Original Article

What Do the Worldwide Governance Indicators Measure?

M.A. Thomas

The Johns Hopkins University, USA.

Abstract As policymakers and researchers focus more on the question of the impact of governance in economic development, they have required measures of the quality of governance to set policy or to conduct analyses. A number of measures of the quality of governance have been created. Among these are the Worldwide Governance Indicators, which rank countries on six aspects of 'good governance'. Critics have focused on problems of bias or lack of comparability that raise questions about the utility of these indicators. However, a more fundamental question is that of whether the indicators have 'construct validity' – whether they measure what they purport to measure. This paper considers the construct validity of the indicators and raises the question of whether researchers and policymakers are relying on wrong data, rather than poor data.

Les responsables politiques et les chercheurs ont besoin de mesures concrètes de la qualité de la gouvernance afin de pouvoir déterminer l'impact de celle-ci, en particulier par rapport au développement économique. Un certain nombre d'indicateurs ont récemment été créés, parmi lesquels les Indicateurs de gouvernance dans le monde de la Banque Mondiale, qui classent les pays à partir de six critères de « bonne gouvernance ». L'utilité de ces indicateurs a été mise en question pour des raisons de distorsion ainsi que des problèmes de manque de comparabilité. Cependant, une question plus fondamentale est celle de la validité théorique de ces indicateurs, c'est-à-dire, s'ils mesurent ce qu'ils prétendent mesurer. Cet article considère la validité conceptuelle de ces indicateurs et cherche à déterminer dans quelle mesure les chercheurs et les responsables politiques ne sont pas en train de se baser sur des données fausses, plutôt que des données insuffisantes.

European Journal of Development Research (2010) **22,** 31–54. doi:10.1057/ejdr.2009.32
Published online 16 July 2009

Keywords: governance; corruption; rule of law; measurement; construct validity

Introduction

Since the 1990s, development researchers and practitioners have focused on 'good governance' as both a means of achieving development and a development objective in itself. The World Bank has defined 'good governance' as 'epitomized by predictable, open and enlightened policy making; a bureaucracy imbued with a professional ethos; an executive arm of government accountable for its actions; and a strong civil society participating in public affairs; and all behaving under the rule of law' (World Bank, 1994, p. vii).

In response to the growing demand for measures of the quality of governance, a number of governance indicators have been produced, such as the World Bank's Worldwide Governance Indicators ('WGI').[1] The WGI rank countries with respect to six aspects of good governance: Voice and Accountability, Political Stability and Violence, Government Effectiveness, Rule of Law, Regulatory Quality, and Control of Corruption. These indicators have been used by researchers as well as foreign aid donors such as the United States, who use the indicators to allocate aid packages of hundreds of millions of dollars.

The authors of the indicators have tried to draw attention to the large standard errors of the estimates – caveats that have been largely ignored. Researchers and policymakers should not only be concerned with large standard errors, however. Before they rely on these indicators, they should ask much more fundamental questions. What do these indicators measure, if anything, and how would we know? This paper describes the WGI, their use, and the methodology of their construction. It then discusses how abstract concepts ('constructs') are measured using observable variables, and how such measures are tested for validity. It analyzes the methodology in light of the need for evidence of construct validity. Finally, it concludes that the indicators stand as an elaborate and unsupported hypothesis. Accordingly, reliance on the indicators is premature.

The Worldwide Governance Indicators

In the 1990s, scholars focused on the role of good governance in economic development. North and Thomas (1996) argued that there are institutional prerequisites for economic growth. An influential report by World Bank researchers claimed that foreign aid could only be used effectively in countries whose governments had good policies, although the results are now disputed (World Bank, 1998; Burnside and Dollar, 2000, 2004; Easterly, Levine and Roodman, 2004).

In turn, aid donors began to condition aid eligibility on the quality of governance. Former President of the World Bank James Wolfensohn brought the issue of corruption into the mainstream of the Bank's dialogue with its borrowers, a dialogue that was continued and deepened by his successor, Paul Wolfowitz (World Bank, 2006a). The World Bank now considers the quality of governance in determining eligibility for loans (International Development Association, 2004). United States President George W. Bush launched a major foreign aid initiative to direct grants to countries with good governance (Millennium Challenge Account, 2009 http://www.mcc.gov/documents/mcc-fy-09-guidetotheindicators.pdf).

In order to test claims about the importance of good governance, or to implement policies that aim either to strengthen governance or target aid to well-governed countries, valid measurements of the quality of governance are needed. Although some data are available, they are problematic because they are not always good quality, their coverage is spotty and they are not comparable. Data come from expert assessments, polls of experts, and surveys of government officials, businesses and households. A few sources aim at global coverage, but the coverage of most sources is much more limited. The surveys and polls from various sources do not share a common methodology, definition of terms, set of questions or measurement scale of responses.[2]

World Bank researchers attempted to address these problems by developing indicators that rank countries according to quality of governance by aggregating data from many available sources. As of this writing, the WGI are based on 340 variables produced by 32 different sources, including commercial information providers, surveys of firms and households, non-governmental organizations and public sector organizations (Kaufmann *et al*, 2008). While the authors have drawn explicit attention to the large standard errors associated with the governance estimates, they argue that the methodology employed for developing the indicators has two important strengths. First, the aggregation methodology makes the WGI more informative than any individual data source. Second, it allows calculation of the margins of error of the estimated indicators. To these should be added a third advantage of the methodology, which is that it creates a data set that is global in coverage, albeit with some missing values. The WGI cover 212 countries and territories (Kaufmann *et al*, 2008).

The indicators are defined to correspond to what the authors consider to be 'fundamental governance concepts' (Kaufmann *et al*, 1999b, p. 1). The definitions of the indicators have changed over time since the indicators were first introduced 10 years ago. Most recently (Kaufmann *et al*, 2008), the indicators were defined as follows:

1. *Voice and accountability* (VA) – measuring perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association and a free media.
2. *Political stability and absence of violence* (PV) – measuring perceptions of the likelihood that the government will be destabilized or overthrown by unconstitutional or violent means, including political violence and terrorism.
3. *Government effectiveness* (GE) – measuring the quality of public services, the quality of the civil service and the degree of its independence from political pressures, the quality of policy formulation and implementation, and the credibility of the government's commitment to such policies.
4. *Regulatory quality* (RQ) – measuring perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.
5. *Rule of law* (RL) – measuring perceptions of the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, the police and the courts, as well as the likelihood of crime and violence.
6. *Control of corruption* (CC) – measuring perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as 'capture' of the state by elites and private interests.

The indicators have been produced for 1996, 1998, 2000, 2002 and 2004–2007. Eight papers have been written presenting the data sets (Kaufmann *et al*, 1999a, b, 2002, 2004, 2005, 2006a, 2007, 2008).

The careful construction of the indicators, their global coverage and the claim that they achieve maximum precision make the indicators attractive to researchers. In addition, although the World Bank itself has decided not to rely on the indicators for its operational work and the indicators have no official standing,[3] the fact that they are produced by World Bank researchers and are financed and publicized by the World Bank lends them the authority of that powerful institution. A number of studies have used the indicators as explanatory variables; their results thus depend on the indicators (see, for example, Neumayer, 2002; Apodaca, 2004; Hart, Atkins and Youniss, 2005; Llamazares, 2005; Andres, 2006; Das and Andriamananjara, 2006; Jung, 2006; Liu and San, 2006). For example, economists, including the authors of the indicators, have used the indicators to explore possible relationships between aspects of governance and growth (see, for example, Kaufmann and Kraay, 2002; Dollar and Kraay, 2003; Kaufmann and Kraay, 2003; Naude, 2004; Méon and Sekkat, 2005).

