

The Worldwide Governance Indicators Project: Answering the Critics

Daniel Kaufmann, Aart Kraay, and Massimo Mastruzzi¹
The World Bank

September 2006

In this note we summarize some recent critiques of our project to develop Worldwide Governance Indicators, as described in our series of papers (Kaufmann, Kraay and Zoido-Lobaton (1999a,b) and (2001), and Kaufmann, Kraay and Mastruzzi (2004, 2005, and 2006)). The latest round of these governance indicators reports on six dimensions of governance every two years since 1996, and annually between 2002 and 2005, for over 200 countries. For brevity we will refer to them here as the WGI. The WGI have in recent years become widely-used indicators by policymakers and academics.² The data and accompanying descriptive papers, together with a Web-based interactive graphical tool, are available at www.worldbank.org/wbi/governance/govdata.

We address nine specific criticisms of the WGI that are made in three recent publications (Arndt and Oman 2006 (AO), Knack 2006 (K), and Kurtz and Shrank 2006 (KS)). AO provide an extensive and very useful survey of the many different types of governance data available, and in fairness we note that AO have many nice things to say about the WGI, kindly referring to them as "probably the most carefully constructed governance indicators". Here we focus only on addressing their main criticisms, contained in Section 4 of their paper. Similarly, K's focus is on interpreting the available data on trends in corruption in countries in Europe and Central Asia (ECA) between 2002 and 2005, and also contains in our view a very useful and thorough review of the many different types of data available to measure corruption in these countries. However, along the way he raises several criticism of the WGI as well as the underlying

¹ 1818 H Street NW, Washington, DC 20433, dkaufmann@worldbank.org, akraay@worldbank.org, mmastruzzi@worldbank.org. The views expressed here are the authors' and do not reflect those of the World Bank, its Executive Directors, or the countries they represent.

² For example, the United States Millennium Challenge Account aid program prominently relies on five of the WGI in its procedures for determining country eligibility, see www.mcc.gov for details.

data that we do not agree with. Finally, the paper by KS, which is forthcoming in the Journal of Politics, is primarily focused on critiquing the WGI. We have prepared a fuller response to the issues they raise for publication in the same journal, and we refer the interested reader to that article for details. Since some of the main points raised by KS are also raised in various forms by AO and K, and so we address all three papers together.

In this note we confine ourselves to responding to the criticism raised in these three papers, and do not address a larger set of questions that are sometimes raised about the WGI. For a non-technical overview of some of these questions, and our responses, please visit the "Frequently-Asked Questions" section of the WGI website (www.worldbank.org/wbi/governance/govdata), and refer also to our discussion of myths and realities in measuring corruption, in Kaufmann, Kraay, and Mastruzzi (2006c). In short, we do not find the criticisms raised in these three papers to be particularly compelling. As we argue below they are usually based on misinterpretations of our indicators, or of the empirical evidence involving these indicators. We have privately responded to the authors of each of the papers, presenting our views on these issues along the lines presented below. The purpose of this note is to place our responses to these criticisms in the public domain.

We organize the points made in these papers into nine related critiques, and provide our responses. The first four critiques call into question the usefulness of the Worldwide Governance Indicators for making comparisons of governance over time and across countries. Critiques 5 and 6 allege various sorts of biases in the individual indicators underlying our aggregate governance indicators. Critiques 7 and 8 concern the independence of the assessments of governance provided by our different data sources, and the consequences for the aggregate governance indicators. Finally, we respond here a final critique raised by AO, not specifically about the WGI, but about a paper of ours on the causality between governance and growth that used data from the WGI.

Critique 1: Governance cannot be compared over time using the WGI since they are scaled to have the same global averages in every period

Variants on this critique are raised by both AO and K. AO first correctly point out that our aggregate governance indicators are scaled to have a zero mean and unit standard deviation in each period. They then go on to assert that this means that the WGI "...cannot reliably be used for monitoring changes in levels of governance over time, whether globally, in individual countries, or among specific groups of countries" (AO 2006, p. 61). With the exception of global averages, this statement is incorrect. We have clearly acknowledged in our past work that by setting the world average of governance to zero in each period, our aggregate indicators are obviously not informative about trends in global averages of governance by definition. Recognizing this, we have in the last three updates of our indicators also provided whatever evidence we could from a selection of our individual underlying sources that are consistently available for longer periods of time about trends in world averages (see for example Kaufmann, Kraay and Mastruzzi 2004, Table 7; 2005, Table 6; and 2006, Table 5). These exercises have turned up little evidence of significant trends in world averages of governance, and so our choice of units for governance which sets the world average to zero in each period is innocuous.

This evidence from our individual sources that world averages of governance are not changing much is crucial, because it allows us to interpret the *relative* changes in country scores on our aggregate indicators, or groups of countries' scores, as *absolute* changes. In particular, if world averages do not change, then it is appropriate for us to rescale our governance indicators to have the same mean in each period, and there is no difference between changes in countries' relative positions on our indicator, and their absolute changes. This point has been made in Kaufmann, Kraay, and Mastruzzi (2004, 2005, and 2006).

