* 6(2):171-195.
Hazlewood, Leo A., and G. T. West (1974): Bivariate Associations, Factor Structures, and Substantive Impact: The Source Coverage Problem Revisited. *International Studies Quarterly* 18(3):317-337.
Herrera, Yoshiko M., and Devesh Kapur (2007): Improving Data Quality: Actors, Incentives, and Capabilities. *Political Analysis* 15(4):365-386.
Hug, Simon (2003): Selection Bias in Comparative Research: The Case of Incomplete Data Sets. *Political Analysis* 11(3):255-274.
Hyden, Goran, and Julius Court (2002): Governance and Development. *World Governance Survey Discussion Paper 1*.
Jackman, Robert W., and William A. Boyd (1979): Multiple Sources in the Collection of Data on Political Conflict. *American Journal of Political Science* 23(2):434-458.
Jacobs, Colin (2005): Evaluating the Comprehensive Development Framework in Kyrgyz Republic, Central Asia: Magic Bullet or White Elephant? *Evaluation* 11(4):480-495.
Kaufmann, Daniel , Aart Kraay, and Massimo Mastruzzi (2007a): Governance Matters VI: Governance Indicators for 1996-2006. *World Bank Policy Research Paper 4280.*
Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi (2004): Governance Matters III: Governance Indicators for 1996, 1998, 2000, and 2002. *World Bank Economic Review* 18(2):253-287.
Kaufmann, Daniel, Aart Kraay, and Massimo Mastruzzi (2007b): The Worldwide Governance Indicators Project: Answering the Critics. *Working Paper*.
King, Gary (1995): Replication, Replication. *Political Science & Politics* 28(3):444-452.
King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve (2001): Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 95(1):49-69.
Knack, S., M. Kugler, and N. Manning (2003): Second-Generation Governance Indicators. *International Review of Administrative Sciences* 69(3):345-364.

Knack, Stephen (2006): Measuring Corruption in Eastern Europe and Central Asia: A Critique of the Cross-Country Indicators. *World Bank Policy Research Paper* 3968.
Knack, Stephen, and Nick Manning (2000): Towards Consensus on Governance Indicators. *Working Paper*.
König, Thomas, Daniel Finke, and Stephanie Daimer (2005): Ignoring the Non-Ignorables? Missingness and Missing Positions. *European Union Politics* 6(3):269-290.
Kurtz, Marcus J., and Andrew Schrank (2007): Growth and Governance: Models, Measures, and Mechanisms. *The Journal of Politics* 69(2):538-554.
Ledet, Richard (2006): The Quality of Government Institute's Cross-Sectional and Cross-Sectional Time-Series Dataset. *APSA Comparative Politics Newsletter* 17(2):29-31.
Lustick, Ian S. (1996): History, Historiography, and Political Science: Multiple Historical Records and the Problem of Selection Bias. *American Political Science Review* 90(3):605-618.
Munck, Gerardo L., and Jay Verkuilen (2002): Conceptualizing and Measuring Democracy - Evaluating Alternative Indices. *Comparative Political Studies* 35(1):5-34.
Petrova, Velina P. (2007): Civil Society in Post-Communist Eastern Europe and Eurasia: A Cross-National Analysis of Micro- and Macro-Factors. *World Development* 35(7):1277-1305.
Przeworski, Adam, and Henry Teune (1966): Equivalence in Cross-National Research. *Public Opinion Quarterly* 30(4):551-568.
Rohlfing, Ingo (2008): What Is Good Macrodata? A Unified Framework for the Assessment of Data Quality. *Working Paper*.
Rothstein, Bo, and Dietlind Stolle (2007): A Theory of Political Institutions and Generalized Trust. *Quality of Government Working Paper Series, 2007, no. 2*.
Rothstein, Bo, and Jan Teorell (2005): What Is Quality of Government? A Theory of Impartial Political Institutions. *Quality of Government Working Paper Series 2005, no. 6*
Sartori, Giovanni (1970): Concept Misformation in Comparative Politics. *American Political Science Review* 64(4):1033-1053.
Scharpf, Fritz W. (1999): *Governing in Europe: Effective and Democratic?* Oxford: Oxford University Press.
Smith, Raymond F. (1969): On the Structure of Foreign News: A Comparison of the New York Times and the Indian White Papers. *Journal of Peace Research* (1):23-36.
Sommer, H., and J. R. Scarritt (1999): The Utility of Reuters for Events Analysis in Area Studies: The Case of Zambia-Zimbabwe Interactions, 1982-1993. *International Interactions* 25(1):29-59.
Teorell, Jan, Sören Holmberg, and Bo Rothstein (2007): *The Quality of Government Dataset, Version 1, July 2007*. Göteborg University: The Quality of Government Institute (http://www.qog.pol.gu.se).
Thomson, Robert (2007): Time to Comply: National Responses to Six EU Labour Market Directives Revisited. *West European Politics* 30(5):987-1008.
Treisman, Daniel (2007): What Have We Learned About the Causes of Corruption from Ten Years of Cross-National Empirical Research? *Annual Review of Political Science* 10(1):211-244.
van Deth, Jan W. (1998): "Equivalence in Comparative Political Research." In *Comparative Politics: The Problem of Equivalence*, edited by Jan W. van Deth, pp. 1-19. New York: Routledge.
Winship, Christopher (2007): Introduction to the Special Section on Replication and Data Access. *Sociological Methods & Research* 36(2):151-152.