The indicators are also used by policymakers. In the United States, the Millennium Challenge Account ('MCA') was launched by the George W. Bush administration to target grants of hundreds of millions of dollars of foreign aid to countries that are well governed compared to their income peers. To make the determination of eligibility, the MCA relies on 17 third-party indicators of the quality of governance divided into three categories: Ruling Justly (six indicators), Encouraging Economic Freedom (six indicators) and Investing in People (five indicators). The MCC Board 'considers whether countries perform above the median in their income peer group … on at least half of the indicators in each of the three policy categories and above the median on the Control of Corruption indicator' (Millennium Challenge Account, 2009, p. 2, http://www.mcc.gov/documents/mcc-fy-09-guidetotheindicators.pdf). Of the 17 indicators, five are from the WGI, including the majority of the indicators in the category 'Ruling Justly' and the make-or-break corruption indicator.

In addition to the US government, according to the World Bank, '[o]ther donor governments, such as the Netherlands, also rely on the Worldwide Governance Indicators to monitor the quality of governance in aid recipient countries. Risk rating agencies as well as [non-governmental organizations] also use them' (World Bank, 2006b).

The suitability of the indicators for shaping policy towards individual countries is questionable, given the large standard errors of the estimates. Kaufmann *et al* (2005) report that of the 70 countries identified as potential MCA beneficiaries for the 2005 fiscal year, about 40 countries could be placed above or below the median with 90 per cent confidence; the remaining 30 fell into a 'zone of uncertainty.' Relaxing the confidence level to 75 per cent still left 20 countries in this zone. For the 2007 Control of Corruption indicator, the authors report that 35 per cent of countries still fell into this zone at the 90 per cent confidence level and 26 per cent fell into this zone at the 75 per cent confidence level (Kaufmann *et al*, 2008, p. 19), although, as the authors cautioned, these estimated standard errors are likely to be 'substantially understated' (Kaufmann *et al*, 1999a, p. 21). Aid allocation by the MCA is based on a determination of a country's placement compared to the median in its income group; where the data cannot provide support for such a placement, there is no basis for a determination.

In addition to the standard errors, however, scholars and development practitioners have begun to raise other concerns about the indicators and their use (see, for example, Arndt and Oman, 2006; Knack, 2006; Kurtz and Shrank, 2006; Razafindrakoto and Roubaud, 2006; Kurtz and Shrank, 2007; Iqbal and Shah, 2008). Kaufmann *et al* (2007) have categorized some of these critiques as concerns about the comparability of the indicators across countries and across time; concerns about bias in expert polls or in particular sources; and concerns about the independence of the different data sources and the consequences for the aggregate indicators.

The indicators may not be good data, but policymakers often do not have the luxury of waiting for good data to make decisions. A more fundamental question is that of whether the indicators are 'wrong' data. What does it mean to measure an abstract concept like 'government effectiveness' or 'rule of law'? Do these indicators measure what they claim to measure, or do they measure anything at all? What evidence would we need and how would we know? These are questions about the construct validity of the WGI.

Method of the Construction of the WGI

The estimates of governance produced by the WGI Project are the result of up to three levels of aggregation of underlying variables. The methodology is described most fully in the seminar paper by Kaufmann *et al* (1999a) and is nicely summarized in a study

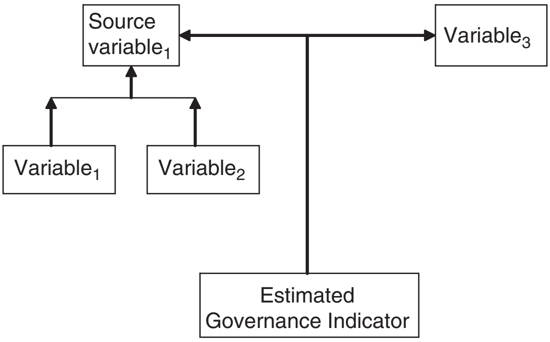

Figure 1: Model of governance underlying the WGI. Variables produced by the same data source (here, variables 1 and 2) are averaged to produce a source variable.

by Arndt and Oman (2006). Figure 1 illustrates the model underlying the WGI. The authors first identify interesting variables produced by third-party sources. While some of the variables are responses to survey questions or to questions asked of experts, some of these variables are themselves aggregates produced by third parties, and the methodologies for their construction – even their definitions – are not publicly available.

The authors rescale the variables for comparability. They then sort the variables into six clusters, each of which corresponds to what the authors believe is a ‘fundamental concept’ of governance. Each variable is assigned to only one cluster. In order to eliminate correlated errors among the variables arising from source-specific perception errors, where more than one variable from the same data source is used in constructing an indicator these variables are averaged and this simple average (referred to in this paper as a ‘source variable’) is used in the regressions in place of the variables themselves. The authors then reduce the variables and the source variables in each cluster to a single number for each country, which is the governance indicator.

A key assumption of the model is that each variable and source variable in a cluster is assumed to be a ‘noisy’ observation of the corresponding unobservable governance construct for that cluster. The problem is then one of signal extraction:

> We use an extension of the standard unobserved components model which expresses the observed data in each cluster as a linear function of the unobserved common component of governance, plus a disturbance term capturing perception errors and/or sampling variation in each indicator. In particular, we assume that we can write the observed score of a country j on indicator k, $y(j,k)$, as a linear function of unobserved governance, $g(j)$, and a disturbance term, $\varepsilon(j,k)$, as follows:
>
> $$y(j,k) = \alpha(k) + \beta(k) \cdot (g(j) + \varepsilon(j,k))$$
>
> where $\alpha(k)$ and $\beta(k)$ are unknown parameters that map unobserved governance $g(j)$ onto the observed data $y(j,k)$.

(Kaufmann *et al*, 2005, pp. 131–132). Accordingly, each variable and source variable is assumed to be a linear function of one, and only one, independent variable, namely the unobserved governance component for that cluster.

The model depends on a number of important assumptions regarding the error terms, which capture both measurement errors due to perception errors or sampling errors and errors due to imperfect relationships between the indicator and the governance concept

(Kaufmann *et al*, 2005). The error terms are assumed to have a zero mean and to be independent across source variables and countries (Kaufmann *et al*, 1999a). The variance of the error terms is assumed to differ across indicators but to be the same across countries (Kaufmann *et al*, 1999a). Governance and error terms are assumed to be jointly normally distributed. Finally, the distribution of unobserved governance and the mean of world governance are assumed to be the same in every period; in other words, world governance never gets better or worse over time.

To produce estimated governance indicators for a particular governance component for a country, the authors produce a weighted average of the governance estimates for the variables and the source variables in the cluster for that country. The weights are inversely proportional to the variance of the error term of the source (Kaufmann *et al*, 2005). According to the authors, this method '"rewards conformity," in the sense that indicators that are highly correlated will have low estimated variances and hence will be perceived as more precise' (Kaufmann *et al*, 1999a).

Measuring perceptions, or the thing itself?

The definitions of the indicators, and, therefore, the authors' hypotheses about what the indicators measure, have changed over time. Perhaps the most important change is that in 2006, the indicator 'Political stability and absence of violence' was redefined to measure 'perceptions of the likelihood that the government will be destabilized', rather than the likelihood itself (Kaufmann *et al*, 2007), and in 2008 the remaining indicators were redefined as measures of perceptions, rather than as measures of the underlying phenomena (Kaufmann *et al*, 2008).

Measurements of perceptions (as opposed to realities) can be of interest, as when, for example, Olken (2006) evaluates the effectiveness of grassroots monitoring of the corruption of local officials. However, there is a substantial difference between measuring a thing and measuring perceptions of it. In the context of governance, for example, perceptions of crime risk have been shown to be quite different than actual crime levels (see, for example, Forgas, 1980; Pfeiffer, 2005); perceptions of corruption have been shown to differ from actual corruption levels (see, for example, Olken, 2006; Seligson, 2006); and trust in government has been shown to differ from administrative performance (Van de Walle and Bouckaert, 2007).