K raises a more sophisticated objection to our normalization of a zero mean and unit standard deviation. Referring to our corruption indicator, he notes that the country coverage of this indicator has increased substantially over time (in fact, from 152 countries in 1996 to 204 countries in 2005). He then correctly notes that adding new countries can in principle change the ranks of existing countries in a relative ranking like our corruption indicator. If for example we add a country with very low corruption, i.e. a very high score on our Control of Corruption indicator, then this will reduce the rank of all

of the other countries in the sample by construction. While technically this point is correct, practically it turns out to not to matter much, for three reasons:

- As a minor point, K's primary interest is in trends in corruption in the ECA region between 2002 and 2005. During this period the country sample covered by our Control of Corruption indicator changes inconsequentially, from 197 to 204 countries worldwide. Even if all seven countries added had the lowest corruption in the world (and they do not), this would lower the percentile rank of the remaining countries by only about 3 percent, which is well within the margins of error for changes in country scores that we continually encourage users to take into account when making comparisons of changes over time.
- If users are interested in comparing the relative ranks of a set of countries over time (for example, the ECA countries relative to the rest of the world as is the case in K), then this problem can trivially be circumvented by simply looking at country ranks in a common set of comparator countries in both periods. This requires nothing in the way of sophisticated statistical tools, nor does it require access to the underlying data. This point is also noted by K.
- The extent to which the addition of new countries affects the relative ranks of countries already in the sample depends on how different the "entrants" to the sample are relative to the "incumbents". In the case of corruption between 1996 and 2005, the difference appears not to be very large. One way to check this is to look at the mean value of our 2005 Control of Corruption measure for the "entrants" between 1996 and 2005. The mean score of the entrants is 0.06, which is only slightly, and not significantly, above the world average which by construction is set equal to zero. In simple terms, this means that the new countries added to the sample included some with very little corruption, and some with a lot of corruption, and so the ranks of the remaining countries are not systematically affected by the addition of these countries.

In fairness, however, it does turn out that for some of our other indicators, there are bigger -- but never very large -- differences in the mean scores of the "incumbents" and "entrants" over different subperiods. In the 2005 update of the governance indicators,

we have provided small adjustments to the mean and standard deviation of the aggregate governance indicators for earlier periods to recognize this changing sample composition. In short, we maintain our assumption (as a choice of units) that governance has a zero mean and unit standard deviation across all countries in the world in each period. However, recognizing that in earlier years we do not have data for all countries, we allow the mean and standard deviation of the governance indicators to be slightly different from zero and one *in the sample of countries for which we actually have data*. For details refer to Section 2 of Kaufmann, Kraay and Mastruzzi (2006).

Critique 2: Governance cannot be compared across countries or over time with the WGI since the estimates for governance for different countries or periods may be based on different underlying data sources.

K and AO also both raise, in varying detail, this issue that arises in making comparisons over time and across countries. They both correctly note that when comparing two countries (or periods) using one of our governance indicators, the estimates of governance for the two countries (or periods) might be based on two very different, and even non-overlapping, sets of underlying data sources. While this point is factually correct of course, we do not find it to be a serious criticism, for five reasons:

- In the extreme case where two countries of interest do not appear in any single common data source, we actually would argue that one of the strengths of our aggregate indicators is that they do in fact make it possible to compare these countries, despite the lack of common sources. After all, one way to think about our aggregation methodology is that it provides a reasonably sophisticated way of placing very different underlying data sources into common units, and this is precisely what permits comparisons across countries that do not appear in the same sources. To take a specific example, suppose one wanted to compare corruption in the Bahamas with Saint Kitts and Nevis in 2000. These two small countries each appear in only one of our data sources, the Bahamas in the ICRG ratings produced by Political Risk Services, and Saint Kitts and Nevis in the World Bank's CPIA ratings. The virtue of our aggregate indicator is that it provides a way of putting the scores from these very different agencies into common units and permits comparison between them despite the absence of a

common data source, subject of course to the margins of error that we report, and that would be large for such countries that unusually appear in only one data source each. Admittedly this is an extreme example, but a more general point holds: if we want to make comparisons between countries based on a common set of data sources, this limits the information set we have available on which to base our judgements (in the extreme case of these two Caribbean states, it would eliminate the information set completely and prevent any comparison).

- Related to the previous point about units, we disagree with K's argument that since each underlying data source measures a somewhat different concept of corruption, the implicit definition of corruption is different when we compare two countries with different sets of underlying data sources. Again, it is useful to think about our aggregation method as a way of putting different data sources in common units. Suppose one data source measures corruption in procurement, while another measures corruption in the judiciary, and suppose once again we want to compare one country that appears only in the one indicator with another that appears only in the other indicator. Our aggregate indicator extracts the common component of these (and all our other data sources), which we label as overall "Control of Corruption". That is, we have just one implicit definition of corruption, which comes from the aggregation of these many data sources across many countries. Using the aggregate indicator we of course cannot distinguish between these particular dimensions of corruption, and for policy purposes in a particular country this distinction may very well be useful.³ But what our indicator does do is allow us to compare the extent of overall corruption in the two countries, based on the imperfect information both particular indicators provide about overall corruption.
- Whether this criticism is practically important or not depends a lot on whether the different nuances of corruption measured by different sources really differ a lot across countries. If some countries have very "clean" judiciaries but high

