WPS6526

Policy Research Working Paper 6526

It's Only Words

Validating the CPIA Governance Assessments

Stephen Knack

The World Bank
Development Research Group
Human Development and Public Services Team
June 2013

Policy Research Working Paper 6526

Abstract

This study analyzes the validity of the World Bank's Country Policy and Institutional Assessments governance ratings, an important factor in allocating the Bank's concessionary International Development Association funds. It tests for certain biases in the ratings, and examines the quality of the written justifications that accompany the ratings. The study finds no evidence of bias in favor of International Development Association-eligible countries, despite a potential moral hazard problem inherent in the ratings process. However, there is some evidence of an upward bias in ratings for one region, relative to the other five regions. The study finds significant regional differences in the quality of the written justifications accompanying the six World Bank regions' proposed ratings. The length of these write-ups has exploded over time. Although higher-quality write-ups are also longer on average, there is wide dispersion in the word count at any given quality level, and some long write-ups provide little relevant information. Higher quality write-ups are associated with a lower likelihood that central unit reviewers will either disagree with proposed ratings, or request additional information to assess the proposed rating. Controlling for quality, longer write-ups are associated with a greater probability that central reviewers will disagree with a proposed rating. Although checks and balances built into the process appear to work reasonably well, the author concludes that a more proactive role for central unit reviewers and regional chief economists' offices could further enhance the quality of write-ups and reduce regional bias.

This paper is a product of the Human Development and Public Services Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org.
The author may be contacted at sknack@worldbank.org.

It's Only Words:
Validating the CPIA Governance Assessments

Stephen Knack [1]

JEL: O10, O17, O19

Keywords: Aid effectiveness, aid allocation, governance, corruption, rule of law

[1] Lead Economist, DECHD & PRMPS, World Bank (sknack@worldbank.org) The conclusions of this paper are not intended to represent the views of the World Bank, its Executive Directors, the countries they represent, or the members of the Public Sector Governance Board. Claudia Berg provided able research assistance. The paper benefited from valuable comments by Rui Coutinho, David Gould, Roumeen Islam, Kimberly Johns, Markus Kitzmuller, Jana Kunicova, Nick Manning, Shilpa Pradhan, Nadeem Rizwan, Halsey Rogers, Smriti Seth, and Hernan Winkler. However, the author assumes full responsibility for the analysis and interpretation, including any remaining errors.

1. Introduction

The World Bank's Country Policy and Institutional Assessment (CPIA) ratings are the primary factor in determining allocations of its concessionary IDA (International Development Association) funds across recipient countries. The premise is that the development effectiveness of aid and other resources is conditional on the quality of macroeconomic and other policies, and on the quality of public sector management (including budgetary and legal systems). Under IDA's "performance-based allocation" (PBA) formula, countries with lower per capita incomes receive higher allocations per capita, but CPIA ratings are weighted more heavily than income.

The African and Asian Development Banks have their own CPIA ratings, used in their own PBA systems. Collectively, these two regional development banks and IDA annually account for roughly $17 billion in gross ODA disbursements (or about $11 billion net) in recent years, with the majority allocated partly on the basis of their respective CPIA ratings. Over two-thirds of these funds are from IDA. Each of these three development banks produces its own CPIA ratings, based on assessments of their own staff and implementing a "questionnaire" with detailed criteria on 16 policy areas, grouped into 4 "clusters" (World Bank, 2010).

The purpose of this study is to contribute to our understanding of the CPIA process in the World Bank, and thereby of the likely accuracy of the ratings. It analyzes the length and quality of the written justifications that accompany the regions' proposed ratings, as well as the responses of central unit reviewers to these write-ups and ratings. We also test for certain biases in the ratings, and find some evidence of bias in favor of one region, but no evidence of any bias toward IDA countries. Finally, we recommend reforms to strengthen the process where apparent weaknesses are identified.

This study is not designed to be a comprehensive treatment of the CPIA as an appropriate mechanism for allocating aid. In particular, it does not address issues regarding the content of the CPIA questions. Nor does it address the weights that the IDA donors assign to the different CPIA questions and other variables in the IDA allocation formula. Moreover, it focuses exclusively on the questions in one of the four clusters. The choice of cluster D, on quality of public sector management and institutions, is motivated by the fact it has by far the largest weight in determining IDA allocations (World Bank, 2010).

The validity of CPIA governance ratings matters for other purposes, in addition to allocating IDA funds to countries where donors believe they will be used most effectively. The CPIA ratings are often used in research studies (e.g. Collier and Dollar, 2002; Dollar and Levin, 2006; Knack, 2009) on the largely untested assumption that the ratings are valid and reliable. They are sometimes used to monitor performance of individual countries or groups of countries. For example, question 13 on quality of budgetary and financial management was used as one of the Paris Declaration monitoring indicators (OECD, 2011; Knack 2013).

The remainder of the study is organized as follows. Section 2 briefly reviews the (sparse) related literature, and places this contribution in the context of that literature. Section 3 describes the CPIA ratings process, and section 4 describes the data and hypotheses to be tested. Results are

presented in sections 5-8, and section 9 summarizes and concludes with several recommendations for strengthening the process, as well as suggestions for additional research.

2. Related Literature

Only a few studies have examined the validity of the CPIA ratings, despite the CPIA's importance in determining aid allocations. They have mostly focused on the content of the CPIA questions, and on how the IDA donors weight the different questions in allocating IDA funds.[2] The scope of this study is limited to analyzing the validity of the ratings themselves, and not how the IDA donors choose to use them in allocating aid among recipient countries.

Several observers have criticized the CPIA content for reflecting a neoliberal or "Washington Consensus" view of what policies matter for development (e.g. Cage 2009). Kanbur (2005) proposes that the CPIA focus more on measuring outcomes. He argues that if countries are generating favorable development outcomes there is good reason to believe they will use aid funds effectively, even if their policies do not conform closely to the CPIA's prescriptions. Steets (2008) recommends adding more content on infrastructure, "opportunities for participation" and "empowerment of communities." Steets (2008) and IEG (2010) advocate an increased emphasis on protection of human rights.

The validity of the CPIA ratings has rarely been explicitly addressed in empirical studies. The IEG (2010) evaluation of the CPIA finds moderate to high cross-country correlations between ratings on various questions and other conceptually related indicators. For example, it reports a .86 correlation between CPIA question 16 (on accountability, transparency and corruption in the public sector) and Transparency International's Corruption Perceptions Index. The IEG report does not test for possible biases, however, despite noting that the use of CPIA scores for IDA allocations creates incentives for country teams to inflate ratings.

Gelb, Ngo and Ye (2004) provide evidence of the CPIA's external validity, showing it predicts income growth in the medium term. They explicitly test for a pro- or anti-Africa bias in the ratings, and find that ratings for countries in the region are neither significantly higher nor lower than predicted by their scores on related indicators.

Other empirical analyses making use of the CPIA implicitly support the external validity of the ratings. At least two World Bank (2011a, 2007) reports find that high or increasing CPIA ratings are associated with a stronger likelihood of achieving several of the Millennium Development Goals. Performance of World Bank projects is also stronger in countries with higher CPIA ratings (Denizer, Kaufmann and Kraay, 2011; IEG, 2010: Appendix G). Donors are more likely to use recipient countries' public financial management systems, rather than managing their aid through parallel systems, when ratings on CPIA question 13 (on strength of budgetary systems) are higher (Knack, 2013). If the CPIA ratings contain a large random-error component, that

[2] Steets (2008), IEG (2010) and an external review panel's report (World Bank, 2004) conclude that there is insufficient empirical evidence to justify a higher weight on cluster D on grounds of development effectiveness. Rather, the extra weight on governance is likely motivated instead by donors' fiduciary concerns (IEG, 2010: 60).

would tend to bias empirical tests against finding correlations such as the ones in these three studies.[3]

This study contributes to the evidence on the validity of the CPIA ratings in several ways. It uses related indicators from other sources to test for systematic ratings bias favoring the IDA-eligible countries. We find no evidence for such a bias, despite the fact that their ratings are linked to aid allocations, unlike the case with non-IDA countries. We also test for regional biases in the data, and confirm the finding by Gelb, Ngo and Ye (2004) on the absence of an African bias. Instead, we find some evidence of ratings bias in favor of countries in the Eastern Europe and Central Asia (ECA) region. By comparing the original regional proposals with the final ratings, we show that the network review process has only a marginal impact in curbing this bias, despite the fact that ratings disagreements are usually resolved in favor of the network's recommendations over the region's proposals.

The study also provides the first quantitative analysis of the written justifications ("write-ups") accompanying the region's proposed ratings. We find significant differences among the regions in the length of these write-ups, and more importantly in their quality, with East Asia (EAP) the leader and Latin America (LCR) the laggard, relative to the other four regions. On average, longer write-ups are higher in quality (as measured by how many of the criteria in the question are addressed in the write-up). Despite this overall positive relationship, many long write-ups actually cover fewer of the criteria than many short write-ups. This fact suggests that in some cases much of the information contained in them is of limited relevance in justifying the proposed ratings.

The Bank's central units, including the network anchors, have a key quality control role in the CPIA. We therefore also analyze responses of network reviewers to the regions' proposed ratings as a function of write-up characteristics and other factors. The reviewers were more likely to request additional information to assess a proposed rating if (1) it was accompanied by a shorter write-up; (2) the write-up addressed fewer of the criteria in the question; (3) it represented an increase or decrease from the previous year; or (4) it belonged to an IDA-eligible country. Finally, network reviewers were more likely to disagree with a region's proposed rating when the write-up addressed fewer of the criteria in the question, or when it represented an increase. Controlling for ratings levels and increases, there were no significant regional differences in the likelihood that reviewers disagreed with a proposed rating. These findings indicate that the network reviews are serving a valuable quality-control function in the CPIA process - and likely improving accuracy of the ratings – although this role could be strengthened further.

[3] Most sources of systematic bias in the ratings would also tend to weaken relationships with development outcomes. However, if quality of policies and institutions as measured by CPIA were merely inferred from observing development outcomes (e.g. donors use country systems in X, so X must have sound budgetary systems), then correlations would be biased upwards.