If these changed definitions of the indicators are taken at face value, they represent the discontinuation of the previous series of governance indicators, which claimed to measure governance itself (see Kaufmann *et al*, 1999b), and the launch of a new set of indicators that, confusingly, bear the same names. However, it does not appear that this is what was intended. The changes in the definitions were not discussed by the authors. The methodology of construction of the indicators as described by Kaufmann *et al* (2004) did not change. This methodology assumes that variables are noisy signals of unobserved governance. If the perceptions themselves were being measured, there would be no reason to interpret variables measuring perceptions as noisy signals of something else. The data used to construct the indicators continue to include both variables that claim to measure subjective perceptions of governance, such as confidence in the honesty of elections, and variables that claim to measure other qualities not based on perceptions, such as the existence of domestic and foreign travel restrictions (Kaufmann *et al*, 2008, p. 72). The latter would not have any obvious relationship with perceptions. Finally, the authors continue to interpret changes in their data as reflecting changes in governance itself, rather than changes in perceptions of governance (Kaufmann *et al*, 2008, p. 20).

Accordingly, this paper follows the authors' lead by ignoring these changes to the definitions, and instead assumes that the 2007 indicators are intended to be measures of governance itself as set out in previous papers, rather than a new set of measures intended to measure perceptions of governance. The remainder of the paper considers whether the indicators are valid measures of governance.

Construct Validity

Constructs are abstract ideas, unobservable, and so cannot be counted or measured directly. Governance indicators are proposed measurements of governance constructs such as 'rule of law'. The first question that should occupy potential users of any governance indicator is not the size of the margins of error, but whether the indicators are valid measurements of what they purport to measure. Evidence must be provided to show that a purported measure of a construct is valid both in its conceptualization and its operationalization, by exploring predicted relationships between the measure and other observable variables and ensuring that the measure performs as the theory predicts.

Social scientists have been developing measures of constructs ever since psychologists attempted to define and measure intelligence in the early twentieth century (see, for example, Bartholomew, 1995; Williams *et al*, 2003). Quantitative measures of constructs have been used in psychometrics, to measure psychological attributes and aptitudes (see, for example, Patterson, 1990); in education, to measure educational achievement and ability (see, for example, Forsythe *et al*, 1986); in public health, to measure variables such as cognitive health (see, for example, Wallace and Herzog, 1995); in public administration, to measure public service motivation (see, for example, Perry, 1996); in economics, to measure contingent valuation and averting costs (see, for example, Laughland *et al,* 1996); and in political science and sociology to measure concepts of governance such as democracy, corruption, and political attitudes and perceptions (see, for example, Faber, 1987; Elkins, 2000; Johnston, 2000). All of these have been concerned with the issue of construct validity: the question of whether the measures in fact measure what they claim.

When direct measurement of an observable variable is impractical – for example, if it is too costly – social scientists often use a proxy. For example, economists are interested in measuring household consumption as an indicator of economic status. However, many demographic surveys do not include information on household consumption, which is difficult, costly and time-consuming to collect. Accordingly, studies have used proxies of consumption expenditure based on the household's ownership of assets such as a bicycle or a radio (Bollen *et al*, 2002). Collecting data on ownership of these assets is much simpler. In doing so, economists are acting on the hypothesis that ownership of specific assets reliably predicts overall consumption levels. Before acting on this hypothesis, however, economists tested that hypothesis and evaluated the ability of the proposed proxy to predict consumption expenditure by comparing the predictions to data on consumption expenditure where both were available (Filmer and Pritchett, 2001; McKenzie, 2005).

A proposed measure of a construct, an inherently abstract concept like the 'rule of law', is like a proxy measure in that it is a hypothesis about measurement. The hypothesis is that the proposed measure correctly measures the construct. Like proposed proxy measures, not all proposed measurements of constructs are equally valid. The hypothesis must be tested, and evidence supplied of the validity of the measure before the measure is used. However, whereas proposed proxy measures can be compared to the observable variable for which they are supposed to proxy to ensure that they are reliably correlated, proposed measures of constructs cannot be validated in this way, as constructs are inherently unobservable. Instead, a measure of a construct is validated, first by showing that it correctly represents the theoretical definition of the construct ('content validity' or 'face validity'), and second by seeing whether the proposed measure has the same relationships with observable variables that the theory predicts the construct itself to have ('convergent and discriminant validity'). Construct validity requires content validity, convergent validity and discriminant validity.[4]

Content validity is concerned with whether the proposed operationalization or model captures the entire domain of a construct and includes nothing extraneous (Carmines and Zeller, 1979). To assure the content validity of a proposed measure of a construct requires, first, a mapping between a theory about the construct and a more specific definition of the construct that is a description of the thing to be measured. Second, it requires a mapping between the description and a specific operationalization of that idea, which is a model based on observable variables that is used to estimate a measure of the construct. Figure 2 illustrates these relationships, while making the point that there may be any number of equivalent operationalizations or models of the same construct (Adcock and Collier, 2001).

'Convergent validity' is concerned with the extent to which a measurement is correlated with other variables with which theory predicts that the construct should be correlated. 'Discriminant validity' is concerned with the extent to which the measurement is uncorrelated with other variables with which in theory the construct should not be correlated (see Campbell and Fiske, 1959; Carmines and Zeller, 1979; Adcock and Collier, 2001). The proposed measure of a construct would be supported by evidence showing that its correlations with observable variables are consistent with the theoretical predictions about those relationships. If the predicted relationships are not found – if, for example, in defiance of theory, democracy is not found to be positively correlated with freedom – then this lack of convergent validity must lead the researcher to question either the theory that predicted such a relationship or the measures used for the constructs 'democracy' and 'freedom.' Validation of the measure must be carried out before the measure is used to explore relationships with other variables for which there are no theoretical predictions, because in the absence of evidence that the measure is meaningful, the results of such explorations can neither be interpreted nor confirmed.

An atheoretical measure of a construct is a contradiction in terms because the construct itself is embedded in theory, which imbues the construct with meaning. First, the researcher draws on theory about the construct to derive a definition of the construct

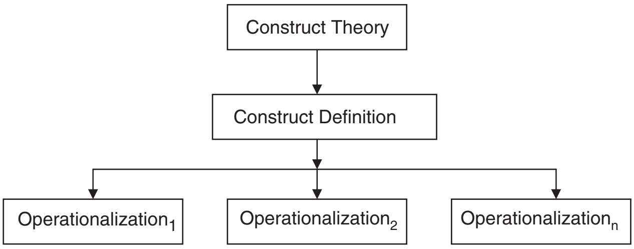

Figure 2: Measuring constructs.

(Campbell and Fiske, 1959). This definition is not a one-line description, but rather a 'fleshed-out account' of the concept that is within the matrix of meanings usually assigned to the construct by scholars (Adcock and Collier, 2001). Second, the researcher draws on that theory to describe how the construct, assuming that it is correctly measured, should relate to other observables (Cronbach and Meehl, 1955 see also; Carmines and Zeller, 1979; Smith, 2005).[5]

The construct, its proposed measure, and the theory from which they spring, gain scientific acceptance through a public iterative process of evaluation and theory revision. Accordingly, 'each component of a research program, or each component of theory derivation, hypothesis formation, and empirical test, must be open to criticism' (Smith, 2005, p. 398).

A number of techniques and tools have been brought to bear on the process of testing and providing evidence of construct validity. Smith (2005) provides a recent survey of quantitative approaches. To check for convergent and discriminant validity, Campbell and Fiske (1959) proposed an examination of the correlation matrix showing correlations among measures of more than one trait, measured by independent methods. Using this Multitrait–Multimethod Matrix ('MTMM'), researchers check that measures of the same trait are more highly correlated with each other than with measures of different traits. Other researchers have used structural equation modeling to isolate different sources of variance (see, for example, Hammond *et al,* 1986; Bollen, 1989, 1993; Eid *et al*, 2003), or to measure the goodness-of-fit of the proposed model (Westen and Rosenthal, 2003).

Kaufmann *et al* (2007) assert that ideas of construct validity are discipline-specific, and have not been imported into political science and economics. Certainly some disciplines, such as psychology and sociology, are more frequently concerned with the development of measures of constructs than others. But the need for evidence to support a hypothesis – in this case, a hypothesis about measurement – is not discipline-specific. Science always requires evidence to support claims. Many political scientists are familiar with the concept, and a key article on measurement validity was published in the *American Political Science Review* in 2001 (Adcock and Collier, 2001). Economists who deal with the development of measures of constructs have also been concerned with construct validity, particularly in work on measures of contingent valuation and economic education (see, for example, Meyer, 1995; Laughland *et al*, 1996; Carson *et al*, 1998; Kealy *et al*, 1988; Stolk and Busschbach, 2001; Girjalva *et al*, 2002).