³ This is something we have long acknowledged. For example, in our very first paper, Kaufmann, Kraay and Zoido-Lobaton (1999a) we write in the conclusion that "There is therefore a need to improve the quality and quantity of governance data, both by improving and extending cross-country survey work of governance perceptions, as well as employing country-specific in-depth governance diagnostics", and similar statements can be found in our subsequent updates of the governance indicators.

administrative corruption, while other countries are the other way around, then a measure of overall corruption based on measures of administrative and judicial corruption would not be very informative. In fact, this very common-sense observation is explicitly built into our statistical methodology: if we were to construct an aggregate indicator of corruption based on on three measures of corruption that were quite uncorrelated across countries, the resulting aggregate indicator would have very large margins of error. It turns out however, that, especially for corruption, most of our individual indicators are quite highly correlated with each other. In simple terms, it seems unlikely that there would be many countries with high judicial corruption but low administrative corruption, and vice versa. And this is what the data tells us. For example, for Control of Corruption in 2005, the median correlation of our 18 underlying data sources with the aggregate indicator is 0.85, and only two data sources are correlated at less than 0.5.⁴

- Related to the previous point, in past work we have empirically documented the extent to which changes over time in our aggregate governance indicators for individual countries are influenced by the addition of data sources. In principle, of course, the addition of a new data source for a country that provides a very different assessment than other data sources can result in a large change in the aggregate indicator for that country. In practice, however, this effect does not appear to be very important. In KKM (2005) for example, we looked at all "large" -- in the sense of being statistically significant -- changes in each of our six governance indicators between 1996 and 2004. We first computed what the change in our estimate of governance would have been based on a common set of indicators, and then isolated the remaining component of the change which reflected the addition of data sources for each country. On average, we found that the addition of data sources accounted for only about 9 percent of the

⁴ These two are the Latinobarometro survey of countries in Latin America, and Freedom House's Countries and the Crossroads report. One other source with a low correlation for corruption is the BEEPS survey (correlation with aggregate indicator of only 0.55. Since one of the main interests of K is to account for differences between the BEEPS survey and other measures of trends in corruption in the ECA region, we are sympathetic with K's emphasis on the differences among data sources for this particular indicator and region. However, we do not think the point is more generally true for the majority of our data sources, which tend to be very much in agreement with each other.

variation in changes in our aggregate indicators, for countries with large changes in governance.

- Fifth and finally, a simple practical point. For many purposes we do recognize that it can be of interest for some users to make comparisons of governance based on particular individual data sources. To facilitate this, with the 2005 release of the aggregate governance indicators, we are making all but three of our underlying data sources available to users on our website.⁵

Critique 3: Changes over time in some of the individual indicators underlying the KKM aggregate indicators reflect corrections of past errors rather than actual changes.

This critique of several of our underlying indicators is made by K in the context of his discussion of trends in governance in countries in the ECA region between 2002 and 2005. While this is not a criticism that we have tried to address in our previous work, here we simply explain why we think that the evidence provided by K does not support his claim. K argues that in many cases individual data sources change their assessments of governance in a country simply to correct past errors in their assessments: countries that were rated "too high" in the past get lower scores, and vice versa. K then goes on to document that among ECA countries there is "regression to the mean", in the sense that changes in governance tend to be negatively correlated with initial scores, and interpret this as evidence that risk rating agencies change their scores to correct past "errors".

We believe that this is an overinterpretation of a regression that has a much simpler explanation. To be specific, let $y(j,t)$ be the rating of country j in year t . K regresses $y(j,t)-y(j,t-1)$ on $y(j,t-1)$, which is of course mathematically identical to regressing $y(j,t)$ on $y(j,t-1)$: the coefficient on the initial value in the first regression will just be the coefficient in the second regression, minus one. But the coefficient in the

⁵ These three sources are the Country Policy and Institutional Assessments produced by the World Bank, the African Development Bank, and the Asian Development Bank, which for the most part are treated as confidential by these organizations. Only in the past few years has limited, but growing disclosure of this data been made by these organizations, and full public access to the detailed disaggregated and historical data on which we rely is still not permitted..

second regression, which is just the autocorrelation coefficient of the governance rating, is for most of these data sources a positive number between 0 and 1. Suppose next that the governance rating is a noisy proxy for true governance, i.e. that $y(j,t) = g(j,t) + e(j,t)$ where $g(j,t)$ is true governance and $e(j,t)$ is the error made by a particular source. It seems quite plausible to us that governance on average changes rather slowly over time, indicating that $g(j,t)$ would have strong persistence, which would be reflected in strong persistence in $y(j,t)$. Thus, as long as true governance is persistent over time, we should expect to find a negative coefficient in K's regression reflecting nothing more than the persistence in governance itself.