3. CPIA Process

The CPIA's 16 questions are intended to assess a recipient country's ability to make effective use of aid resources in furthering development and poverty reduction. The set of questions and their criteria have evolved over time, and are revised periodically to reflect changes in the collective knowledge of practitioners and specialists – both inside and outside the World Bank - regarding policies and public sector management institutions that matter for these outcomes (IEG, 2010; World Bank, 2004). The questions are grouped into four "clusters" as follows:

A. Economic Management
   1. Macroeconomic Management
   2. Fiscal Policy
   3. Debt Policy

B. Structural Policies
   4. Trade
   5. Financial Sector
   6. Business Regulatory Environment

C. Policies for Social Inclusion/Equity
   7. Gender Equality
   8. Equity of Public Resource Use
   9. Building Human Resources
   10. Social Protection and Labor
   11. Policies and Institutions for Environmental Sustainability

D. Public Sector Management and Institutions
   12. Property Rights and Rule-based Governance
   13. Quality of Budgetary and Financial Management
   14. Efficiency of Revenue Mobilization
   15. Quality of Public Administration
   16. Transparency, Accountability, and Corruption in the Public Sector

The questions are designed to assess government policies and institutions to the extent possible, rather than outcomes. Some of the sub-criteria in the questions are quantitative, but a certain amount of expert judgment is required in determining the scores for all 16 questions.

Ratings originate with the country teams in the Bank's regional departments. This can be viewed as a strength of the CPIA ratings process, relative to cross-country assessments produced by some other organizations: "closeness to the client is needed to provide an in-depth knowledge of the policies and institutions in a given country" (Gelb, Ngo and Ye, 2004: 2). However, it also creates a potential conflict of interest, as "having staff rate the countries on which their work programs depend could lead to rating biases" (IEG, 2010: xv). For this reason, the Chief Economist's office in each of the six World Bank regional departments has a quality control function in the process. These regional chief economists' (RCE) offices have a lot of discretion over how they conduct their respective regional-level reviews, and procedures vary. For example, quantitative checks for ensuring intra-regional consistency are more feasible and useful for regions with numerous countries such as Africa (AFR) than for others such as South Asia (SAR). Gelb, Ngo and Ye (2004) report that in the Africa region "intensive discussions take

place between the country economists" and sector specialists, and ratings are debated at several meetings convened by the Chief Economist's office.

Following the intra-regional reviews, the six RCE offices then forward their proposed ratings to the World Bank's central units for a cross-regional review coordinated by the Vice-President's Office in the Operational Procedures and Country Services (OPCS) department. Each of the 16 questions is assigned to a particular network anchor with primary responsibility for the review, although there is no one-for-one mapping: multiple central units comment on each questions, and some network anchors comment on multiple questions. The Public Sector Governance anchor in the PREM network (PRMPS) has primary responsibility for four of the five questions in cluster D, all but question 13.[4] These are the four questions that will be the subject of the analyses in the remainder of the study.

The main value added of the network review is to strengthen cross-country comparability of ratings across regions. The central units are usually better positioned than the RCE offices to conduct comparative statistical analyses using a range of other indicators pertaining to the content of the CPIA questions. Comments from network reviewers are collected by OPCS and submitted to the regions for their responses. Many comments take issue with a particular ratings proposal, in the majority of cases (but by no means all) recommending a lower rating than the one proposed by the region. Some comments disagree only with a proposed sub-rating that has no implications for the overall question score.

Many other comments do not express disagreement with a rating or sub-rating, but instead point out that some of the criteria contained in the question are neglected in the written narrative submitted by the regions in support of each proposed rating. These "write-ups" were introduced into the CPIA process in 2001, and are of enormous importance in helping regional and network reviewers – and government officials - to understand the reasoning behind the ratings proposed by the country teams. To encourage candid input from staff, the write-ups are not publicly released. Government officials can see their own country's write-up (but not those of other countries) and discuss it with the Bank country team. In contrast to some other expert-based subjective governance ratings, the CPIA's write-ups therefore can provide more transparent indications of how governments can improve their ratings. Some write-ups are more thorough than others, however. Omissions in the write-ups often prompt requests by network reviewers for additional information to be provided in revised write-ups, to enable a more informed network review.

The write-ups were extremely brief when first introduced in 2001, but have expanded steadily over the years, with little sign of convergence towards a finite length. Figure 1 shows the trend in total word count for the largest region's (AFR) write-ups. The average annual increase in the word count over the period is 26.3%. The rate of increase slowed from an average 47.7% between 2001 and 2007 to 13.2% from 2007 to 2012. In the most recent two years (2010-2012), however the annual growth rate was 13.8%. The five other regions have experienced similar increases.

[4] In recent years the Financial Management Board in OPCS has taken over lead responsibility from PRMPS for question 13 (on Quality of Budgetary and Public Financial Management). This change allows PRMPS to focus more resources on reviewing the remaining four questions.

In the IDA country allocation formula, cluster D is weighted much more heavily than the other three clusters combined (World Bank, 2010).[5] Participants in the CPIA process also regard this cluster as the most difficult to assess, because of the relative scarcity of hard data and hence greater reliance on expert judgment (IEG, 2010: 50). If staff members are trying to raise a country's ratings in order to increase its IDA allocation, the IDA formula provides a motive for focusing on the cluster D questions, and the subjective nature of some of its questions provides the opportunity. It is therefore particularly important to investigate the process and validity of the cluster D ratings, although we welcome future research into the other three clusters.

4. Data and Hypotheses

In this study the unit of analysis is the country-question; with 4 CPIA questions and 136 countries there are 544 observations in total. Two dependent variables were constructed from the write-ups accompanying the regions' 2011 proposed ratings for four of the five cluster D questions.[6] The first variable is a simple word count, measuring the length of each write-up. The second variable assigns a grade to each write-up based on the proportion of the criteria contained in the question that are addressed at all. A grade of "A" is assigned if all or nearly all (more than roughly 85%) of the criteria are addressed, and a "B" is assigned to other write-ups that cover at least one half of the criteria. The lowest grade of "D" is assigned to the relatively few cases in which all or nearly all (again, roughly 85%) of the criteria are *not* addressed. A grade of "C" applies to the remaining cases, where fewer than half of the criteria are covered, but not so few as to merit a "D" grade.

Credit is given for covering a particular criterion in the question even if the write-up addresses it only in a perfunctory manner. Moreover, no attempt was made to grade write-ups with respect to accuracy or impartiality of the information they contain. The write-up grade is therefore only a simple and partial measure of write-up quality. A more comprehensive measure of quality would not only have been far more time consuming to construct, but would have further increased the subjectivity in determining quality grades.

The mean word count is 735, with a standard deviation of 396 words. The shortest write-up is only 159 words, and the longest is 4539. Although word count is highly skewed, the log of word count approximates a normal distribution quite well. We therefore use log of the word count in all of the analyses below.

[5] The four clusters are weighted equally in computing an overall CPIA score, termed the "IRAI" (IDA Resource Allocation Index). In the Country Performance Rating (CPR) that actually determines IDA allocations, cluster D is assigned a weight of .68, the average of clusters A, B and C is assigned a weight of .24, and a portfolio performance rating receives the remaining .08 weight. These weights have been adjusted several times in determining the CPR, and the weighting of the CPR relative to per capita income in determining IDA allocations is also occasionally adjusted. However, in the CPR the CPIA's governance questions have been weighted more heavily than the other questions since 1998 (World Bank, 2010).

[6] These four include all of those for which the PREM Public Sector Governance Group has primary responsibility for the network review. The exception is question 13.

The median quality grade is a "B". Nearly 65% of write-ups covered one half or more of the criteria (i.e. received a grade of either "A" or "B"), but only 19% received an "A" grade. Only 5.5% of write-ups received a grade of "D", for addressing less than 15% of the relevant criteria. As discussed below, however, there is substantial variation across regions in quality grades.

Three other dependent variables in the analysis are constructed from the network review outputs and finalized CPIA ratings. First, we code each country-question observation for whether or not at least one network reviewer *requests more information* from the regions; such requests typically reflect a perception that the write-up is inadequate in one or more respects for making a confident assessment of the proposed rating. Second, we code for whether or not at least one network reviewer *disagrees* with the proposed rating and recommends a different rating. Third, we code for ratings adjustments, or cases where the final "official" rating differs from the one originally proposed by the regions.

IDA countries

The CPIA ratings are produced for purposes of allocating IDA funds. However, the CPIA assessments are conducted for IBRD-eligible countries as well as for IDA-eligible countries. Ratings are publicly disclosed only for the IDA countries (and only beginning in 2005). Intuitively, one would expect staff to devote extra effort to the CPIA assessments for the IDA countries, because more is at stake, and because their ratings (if not the write-ups) are publicly released so potentially subject to wider scrutiny and criticism. We therefore hypothesize that write-ups will be longer and cover more of the question criteria for IDA than for IBRD countries. If this increased effort produces more accurate and well-justified ratings, then network reviewers – other things equal – should disagree less often with the regions' proposals for IDA countries, and request additional information less often. Other things may not be equal, however. Network reviewers in turn may devote more attention to IDA country ratings, because of their importance for allocations and because they are public. Moreover, network reviewers may scrutinize IDA country ratings more closely, recognizing a certain degree of moral hazard inherent in the process.

> Although the expert judgment of Bank staff is clearly an asset in the CPIA exercise, at the same time there is a potential conflict of interest in having staff provide ratings, particularly for IDA countries. This potential for conflict of interest arises from the fact that ratings produced by staff are in turn used for allocating IDA resources for the same countries on which the work programs of those staff depend. Therefore, staff may potentially be upwardly biased in assigning ratings for their countries. (IEG, 2010: 50)

This reasoning suggests that any extra effort devoted to IDA ratings by country teams may not necessarily improve the accuracy of scores. Accordingly, network reviewers might disagree more often with IDA than with non-IDA ratings even if the write-ups for the former are lengthier and cover the criteria in the questions more completely. Most notably, the moral hazard issue introduces a strong asymmetry into the network reviews; it implies that a majority of disagreements will take the form of network reviewers recommending a rating lower than the one proposed by the region.

Benchmark countries

The CPIA process is implemented in two phases. In the benchmark phase about 20 countries are assessed. The benchmark countries change somewhat from year to year, but the sample is designed to include at least one IDA country in each region, and to represent a mix of higher- and lower-performing countries. The remaining 115 (approximately) countries are assessed in a second phase after ratings for the benchmark sample are finalized.

The timetable for the benchmark assessments is relatively long, given that there are only about 20 countries. For example, the time allotted for the network review is usually about two weeks for the benchmark phase and about three weeks for the second phase. It is doubly important to set ratings at the appropriate level for the benchmark countries, if those ratings are often used as comparators by the regions and reviewers in the second phase. In this context, the relatively generous timetable for the benchmark phase is appropriate.