Are the WGI Valid Measures of Governance?

The methodology raises several concerns. The first is that some of the constructs themselves are poorly defined and may be meaningless. The second is that the proposed measures depend on undefended and unlikely assumptions about the nature of governance. The last is that no evidence for construct validity has been presented; indeed, given the methodological choices, it is doubtful that it could be.

What Are the Constructs to be Measured?

The definition of the Control of Corruption indicator is 'the extent to which public office is exercised for private gain.' This reflects a widely used and accepted definition of corruption, although that definition itself is not without limitations. However, the remaining indicators either purport to measure constructs for which there is no theory or are divorced from existing theory. The indicators Regulatory Quality and Government Effectiveness are not associated with established theoretical literatures, and the authors do not explain how in theory they should be related to observables. There is theoretical literature dealing with voice, accountability and rule of law, but the WGI neither refer to nor correspond with the theoretical literature defining these concepts.

The indicator ‘Voice and Accountability’ is defined as ‘the extent to which a country’s citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association and free media’ (Kaufmann *et al*, 2005). In political science, the term ‘voice’ derives from Hirschman’s classic work ‘Exit, Voice and Loyalty.’ Hirschman (1970, p. 30) gives the following definition of ‘voice’:

> any attempt at all to change, rather than escape from, an objectionable state of affairs, whether through individual or collective petition to the management directly in charge, through appeal to a higher authority with the intention of forcing a change in management, or through various types of actions and protests including those that are meant to mobilize public opinion.

Accordingly, ‘voice’ is not synonymous with accountability, freedom to select government, or other political freedoms. Nor are there well-known or well-documented relationships among them.

The definition of the term ‘rule of law’ is the subject of wide debate in a post-natural law world. Fallon and Richard’s meta-analysis (1997, p. 8) of the many current definitions concludes that they refer to five constituent elements. These are (1) ‘the capacity of legal rules, standards or principles to guide people in the conduct of their affairs’; (2) efficacy; (3) stability; (4) ‘the supremacy of legal authority’ for both citizens and government actors; and (5) the availability of impartial institutions of enforcement. The definition of the WGI Rule of Law governance indicator ‘the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, the police and the courts, as well as the likelihood of crime and violence’ captures some elements of the rule of law, includes elements not traditionally incorporated in the concept, and omits others. As such, this construct does not have ‘content validity’ – it does not map to conceptions of the rule of law.

Kaufman *et al* (2007) defend their definition, arguing that it corresponds to a definition offered subsequently by Dakolias (2006, p. 1117):

> When the rule of law is in effect, there are: meaningful and enforceable laws where decisions are transparent, fair, and predictable; enforceable contracts that promote business and commerce; basic security with personal safety; protection of individual and property rights, and an independent judiciary that safeguards both; and access to justice with concrete ways to invoke rights and protect them.

But the WGI definition of rule of law again omits elements listed by Dakolias, including the existence of meaningful and enforceable laws, the transparency, fairness and predictability of decisions, and access to justice. Moreover, this list is arguably a list of consequences that Dakolias believes flow from the rule of law. Dakolias (2006, p. 1122) offers a definition of the rule of law earlier in the article:

> Rule of law may be defined as a system (1) in which the government itself is bound by the law; (2) in which all in society are treated equally under the law; (3) where the government authorities, including the judiciary, protect citizens’ aspirations for human dignity, and (4) which is accessible to its citizens.

This definition does not include crime levels or contract enforcement, and does include elements not captured by the WGI Rule of Law indicator such as equality under the law, the protection of citizen aspirations of dignity, and the accessibility of justice. The WGI definition of the Rule of Law does not correspond to this one either.

An examination of the underlying variables suggests that the WGI construct definitions, which have changed over time, are merely summary descriptions of the variables that the authors placed in the cluster, deriving from the act of clustering rather than guiding it. If this is the case, then the authors have not defined the underlying construct at all. The 2008 definitional changes by which the indicators were redefined to be measures of perception rather than of governance make the intended targets of measurement less clear.

The authors argue that a demand for content validity constitutes 'playing a game of definitional gotcha' that means that no measurement of a construct can take place in the absence of 'definitional consensus'' (Kaufmann *et al*, 2007). Content validity does not require a definitional consensus, but it does require that researchers rigorously define what it is they wish to measure before they set out to measure it, and that the definition has as much in common as possible with the way the construct is typically defined and used in theory. This process of definition allows an assessment of the validity of a proposed measure by checking the behavior of the measure against the predicted behavior of the construct. If the measure does not match the construct, or if the construct or measure has been invented in the absence of theory, there are no existing predictions of the construct's behavior against which to check the measure's validity. Moreover, if there are no predicted relationships with observables, it is not then clear why the new construct is meaningful, why one would develop a measure of it, or what correlations between such a measure and anything else would mean.

Does the Operationalization Capture the Construct to be Measured?

Once the construct is well defined, the researcher must specify an operationalization of the construct – a way of measuring it – in terms of a model describing the relationship between the construct and observable variables. It is critical that researchers explain and justify the assumptions that underlie their models, which are presumed to be informed by theory. Studemund's introductory economics text (1997) includes the caution 'Of all the kinds of mistakes made in applied regression analysis, specification error is usually the most disastrous to the validity of the estimated equation.'

The WGI rely on a large number of assumptions that could, if erroneous, make the WGI not noisy or imprecise data, but wrong data. Two sets of assumptions underlie the WGI model. The first assumptions are the clustering decisions in which particular variables are identified as noisy signals of a particular unobserved governance construct. The second set of assumptions consists of those that specify how those variables are related to unobserved governance. The assumptions of the model are neither intuitive nor justified by the authors. Indeed, where the constructs to be measured are poorly defined, it may be impossible to justify any of these modeling choices.

Clustering

Each variable is assumed to be a function of one, and only one, unobserved governance construct. Accordingly, the assignment of a variable to a cluster of variables used to derive an indicator of unobserved governance is a hypothesis about the nature of that variable and its relationship to unobserved governance. For example, the assignment of the variable 'when deciding upon policies and contracts, government officials favor well-connected firms' to the cluster of variables used to derive a measure of Voice and Accountability represents a hypothesis that the variable is determined by Voice and Accountability but not by Control of Corruption.

Clustering variables determined by different unobserved governance constructs would make the estimates meaningless. Separating variables determined by the same underlying governance construct into multiple clusters would result in unnecessarily large margins of error. Depending on the model and the data, clustering decisions could conceivably affect ultimate rankings. Box 1 illustrates how clustering decisions can determine relative country rankings in the case of a simple aggregate governance indicator. The impact of clustering decisions on the rankings of the WGI is unknown, but the potential shows the need for transparent and theoretically defensible clustering decisions.

Table 1 shows the cluster of underlying variables used to calculate the Voice and Accountability indicator. (For a list of all the variables for all the indicators, see Kaufmann *et al*, 2008.) However, the rationale for the assignment of variables to clusters is not explained or theoretically defended. For example, if Voice and Accountability

Box I. Clustering and Country Ranking with a Simple Indicator

In the example below, two countries, Blue and Green, have been rated with respect to four variables, G_1-G_4, scaled so that higher numbers mean better governance. An aggregate governance indicator is built that is a simple average of all variables in the cluster.

Raw Data

	G_1	G_2	G_3	G_4
Blue	8	1	10	3
Green	4	2	1	7

In Example 1, researchers have decided to cluster all four variables together in a single cluster, and to produce an aggregate governance indicator (Ind1) that is the average of the four variables. With this clustering decision, Blue ranks higher than Green.