We thus argue that K provides no evidence in support of his claim that changes in governance estimates reflect "correction" of past mistakes. We do think however that it is plausible *a priori* that similar kinds of corrective mechanisms might be at work in some of our indicators. For example, in our latest update of the governance indicators we have devised a test of the hypothesis that data sources update their scores in order to reduce past discrepancies between themselves and other data sources (Kaufmann, Kraay, and Mastruzzi (2006), Section 3). The simple intuition is that if data sources update their scores to reduce the past differences between them and other data sources, we should expect to see that the different data sources become more correlated with each other over time. K provides some evidence that this is the case, but only for measures of corruption in ECA countries over the past few years which is his primary interest. But it would be wrong to conclude from this that it is a general pattern. We have examined trends over time in the pairwise correlations between our sources for all countries, between 1996 and 2005, and find no systematic evidence of increased correlations (Kaufmann, Kraay and Mastruzzi 2006, Table 7, and accompanying discussion). The median change in correlation is only 0.03, and roughly the same number of pairs of sources exhibit increased and decreased correlations over time. We therefore do not think that this kind of updating is empirically very important.

Critique 4: The WGI are too imprecise to permit meaningful comparisons of governance over time or across countries

AO argue at some length that the WGI "...do not allow for a reliable comparison of levels of governance over time..." (AO 2006, p. 68). The gist of their critique is that,

since only a relatively small number of countries experience changes in governance that are large enough to be considered statistically significant, the indicators cannot be used to make comparisons over time. We find this critique peculiar, and without basis. It is not clear to us how the fact that many countries do not experience significant changes in governance according to our measures is a drawback of the governance indicators.⁶ Absent other information that governance in such countries is indeed changing but our indicators miss the changes, or conversely, without evidence that governance is indeed not changing in countries where our indicators show significant changes, AO's assertion is purely speculative. The presence of margins of error in our (and all other indicators) does not obviate the ability to make comparisons over time -- rather it enhances the ability of the user to make comparisons over time, by providing guidance as to which observed changes are likely to be meaningful, and which are not.

With respect to cross-country comparisons, we have always encouraged users of the governance indicators to take margins of error into account when making comparisons across countries. But this encouragement does not mean that no significant comparisons can be made. Consider for example our Control of Corruption indicator in 2005 which covers 204 countries, so that it is possible to make 20,706 pairwise comparisons of corruption across countries using this measure. For 64 percent of these comparisons, 90% confidence intervals do not overlap, signalling quite highly statistically significant differences across countries. And if we lower our significance level to 75 percent, which may be quite adequate for many applications, we find that 74 percent of all pairwise comparisons are statistically significant. While we continue to emphasize to users that many of the small differences between countries may well be neither statistically or practically significant, we also emphasize that a great many significant differences between countries can in fact be established using our aggregate indicators.

Critique 5: The individual indicators underlying the WGI are biased towards the views of business elites, and thus so are the aggregate indicators.

⁶ Even the statement that "many" countries do not have significant changes in governance is very much a matter of degree. Between 1996 and 2005, we find that roughly one country in three experienced a significant change (at the 90 percent confidence level) in at least one of our six governance indicators.

This concern is raised by both AO and KS. It is apparently based on the observation that several of our data sources are commercial risk rating agencies, whose main customers are businesses, as well as a number of cross-country surveys of firms. In short, they argue, businesspeople like low taxes and minimal regulation, while the public interest demands reasonable taxation and appropriate regulation. Estimates of governance based on the perceptions of businesspeople will therefore necessarily be biased. We do not think this criticism is particularly valid, for two broad reasons.

- First, we note that our indicators rely on much more than just the views of businesspeople. In the latest 2005 update of our governance indicators, our data sources include four cross-country surveys of firms, as well as seven commercial risk rating agencies, which one might expect to reflect narrower business interests. But we also rely on three cross-country surveys of individuals, six sets of ratings produced by government and multilateral organizations (such as the World Bank, the African Development Bank and the US State Department), and finally another 11 data sources produced by a wide range of non-governmental organizations (such as Freedom House, Reporters Without Borders, and many others). It is therefore simply incorrect to dismiss our indicators as reflecting solely the narrow interests of the business elite.
- Second, the extent to which this critique is valid depends crucially on the extent to which there are fundamental differences in the perceptions of business people versus other members of society as to what constitutes good governance. If this is true, then the responses of firms (or commercial risk rating agencies who serve mostly business clients) to questions about governance should not be very correlated with ratings provided respondents who are more likely to sympathize with the common good, such as individuals, NGOs, or public sector organizations. In fact, overall there are quite strong correlations among most of our different types of data sources. As an example, consider the ingredients of our Government Effectiveness indicator for 2005. The correlation between two of our major cross-country firm surveys is 0.74, and the correlation of these firms surveys with a survey of households in Africa is very similar at 0.7. More systematically, as we discuss further below, the rankings provided by our aggregate indicators are quite robust to alternative weighting schemes. This

robustness reflects precisely the fact that on average our different types of data sources provide highly correlated assessments. This in turn suggests to us that it is implausible that the interests of businesspeople differ so dramatically from those of other types of respondents.