The extra time available for assessing the benchmark countries, and the added importance of getting their ratings right, should affect the write-ups and review outputs. Other things equal, we would expect write-ups for the benchmark countries to be longer and to cover more of the criteria in each question. If a longer and more thorough benchmarking process produces more accurate ratings and better write-ups, then network reviewers should be less likely to request additional information from the regions and to disagree with the proposed ratings. However, network reviewers have more time available to investigate each proposed rating in the benchmark phase. Therefore, controlling for the more thorough process (using our indicator of how many of the criteria are addressed in the write-ups), reviewers might disagree and request additional information more often in the benchmark phase.

Ratings increases and decreases

Ratings changes should generally require more elaborate justifications than unchanged ratings. Country teams are perceived as having stronger incentives to increase than to decrease ratings, so they may anticipate more skeptical reactions from regional and network reviewers to proposed increases by providing lengthier and more comprehensive justifications. If it is both easier, and less important, for country teams to convince reviewers to go along with a ratings decrease than with an increase, write-ups accompanying proposed decreases may differ little from those for unchanged ratings.

Openness

More information relevant to the CPIA governance questions is likely to be available in countries with more open governments that encourage or at least tolerate freedom of the press. We measure openness using the Freedom House index of press freedoms, and hypothesize that more openness will be associated with longer and higher-quality write-ups.

Country size and income

Less information is generally available for smaller countries, particularly the "micro-states" (IEG, 2010: 46; Gelb, Ngo and Ye, 2004). For this reason write-ups are hypothesized to be longer and cover more of the question criteria in larger countries, as measured by log of population. We also control for log of per capita income. More information on governance and public sector institutions, particularly at sub-national levels, is likely to be available for higher income countries, which tend to have better communications and transportation infrastructure and more foreign investment. Among IDA-eligible countries, however, the poorer ones receive higher IDA allocations. More intensive engagement associated with higher aid levels is likely to provide country teams with more information. The net impact of per capita income is therefore theoretically ambiguous.

CMU in country

For 32 of the 136 countries, the World Bank's Country Management Unit (CMU) is based in the country. For these cases, the country team may have access to more relevant information for the CPIA. Other things equal, therefore, we would expect lengthier write-ups that cover more of the criteria, and fewer requests from reviewers for additional information. If the added information produces more accurate ratings, we might also expect to observe fewer disagreements by network reviewers with the proposed ratings. On the other hand, country teams based in Washington (or, as is sometimes the case, in a neighboring country in the region) may find it easier to maintain objectivity. Where staff are based in the country, there may be more subtle pressures to produce higher ratings, either to increase country allocations or to maintain friendly relations with government counterparts. If so, network reviewers may disagree more often with proposed ratings for these countries. The net effect of these counteracting influences may be either positive or negative.

Regional effects

All IDA-eligible and IBRD-eligible countries in the World Bank are administratively assigned to one of six regions: Sub-Saharan Africa (AFR), East Asia and Pacific (EAP), Eastern Europe and Central Asia (ECA), Latin America and Caribbean (LCR), Middle East and North Africa (MNA), or South Asia (SAR). The regional chief economists' offices have substantial discretion in how they manage the CPIA process and regional reviews. Some may encourage more input from sector specialists than others, or conduct more intensive comparative analyses, or debate more vigorously with their country teams on specific ratings. Any regional effects emerging from the tests here may reflect such differences in how the process is conducted. There are no strong theoretical reasons to expect a particular region to produce more thorough write-ups. However, one plausible argument is that the CPIA process is given higher priority in regions where most countries are IDA eligible. If so, then even when we control for IDA status of individual countries we might observe longer and more thorough write-ups on average for countries in AFR than in ECA, LCR and MNA.

In the 2007 CPIA, the network reviewers disagreed with ECA's proposed ratings most often (IEG, 2010). For all 16 CPIA questions, reviewers disagreed with 17.5% of ECA's ratings. For the other regions this figure varied from 8.6% for SAR to 12.3% for LCR. In the absence of any

substantive explanation for these regional differences, it is unclear whether these results should generalize to our analysis of the 2011 CPIA, which is limited to the cluster D ratings.

Question effects

Some questions cover more extensive issues or sub-criteria than others, so we would expect their write-ups to be longer on average. Questions have multiple components that are each assigned a sub-rating, and each component in turn lists several sub-criteria to address in the write-ups. Questions 12 and 15 each have three components, while question 16 has four and question 14 has only two. We should therefore observe longer write-ups for question 16 and shorter ones for question 14, relative to questions 12 and 15.

5. Write-ups: Word Counts

Regressions presented in Table 1 show partial correlations of write-up length with the variables described in section 4. The baseline specification in equation 1.1 shows results for the full sample of 136 countries and 544 questions. Standard errors (in this and subsequent tables) are adjusted for clustering by country, as errors for the four observations for each country are not likely to be independent.

As hypothesized, write-ups for IDA countries are significantly longer than for non-IDA countries. The IDA dummy coefficient estimate of 0.15 implies that IDA country write-ups are about 16% longer on average, other things equal. Length of write-ups for benchmark and non-benchmark countries do not differ significantly, although the coefficient for the benchmark dummy is positive as hypothesized.

Write-ups for ratings increases are also about 16% longer, other things equal. Ratings decreases, on the other hand, are not accompanied by longer justifications. The index of press freedoms, population, per capita income, and the dummy for in-country CMU are also not associated with significantly longer (or shorter) write-ups, contrary to predictions.

Coefficients for the regional dummies are interpretable relative to the omitted category, LCR. Other things equal LCR has the shortest write-ups on average. Coefficients in equation 1.1 are positive for the other five regions, and statistically significant for AFR, EAP and ECA. Relative to LCR, write-ups are longer on average by 19% for AFR, 32% for EAP, and 24% for ECA. Other than LCR, the next-shortest write-ups are for MNA, but the difference between MNA and the other four regions is significant only for EAP.

Question effects in equation 1.1 are consistent with intuition. Relative to the omitted category of question 12, write-ups are significantly shorter on average (by about 13%) for question 14, and longer for questions 15 and 16 (by about 15% and 35% respectively). Question 12 has more sub-questions (three) than question 14 (two), but fewer than question 16 (four). Question 15 also has three sub-questions, consistent with the finding that its write-ups are longer than question 12's but shorter than question 16's. Until it was revised for the 2011 CPIA, question 15 had four sub-questions. The write-ups each year are not entirely re-written in most cases, but edited from the

previous year. Many of the question 15 write-ups in 2011 still contained information applicable to some of the obsolete sub-questions. This perhaps explains why write-ups for question 15 in 2011 were significantly longer than for question 12, despite the fact both now include three sub-questions.

If equation 1.1 is run separately for each of the four CPIA questions, the IDA dummy is highly significant in question 16, and the dummy for ratings increases is significant for questions 14 and 16. Relative to LCR, 18 of the 20 regional coefficients are positively signed. Six of these 18 positive coefficients are statistically significant at the .05 level. The two negative coefficients belong to MNA, in the regressions for questions 12 and 14, but neither of these two differences with LCR is statistically significant. Differences between LCR and other regions in write-up length are most pronounced for question 15.[7]

Equations 1.2 and 1.3 in Table 1 report results for sub-samples of IDA and non-IDA countries respectively. Surprisingly, the marginal effect of ratings increases on word count is more significant and larger for non-IDA (26%) than for IDA countries (11%), although ratings affect resource allocations only for the latter. Differences between LCR and other regions are much smaller for the non-IDA than for the IDA sample. So are question effects. The explanatory power of the model is therefore much smaller for the non-IDA (R^2=.18) than for the IDA (.39) sample.

6. Write-ups: Quality Grades

As expected, longer write-ups tend to address more of the criteria in the CPIA questions. Figures 2-5 show the mean, maximum and minimum word counts for each question and quality grade combination. The average word count is consistently higher for higher grades, but there is huge variation in word counts among write-ups for a given question and a given quality grade. For question 15 (see Figure 4), the average word count is about 560 for write-ups graded "D", increasing to 690 for "C" grades, 840 for "B", and 920 for "A". Despite the positive correlation between write-up length and quality, it is clear that a lengthy write-up is neither necessary nor sufficient to address most of the criteria in the questions. The shortest write-up for question 15 with a grade of "A" is only 380 words, about two-thirds of the average length and one-third of the maximum length of "D"-graded write-ups. No write-up of 1000 or more words for questions 12, 14 or 16 was graded as low as "D", but at least one write-up of that length was graded "C" for each question.

Those write-ups that address all of the criteria in at least a cursory manner can receive the same "A" grade as a 1000-word write-up that addresses all of them in greater depth. However, the message from Figures 2-5 that lengthy write-ups sometimes neglect most of the criteria in the question is consistent with anecdotal accounts and impressions of reviewers that write-ups are often not sufficiently focused or even pertinent. Many write-ups contain extensive descriptions of ongoing efforts to pass legislation or reform procedures in ways that might eventually improve

[7] Results in this paragraph are not shown in tables for space reasons, but are available on request from the author. Disaggregating by region instead of by question adds little of interest to the analysis, in part because some regions (MNA, SAR) have very few countries.

performance on some of the criteria in the question, but without providing any indication of the current level of performance.

Average write-up grade differs by question: they tend to be highest for question 16 and lowest for question 14. Note that, unlike the case with write-up length, these differences cannot be attributed to the number of sub-ratings or criteria contained in the questions. Relatively low grades for question 15 can be attributed at least in part to major revisions in the question prior to the 2011 CPIA exercise. Many question 15 write-ups for 2011 still addressed the old criteria better than they did the new criteria. There is no obvious explanation for the low grades for question 14, on revenue mobilization. Fewer than 1 in 13 write-ups for question 14 received a grade of "A", compared to more than 1 in 4 for question 16.

Regional variations portrayed in Figure 6 are even larger than variations across questions. Only about 1 in 10 grades in EAP countries are "C" or lower, compared to more than one half in LCR. No EAP write-ups were graded "D", compared to 1 in 8 of LCR's write-ups. More than half of EAP write-ups received an "A" grade, compared to only 5.4% in LCR.

Table 2 presents multivariate tests of the association of write-up grades with word count and other variables. Equation 2.1 reports results from an ordered probit regression, while equation 2.2 reports an OLS regression for the same model specification. The dependent variable is an ordinal scale with only four categories, so ordered probit is the preferred method. We run OLS as well, however, because its coefficients (unlike the case with probit) are directly interpretable as marginal effects. Those coefficients could be misleading if the two methods generate very different results. However, the t-statistics are remarkably similar in equations 2.1 (ordered probit) and 2.2 (OLS), and the same set of variables are statistically significant in both tests. Moreover, marginal effects computed from binary probit regressions on sub-samples of observations with adjacent grades[8] indicate that little information is lost by assuming linearity and using OLS. For simplicity, the remainder of this section will therefore focus on results from OLS tests.