Example 1. One Cluster

	Ind1
	$G_1G_2G_3G_4$
Blue	5.5
Green	3.5

In Example 2, researchers have decided to divide the four variables into two clusters to produce two aggregate governance indicators, Ind1 and Ind2. They must then decide which variables to assign to each cluster. Clustering 1 through Clustering 3 show three different possible assignments of variables to clusters. For example, in Clustering 1, variables G1 and G3 are clustered to produce aggregate governance indicator 1, while variables G2 and G4 are clustered to produce aggregate governance indicator 2. *Each assignment produces a different ranking.* This example illustrates the importance of clustering decisions. It is not known whether the clustering decisions made by the WGI Project – which uses a different model and different data -- affect the country rankings produced. However, because of the potential impact, special care must be taken to ensure that the clustering decisions are transparent and theoretically justified.

Example 2. Two Clusters

	Clustering_1		Clustering_2		Clustering_3	
	Ind1	Ind2	Ind1	Ind2	Ind1	Ind2
	G_1G_3	G_2G_4	G_1G_2	G_3G_4	G_1G_4	G_2G_3
Blue	2	2	4.5	6.5	5.5	5.5
Green	2.5	4.5	3	4	6.5	1.5

Table 1: Variables used to estimate voice and accountability for 2007

Voice and accountability	
Representative sources	
Source	*Concept measured*
EIU	Orderly transfers Vested interests Accountability of public officials Human rights Freedom of association
FRH	Civil liberties: Freedom of speech, assembly and demonstration, religion, equal opportunity, excessive governmental intervention Political rights: free and fair elections, representative legislative, free vote, political parties, no dominant group, respect for minorities
FRP	Freedom of the press
GCS	Newspapers can publish stories of their choosing without fear of censorship or retaliation When deciding upon policies and contracts, Government officials favor well-connected firms Effectiveness of national Parliament/Congress as a law making and oversight institution Passive voice
GWP	Confidence in the honesty of elections
HUM	Travel: domestic and foreign travel restrictions Freedom of political participation Imprisonments: Are there any imprisoned people because of their ethnicity, race, or their political, religious beliefs? Government censorship
IPD	Political rights and functioning of political institutions Freedom of the press Freedom of association Freedom of assembly and demonstration Respect for minorities (ethnic, religious, linguistic, and so on) Transparency of public action in the economic field Transparency of economic policy (fiscal, taxation, monetary, exchange-rate, and so on) Award of public procurement contracts and delegations of public service Free movement of persons, information, and so on.
PRS	Military in politics: The military are not elected by anyone, so their participation in government, either direct or indirect, reduces accountability and therefore represents a risk. The threat of military intervention might lead as well to an anticipated potentially inefficient change in policy or even in government. Democratic accountability: Quantifies how responsive government is to its people, on the basis that the less response there is

Table 1: *continued*

Voice and accountability

Representative sources

Source	*Concept measured*
	the more likely it is that the government will fall, peacefully or violently. It includes not only if free and fair elections are in place, but also how likely it is that the government will remain in power.
RSF	Press freedom index
WMO	Institutional permanence: An assessment of how mature and well-established the political system is. Representativeness: How well the population and organized interests can make their voices heard in the political system.
Non-representative sources	
AEO	Hardening of the regime
AFR	Elections are free and fair
BTIS	Stateness Political participation Institutional stability Political and social integration
CCR	Civil liberties Accountability and public voice
GII	Civil society organizations Media Public access to information Voting and citizen participation Election integrity Political financing
IFD	Policy and legal framework for rural organizations Dialogue between government and rural organizations
LBO	Satisfaction with democracy Trust in parliament
MSI	Media sustainability index
OBI	Open Budget Index
VAB	Trust in parliament Satisfaction with democracy
WCY	Transparency of government policy

This table is reproduced with minor revisions from Kaufmann *et al* (2008).

measures ‘the extent to which a country’s citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association and free media,’ what is the hypothesized relevance of the indicator ‘when deciding upon policies and contracts, government officials favor well-connected firms’ (Kaufmann *et al*, 2008)? Would it be better assigned to the Control of Corruption cluster? What is the hypothesized relevance of the indicator ‘”Institutional Stability” and what does it mean? Would this variable be better assigned to the Government Effectiveness’ cluster?

Some researchers have begun to question the way in which variables are clustered. For example, Kurtz and Schrank (2007) agree with the WGI’s definition of the government effectiveness construct, but are not convinced by the operationalization. The indicator was defined as 'the competence of the bureaucracy and the quality of public service delivery' (Kaufmann *et al*, 2005). Kurtz and Schrank (2007) argue that the variables included in the cluster included measures of specific policy choices, which were extraneous to the definition and introduce systematic bias into the measures. Similarly, with respect to the Rule of Law indicator, Tamanaha (2008, p. 29) points to the heavy weighting of variables concerned with the protection of property rights. 'This makes sense from the standpoint of the business community wishing to conduct commerce in developing countries – which matters to economic development – but nonetheless it is an exceedingly narrow view of the rule of law, inconsistent with the multifaceted definition set forth elsewhere in this report'.

There is reason to ask whether the clusters succeed in isolating different aspects of governance. In some cases, the correlations between the governance indicators are so high that they raise the question of whether the indicators are measuring distinct constructs. For the year 2007, for example, the Government Effectiveness and Regulatory Quality indicators are correlated at 0.95; the Rule of Law and the Control of Corruption indicators are correlated at 0.94; the Government Effectiveness and Control of Corruption indicators are correlated at 0.93; and the Government Effectiveness and the Rule of Law indicators are correlated at 0.92 (see Kaufmann *et al*, 2008).

The assumptions underlying the clustering methodology must be clearly articulated and theoretically defended. However, the assumptions are not theoretically justified by the authors; they reflect an unarticulated and purely personal conception of governance: 'This classification of indicators into clusters corresponding to this definition of governance is not intended to be definitive. Rather, it simply reflects our views of what constitutes a consistent and useful organization of the data that is concordant with prevailing notions of governance' (Kaufmann *et al*, 2005, p. 130).

Model specification

The model used to construct the indicators assumes that error terms are uncorrelated across data sources (Kaufmann *et al*, 1999a, p. 2; 2008, p. 97), that is, across variables and source variables. The authors recognize this assumption as 'a strong one' that represents 'a best case scenario,' and considered several reasons why errors might not be independent across variables and source variables. (Kaufmann *et al*, 1999a, p. 10) They cautioned that 'it is useful to realize that the estimated standard errors associated with point estimates of governance are likely to be substantially understated under the assumption of independent errors,' and noted that 'the relative rankings of the countries could be affected' (Kaufmann *et al*, 1999a, p. 21). Several possible sources of correlated errors have been flagged by researchers (see Kaufmann *et al*, 2007), and the authors have addressed the possibility that correlated errors could be created if experts shared common preconceptions or prejudices (Kaufmann *et al*, 2006a, p. 19).

There is, however, another possible source of correlated errors – specification error, which can lead to serial correlation (Studemund, 1997, p. 333). The model assumes that each variable and source variable in a cluster is a function only of the unobserved governance construct for that cluster. Accordingly, just as the act of assigning a variable to a cluster represents a hypothesis that the variable is determined by the unobserved governance component for that cluster, the assumptions of the model represent hypotheses that the variable or source variable is determined *only* by the unobserved governance component. If this is not true – if, for example, two variables or source variables share a common omitted independent explanatory variable – this would violate the assumptions of the model, leading to serial correlation. In addition to larger standard errors, such a specification error could lead to inconsistent parameter estimates (White, 1982), which in turn would nullify any claimed advantages of including more data in the model.[6]

How likely is it that each variable and source variable are a function of only a single explanatory variable – the unobserved governance component? An examination of the variables in each cluster shows a number of theoretically plausible functional relationships among variables from different sources.[7] Is it likely that 'freedom of the press' (Freedom House), 'freedom of association' (Economist Intelligence Unit) and 'democratic accountability' (Political Risk Services International Country Risk Guide) are determined by unobserved 'Voice and Accountability' (Kaufmann *et al*, 2008) but that freedom of the press and freedom of association do not contribute to democratic accountability? Is it likely that 'revenue mobilization' (African Development Bank Country Policy and Institutional Assessment), 'management of external debt' (CPIA) and 'government handling of education' (Afrobarometer Survey) are determined only by unobserved 'Government Effectiveness' but that revenue mobilization does not have an impact on management of external debt and the government handling of education?