Critique 6: *The data sources underlying the WGI are overly influenced by recent economic performance, and/or the level of development of a country -- rich, or fast-growing countries get better scores simply because they rich or growing fast.*

This critique is a common one, and is made at length by KS, and also in another widely cited paper, Glaeser et. al. (2004). The gist of the argument is simple. Governance ratings, especially those produced by commercial risk rating agencies, assume that governance must be good in countries that are rich or enjoying recent strong economic performance, and so these countries receive ratings that are better than they deserve. This phenomenon is sometimes referred to as "halo effects", and is something that we have studied in our earlier work with these indicators.

- In Kaufmann, Kraay, and Mastruzzi (2004) we look for evidence of halo effects associated with levels of development. Glaeser et. al. (2004) argue that much of the observed correlation between governance and levels of development can be explained by such halo effects. We develop a simple statistical model to investigate the empirical importance of this claim (which Glaeser et. al. do not), and show that there is in fact a tradeoff. Halo effects can be thought of as measurement error. By itself, greater measurement error in governance actually *lowers* the correlation between governance and per capita incomes, while measurement error that is correlated with per capita incomes *increases* it. Given this tradeoff, we provide calibrations that show that halo effects would have to be implausibly strong in order to account for the observed correlation between governance and per capita incomes.
- In contrast, KS do claim to provide direct empirical evidence of halo effects, showing that one of our six governance indicators, Government Effectiveness, tends to have a significant partial correlation with two-year average growth rates prior to the date of the governance indicator in a limited set of regressions that

they report. In our fuller response to KS we document in detail how the evidence they report is not robust, and in any case is misinterpreted by KS. In brief, we show that very minor changes to their empirical specification entirely overturn their results. We also show that after controlling for long-run economic performance of countries, the short-term growth that KS claim is driving halo effects is also no longer significant. Based on this we argue that the short-run growth variable is simply proxying for longer-run growth, and that the KS regressions could just as well be interpreted as picking up an entirely reasonable causal effect of good governance on long-run economic performance. Consistent with this, we show that a very careful measure of government effectiveness that KS -- likely correctly -- hold up as a model indicator untainted by "halo effects" exhibits the same partial correlations with long- and short-run growth as do the WGI. We therefore do not find their evidence of alleged "halo effects" to be at all compelling.

Critique 7: The individual data sources underlying the WGI, particularly those from commercial risk rating agencies, make correlated errors in their assessments of governance and so are less informative about governance than they appear.

This criticism (together with Critique 8 below) are discussed at length in AO (pp. 65-67), as well as in K (pp 21-27)). The point here is a simple one. Suppose that one cross-country rating agency "does its homework" and comes up with an assessment of governance for a set of countries based on its own independent research, but a second rating agency simply reproduces the assessments of the first. Then in reality we would only have one data source, not two, and inferences about governance based on the two data sources would be no more informative than inferences based on just one of them. In short, the rationale for constructing an aggregate governance indicator would disappear since we really only have just a single data source. K goes so far as to assert that "...this unknown but substantial degree of interdependence among many of the sources also obviates any claims regarding the "precision" of these indicators." (p. 23).⁷

⁷ Of course this raises a logical puzzle -- if the degree of correlation in errors across sources is unknown, how can K know that it is "substantial"? Below we discuss in more detail other work we have done which allows us to identify empirically the degree of error correlation across sources,

This example is of course contrived because it makes the implausible assumption that the two data sources make perfectly correlated measurement errors when they assess governance across countries. A first important point to note is that any deviation from this assumption of perfectly correlated errors means that there are in fact gains in precision to be had from aggregation. Even if the errors made by the two data sources are highly, but not perfectly, correlated, an aggregate indicator averaging the two will be at least somewhat more informative than either individual indicator. In short, the presence of correlated errors among sources does not eliminate the benefit of constructing an aggregate governance indicator, although it does of course reduce it. This critique is not news to us. In fact, in our very first methodological paper on the aggregate governance indicators (Kaufmann, Kraay and Zoido-Lobaton 1999a) we devoted an entire section of the paper to this possibility, and showed how the estimated margins of error of our aggregate governance indicators would increase if we assumed that the error terms made by individual data sources were correlated with each other. We also note that even if two data sources make correlated errors, it does not mean that we should discard them entirely from the aggregate indicator -- they jointly still might well contain useful information, just not as much information as they would if they were truly independent.

The more important empirical question is whether this correlation of errors across sources is large or not. Both AO and K offer anecdotal evidence of cases where some of our specific data sources have access to other of our data sources when formulating their assessments. We note however that the mere fact that data sources "look at each other" does not by itself constitute evidence that these data sources will therefore make correlated errors. It is useful to think of the assessment of any data source as providing some "signal" of governance, combined with an error term capturing the idiosyncracies of that particular data source. Suppose that one commercial risk rating agency decides to look at the necessarily noisy estimate of governance produced by another rating agency. Surely the first agency, which is in the business of providing informative estimates of governance to its customers, has every incentive to try to filter out the measurement error from the other data source that it is looking at, before incorporating it

making it "known" -- at least conditional on identifying assumptions -- and also showing that it is not "substantial".

into its own estimates. Moreover, it would make sense for the rating agency to "look at" the assessments of as many other agencies as possible in order to try to extract all of the information about governance embodied in those other data sources. While we do not pretend to know how exactly all of our individual data sources process the information at their disposal, it seems strange to us to suggest -- as implicitly do AO and K -- that they blindly copy each other and so make correlated errors.