Lengthier write-ups receive significantly higher grades, but quantitatively the average effect is rather modest. A 1.5-unit increase in the log of word count (more than three standard deviations, or an increase from the mean of 735 words to about 3000 words) is required to produce an increase of one grade level, e.g. from "C" to "B".

Grades are not significantly different for IDA or benchmark countries, or for proposed ratings increases or decreases from the previous year. Grades are significantly higher in larger countries and in those with more press freedoms. In-country CMUs are associated with lower average grades. Grades are lower for LCR than for other regions, and differences between LCR and three regions (EAP, ECA and MNA) are statistically significant. Other things equal, grades are nearly a full level higher on average for EAP than for LCR countries. The mean grade (calculated by assigning scores of 4, 3, 2 and 1 respectively to grades A, B, C and D) is 3.41 for EAP and 2.40 for LCR, for a difference of 1.01. The EAP regression coefficient of .835 in

[8] Specifically, three probit regressions were run on observations with "A" and "B", "B and "C", and "C" and "D" grades respectively. Marginal effects for most variables (calculated at the mean values of all other regressors) are very similar across the three sub-samples.

equation 2.2 implies that differences in write-up length and other variables in the model account for less than one-fifth of this full-grade difference between average write-up grades in EAP and LCR.

Grades for questions 14 and 15 are significantly lower, other things equal, than for questions 12 and 16. The largest negative coefficient belongs to question 15, the one with criteria that were substantially revised between 2010 and 2011.

Equations 2.3 and 2.4 report similar regressions, but for the IDA and non-IDA sub-samples respectively. Results are broadly similar. The word count coefficient is somewhat larger in the non-IDA sample, and the EAP coefficient is somewhat larger for the IDA sample. The negative effect of in-country CMU is significant only for the non-IDA sample. The large difference in write-up quality between question 12 and questions 14 and 15 observed for the full sample widens further in the IDA sample, but narrows in the non-IDA sample.

To summarize, two regions stand out from the other four with respect to average write-ups grades, EAP positively and LCR negatively. Although length of write-ups is positively related to quality, it accounts for very little of the difference between EAP and LCR. Somewhat surprisingly, grades are not significantly higher for IDA or benchmark countries, or for ratings changes.

7. Network Review

The network review phase of the CPIA process generates a set of comments on some of the proposed ratings. Many comments express disagreement with proposed ratings, but others simply request a revised write-up that better addresses the criteria in the question. This section analyzes the determinants of both of these types of responses, information requests and disagreements with ratings. It also looks at factors associated with ratings adjustments, i.e. the subset of disagreements resolved in favor of network recommendations.

In the 2011 CPIA process, network reviewers requested additional information for 8.1% of all proposed ratings. As shown in Figure 7, there is substantial regional variation in this figure, ranging from a minimum of 1.3% for EAP to a maximum of 10.2% for AFR. Much of these regional differences can be attributed to write-up volume and quality grades.

Equation 3.1 in Table 3 reports results from a probit regression, where the dependent variable is coded 1 for the 44 country-question observations (8.1% of the sample) where network reviewers requested additional information, and 0 for the other 500 (91.9%). The variable coefficients reported have been transformed to represent the marginal effects of a one-unit increase, evaluated at the mean value of all other independent variables.

Each letter-grade increase in quality of the write-up is associated with a highly significant three percentage point drop in the likelihood that a network requests more information. Controlling for the write-up grade, higher word counts also significantly reduce the probability of a request for more information. Presumably this result reflects the fact that write-up grade is an

incomplete measure of quality, and that longer write-ups not only tend to address more of the criteria in the question but also (more often than not) to address them in greater depth. Equation 3.2 presents the reduced-form estimate of the word count effect. When write-up grade is not controlled for in equation 3.2, the word count coefficient is four times as large as in equation 4.1. Note that the explanatory power of the model drops by one third when grade is omitted: the R^2 is .39 in equation 3.1 but only .26 in equation 3.2.

Network reviewers appear to devote extra attention to ratings for IDA countries. Other things equal, the probability of an information request is 8 percentage points higher if the proposed rating belongs to an IDA country. For proposed increases in a rating, the impact is an even larger 18 percentage points. Reviewers were also significantly more likely to request additional information when a ratings *de*crease was proposed, but this marginal effect is only 6 percentage points, one-third as large as for a ratings increase.

While proposed *changes* generate more reviewer requests, higher ratings *levels* do not. Conceivably, reviewers might want to scrutinize more closely any proposals for a high rating, whether or not it represents a change. However, no support is found in the data for this conjecture.

Benchmark countries are also not associated with more frequent information requests. Reviewers may have more time in the benchmark phase to identify informational shortcomings in the write-ups, but country teams may also have more time in that phase to produce more complete write-ups. Given these countervailing forces the absence of a significant benchmark effect is not surprising.

Regional differences in information requests are small in equation 3.1, controlling for write-up grades. In equation 3.2, where write-up grade is not controlled for, the likelihood of information requests differs trivially among AFR, ECA, LCR and MNA, but relative to those four is significantly lower in EAP and SAR (by about 4 percentage points). Information requests are significantly more frequent for question 15 (in equation 3.2), the question which experienced the most substantial revisions to its criteria leading into the 2011 CPIA exercise. Controlling for the lower average quality grade of question 15 write-ups in equation 3.1, however, this difference is not significant.

Network reviewers disagreed with the regions' proposals in 7.9% of cases. As Figure 8 shows, there is again large variation across regions, from a minimum of 2.6% for EAP to a maximum of 17.4% for ECA. Most of this variation turns out to be attributable to differences in the number of proposed ratings increases; the frequency of ECA's proposed increases (19.6%) is more than double that of the five other regions collectively (9.1%).

Table 4 analyzes in detail the determinants of networks' propensity to disagree with proposed ratings. The dependent variable, "disagreement," is coded 1 for the 42 cases where network reviewers expressed disagreement with a proposed rating and recommended a different rating, and is coded 0 for the other 502 observations. Of these 42 cases, 31 represent proposed increases, 1 a proposed decrease, and 10 were unchanged from the previous year. There were 59

proposed increases in total, so networks disagreed with a slight majority of them. They disagreed with only 3% (1 of 32) of decreases, and 2% (9 of 453) of unchanged ratings.[9]

Network reviewers are less likely to take issue with a proposed rating when it is accompanied by a more thorough write-up. As shown in equation 4.1, each one-grade increase in write-up grade is associated with a (statistically significant) reduction of 2.2 percentage points in the probability of a disagreement. By far the most important predictor of disagreement is the ratings increase dummy. The likelihood of disagreeing rises by 45 percentage points, other things equal, when a ratings increase is proposed.

Regional differences are small. The omitted category is ECA. Coefficients for the other 5 regions are negative, but small in magnitude (1.5 percentage points or less). Only the MNA coefficient is (borderline) significant. Question effects are more substantial. Other things equal, reviewers were significantly less likely to disagree with ratings for questions 14 and 15 than for question 12. The substantial revisions to the question 15 criteria leading into the 2011 CPIA exercise may be responsible for lower write-up quality grades (Table 2) and thus for more requests for additional information in the review process (Table 3, equation 3.2). But results in Table 4 suggest that they do not appear to have produced more frequent disagreements over ratings.

Equation 4.2 excludes the ratings increase dummy from the model. Most notably, the explanatory power of the model plunges from .52 to only .23. Also, regional effects become much more strongly negative (relative to ECA) and more significant in equation 4.2, where we are not controlling for the fact that ECA proposes a far greater number of ratings increases than the other regions. This finding is consistent with IEG (2010), which reports that network reviewers disagreed more often with ECA's proposed ratings, in its analysis of all 16 questions for the 2007 CPIA.

Word count of write-ups is not significant in equation 4.1, but in equation 4.2, where ratings increases are not controlled for, it is associated with an increase in the likelihood of disagreement over ratings. Regional staff may anticipate network disagreement with more dubious proposals for an increase, and provide longer write-ups in attempting to justify them.

Ratings level is also significant in equation 4.2, but not in equation 4.1 where proposed increases are controlled for. When ratings level is dropped from the model in equation 4.3, the R^2 declines even further (from .23 to .17), and regional differences are further accentuated. Taken together, results in equations 4.1-4.3 indicate that the greater propensity of network reviewers to disagree with ECA ratings is mostly attributable to the fact that this region compared to the other five proposes more high ratings, and more increases.

Results are broadly similar for the sub-sample of IDA countries. In equation 4.4, which replicates equation 4.1 for that sub-sample, the coefficient on proposed increases remains highly significant, but is somewhat smaller in magnitude. For the IDA sample, in-country CMUs are associated with a significantly lower probability of disagreement.

[9] In several other cases, network reviewers recommended increasing sub-ratings, but with no implications for the overall question rating.

Table 5 analyzes ratings "adjustments." For these probit regressions, country-question observations are coded 1 if the final, official rating for 2011 set following the review process differs from the one proposed by the regions and sent to the networks and central units for review. "Adjustment" is coded 0 if and only if the final rating is the same as the one proposed by the region.

There were 36 adjustments, out of the 42 instances in which network reviewers had disagreed with the proposed rating. As in the IEG's (2010) analysis of 2007 ratings, therefore, in most cases of disagreement the network recommendations prevailed in 2011. By this measure, the network review appears to have a significant influence on the ratings. Presumably, the neutral role and cross-regional perspectives of the networks serve to improve the accuracy of the ratings overall, even if there is no way of knowing what the "true" rating should be in each instance of ratings disagreement or adjustment. Because most adjustments are downward rather than upward, the network reviews also help counteract ratings inflation over time, although again there is no way of knowing the "true" time trend of average ratings for each question.

The probability of an adjustment is about 5 percentage points higher for ECA's ratings than for the other regions. Equation 5.1 of Table 5 presents regional and question effects, without controlling for any other variables. Ratings for question 16 are significantly more likely (again, by about 5 percentage points) to be adjusted than those for the other three questions.