Each variable or source variable used in the construction of the indicators represents a hypothesis about the determinants of that variable or source variable; and because of the assumption of independent error terms, its inclusion also represents a hypothesis about how that variable or source variable relates to every other variable and source variable already used in estimation. While researchers are usually urged to be parsimonious in the development of models and theories, the large number of variables used in constructing the governance indicators means that there are many such assumptions, and none of them have been theoretically or empirically justified by the authors. The 2007 Voice and Accountability governance indicator, for example, is derived from 54 underlying variables. As more variables are included, the number of underlying assumptions increases along with the risk of specification error. The result is a complex set of hypotheses about both the nature of governance and about measurement.

Lack of Evidence of Construct Validity

What evidence supports the WGI hypothesis? Kaufmann *et al* (2007) argue that the indicators show convergent validity because 'the various individual components of our governance indicators are quite highly correlated with each other within each of the six governance indicators.' However, for a measure to show convergent validity, the proposed measure must be correlated with variables with which theory predicts *the construct* would be correlated. They say nothing about correlations among variables used to construct the measure. The question is, how does the governance indicator itself behave, compared to theoretical predictions?

Kaufmann *et al* argue that the concept of discriminant validity is not relevant to their work because they are attempting to develop several measures of constructs that they assume are likely to be correlated. However, this argument misunderstands the nature of discriminant validity. A measure of a construct demonstrates discriminant validity when it is shown not to be correlated with variables that *theory predicts* are not correlated with the construct.

Neither convergent validity nor discriminant validity can be shown without an appeal to theory, and the authors offer no other argument for the construct validity of the indicators. The WGI model stands as an elaborate unsupported hypothesis about the nature of governance.

Evidence of the construct validity of the WGI is likely to be very difficult to produce, for several reasons. The first is that any showing that a construct has been meaningfully operationalized is an exercise rooted in theory about the construct. Most critically, researchers need to be able to make predictions derived from the theory and test their operationalization against those predictions. But most WGI constructs are not based in theory, even where theory exists. No predictions have been made for the relationship between these constructs and observables, and therefore no evidence of convergent or discriminate validity can be advanced.

While the authors of the WGI have offered no evidence in support of their indicators, and do not address their construct validity, Kurtz and Schrank (2006) have attempted to explore the convergent and discriminant validity of the Government Effectiveness indicator. They consider the correlation between the Government Effectiveness indicator and a number of other variables, such as the corruption indicator produced by Transparency International, the 'country risk' indicator produced by the International Country Risk Guide, a measure of 'Weberianess' produced by Evans and Rausch, the level of adult education, the size of the population, and the rate of recent antecedent economic growth. They interpret the strength or weakness of the correlations between the Government Effectiveness indicator and these variables, but the origin of their predictions of the relationship between the indicator and these variables is not clear, because the authors of the WGI have not situated the 'Government Effectiveness' construct in any body of theory.

A second reason why it may be difficult to produce evidence of construct validity is the authors' decision to encompass as much available data as possible in the construction of the governance indicators. This approach leaves few independent sources that could be used either for comparison or model-testing.

A final reason why it may be difficult to produce evidence of construct validity is that many of the underlying variables used to produce the indicators are not in the public domain. Table 2 shows the availability of the data sources and data sets used to construct the WGI as of this writing. The WGI relies on three data sets that are partially confidential and variables from 11 other proprietary data sets for which in most cases it publishes only the source variables. However, without access to the entirety of the raw data set, neither the averaging nor the clustering decisions can be reviewed.[8] Without full public data access, it is impossible for other researchers to critique, improve or build on the indicators in the iterative process of theory justification and refinement described by Smith (2005). Without permission to publish the variables, it would be difficult for researchers to offer evidence if they had it.

Because the operationalization of the indicators depends on a very large number of undefended assumptions about the nature of the data and the relationship among variables, there is reason to seek reassurance of the validity of the indicators. However, the constructs that the WGI seek to measure are neither defined nor rooted in theory; the indicators have not been tested against the behavioral predictions of theory and so the indicators are meaningless without more.

Conclusion

This paper has focused on the WGI because of rising third-party use and, in particular, the reliance on the WGI for important policy decisions regarding foreign aid. But the concerns

Table 2: Data sources and data sets, 2007

Data sources and data sets 2007	*Confidential*	*Freely and publicly available*	*Variables used published by WGI*	*Average weight*
African Development Bank Country Policy and Institutional Assessments (ADB)	Partial	N	N	0.085
OECD Development Center African Economic Outlook (AEO)				0.017
Afrobarometer (AFR)				0.026
Asian Development Bank Country Policy and Institutional Assessments (ASD)	Partial	N	N	0.045
Business Environment & Enterprise Performance Survey (BPS)				0.006
Business Environment Risk Intelligence (BRI)		N	N	0.056
Business Environment Risk Intelligence (QLM)		N	N	0.075
Bertelsmann Transformation Index (BTI)				0.061
Global Insight Global Risk Service (DRI)		N	N	0.043
European Bank for Reconstruction and Development Transition Report (EBR)				0.086
Global E-Government Index (EGV)				0.008
Economist Intelligence Unit (EIU)		N	N	0.083
Freedom House (FRH, CCR)				0.155
Transparency International Global Corruption Barometer Survey (GCB)				0.006
World Economic Forum Global Competitiveness Survey (GCS)		N	N	0.041
Global Integrity Index (GII)				0.003
Gallup World Poll (GWP)		N	N	0.004
Heritage Foundation Index of Economic Freedom (HER)				0.005
Cingranelli Richards Human Rights Database & Political Terror Scale (HUM)				0.038

IFAD Rural Sector Performance Assessments (IFD)				0.018
iJET Country Security Risk Ratings (IJT)		N	Y	0.094
Institutional Profile Database (IPD)				0.071
Latino-Barometro (LBO)		N	N	0.001
Merchant International Group Gray Area Dynamics (MIG)		N	N	0.066
International Research and Exchanges Board Media Sustainability Index (MSI)				0.048
International Budget Project Open Budget Initiative (OBI)				0.029
World Bank Country Policy and Institutional Assessments (PIA)	Partial	N	N	0.055
Political Economic Risk Consultancy Corruption in Asia (PRC)		N	Y	0.064
Political Risk Services International Country Risk Guide (PRS)		N	N	0.045
Reporters without Borders Press Freedom Index (RSF)				0.032
US State Department's Trafficking in People Report (TPR)				0.004
Vanderbilt University Americas Barometer (VAB)		Partial	Y	0.017
Institute for Management Development World Competitiveness Yearbook (WCY)		N	N	0.046
Global Insight Business Conditions and Risk Indicators (WMO)		N	N	0.095

The data sources and data sets are listed in the left column. Data availability is indicated in the 'Confidential' and 'Freely and Publicly Available' columns. Unless otherwise noted, data are freely and publicly available and the variables used by the WGI are published by the WGI project. The WGI Project does not publish confidential data. The average weight accorded to the data source in the construction of the WGI is given under 'Average Weight'. The two data sets from Business Environment Risk Intelligence are given separate entries to preserve information about their average weight in the construction of the indicators. Variables that are not freely and publicly available and are not published by the WGI are more heavily weighted under the WGI aggregation methodology. *Source*: Kaufmann *et al* (2008), Tables A1–A33 and Table 3.

raised here about the WGI apply equally to other current governance indicators. Of this group, the WGI are notable because of the authors' efforts to document and make public the methodology for producing the indicators, to employ a statistically sound approach, to seek precision, and to both calculate and emphasize the importance of margins of error. Moreover, since the initial release of the WGI, the authors have made much of their underlying data easily available. On the whole, the authors have done a large service by focusing the development community on the governance issue, by moving the discussion in an empirical direction, and by highlighting the challenges of measurement and the importance of attention to random error.