AO and K both also observe that different data sources might be influenced by the same media reports about a country, and argue that this justifies their claim that individual data sources make correlated errors. Logically this does not follow, as it depends on whether the media reports are accurate or not. If the media reports are accurate, then all the individual data sources that rely on this common media report will both be more accurate, and also more correlated with each other, as a result. Of course, some media reports are more accurate than others, but AO and K do not enter into this crucial part of the argument.

While AO and K both provide some anecdotal evidence of data sources making correlated errors, only K attempts to provide some empirical evidence, for one measure of corruption.⁸ K first documents convincing evidence of a methodological break in one widely-used expert assessment of corruption, the International Country Risk Guide (ICRG), in October of 2001, noting that in this particular month an extraordinarily large number of countries in the sample have their scores change compared with typical other months. He then goes on to point out that, compared with earlier dates, the ICRG corruption ratings become more correlated with the Transparency International (TI) corruption ratings, with the correlation increasing from 0.71 to 0.92. He concludes that

⁸ We note in passing that two of the four examples offered by AO are either incorrect or exaggerated. The incorrectly assert that we use the Cingranelli and Richards Human Rights database and the Political Terror Scale, which both rely on numerical coding of information in the US State Department's Human Rights Report, as separate data sources in the same indicator. We do indeed use data from both these sources, but recognizing their common origin in the State Department reports, we average them together and use them as a single indicator in the aggregate indicators. They also suggest that we use three different data sources from Freedom House in the same aggregate indicators. This is in fact the case only for Voice and Accountability. For Rule of Law and Control of Corruption we do rely on two data sources from Freedom House, Nations in Transit, and Countries at the Crossroads. However, in 2005 these two data sources, covering 28 and 30 countries respectively, have just two countries in common, Russia and Tajikistan. This means that there is practically no opportunity for correlated errors between these two sources to have any effect.

this provides evidence that the ICRG corruption ratings are not independent of the other data sources embodied in the TI ratings, and in particular suggests that the reason for the methodological break was to camouflage a move to greater consistency with the TI ratings. While interesting, we do not find this tidbit of evidence to be compelling, for two main reasons.

- First, the mere fact that the ICRG correlation with TI increases does not provide any evidence at all that the errors made by ICRG are correlated with the errors made by the other sources embodied in TI. Nor does it even provide evidence that the ICRG scores are based on the TI scores. The increased correlation with TI following the methodological break could logically also reflect the fact that ICRG had improved the quality of its own assessments, and improving the signal to noise ratio in its own assessments made it more correlated with other assessments. A purely "home-grown" improvement in quality by ICRG could therefore also account for the higher correlation with TI.
- Second, and perhaps more important, this pattern of increased correlation with other sources following methodological breaks is not in fact a systematic feature of the ICRG ratings. From the standpoint of analysis, it is fortunate that ICRG has in fact made methodological changes to several, but not all, of its indicators, in two different years. In Kaufmann, Kraay and Mastruzzi (2006) we have systematically looked at the 10 ICRG indicators we use in the WGI, and identified two series with methodological breaks in 1997, and five in 2001. If the objective of such breaks really were to generate new ratings that are more correlated those of with other experts, as argued by K, then we should systematically expect to see increases in correlations with other expert assessments when comparing the period before and after the methodological break. In contrast, we should see no change in the correlation of ICRG with other expert assessments for series that did not have methodological breaks in the same year. It turns out that this simply is not true in the data. We do find, consistent with K, that the correlation of the ICRG corruption rating with other expert assessments increases after the break in 2001. But when we compare the change in correlations with other expert assessments in the set of five ICRG indicators with breaks in 2001, with the change in correlations of the remaining

five ICRG indicators without breaks, we find virtually no difference. In fact, the typical change in correlation of indicators with breaks is just 0.01, while the correlation of indicators without breaks is almost identical at -0.01. If, as suggested by K, ICRG has used methodological breaks to camouflage a greater correlation with other sources, then it is very puzzling why they should not do so systematically. We therefore do not find K's isolated evidence for just one of the many ICRG indicators, to be at all compelling.