Equation 5.2 controls for the effects of proposed increases and other variables on the likelihood of ratings adjustments. Proposed increases are the most powerful predictor of ratings adjustments. Higher ratings levels and lower write-up grades are also associated with an increased likelihood that the final rating is different from the one proposed by the region. The regional effects in equation 5.1 largely disappear in equation 5.2: controlling for ECA's propensity to propose more ratings increases, its ratings are no more likely to be adjusted than those of other regions. Ratings for questions 14 and 15 are significantly less likely to be adjusted than those for questions 12 and 16.

8. Testing for Regional or IDA Bias

As IEG (2010: 50) notes, the reviews conducted by the six regional Chief Economists' offices are intended mostly to correct for over-exuberance regarding individual countries' ratings, but "there could still be issues" regarding *inter*-regional comparability "even if the relative rankings of countries are adjusted" appropriately within each *intra*-regional review.

Gelb, Ngo and Ye (2004) test for an Africa bias in the CPIA, but do not test for any other regional biases. Specifically, they regress CPIA cluster D scores on a simple average of the six Worldwide Governance Indicators (WGI) indexes and on a dummy for Sub-Saharan African countries. The CPIA-WGI relationship is very strong and significant, but the Africa dummy is insignificant. Similarly, they report no Africa bias in regressing either cluster D or cluster B scores on the Heritage Foundation's Economic Freedom Index, or in regressing cluster C scores on the UNDP's Human Development Index.

In attempting to shed light on the validity of CPIA ratings, IEG (2010) reports correlations between some of the CPIA ratings and other related indicators. It does not test for possible regional biases, however. Nor does IEG (2010) test for an IDA country bias, despite noting repeatedly that the use of CPIA scores for IDA allocations created incentives for country teams to inflate ratings. If it is common for regional staff to inflate ratings proposals for the purpose of increased IDA allocations, then we would expect to observe a positive, significant coefficient on an IDA country dummy included in regressions of CPIA ratings on related indicators.

There is never perfect conceptual overlap between any single CPIA question (or cluster of questions) and related indicators produced by other organizations. The lack of a perfect or even close relationship empirically does not necessarily indicate that the CPIA (or the related indicator) is invalid. For any individual country, a higher (or lower) ranking in the CPIA than on the related indicator could reflect less-than-perfect conceptual overlap, or measurement error in the CPIA, or measurement error in the related indicator. For a large group of countries, however, it is more difficult to dismiss systematic discrepancies between the CPIA and related indicators that tend to favor IDA countries, or countries from a particular region. If CPIA ratings are significantly higher than predicted – based on the values of a related indicator – for IDA countries or for an entire region, it cannot plausibly be attributed to random measurement error in the full sample of more than 130 countries covered by the CPIA. Conceivably, a source of comparator data may contain its own regional biases that would contaminate tests for bias in CPIA. For this reason our tests use a range of related indicators, from several largely independent data sources. It is highly unlikely that these sources would all exhibit the same regional biases.

Table 6 presents a pair of OLS regressions for each of the four cluster D questions (12, 14, 15 and 16). The dependent variable in the first of each pair is the region's proposed rating in the 2012 CPIA exercise, and in the second of each pair is the final 2012 rating. Independent variables include a conceptually related "comparator" indicator, log of per capita income, an IDA dummy, and an ECA dummy. Preliminary tests showed that ECA's coefficients were consistently larger (more positive) than those for other regions and variations among the other five regions were small. We therefore test only one regional dummy in Table 6, focusing on the issue of whether ECA ratings are significantly higher than those of countries in the other regions combined, controlling for their ratings on comparator indicators and per capita incomes.

The regressions within each pair are identical other than using proposed ratings and final ratings as alternative dependent variables. Reporting these tests side-by-side shows the extent to which any IDA or regional biases in proposed ratings are dampened or eliminated by the network review.[10]

Equations 6.1 and 6.2 analyze question 12 ratings. The related comparator indicator is an index of "guidepost indicators" constructed for purposes of the network review, and designed to match

[10] This method captures only the immediate direct effects of the network review. From a longer-term perspective the existence of the reviews can deter regions from proposing ratings that are higher than justified.

the criteria in question 12 as closely as possible.[11] The guidepost index is positively and significantly related to question 12 ratings. The IDA dummy coefficient is very small and insignificant in both equations. Based on this finding, the large weight assigned to CPIA governance questions in the IDA allocation formula does not appear to be distorting the ratings.

The ECA coefficient is positive and highly significant in both equations 6.1 and 6.2. Question 12 proposed ratings for ECA countries are more than one-fourth of a point higher than predicted from their incomes and guidepost index scores (equation 6.1). Following the network review, this positive ECA effect is basically unchanged, increasing from .275 to .276.

Equations 6.3 and 6.4 analyze question 14 ratings, on revenue mobilization. The two related regressors are from the "Paying Taxes" component of the Doing Business project. Tax rates paid by a "typical firm" pertain to the first sub-rating in question 14 on tax policy. Number of distinct tax payments that a "typical firm" must pay pertains mostly to the second sub-rating on tax administration. Both variables are significantly and negatively associated with question 14 ratings, as expected. The effects are quantitatively small, however: a two standard deviation increase in both tax rate (equal to 80% of firm profits) and number of tax payments (equal to 40) would be required to increase the question 14 rating by one-half point. Moreover, these two indicators do not address many of the criteria in question 14. Accordingly, the R^2 in equations 6.3 and 6.4 is much lower than in the other regressions reported in Table 6. As in equations 6.1 and 6.2, the IDA coefficient is negative but insignificant.

The ECA coefficient is positive and marginally significant in equations 6.3 and 6.4. Proposed ratings for ECA countries are nearly one-quarter of a point higher than predicted on average. For the final ratings, this effect shrinks to one-fifth of a point. The review process appears to reduce the bias somewhat for question 14.

Question 15's proposed and final ratings are the dependent variables in equations 6.5 and 6.6, respectively. These regressions control for an index of two related indicators from the Economist Intelligence Unit's (EIU) country risk ratings. One is on bureaucratic quality (including meritocracy), and the other is on red tape encountered in dealing with the government bureaucracy. This EIU bureaucracy index is very strongly related to question 15 ratings. The IDA dummy coefficient is again very small and insignificant in equations 6.5 and 6.6. The ECA coefficient is positive and significant at the .01 level in both regressions. The average question 15 proposed ratings for ECA countries are more than one-third of a point higher than predicted. The network review process has only a minimal impact on this bias, as the coefficient declines from .363 in equation 6.5 to .355 in equation 6.6.[12]

Finally, equations 6.7 and 6.8 test for IDA and regional biases in question 16 ratings. The comparator variable is an index of guidepost indicators, constructed to match as closely as

[11] The index includes 19 variables from five sources: the World Economic Forum's (WEF) "Executive Opinion Survey," the International Country Risk Guide (ICRG), Economist Intelligence Unit (EIU), Freedom House, and the Heritage Foundation. Data are all from 2012. The ICRG and EIU ratings are updated monthly, and the others annually.

[12] Results are very similar if we replicate this exercise substituting the bureaucratic quality indicator from International Country Risk Guide for the EIU indicators.

possible the criteria in question 16.[13] Again, the IDA dummy is not significant, but the ECA dummy is positive and significant at the .05 level in both regressions. The average ECA country proposed rating (equation 6.7) is about one-fifth of a point higher than predicted. Following the network review process, the magnitude of the bias is reduced slightly, from .192 to .182.[14]

The network review has the effect of strengthening (albeit modestly) the relationship between CPIA ratings and comparator indicators. The coefficient on the comparator indicator is slightly greater in the second regression within each pair, for questions 12, 15 and 16. This result is not surprising, as the network reviewers make use of these indicators (as well as other quantitative and qualitative information) in formulating their ratings recommendations.

Table 7 reports regressions similar to those for the even-numbered equations in Table 6, i.e. using the final rather than proposed ratings as dependent variables. They differ however by specifying dummy variables for the other 5 regions, with ECA as the base category. The regional dummy coefficients in this table thus show pair-wise comparisons between ECA and any other single region. Of the 20 regional dummy coefficients in Table 7, 19 are negative (all but the one that is smallest in absolute value), and 8 are statistically significant. Differences with ECA are smaller for SAR than for the other four regions. The largest single coefficient is for the MNA dummy in equation 7.3: ECA ratings on question 15 are more than ½ point higher than SAR's, controlling for the EIU comparator indicator and other factors.

The results in Table 6 are reassuring in finding no evidence at all that the CPIA governance ratings incorporate an IDA country bias. These findings do not definitively reject the possibility of an upward bias in the ratings, but any such bias must apply equally to non-IDA countries. An IDA bias does not even show up in the regions' proposed ratings, suggesting that the regions' own internal review procedures are successful in deterring or purging any IDA bias in their ratings before they are forwarded to central units for review.

There is evidence, however, of a significant regional bias. Coefficients for the ECA dummy are positive and at least marginally significant in all eight regressions. A comparison of proposed and final ratings indicates that the network review has only a modest impact in curbing this bias. The mean bias over the four questions is slightly more than one-fourth of a point (0.265) in the proposed ratings, and it is reduced only by about 4% (to 0.254) in the final ratings.

If reform progress is spatially correlated and the guidepost indexes reflect lagged information, then a positive and significant regional dummy coefficient in these tests would not necessarily reflect ratings bias. The CPIA ratings might simply reflect more up to date information about progress in one region relative to the other five. However, the network reviews conduct regional comparisons every year similar to those reported in Table 6, and a similar "ECA effect" has been

[13] The index includes 20 variables from five sources: the World Economic Forum's (WEF) "Executive Opinion Survey," the International Country Risk Guide (ICRG), Economist Intelligence Unit (EIU), Freedom House, and Reporters Without Borders. Data are all from 2012. The ICRG and EIU ratings are updated monthly, and the others annually.

[14] Results for questions 12 and 16 in the table are very similar if the 2011 Worldwide Governance Indicators (WGI) "Rule of Law," "Control of Corruption" and "Voice and Accountability" indexes are substituted for the guidepost indexes. The WGI indexes include more sources than the guidepost indexes, but are not designed to match the content of the CPIA criteria and contain less up to date information.

present for at least several years. It seems unlikely that an information lag for the EIU, WEF and other sources relative to the CPIA would persist for more than a few years at most.