Since the 1950s, however, public evidence of construct validity has been required for proposed measurements of constructs. In the absence of the rigorous examination of the inferences involved in creating a measurement, authors such as Underwood (1957, p. 55) and Campbell and Fiske (1959, p. 101) recognized 'the danger … that the investigator will fall into the trap of thinking that because he went from an artistic or literary conception … to the construction of items for a scale to measure it, he has validated his artistic conception'. This is the very danger presented by the WGI, which are admittedly based on personal and untested notions of governance. The WGI claim to measure governance; as yet no evidence has been offered that this is true. The WGI represent a complex atheoretical and as yet poorly articulated hypothesis for which no evidence has been advanced.

Recognition that a governance indicator is a hypothesis about measurement and about the nature of governance is a prerequisite for validation and for improvement of the measures. Developing a meaningful measurement of a construct is an iterative process that involves a theoretical specification of the construct and its relationship with observable variables, model-testing as against predictions, and refinement. The process of accumulating evidence to support a measurement hypothesis is one that involves the scientific community as a whole, and, as such, an investigator must provide evidence of construct validity and make available the raw data, models and results that would allow the community to make independent judgments. In the words of Cronbach and Meehl (1955, p. 296), 'Defending a claim of construct validity is a major task, not to be satisfied by a discourse without data … . A claim is unsubstantiated unless the evidence for the claim is public, so that other scientists may review the evidence, criticize the conclusions, and offer alternative interpretations.' The next generation of governance indicators must be theoretically defensible, supported by evidence of validity, and based on freely and publicly accessible data to allow broad review by the scientific community.

Unfortunately, meaningful measures of governance require as a prerequisite specific definitions of governance that draw from available theory. Developing such definitions is technically challenging because the theory is not yet well developed, but it is made even more difficult because in the foreign policy arena the definition of good governance is highly politicized (Thomas, 2007). A liberal democracy and an authoritarian dictatorship can both agree on the importance of the rule of law, provided that the former means 'a state constrained by rules' and the latter means 'citizen obedience to government edicts'. The consensus on the importance of good governance may not survive if it requires agreement on the details, and measures derived from specific and idiosyncratic definitions may not find wide audiences.

Despite the evident demand for global measures of the quality of governance, the work is still in its infancy, and usage is premature. But while policymakers may sometimes be obliged to rely on the best information available, there is a difference between bad or 'noisy' data and wrong data. Both researchers and policymakers should require evidence that governance indicators are valid before employing them. In the absence of such evidence, research results obtained using such indicators are uninterpretable and should not survive peer review. For policymakers, reliance on such indicators would be arbitrary.

Acknowledgement

The initial draft of this paper, under the same title, was written and posted in 2006; a revision was posted in 2007. The author thanks Ray Koopman, Francis Fukuyama, David Armor, Daniel Kaufmann, Aart Kraay, Richard Messick, Anwar Shah, Michael Johnston, Omar Azfar, Stephen Knack, Simeon Djankov, Alice Fredericks and the anonymous referees of this journal for their thoughtful comments and insights, as well as her research assistant, Scott Merrill, for his able assistance.

Notes

1. The indicators have had various names over the years, but this paper uses the name used on the World Bank's website and most recently by the authors.
2. For a description of available data sources, see Gray (2007) Governance for economic growth and poverty reduction: Empirical evidence and new directions reviewed. http://www.gsdrc.org/docs/open/RET422.pdf.
3. See the disclaimer at www.govindicators.org that reads 'Disclaimer: The aggregate indicators do not reflect the official views of the World Bank, its Executive Directors, or the countries they represent. The WGI are not used by the World Bank Group to allocate resources or for any other official purpose.'
4. Another necessary property of a measure of a construct is reliability. Validity and reliability are usually discussed together. A measure is 'reliable' 'to the extent to which an experiment, test, or any measuring procedure yields the same results on repeated trials' (Carmines and Zeller, 1979). The WGI will be reliable to the extent that the underlying individual indicators are reliable. Accordingly, the reliability of the WGI is beyond the control of the authors and will not be discussed here, except to note that the authors recognize the importance of measurement error and have paid considerable attention to minimizing the impact of such error.
5. These predicted relationships were dubbed the 'nomological network' by Cronbach and Meehl (1955).
6. A consistent estimator is one that has a greater probability of converging on the population parameter the greater the sample size. (Wooldridge, 2003). An inconsistent estimator does not; increasing the sample size is no guarantee of improved accuracy.
7. A list of the underlying variables used to construct each governance indicator appears in the appendices of the study by Kaufmann *et al* (2008). Table 1 in this paper reproduces for convenience the list of variables that are used to construct the Voice and Accountability indicator.
8. Confidential data are not available to researchers. While in theory researchers could purchase the proprietary data, many are likely to find it to be cost-prohibitive. For example, a one-year subscription to the Gray Area Dynamics data set produced by Merchant International Group for an individual academic researcher would cost $10 000 (private communication with Paddy Breiner, 20 February 2009); this is just one of the proprietary data sets. Further, as Schrank points out, the definitions and methods of collection of the underlying variables should also be available; where they themselves are aggregates, the raw data and the method of aggregation should be made available (private conversation with Andrew Schrank (8 March 2007)).

References

Adcock, R. and Collier, D. (2001) Measurement validity: A shared standard for qualitative and quantitative research. *The American Political Science Review* 95(3): 529–546.

Andres, A.R. (2006) Software piracy and income inequality. *Applied Economics Letters* 13(2): 101–105.

Apodaca, C. (2004) The rule of law and human rights. *Judicature* 87(6): 292–299.

Arndt, C. and Oman, C. (2006) *Uses and abuses of governance indicators*. Paris: OECD Development Centre.

Bartholomew, D.J. (1995) Spearman and the origin and development of test theory. *British Journal of Mathematical and Statistical Psychology* 48: 211–220.

Bollen, K.A. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K.A. (1993) Liberal democracy: Validity and method factors in cross-national measures. *American Journal of Political Science* 37(4): 1207–1230.

Bollen, K.A., Glanville, J.L. and Stecklov, G. (2002) Economic status proxies in studies of fertility in developing countries: Does the measure matter? *Population Studies* 56: 81–96.

Burnside, C. and Dollar, D. (2000) Aid, policies, and growth. *American Economic Review* 90(4): 847–868.

Burnside, C. and Dollar, D. (2004) Aid, policies, and growth. *American Economic Review* 94(3): 781–784.

Campbell, D.T. and Fiske, D.W. (1959) Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* 56: 81–105.

Carmines, E.G. and Zeller, R.A. (1979) *Reliability and Validity Assessment*. Beverly Hills, CA: Sage Publications.

Carson, R.T. *et al* (1998) Referendum design and contingent valuation: The NOAA panel's no-vote recommendation. *The Review of Economics and Statistics* 80(2): 335–338.

Cronbach, L. and Meehl, P. (1955) Construct validity in psychological tests. *Psychological Bulletin* 52(4): 281–302.

Dakolias, M. (2006) Are we there yet? Measuring success of constitutional reforms. *Vanderbilt Journal of Transnational Economics* 39(4): 1117–1231.

Das, G.G. and Andriamananjara, S. (2006) Hub-and-spokes free trade agreements in the presence of technology spillovers: An application to the western hemisphere. *Review of World Economics* 142(1): 33–66.

Dollar, D. and Kraay, A. (2003) Institutions, trade and growth. *Journal of Monetary Economics* 50(1): 133–162.

Easterly, W., Levine, R. and Roodman, D. (2004) Aid, policies and growth: Comment. *American Economic Review* 94(3): 774–780.

Eid, M., Lischetzke, T., Nussbeck, F.W. and Trierweiler, L.I. (2003) Separating trait effects from trait-specific method effects in multitrait–multimethod models: A multiple-indicator ct-c(m-1) model. *Psychological Methods* 8: 36–60.

Elkins, Z. (2000) Gradations of democracy? Empirical tests of alternative conceptualizations. *American Journal of Political Science* 44(2): 293–300.

Faber, J. (1987) Measuring cooperation, conflict and social network of nations. *The Journal of Conflict Resolution* 31(3): 438–464.

Fallon, J. and Richard, H. (1997) "The rule of law" as a concept in constitutional discourse. *Columbia Law Review* 97(1): 1–56.