Finally, in our latest paper on the governance indicators (Kaufmann, Kraay and Mastruzzi 2006) we have provided new empirical evidence on the possible correlation of errors across data sources. As we discuss at length in that paper, empirically identifying correlations in errors across sources is difficult. Simply observing that two data sources provide assessments that are highly correlated is not enough, since the high correlation could reflect either (i) the fact that both sources are measuring governance accurately and so are highly correlated, or (ii) the fact that both sources are making correlated measurement errors in their assessments of countries. In order to make progress we need to make assumptions, and in that paper we detail two sets of assumptions that allow us to disentangle potential sources of correlation in the errors. One assumption is related to the plausible argument of K that surveys of firms or individuals are less likely to make errors that are correlated with other data sources than, for example, the assessments of commercial risk rating agencies. If this is the case, however, we would expect that the assessments of commercial risk rating agencies be very highly correlated with each other, but less so with surveys. This turns out not to be the case. For example, the average correlation among our five major commercial risk rating agencies for corruption in 2002-2005 was 0.80. The correlation of each of these with a large cross-country survey of firms, was actually slightly higher at 0.81, in contrast with what one would expect if the rating agencies had correlated errors. We do this exercise for components of all six of our aggregate governance indicators, and find at most quite modest evidence of error correlation.

Critique 8: If some data sources make correlated errors, the aggregation procedure used by the WGI gives too much weight to such indicators.

The WGI are constructed using a statistical methodology known as an unobserved components model, which in effect estimates governance for each country as a weighted average of the underlying indicators available for that country. The premise for the weighting of indicators is simple. We think of each underlying data source as providing a noisy or imperfect signal of governance. If the errors made by individual sources are uncorrelated with each other, then data sources that produce highly correlated ratings must be more informative about governance than sources that are not highly correlated with each other. Accordingly, these more highly-correlated data sources should receive a greater weight in the aggregate indicator. This neat logic however breaks down if the correlation between data sources is due to their making correlated errors. If this were the only reason for the correlations among sources, then more highly correlated sources should receive less weight, not more. This point is made by both AO and K. Specifically, given their concerns that the errors made by commercial risk rating agencies are likely to be highly correlated (a claim we have argued above not to be very compelling), they conclude that the WGI place excessive weight on such data sources.

As we have discussed above, isolating the effect of correlated errors in driving the observed correlation among sources is extremely difficult, and so a general and robust solution to this potential problem is not yet within our reach. We acknowledge the point that our weighting scheme could in principle be giving more weight to particular sources for the "wrong" reasons, although as we have discussed at length and in KKM (2006), we have not yet found compelling evidence of substantially correlated measurement errors in our sources. Nevertheless, we do accept the point that it is useful to explore how different our aggregate indicators would look if we used alternative weighting schemes.

The simplest possibility is to just construct an unweighted average of our data sources. This in practice substantially reduces the weights applied to expert assessments from commercial risk rating agencies, the main concern of AO and K. We have done this for all six of our indicators for the past seven years, and it turns out that the equally-weighted indicators are extremely highly correlated with the benchmark aggregate indicators. The average (across the 42 indicators) correlation between the equally-weighted indicators and our benchmark indicators is 0.99. In only three cases is

the correlation less than 0.99, and the minimum correlation is 0.97. This clearly shows that equal weighting of our underlying data sources does not practically affect our estimates of governance in the vast majority of cases.⁹ The reason for this very high correlation between the weighted and unweighted aggregate indicators is that by and large, all of our individual data sources tend to agree quite strongly with each other in their assessments of governance.

In fairness, we do think that K suggests a number of reasonable alternative weighting schemes. Motivated by his suggestions, we here consider an alternative weighting scheme that organizes data sources by type and constructs an aggregate indicator weighting each *class* of indicators equally. We consider four groups: (1) cross-country surveys of firms and individuals, (2) commercial expert assessments, (3) expert assessments produced by NGOs, and (4) expert assessments produced by multilateral organizations. We simply average our indicators within each group, and then construct an aggregate indicator based on the four groups alone. We find that the correlation of the resulting indicators with our benchmark indicators is again very high, averaging 0.95 for five of the six indicators in 2005. The only exception to this pattern is for our Regulatory Quality measure, which has a correlation of only 0.44 with the benchmark measure. However, a closer look at the data reveals that this is entirely due to one data source, the Business Environment and Enterprise Performance Survey (BEEPS) carried out by the World Bank in transition economies. This source alone among the very many we use, for this particular indicator alone, gives quite different assessments from other types of data sources. If we drop the BEEPS from the exercise, we find that our revised indicator based on four groups of data has a correlation of 0.96 with the benchmark indicator.

To sum up, this critique and the previous one deal with the tricky issue of whether individual data sources on governance do in fact provide fully independent estimates of governance. We have argued that this is in general a difficult problem to sort out, but we can make progress in some cases, and in these cases the data do not provide any clear evidence of strong dependence in the errors made by individual data

⁹ The main benefit however of weighting sources by their precision is that it yields somewhat smaller standard errors, allowing for more precise inference about cross-country differences and changes over time in governance. On average, the standard error of our equally-weighted indicators is about 10 percent higher than in our benchmark indicators.

sources. In principle, the main reason to be concerned about the independence of data sources is that it matters for the appropriate weighting of the individual indicators in the aggregate indicator. It turns out, however, that since all of our data sources on average do tend to agree quite strongly with each other, the particular weights used to construct the aggregate indicators do not matter very much for the overall estimates of governance.

