Strictly speaking, the numerical title of this note merely refers to the number of issues raised in it. But the use of a mere number in the title is also intended to signal concern at the high current ratio of text words per data points (or the ratio of words per statistical observation or finding) present in a number of critiques and discussions of governance indicators (many of the memo contributions to this symposium are exceptions, however). Thus, after the end of the text, this note also includes some evidence-based charts, aiming at empirically illustrating some of the tenets suggested here.

1. *The ‘True’ Quality of Governance is Inherently Unobservable.* What is the ‘quality of democracy’ in a country? Or its actual extent of corruption? Or the effectiveness of its government and of its rule of law? The ‘true’ level of governance cannot be directly observed at the national or sub-national level due to its very complex and multi-dimensional nature, further compounded by the incentives of agents to hide the actual state of governance within their own midst. The problem is magnified further by the lack of definitional consensus about what constitutes governance and its key components, as well by measurement challenges (and at times the political manipulation of data from official bodies, etc.).

If we start from this assumption on the inherent statistical unobservability of the true level of governance, important corollaries follow. At a trivial level, it takes care of the critique that somebody else’s index is a mere proxy while one’s own is the real (or ‘true’) thing. And it thus permits to move to the more fundamental issue, namely margins of error.

2. *Transparency and Precision about the Imprecision in Governance Indicators: taking margins of error seriously.* We can only observe proxies for governance, and that means that each proxy will by definition be an imprecise approximation of the true value.

This implies that each measured governance variable has a margin of error associated with its country estimate. Although rarely done, such margin of error can be estimated and disclosed. Doing so is of paramount importance for appropriate use and interpretation by researchers, advocacy groups, policy analysts and policy-makers. Margins of error estimates associated with each point estimate (and thus confidence intervals) provide rich information as to what kind of inferences can and cannot be made; which cross-country and over time comparisons are statistically meaningful, and which ones are not. See Exhibits 1a-d after the text.

---

1 Senior Fellow, Brookings Institution. The assistance of V. Penciakova is appreciated. Full disclosure upfront: this note is not directly focused on the Worldwide Governance Indicators project (WGI, at [www.govindicators.org](http://www.govindicators.org)) that A. Kraay and I have been involved with for a dozen years. Nor it attempts to provide with a comprehensive survey of governance indicators (a survey article is at [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1156202](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1156202)). Instead the aim here is to contribute to helping frame a debate on a number of particular challenges related to index choice, construction, and interpretation. The style of the text headings may give the impression that definitive assertions are being made. The actual intention is to put forth in brief some tentative tenets for debate.
Typically, the margins of error involved in estimates of governance are large enough to make precise rankings among countries with similar ratings meaningless (since they are locked in a ‘statistical tie’). Yet in many cases such margins of error are not large enough to render the indicator useless. In worldwide governance indices, one can usually find at least three or four clusters of countries which are meaningfully different from each other in the statistical significance sense.

3. Benefits of Aggregation and of Composite Indicators. While it is common to uncritically extol the virtues of disaggregated indicators, the benefits of aggregation are often underestimated. For starters, in a field where all individual indicators are an imperfect proxy of the true phenomena, the aggregation of various independent proxies for the same concept (say, corruption) can significantly reduce the margins of error (and the likelihood of obtaining nonsensical outliers). Drawing from four independent sources reduces the standard errors by one-half as compared to drawing from an individual source.

Composite indicators can also serve as an empirically important organizing framework for what often is a disparate array of anarchic data. This is particularly the case when the aim is to have a proxy for a relatively broad concept. If we want a proxy for the quality of democracy or rule of law in a country, a composite that combines the various key dimensions of such broad notions would tend to do better than a narrow proxy such as whether elections took place, or the level of trust in the police.

Obviously, if the scholar or policy maker is interested in a very narrow and specific manifestation of a governance component a disaggregated indicator would often do better, although one can argue that if two separate surveys happen to cover the same specific issue, the composite of those two may be less imprecise than relying only on one of those variables.