Not all "expert" ratings are necessarily equal: some sources may have less accurate or up to date information than others on governance and public sector reform in ECA. One source that specializes in ECA countries (with the exception of Turkey) is the annual Freedom House "Nations in Transit" (NIT) report. This report provides detailed country narratives as well as quantitative indicators designed to be comparable over time. The NIT produces ratings on three indicators pertaining to questions 12 and 16 in the CPIA: "judicial framework and independence," "corruption," and "independent media."[15] In its 2012 edition, closely corresponding to the 2011 CPIA in its timing, only 2 of the ratings were upgraded (for countries also included in the CPIA), and 10 were downgraded. In contrast, the region proposed 12 increases and 0 decreases for question 12 and 16 ratings. In its 2013 edition, the NIT upgraded 6 ratings on these questions but downgraded 13 others, for countries covered by the CPIA. The source of comparator indicators that is arguably the best informed about ECA thus appears to disagree markedly with Bank staff on the extent of improvements in governance and public sector institutions.

The ECA region is comprised of several distinct sub-regional groupings, ranging from new EU members to Central Asian republics. We therefore experimented with different sub-regional dummies, to determine which group or groups might be driving the positive ECA effect found in Table 6. The group with the largest positive bias (i.e. CPIA ratings higher than predicted by comparator indicators) turns out to be the EU accession (Croatia, entering in July 2013) and candidate (Macedonia, Montenegro, Serbia and Turkey) countries.

Accession and candidacy should be associated with improvements in governance, so this result may not appear surprising. Although the three ECA countries in the 2012 CPIA that joined the EU earlier (Poland, Bulgaria and Romania) exhibit less of a positive bias, it could be that sources of comparator indicators have had more time to learn about their governance improvements, and catch up with any information advantage temporarily possessed by Bank staff and reflected in their CPIA assessments. Conversely, the comparator indicators may be lagging, relative to the CPIA, with respect to information on improvements in the five accession and candidate countries.

However, the Freedom House NIT reports are the least likely of comparator indicators to contain lagging information, and arguably should be as accurate and up to date as the CPIA on the issues it covers. In the four accession and candidate countries covered by the NIT, it downgraded one rating (Macedonia on "judicial framework and independence") and upgraded none in its 2013 report. In 2012, it upgraded one and downgraded one, and in 2011 it upgraded two and downgraded two. The NIT's ratings are thus inconsistent with the view that governance in these accession and candidate countries is improving so rapidly that assessments by less specialized sources of comparator indicators are hopelessly lagging.

[15] These NIT indicators are not used as components of the "guidepost indexes" for questions 12 and 16 used in Table 6, because the NIT does not cover countries in any of the other 5 regions.

In any event, while the positive "ECA effect" demonstrated in Table 6 is more attributable to the accession and candidates countries than to other sub-groups in the region, it is not solely attributable to them. Even if these 5 countries are dropped entirely from the Table 6 regressions, the ECA coefficient remains positive and is still statistically significant in many of the regressions.

If ratings disagreements are usually resolved in favor of the networks' recommendations, one might ask why the positive "ECA effect" observed in Table 6 diminishes only slightly using ratings from before versus after the network review. There are two likely explanations. First, in borderline cases the networks tend to defer to the regions' views, and ECA is by far the most assertive region in terms of proposed increases, and many of these "extra" increases will represent borderline cases. Second, the ECA coefficient reflects a cumulative effect over many years, and the reviews in any one year mostly (but not entirely) focus on newly-proposed increases.

9. Implications and Recommendations

This analysis of the validity of the CPIA governance questions finds significant regional differences in the coverage of criteria by the written justifications accompanying the regions' proposed CPIA ratings. It shows that the length of the write-ups has steadily increased over time, with little sign of leveling off. Although write-ups with higher quality grades are also longer on average, there is wide dispersion in the word count for any given grade, and some long write-ups provide little relevant information.

The analysis also examined network reviewer responses to ratings proposals, as a function of the quality and length of write-ups. Higher grades are associated with a lower likelihood that central unit reviewers will either disagree with proposed ratings, or request additional information to assess the proposed ratings. Controlling for grades, longer write-ups are actually associated with a *greater* probability that central reviewers will disagree with a proposed rating. If a region wishes to avoid time-intensive back-and-forth exchanges with the central units over its ratings proposals, it should therefore provide relevant and thorough but reasonably concise write-ups to support them.

Using related "comparator" indicators from other sources, we find no evidence of a pro-IDA country bias in the ratings. This is a striking finding, given the central role of country teams in the CPIA process and the fact that aid allocations are highly sensitive to ratings on the governance questions. However, we find a significant upward bias in ratings for one region (ECA), and show that it is only slightly reduced by the network review process.

This study's findings have several implications for improvements in the CPIA process. First, although the write-ups justifying the proposed ratings are a major strength of the process, there is room for further improvements. The write-ups continue to increase in volume, but much of the incremental information may be of only limited relevance. In the most extreme case, one write-up of well over 1000 words neglects to address nearly all of the criteria in the question. Moreover, there are enormous differences across the regions in how well the write-ups cover the

criteria. Only one tenth of EAP write-ups are graded "C" or below (meaning they address less than one half of the criteria in the question), compared to over one half for LCR. Strengthening the guidance provided by central units to regional staff regarding the cluster D questions could help somewhat in steering the write-ups toward provision of more pertinent information. For example, certain data sources (such as the World Bank's own Enterprise Surveys) are under-utilized, while others (such as TI's Corruption Perceptions Index) are often cited in ways that do not provide meaningful comparative information across countries or over time. However, the payoff to strengthening guidance notes may be limited, as it is unclear whether many of the relevant staff members currently make use of the existing guidance.

Second, the network reviews could also more assertively request additional information in a much larger number of cases. The network with the lead role in reviewing question 13 (on budgetary management) employed this strategy with some success over the course of several years. It might be overly ambitious to implement this approach simultaneously for all four cluster D questions for which the Public Sector Governance Group has the lead review role. However, it might be feasible to implement it more gradually, e.g. focusing initially on the question with the lowest average quality grades (14, on revenue mobilization), or starting with the grade "D" cases (where most criteria go unaddressed) for all four questions and moving on the following year to the "C" cases.

Third, the network reviews should be more assertive about questioning proposed ratings even when no increase is being proposed, and in borderline cases. Reviews that are overly focused on rating changes produce a status quo bias in the ratings, and reduce the scope for eliminating regional or other biases when they appear in the ratings.

Finally, the process would benefit from increased checks and balances within each region, including an enhanced role in at least some regions for sector specialists. The most striking results from the analyses above involve regional comparisons. On quality of write-ups, EAP is the positive outlier, and could serve as a useful model for other regions to follow. In contrast, LCR is a negative outlier as the region with the least pertinent write-ups.

On realism of proposed ratings, ECA stands out as the region that appears to be most over-optimistic in its assessments. The network reviews can play a useful role in minimizing any regional or other biases that emerge in the ratings, as well as in alerting regional staff on the need to improve the quality of write-ups in more instances. The regional chief economists' offices arguably can achieve the same or better results, however, at lower cost[16] and via interactions with regional colleagues that country teams might find more credible.

[16] Disagreements by network reviewers typically occur after in-depth analysis of individual outliers identified from cross-country statistical analyses. Conducting these analyses and crafting the written arguments to challenge 42 proposed ratings is a fairly time-intensive task for network reviewers. If other regions were as assertive as ECA in proposing increases, and network reviewers disagreed with them at the same rate, there would have been about 75 disagreements instead of 42.

According to IEG (2010: 49), interviews with regional reviewers suggested that some of these regional chief economist's offices play a much more active role than others.[17] As merely one small example, these offices could provide their own guidance on useful region-specific information sources. More importantly, in those regions where sector specialists have a relatively minimal role in the process, the regional chief economists' offices should take steps to enhance their role.

This study does not intend to represent the last word on validity and reliability of the CPIA ratings. Assessments of the quality of write-ups could also be broadened and deepened. Another important limitation of this study is that it does not analyze changes in write-up length, criteria coverage or ratings over time. Adding a time dimension to the analysis would also make it possible to test other independent variables, such as turnover of key staff involved in the CPIA process. Results of such a study would be useful in assessing the validity of CPIA time series data. Although the CPIA is intended primarily for comparing countries to each other at a point in time, analysts are often tempted to use it for assessing trends (e.g. World Bank, 2006). Use of the data in this way is premised on the untested assumption that year to year changes primarily reflect real progress (or deterioration). Alternatively, a large proportion of changes over time may merely reflect belated corrections to incomplete information, or inferences from changes in economic performance, or even changes in the identity of staff responsible for the ratings.

Finally, further research could also add to the debates on the appropriate content of the CPIA and on the appropriate question or cluster weighting in the IDA allocation formula. Analyses of the validity of the other three CPIA clusters could usefully complement this one. Results for cluster D questions – which are more subjective, and count more for IDA allocations – will not necessarily generalize to clusters A (macro policies), B (structural policies) and C (social sector and environmental policies).

[17] It should be noted however that regional differences likely go well beyond the identity of staff in the RCE offices. E.g., the ECA region has been easily the most assertive region in proposing increases for a period encompassing the tenure of at least two RCEs.

REFERENCES

Cage, Julia, 2009. "Growth, Poverty Reduction and Governance in Developing Countries: A Survey." CEPREMAP Working Papers 0904, CEPREMAP.

Collier, Paul and David Dollar, 2002. "Aid Allocation and Poverty Reduction." *European Economic Review* 46: 1475-1500.

Denizer, Cevdet; Daniel Kaufmann and Aart Kraay (2011). "Good Countries or Good Projects? Macro and Micro Correlates of World Bank Project Performance." World Bank Policy Research Working Paper 5646. Washington DC: The World Bank.

Dollar, David and Victoria Levin, 2006. "The Increasing Selectivity of Foreign Aid, 1984-2003." *World Development* 34(12): 2034-46.

Gelb, Alan, Brian Ngo and Xiao Ye, 2004. "Implementing Performance-Based Aid in Africa: The Country Policy and Institutional Assessment." Africa Region Working Paper Series No. 77. Washington, DC: The World Bank.

IEG (Independent Evaluation Group), 2010. *The World Bank's Country Policy and Institutional Assessments: an IEG Evaluation*. Washington DC: The World Bank.

Kanbur, Ravi, 2005. "Reforming the Formula: A Modest Proposal for Introducing Development Outcomes in IDA Allocation Procedures." *Revue d'Economie du Developpment* 79-99.

Knack, Stephen, 2013. "Building or Bypassing Recipient Country Systems: Are Donors Defying the Paris Declaration?" World Bank Policy Research Working Paper 6423. Washington DC: The World Bank.

Knack, Stephen, 2009. "Sovereign Rents and Quality of Tax Policy and Administration." *Journal of Comparative Economics*, 37(3), 359-71.