Critique 9: Flaws in the evidence on the two-way relationship between governance and growth in Kaufmann and Kraay (2002)

We conclude with a brief discussion of the penultimate section of AO, where they critique an empirical exercise we carried out in Kaufmann and Kraay (2002) using one of the WGI, the Rule of Law measure.¹⁰ In that paper we were interested in providing a first set of estimates of the reverse causation from income levels to institutional quality. In order to identify the causal effect running from institutions to per capita incomes, we borrowed the framework of the famous paper by Acemoglu, Johnson and Robinson (2001), who developed a clever instrument for institutions based on historical settler mortality rates. A puzzling feature of that paper, as well as our replication of it with somewhat different data, is that the estimated causal effect of institutions on per capita incomes was much larger than the slope coefficient in a simple OLS regression. We then showed how with the help of a few minimal assumptions about the degree of measurement error in institutional quality and the role of omitted variables we could pin down the magnitude of the reverse effect, from per capita incomes to governance, in a way that was consistent with the puzzling relationship between the OLS and IV coefficient estimates in Acemoglu et. al.. Our somewhat surprising conclusion was that, for reasonable assumptions, this reverse effect was negative, implying that exogenous shocks to income could worsen governance.

AO present an extension of this exercise, where their main innovation is to propose a new strategy that allows them to directly estimate the causal effect of per capita incomes on institutional quality. Their strategy consists of proposing an

¹⁰ We note in passing that KS also provide a critique of the governance and growth literature that we find to be entirely without merit. Our views on their critique can be found in the response we have prepared for the Journal of Politics.

instrument for per capita incomes. In particular, AO, and we, consider a two equation system in governance and per capita incomes. As usual, identification requires the presence of a variable that affects governance but has no direct effects on incomes (the famous settler mortality instrument), and another variable that affects per capita incomes but has no direct effect on governance other than through its effect on income. Finding such a second instrument is very hard -- sufficiently so that we bothered to write our paper proposing a strategy for avoiding such a difficult task.¹¹

Unfortunately, we do not think that the instrument proposed by AO is especially compelling. They use infant mortality rates as an instrument for per capita incomes, arguing that infant mortality is unlikely to have any direct links to institutional quality. Yet this seems hard for us to believe on *a priori* grounds. Infant mortality rates depend in considerable measure on public health interventions, and it seems plausible to us that the quantity and quality of these in turn depend at least in part on governance. Or mechanically, infant mortality rates are quite likely to be significantly correlated with historical settler mortality rates, and these in turn are correlated with our measure of institutional quality. Thus it seems implausible to us that their proposed instrument satisfies the exclusion restriction.

We certainly do not want to claim that our paper is the "last word" on the magnitude of the feedback from incomes to governance, and we strongly agree that more work is needed in this area. Unfortunately we do not think the particular strategy adopted by AO is a very fruitful one.

References

- Acemoglu, Daron, Simon Johnson, and James Robinson. "The Colonial Origins of Comparative Development. *American Economic Review*. 91(5):1369-1401.
- Arndt, Christiane and Charles Oman (2006). "Uses and Abuses of Governance Indicators". OECD Development Center Study.
- Evans, Peter and James Rauch (1999). "Bureaucracy and Growth: A Cross-National Analysis of the Effects of 'Weberian' State Structures on Economic Growth". *American Sociological Review* 64 (October 1999): 748-765.

¹¹ Rigobon and Rodrik (2004) provide an alternative clever identification strategy of this two-way relationship that also does not rely on searching for a good instrument for per capita incomes.

- Glaeser, Edward, Rafael LaPorta, Florencio Lopez-de-Silanes, and Andrei Shleifer (2004). "Do Institutions Cause Growth?". *Journal of Economic Growth*.
- Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobaton (1999a). "Aggregating Governance Indicators." World Bank Policy Research Working Paper No. 2195, Washington, D.C.
- Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobaton (1999b). "Governance Matters." World Bank Policy Research Working Paper No. 2196, Washington, D.C.
- Kaufmann, Daniel, Aart Kraay and Pablo Zoido-Lobaton (2002). "Governance Matters II – Updated Indicators for 2000/01." World Bank Policy Research Working Paper No. 2772, Washington, D.C.
- Kaufmann, Daniel and Aart Kraay (2002). "Growth Without Governance". *Economia*.
- Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2004). "Governance Matters III: Governance Indicators for 1996, 1998, 2000, and 2002". *World Bank Economic Review*. 18:253-287.
- Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2005). "Governance Matters IV: Governance Indicators for 1996-2004. World Bank Policy Research Working Paper No. 3630. Washington, D.C.
- Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2006). "Measuring Governance Using Perceptions Data", in Susan Rose-Ackerman, ed. *Handbook of Economic Corruption*. Edward Elgar.
- Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2006b). "Governance Matters V: Governance Indicators for 1996-2005. World Bank Policy Research Department Working Paper No. 4012.
- Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2006c). "Measuring Corruption: Myths and Realities". *Development Outreach*. September 2006, pp 37-41.
- Knack, Steven (2006). "Measuring Corruption in Eastern Europe and Central Asia: A Critique of the Cross-Country Indicators". World Bank Policy Research Department Working Paper 3968.