Further, the fact that complementarity among different indicators is both feasible and desirable is often lost in the competition. We note the obvious: to construct composite indicators one relies on individual variables, so it is not difficult to make both available, as they complement each other. And even for the same researcher or policy-maker, these can serve different purposes at different times.

4. To Weight or Not to Weight variables in aggregating?: That is [Not] the Question. Instead, the Question is: What is In, and what is Out? In contrast with common assertions, weights are always assigned to each individual variable used in aggregation. The confusion arises because some indices do not assign explicit weights, but in fact implicitly assign equal weights. Yet this matters less than other considerations, particularly the question of: what was left out altogether?

Often ‘errors of commission’ receive inordinate attention. By contrast, ‘errors of omission’ are overlooked; yet a relevant variable left out implies assigning to it a weight of zero. This makes much more of a material difference than arguments at the margin about weights among included variables. For instance, in its Corruption Perception Index (CPI), Transparency International (TI) excludes any relevant corruption variable from citizen surveys.

Further, inclusion of ‘zombie’ variables can have unintended consequences. Say that two variables inform a corruption aggregate index, one ‘subjective’ (reports from citizens and...
firm managers on their experiences with bribery), and one ‘objective’ (adoption of 
anticorruption legislation by the country). Imagine that the former has a normal distribution, 
with its resulting rich variance around the mean, while the latter has none, because all 
countries have anticorruption laws in their books. Under these circumstances, aggregating the 
two variables ‘without’ weights (read: equal implicit weights) is futile, since only one 
variable, the subjective one, will have any bearing on the composite ranks – it contributes 
100% of the variance, irrespective of the implicit or explicit weights given in adding them up. 
This may be an extreme example for the sake of illustration, but the use of such ‘zombie’ (or 
other ‘dumb’) variables does take place.2

5. Expert Polls vs. Citizen/Firm Surveys vs. ‘Objective’ Data: A Moot Contest?
Recognition that any of these methods serve as valuable efforts to approximate the actual 
state of governance implies that a priori there is no pecking order, since they all have margins 
of error (and different types of biases). Debates on methodology are often colored by the 
particular method the scholar has promoted or used, rather than by concerns arising from 
statistically rigorous tests applied to each method.

For the sake of space, let me make two brief observations here. First, statistical estimations 
using many different sources suggest that the variance in imprecision among individual 
sources within each category (experts, surveys, and ‘objective’ data) is large, and dwarfs any 
average differences in imprecision across these 3 categories.

There are both noisy and more accurate indicators that come from experts, surveys and from 
‘harder’ data. This is not surprising, since each type of indicator has its own set of ‘pros’ and 
‘cons’. For instance, some expert polls may have the advantage (in principle at least) of better 
uniformity in benchmarking cross-sectionally. But surveys, when well designed (in terms of 
sampling, questionnaire instrument, etc.) have the advantage of the much larger number of 
respondents, and also is a more direct way of collecting information on concrete local 
constraints from the ground up.

So called ‘objective’ data, such as the codification of formal constraints to firms in the Doing 
Business project of the World Bank, has the advantage of clarity and ease of monitoring. 
Yet, like most ‘objective’ indicators (other than official statistics, which suffer from other 
drawbacks), they also involve a judgment (by a lawyer in their offices). Further, such data 
refers to the ‘de jure’ codification of rules, ignoring the ‘de facto’ outcomes on the ground. 
We find that the de facto often deviates substantially from the de jure, the result of the role of 
informal rules and institutions in the application of formal rules. Such deviation is 
particularly marked in developing countries. See Exhibits 2a-b and 3 at the bottom.

6. Staying away from Obese and Anorexic Definitions Enables Parallel Progress on 
Conceptual and Empirical Areas. Obese notions in governance, such as those that attempt 
to encompass development outcomes within its definition, are counterproductive. First, they 
fail the falsifiability test of ‘what is in development that it is distinct from governance’.