Filmer, D. and Pritchett, L. (2001) Estimating wealth effects without income or expenditure data – or tears: Educational enrollment in india. *Demography* 38(1): 115–132.

Forgas, J.P. (1980) Images of crime: A multidimensional analysis of individual differences in crime perception. *International Journal of Psychology* 15(1): 287–299.

Forsythe, G.B., McGaghie, W.C. and Friedman, C.P.P. (1986) Construct validity of medical clinical competence measures: A multitrait–multimethod matrix study using confirmatory factor analysis. *American Educational Research Journal* 23(2): 315–336.

Girjalva, T.C., Berrens, R.P., Bohara, A.K. and Shaw, W.D. (2002) Testing the validity of contingent behavior trip responses. *American Journal of Agricultural Economics* 84(2): 401–414.

Gray, H. (2007) Governance for economic growth and poverty reduction: Empirical evidence and new directions reviewed, http://www.gsdrc.org/docs/open/RET422.pdf.

Hammond, K.R., Hamm, R.M. and Grassia, J. (1986) Generalizing over conditions by combining the multitrait–multimethod matrix and the representative design of experiments. *Psychological Bulletin* 100: 257–269.

Hart, D., Atkins, R. and Youniss, J. (2005) Knowledge, youth bulges, and rebellion. *Psychological Science* 16(8): 661–662.

Hirschman, A.O. (1970) *Exit, voice, and loyalty*. Cambridge, MA: Harvard University Press.

International Development Association. (2004) *IDA's performance-based allocation system: IDA rating disclosure and fine-tuning the governance factor*. Washington DC: International Development Association.

Iqbal, K. and Shah, A. (2008) How do the Worldwide Governance Indicators measure up? http://siteresources.worldbank.org/PSGLP/Resources/Howdoworldwidegovernanceindicatorsmeasureup.pdf.

Johnston, M. (2000) The new corruption rankings: Implications for analysis and reform. Paper prepared for Research Committee 24, International Political Science Association World Congress; 2 August, Quebec City, Canada.

Jung, M. (2006) Host country attractiveness for cdm non-sink projects. *Energy Policy* 34(15): 2173–2184.

Kaufmann, D. and Kraay, A. (2002) Growth without governance. *Economia* 3(1): 169–229.

Kaufmann, D. and Kraay, A. (2003) *Governance and growth: Causality which way? – evidence for the world, in brief*. Washington DC: World Bank.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2004) Governance matters III: Governance indicators for 1996, 1998, 2000, and 2002. *The World Bank Economic Review* 18(2): 253–288.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2005) *Governance matters IV: Governance indicators for 1996–2004*. Washington DC: World Bank.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2006a) *Governance matters V: Aggregate and individual governance indicators for 1996–2005*. Washington DC: World Bank.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2006b) *The Worldwide Governance Indicators project: Answering the critics*. Washington DC: World Bank.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2007) *Governance matters VI: Aggregate and individual governance indicators 1996–2006*. Washington DC: World Bank.

Kaufmann, D., Kraay, A. and Mastruzzi, M. (2008) *Governance matters VII: Aggregate and individual governance indicators 1996–2007*. Washington DC: World Bank.

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

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

Kaufmann, D., Kraay, A. and Zoido-Lobatón, P. (2002) *Governance matters II: Updated indicators for 2000/01*. Washington DC: World Bank.

Kealy, M.J., Dovidio, J.F. and Rockel, M.L. (1988) Accuracy in valuation is a matter of degree. *Land Economics* 64(2): 158–171.

Knack, S. (2006) *Measuring corruption in Eastern Europe and Central Asia: A critique of the cross-country indicators*. Washington DC: World Bank.

Kurtz, M. and Schrank, A. (2006) Growth and governance: Models, measures and mechanisms. *Journal of Politics* 69(2): 538–554.

Kurtz, M. and Schrank, A. (2007) Growth and governance: A defense. *Journal of Politics* 69(2): 563–569.

Laughland, A.S., Musser, W.N., Shortle, J.S. and Musser, L.M. (1996) Construct validity of averting cost measures of environmental benefits. *Land Economics* 72(1): 100–112.

Liu, M.C. and San, G. (2006) Social learning and digital divides: A case study of internet technology diffusion. *Kyklos* 59(2): 307–321.

Llamazares, I. (2005) Patterns in contingencies: The interlocking of formal and informal political institutions in contemporary Argentina. *Social Forces* 83(4): 1671–1695.

McKenzie, D.J. (2005) Measuring inequality with asset indicators. *Journal of Population Economics* 18: 229–260.

Méon, P.-G. and Sekkat, K. (2005) Does corruption grease or sand the wheels of growth? *Public Choice* 122(1–2): 69–75.

Meyer, B.D. (1995) Natural and quasi-experiments in economics. *Journal of Business and Economic Statistics* 13(2): 151–161.

Naude, W.A. (2004) The effects of policy, institutions and geography on economic growth in Africa: An econometric study based on cross-section and panel data. *Journal of International Development* 16: 821–849.

Neumayer, E. (2002) Do democracies exhibit stronger international environmental commitment? A cross-country analysis. *Journal of Peace Research* 39(2): 139–164.

North, D. and Thomas, R.P.T. (1996) *Rise of the western world: A new economic history*. Cambridge: Cambridge University Press.

Olken, B.A. (2006) Corruption perceptions v. corruption reality. NBER Working Paper No. 12428.

Patterson, M.L. (1990) On the construct validity and developmental course of rapport. *Psychological Inquiry* 1(4): 320–321.

Perry, J.L. (1996) Measuring public service motivation: An assessment of construct reliability and validity. *Journal of Public Administration Research and Theory: J-PART* 6(1): 5–22.

Pfeiffer, C. (2005) Media use and its impacts on crime perception, sentencing attitudes and crime policy. *European Journal of Criminology* 2(3): 259–285.

Razafindrakoto, M. and Roubaud, F. (2006) *Are international databases on corruption reliable? A comparison of expert opinion surveys and household surveys in Sub-Saharan Africa*. Paris: DIAL.

Seligson, M.A. (2006) The measurement and impact of corruption victimization: Survey evidence from Latin America. *World Development* 34(2): 381–404.

Smith, G.T. (2005) On construct validity: Issues of method and measurement. *Psychological Assessment* 17(4): 396–408.

Stolk, E.A. and Busschbach, J.J.V. (2001) The comparison of the euroqol and the health utilities index in patients treated for congential anomalies. *The European Journal of Health Economics* 2(2): 54–59.

Studemund, A.H. (1997) *Using econometrics: A practical guide*. Reading, MA: Addison-Wesley Educational Publishers.

Tamanaha, B.Z. (2008) The dark side of the relationship between the rule of law and liberalism. *NYU Journal of Law and Liberty*, Vol. 33; St. John's Legal Studies Research Paper No. 08-0096.

Thomas, M.A. (2007) The governance bank. *International Affairs* 83(4): 729–745.

Underwood, B.J. (1957) *Psychological Research*. New York: Appleton-Century-Crofts.

Van de Walle, S. and Bouckaert, G. (2007) Perceptions of productivity and performance in Europe and the United States. *International Journal of Public Administration* 30(11): 1123–1140.

Wallace, R.B. and Herzog, A.R. (1995) Overview of health measures in the health and retirement study. *The Journal of Human Resources* 30: S84–S107.

Westen, D. and Rosenthal, R. (2003) Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology* 94: 608–618.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50(1): 1–26.

Williams, R.H., Zimmerman, D.W., Zumbo, B.D. and Ross, D. (2003) Charles Spearman: British behavioral scientist. *Human Nature Review* 3: 114–118.

Wooldridge, J.M. (2003) *Introductory Economics: A Modern Approach*. US: South-Western.

World Bank. (1994) *Governance: The World Bank's experience*. Washington DC: World Bank.

World Bank. (1998) *Assessing aid – what works, what doesn't, and why*. Washington DC: World Bank.

World Bank. (2006a) *Strengthening Bank Group engagement on governance and anticorruption*. Washington DC: World Bank.

World Bank. (2006b) *World Bank releases largest available governance data source*. Washington DC: World Bank.