Knack, Stephen, F. Halsey Rogers and Nicholas Eubank, 2011. Aid Quality and Donor Rankings. *World Development* 39(19), 1907-17.

OECD, 2011. *Aid Effectiveness, 2005-10: Progress in Implementing the Paris Declaration*. OECD, Paris.

Steets, Julia, 2008. "Adaptation and Refinement of the World Bank's Country Policy and Institutional Assessment (CPIA)." Global Public Policy Institute, on behalf of the German Federal Ministry for Economic Cooperation and Development.

World Bank, 2011a. *Global Monitoring Report 2011: Improving the Odds of Achieving the MDGs*. World Bank, Washington DC.

World Bank, 2011b. "Country Policy and Institutional Assessment 2011: Assessment Questionnaire." World Bank, Washington DC.

World Bank, 2010. "IDA's Performance Based Allocation System: Review of the Current System and Key Issues for IDA16." World Bank, Washington DC.

World Bank, 2007. "Selectivity and Performance: IDA's Country Assessment and Development Effectiveness."

World Bank, 2006. *Global Monitoring Report 2006: Strengthening Mutual Accountability, Aid, Trade and Governance*. World Bank, Washington DC.

World Bank, 2004. “Country Policy and Institutional Assessment: An External Panel Review.” World Bank, Washington DC.

Table 1
Word count in write-ups

Equation	1.1 (All)	1.2 (IDA)	1.3 (non-IDA)
IDA	0.151**		
	(1.97)		
Benchmark country	0.064	0.111	0.020
	(0.88)	(1.24)	(0.22)
Increase proposed	0.153***	0.098	0.228**
	(2.97)	(1.63)	(2.56)
Decrease proposed	0.012	0.091	-0.039
	(0.15)	(0.82)	(-0.36)
Freedom House press	-0.003	-0.002	-0.002
freedoms	(-1.56)	(-1.18)	(-0.65)
Log of population	0.001	0.017	-0.010
	(0.07)	(0.70)	(-0.48)
Log of per capita GNI	-0.035	-0.047	-0.024
	(-0.93)	(-0.85)	(-0.37)
CMU in country	0.075	0.112	0.072
	(0.99)	(0.89)	(0.85)
Sub-Saharan Africa	0.174**	0.201*	0.239**
	(2.30)	(1.71)	(2.28)
East Asia & Pacific	0.274***	0.371***	0.259**
	(4.11)	(3.83)	(2.44)
East Europe & Central Asia	0.213**	0.429***	0.095
	(2.56)	(3.23)	(0.90)
Middle East & North Africa	0.071	0.378**	0.027
	(0.76)	(2.12)	(0.25)
South Asia	0.182	0.199	
	(1.15)	(1.19)	
Question 14	-0.142***	-0.169***	-0.104
	(-3.54)	(-3.56)	(-1.46)
Question 15	0.142***	0.175***	0.096
	(3.36)	(3.48)	(1.29)
Question 16	0.303***	0.413***	0.154***
	(8.19)	(9.07)	(2.66)
Constant	1.910	1.766	2.003
	(3.68)	(2.65)	(2.64)
No. of observations	544	316	228
No. of countries	136	79	57
R^2	.32	.39	.18

Dependent variable is log of number of words in question write-up. Unit of analysis is country-question. Omitted region is Latin America and Caribbean. Omitted question is 12. T-statistics, reported in parentheses below point estimates, are based on standard errors adjusted for non-independence of errors within regional clusters of observations, with *** p<0.01, ** p<0.05, * p<0.1.

Table 2
Coverage of criteria in write-ups

Equation	2.1	2.2	2.3	2.4
Method	Ordered probit	OLS	OLS	OLS
Sample	All	All	IDA	Non-IDA
Word count (log)	1.143***	0.672***	0.655***	0.734***
	(8.02)	(8.44)	(5.21)	(6.28)
IDA	0.082	0.039		
	(0.43)	(0.35)		
Benchmark country	0.061	0.036	0.036	0.130
	(1.45)	(0.42)	(0.30)	(0.87)
Increase proposed	0.005	0.002	0.105	-0.235
	(0.03)	(0.02)	(0.98)	(-1.50)
Decrease proposed	0.225	0.133	0.010	0.161
	(0.92)	(0.92)	(0.05)	(0.89)
Freedom House press	0.009***	0.006***	0.006*	0.005**
freedoms	(2.68)	(2.78)	(1.86)	(2.07)
Log of population	0.139***	0.083***	0.075**	0.074*
	(3.04)	(3.01)	(2.09)	(1.92)
Log of per capita GNI	0.078	0.045	0.002	0.165*
	(0.83)	(0.79)	(0.03)	(1.69)
CMU in country	-0.379**	-0.224**	-0.074	-0.338**
	(-2.22)	(-2.14)	(-0.46)	(-2.56)
Sub-Saharan Africa	0.233	0.158	0.190	0.067
	(1.37)	(1.51)	(1.17)	(0.50)
East Asia & Pacific	1.440***	0.835***	0.984***	0.760***
	(5.73)	(6.06)	(5.06)	(3.81)
East Europe & Central Asia	0.371*	0.233**	0.348*	0.201
	(1.92)	(1.97)	(1.87)	(1.46)
Middle East & North Africa	0.589**	0.359**	0.400	0.372*
	(2.27)	(2.29)	(1.18)	(1.92)
South Asia	0.519	0.313	0.307	
	(1.44)	(1.45)	(1.32)	
Question 14	-0.399***	-0.241***	-0.327***	-0.120***
	(-3.51)	(-3.44)	(-3.41)	(-1.12)
Question 15	-0.509***	-0.311***	-0.538***	0.013
	(-3.44)	(-3.55)	(-5.17)	(0.09)
Question 16	0.102	0.049	-1.08	0.273***
	(0.93)	(0.78)	(-1.31)	(2.66)
Constant		-0.510	0.007	-1.607
		(-0.68)	(0.01)	(-1.50)
No. of obs., countries	544, 136	544, 136	316, 79	228, 136
Pseudo R^2 or R^2	.17	.33	.40	.31

Dependent variable is "Grade" (A, B, C or D) indicating coverage of criteria in question. Unit of analysis is country-question. Omitted region is Latin America and Caribbean. Omitted question is 12. T-statistics, reported in parentheses below point estimates, are based on standard errors adjusted for non-independence of errors within regional clusters of observations, with *** p<0.01, ** p<0.05, * p<0.1.

Table 3
Network review (probit regressions)

Equation	3.1	3.2
Dependent variable	Request info	Request info
Proposed rating (1-6 scale)	-0.001	-0.006
	(-0.14)	(-0.52)
Grade (coverage of criteria)	-0.029***	
	(-5.60)	
Word count (log)	-0.017**	-0.069***
	(-2.03)	(-3.72)
IDA	0.083***	0.113***
	(5.33)	(4.55)
Benchmark country	-0.005	-0.011
	(-0.64)	(-0.88)
Increase proposed	0.179***	0.219***
	(4.61)	(4.52)
Decrease proposed	0.058**	0.067*
	(2.17)	(1.76)
Freedom House press	0.0002	0.0003
Freedoms	(0.97)	(0.60)
Log of population	0.007***	0.006
	(2.65)	(1.42)
Log of per capita GNI	0.010**	0.019*
	(1.96)	(1.70)
CMU in country	0.007	0.036
	(0.77)	(1.61)
Sub-Saharan Africa	-0.002	-0.006
	(-0.22)	(-0.32)
East Asia & Pacific	-0.013	-0.040**
	(-1.44)	(-2.43)
East Europe & Central Asia	0.007	0.001
	(0.57)	(0.05)
Middle East & North Africa	0.015	0.002
	(0.70)	(0.07)
South Asia	-0.014***	-0.035***
	(-2.71)	(-2.99)
Question 14	0.004	0.030
	(0.35)	(1.16)
Question 15	0.013	0.071***
	(1.01)	(2.61)
Question 16	0.013	0.008
	(1.15)	(0.44)
No. of obs., countries	544, 136	544, 136
Pseudo R^2	.39	.26

Dependent variable is "Request," coded 1 if either the PREM or CFP network reviews requested that additional information be provided in a revised write-up, and 0 otherwise. Omitted region is Latin America and Caribbean. Omitted question is 12. Coefficients represent marginal effects evaluated at mean of other independent variables. T-statistics, reported in parentheses below point estimates, are based on standard errors adjusted for non-independence of errors within regional clusters of observations, with *** p<0.01, ** p<0.05, * p<0.1.

Table 4
Network review: Disagreement (probit regressions)

Equation	4.1	4.2	4.3	4.4
Sample	All	All	All	IDA
Proposed rating (1-6 scale)	0.021	0.060***		0.021
	(1.58)	(2.90)		(1.58)
Grade (coverage of criteria)	-0.022***	-0.031***	-0.032**	-0.009**
	(-3.30)	(-2.62)	(-2.10)	(-2.43)
Word count (log)	0.021	0.057***	0.071***	0.006
	(1.61)	(2.82)	(2.83)	(0.90)
IDA	0.005	-0.017	-0.008	
	(0.22)	(-0.55)	(-0.24)	
Benchmark country	0.026	-0.016	-0.016	-0.002
	(1.63)	(-0.93)	(-0.65)	(-0.60)
Increase proposed	0.454***			0.287***
	(8.26)			(6.53)
Decrease proposed	0.035	-0.003	-0.023	0.057
	(0.96)	(-0.10)	(-0.61)	(1.33)
Freedom House press	0.021	-0.001	0.001	-0.0002
Freedoms	(1.58)	(-1.29)	(0.77)	(-1.62)
Log of population	0.006	0.003	0.008	0.007
	(1.11)	(0.40)	(1.06)	(4.28)
Log of per capita GNI	-0.007	-0.039**	-0.024	0.003
	(-0.57)	(-2.25)	(-1.58)	(0.74)
CMU in country	-0.010	0.020	0.017	-0.007**
	(-0.87)	(0.89)	(0.66)	(-2.01)
Sub-Saharan Africa	-0.023	-0.066***	-0.079***	-0.008
	(-1.38)	(-2.83)	(-3.05)	(-1.00)
East Asia & Pacific	-0.014	-0.041**	-0.059***	-0.008**
	(-1.29)	(-2.15)	(-2.64)	(-2.20)
Latin America & Caribbean	-0.007	-0.020	-0.036*	-0.001
	(-0.61)	(-1.05)	(-1.63)	(-0.08)
Middle East & North Africa	-0.015*	-0.032**	-0.043*	-0.002
	(-1.90)	(-2.23)	(-1.92)	(-0.30)
South Asia	-0.012	-0.036**	-0.050**	-0.005
	(-0.79)	(-2.21)	(-2.29)	(-1.32)
Question 14	-0.022**	-0.039**	-0.012	-0.004
	(-2.06)	(-2.07)	(-0.47)	(-0.62)
Question 15	-0.023***	-0.039**	-0.040*	-0.008
	(-2.93)	(-2.46)	(-1.71)	(-1.45)
Question 16	0.008	0.043**	0.051**	0.019**
	(0.84)	(2.14)	(2.16)	(2.00)
No. of obs., countries	544, 136	544, 136	544, 136	316, 79
Pseudo R^2	.52	.17	.17	.57

Dependent variable is "Disagreement," coded 1 if either the PREM or CFP network reviews recommend a different rating from the one proposed by the region, and 0 otherwise. Omitted region is East Europe and Central Asia. Omitted question is 12. Coefficients represent marginal effects evaluated at mean of other independent variables. T-statistics, reported in parentheses below point estimates, are based on standard errors adjusted for non-independence of errors within regional clusters of observations, with *** p<0.01, ** p<0.05, * p<0.1.