2 Even more prevalent than invariance of an indicator across countries (such as anticorruption legislation) is the 
invariance in an indicator over time for a country, while other variables for the same country and subject matter may 
exhibit considerable variance over time. The latter variable will fully drive the resulting changes over time in the 
composite indicator, with no contribution from the former variable, irrespective of the implicit or explicit weighting 
scheme in the aggregation.
Second, it conflates by construction the open question which many researchers are interested in: does governance matter for development outcomes, and if so how much and which dimensions in particular? At any rate, development outcomes are the result of a number of factors, governance being only one. Thus, the presumption of full attribution to governance, which is implicit in conflating development outcomes with governance, is questionable.\(^3\)

At the other extreme of the spectrum, anorexic definitions can miss crucial elements of governance in general, or of a component of governance in particular. For example, in democracy work, it is obviously unwise to focus narrowly on whether elections are held or not, given how meaningless (or unfree, unfair, manipulated, and rigged) many of these elections are. Similarly, official World Bank definitions have historically tended to focus on economic governance and traditional Public Sector Management, often leaving out key aspects of political governance, such as ‘voice’, democratic accountability, political violence, and a free press. In rule of law, some legal experts argue for a rather thin definition, focused on law and order.

If one stays clear of the obese or the anorexic definitions, there would still remain a non-trivial middle range for definitional debate around definitions of governance and their components. And there would still be a debate around the extent to which various governance measures are sufficiently grounded on a priori theoretical considerations. Further progress in theory, definitions, and conceptualizing should be encouraged, obviously.

But at the same time, it is worth noting that more progress in conceptualization in the field has taken place than at times acknowledged, and that the existing indicators do draw from such progress. Further, other theoretical models may not result in vastly different empirics than those that currently exist. Further work in measurement can proceed apace as theorists also continue to contribute to the field in parallel (and mutually reinforce each other).


Rather than fearing statistics, embracing it more fully is warranted, and not only to construct statistically robust indices and carry out empirical analysis. We also emphasize that statistics permits estimation of margins of error, which enable us to address a plethora of concerns regarding the use and interpretation of indicators. Econometrics provides us with powerful tools to deal with noisy indicators, with missing data and non-random samples, and also to test for thorny challenges of causality direction and endogeneity. These tools and tests enable a continuing process of forging ahead on multiple empirical measurement fronts.

Further, statistics permit addressing many a priori (and at times theoretical) concerns brought about by critics of governance indicators, and for testing for the validity and relevance of such critiques. Too often critiques feature long prose. Rare is the critique putting forth concrete and detailed alternatives. Even more rare is to find reviews that carry out empirical tests between the various options ascertaining whether the differences (concern) is of first, second, or third order. We illustrated this with the point that when aggregating the decisions

\(^3\) Furthermore, obese definitions of governance can be subject to thorny challenges (and uncertainty/margins of error) when constructing a composite, since the inclusion of, say, notions such as human rights and macro-economic (or health) outcomes may tend to lump together vastly opposing individual ratings for some country (e.g. Libya being close to the bottom in the former, and close to the top in the latter).
of which variables to include vs. exclude dwarfs the conventional concern about equal or somewhat unequal weights among included variables.\(^4\)

Furthermore, statistics in general, and the judicious use of governance point estimates and margins of error in particular, can empower the policy-maker. Evidence-based performance monitoring and policy-making is superior to performance assessment and policy-making devoid of empirics, particularly when the policy analyst and the decision-makers are provided with a range of estimates based on the margins of error of the composite indicator, and/or the various estimates at a disaggregated level, enabling triangulation.

In concluding, perspective and circumspection is in order: not all roads lead to data and statistics. Conceptual underpinnings are very important as well, obviously. That is why I emphasized some of these issues under item six above, making the case for avoiding definitional obesity or anorexia. Yet that does not suffice. In moving forward, while taking in the progress already attained, we ought to further probe within such ‘reasonable definitional midrange’ in these notions, in order to make further progress and converge on:

i) what ingredients are the crucial a priori ones in notions of democracy, ‘voice & democratic accountability’, government effectiveness, rule of law, and corruption, and,

ii) what is particularly noteworthy, what is amiss, and what can be improved in existing individual and composite measures in these dimensions.

More generally, Albert Einstein’s dictum serves as a useful reminder of the need for perspective:

‘Not everything that counts, can be counted, and not everything that can be counted, counts’.

***

*Charts enclosed below.*

---

\(^4\) Similarly, the common concern about unbalanced number of sources available from one period to the next can also be addressed statistically, permitting under certain conditions over time comparisons among aggregate indices. And empirical tests also suggest that the concern about potential ideological biases by some providers of individual variables is not really borne out empirically as a first order issue.
Exhibit 1a:  Taking Margins of Error Seriously (a)

Taking Margins of Error Seriously (1), Control of Corruption in WGI, selected countries, 2008

Source for data: 'Governance Matters VIII: Governance Indicators for 1996-2008', by D. Kaufmann, A. Kraay and M. Mastruzzi, June 2009, www.govindicators.org. Colors are assigned according to the following criteria: Dark Red: country is in the bottom 10\textsuperscript{th} percentile rank ('governance crisis'); Light Red: between 10\textsuperscript{th} and 25\textsuperscript{th} percentile rank; Orange: between 25\textsuperscript{th} and 50\textsuperscript{th} percentile rank; Yellow, between 50\textsuperscript{th} and 75\textsuperscript{th}; Light Green between 75\textsuperscript{th} and 90\textsuperscript{th} percentile rank; and Dark Green: between 90\textsuperscript{th} and 100\textsuperscript{th} percentile (exemplary governance). Estimates subject to margins of error.

Exhibit 1b:  Taking Margins of Error Seriously (b)


Source for data: ‘Governance Matters VIII: Governance Indicators for 1996-2008’, D. Kaufmann, A. Kraay and M. Mastruzzi, June 2009, http://www.govindicators.org. Colors are assigned according to the following criteria: Dark Red, bottom 10\textsuperscript{th} percentile rank; Light Red between 10\textsuperscript{th} and 25\textsuperscript{th}; Orange, between 25\textsuperscript{th} and 50\textsuperscript{th}; Yellow, between 50\textsuperscript{th} and 75\textsuperscript{th}; Light Green between 75\textsuperscript{th} and 90\textsuperscript{th}; Dark Green above 90\textsuperscript{th}.
Exhibit 1c: Taking Margins of Error Seriously (c)

Uncertainty in Ibrahim Index of African Governance (IIAG), Safety & Rule of Law Component -- Margin of Errors from Missingness and Measurement (M&M) 2007/08

Horizontal lines indicate terciles and median (50.0):
- Top tercile (56.1-62.4): 18 countries included; 3 countries with high confidence
- Middle tercile (50.0-56.1): 17 countries included; no countries with high confidence
- Bottom tercile (45.9-50.0): 18 countries included; 6 countries with high confidence

Exhibit 1d: Taking Margins of Error Seriously (d)

Uncertainty in IIAG Human Development Margin of Errors (M&M) Graph, 2007/08 Period

Horizontal lines indicate terciles and median (49.5):
- Top tercile (53.1-60.0): 18 countries included; 5 countries with high confidence
- Middle tercile (45.9-53.1): 17 countries included; no countries with high confidence
- Bottom tercile (45.9): 18 countries included; no countries with high confidence
Exhibit 2a: De Jure vs. De Facto Indicators (a)

Subjective and Objective Measures of Ease of Business Entry: Developing Country Sample

![Graph showing the relationship between the number of days to start a business and the difficulty of starting a business, with a correlation coefficient of r = 0.24.]

Exhibit 2b: De Jure vs. De Facto Indicators (b)

On Elections: De Jure and De Facto Measures

![Graph showing the relationship between the Global Corruption Barometer and the Global Integrity Index, with two regression lines and R-squared values: y = -3.14x + 85.69 (R² = 0.00) and y = 23.09x + 55.30 (R² = 0.19).]
Exhibit 3: Capturing Informality

Capturing Informality: Bribery vs. ‘Legal Corruption’ (& State & Regulatory Capture), 2004

% Firms report ‘corruption’

Source: Author’s calculations based on EOS 2004.