Table 5
Network review: Ratings Adjustment (probit regressions)

Equation	5.1	5.2
Sample	All	All
Disagreement		
Proposed rating (1-6 scale)		0.021**
		(2.23)
Grade (coverage of criteria)		-0.017***
		(-3.16)
Word count (log)		0.005
		(0.58)
IDA		0.002
		(0.10)
Benchmark country		0.033*
		(1.71)
Increase proposed		0.408***
		(8.13)
Decrease proposed		0.040
		(1.23)
Freedom House press		-0.1001
Freedoms		(-0.60)
Log of population		0.004
		(0.96)
Log of per capita GNI		-0.001
		(-0.09)
CMU in country		-0.006
		(-0.62)
Sub-Saharan Africa	-0.050**	-0.010
	(-2.00)	(-0.72)
East Asia & Pacific	-0.070***	-0.011
	(-2.75)	(-0.83)
Latin America & Caribbean	-0.046*	-0.009
	(-1.87)	(-0.79)
Middle East & North Africa	-0.042*	-0.010
	(-1.58)	(-1.06)
South Asia	-0.036	-0.001
	(-1.23)	(-0.04)
Question 14	-0.003	-0.019***
	(-0.11)	(-2.58)
Question 15	-0.012	-0.013*
	(-0.47)	(-1.76)
Question 16	0.054**	0.001
	(2.04)	(0.07)
No. of obs., countries	544, 136	544, 136
Pseudo R^2	.08	.46

Dependent variable is "Rating Adjustment," coded 1 if the final rating differs from the one originally proposed by the region. Omitted region is East Europe and Central Asia. Omitted question is 12. Coefficients represent marginal effects evaluated at mean of other independent variables. T-statistics, reported in parentheses below point estimates, are based on standard errors adjusted for non-independence of errors within regional clusters of observations, with *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Testing for Regional or IDA Bias

Equation	6.1	6.2	6.3	6.4	6.5	6.6	6.7	6.8
Rating	proposed	final	proposed	final	proposed	final	proposed	final
Question	12	12	14	14	15	15	16	16
IDA	-0.089	-0.135	-0.016	-0.010	-0.036	-0.037	0.017	0.021
	(-0.79)	(-1.23)	(-0.12)	(-0.08)	(-0.23)	(-0.24)	(0.15)	(0.20)
ECA	0.275***	0.276***	0.231*	0.201*	0.363***	0.355***	0.192**	0.182**
	(3.00)	(3.20)	(1.93)	(1.86)	(3.15)	(3.20)	(2.01)	(2.04)
Log per capita GNI	0.012	-0.007	0.156**	0.161**	-0.004	-0.010	0.084	0.082
	(0.19)	(-0.11)	(2.32)	(2.41)	(-0.05)	(-0.13)	(1.50)	(1.57)
Index of guidepost	0.896***	0.907***						
indicators (Q12)	(11.10)	(11.62)						
DB number of tax			-0.004**	-0.005***				
Payments			(-2.05)	(-2.42)				
DB tax rate			-0.003**	-0.002*				
			(-2.20)	(-2.15)				
EIU red tape/quality of					0.415***	0.428***		
Bureaucracy					(7.79)	(8.34)		
Index of guidepost							0.975***	1.002***
indicators (Q16)							(11.32)	(12.81)
Constant	3.159	3.321	2.692	2.664	3.118	3.161	2.555	2.567
	(5.54)	(5.81)	(4.53)	(4.55)	(4.74)	(4.87)	(5.19)	(5.65)
No. of observations	132	132	130	130	115	115	132	132
R^2	.69	.71	.23	.25	.47	.48	.69	.74

Dependent variable is CPIA rating on indicated question. T-statistics, reported in parentheses below point estimates, are based on robust standard errors, with *** p<0.01, ** p<0.05, * p<0.1.

Table 7: Other regions relative to ECA (final ratings)

Equation	7.2	7.4	7.6	7.8
Question	12	14	15	16
IDA	-0.149	-0.031	-0.091	0.013
	(-1.31)	(-0.23)	(-0.58)	(0.13)
AFR	-0.344**	-0.169	-0.387***	-0.193
	(-2.48)	(-0.89)	(-2.83)	(-1.56)
EAP	-0.438***	-0.328**	-0.205	-0.233*
	(-3.26)	(-1.98)	(-1.37)	(-1.68)
LCR	-0.211**	-0.120	-0.342**	-0.228**
	(-2.40)	(-0.96)	(-2.55)	(-2.12)
MNA	-0.217	-0.430***	-0.545**	-0.061
	(-1.63)	(-2.85)	(-2.31)	(-0.44)
SAR	-0.191	-0.272	-0.192	0.010
	(-1.13)	(-1.42)	(-1.27)	(0.06)
Log per capita GNI	-0.043	0.147	-0.026	0.084
	(-0.51)	(1.55)	(-0.30)	(1.35)
Index of guidepost indicators (Q12)	0.943***			
	(11.54)			
DB number of tax Payments		-0.006***		
		(-2.70)		
DB tax rate		-0.003**		
		(-2.23)		
EIU red tape/quality of bureaucracy			0.418***	
			(7.55)	
Index of guidepost indicators (Q16)				1.023***
				(12.89)
Constant	3.915	3.032	3.665	2.747
	(5.13)	(3.47)	(4.73)	(4.87)
No. of observations	132	130	115	132
R^2	.73	.28	.50	.74

Dependent variable is CPIA rating on indicated question. T-statistics, reported in parentheses below point estimates, are based on robust standard errors, with *** p<0.01, ** p<0.05, * p<0.1.

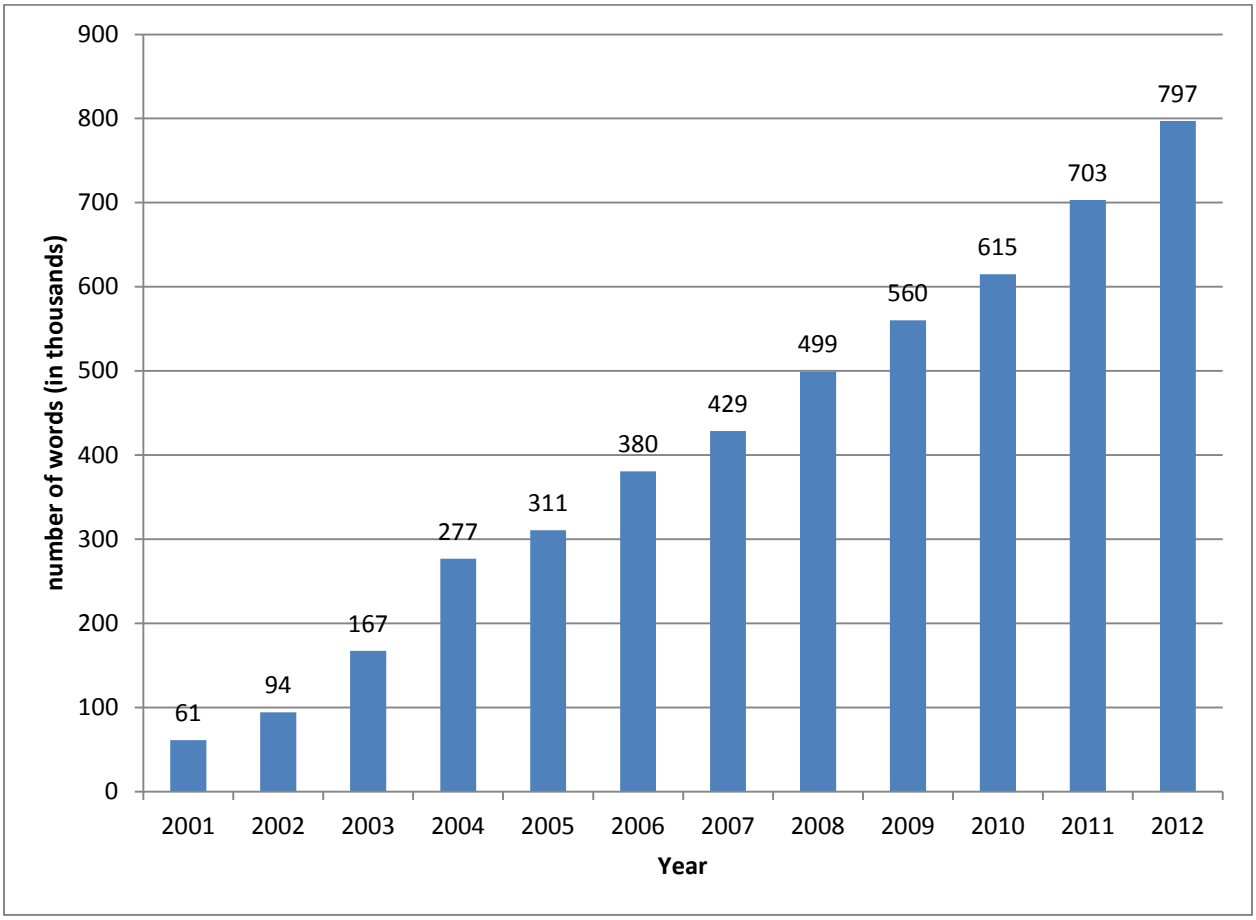

Figure 1
Word Count (in thousands) of AFR Write-ups, 2001-2012

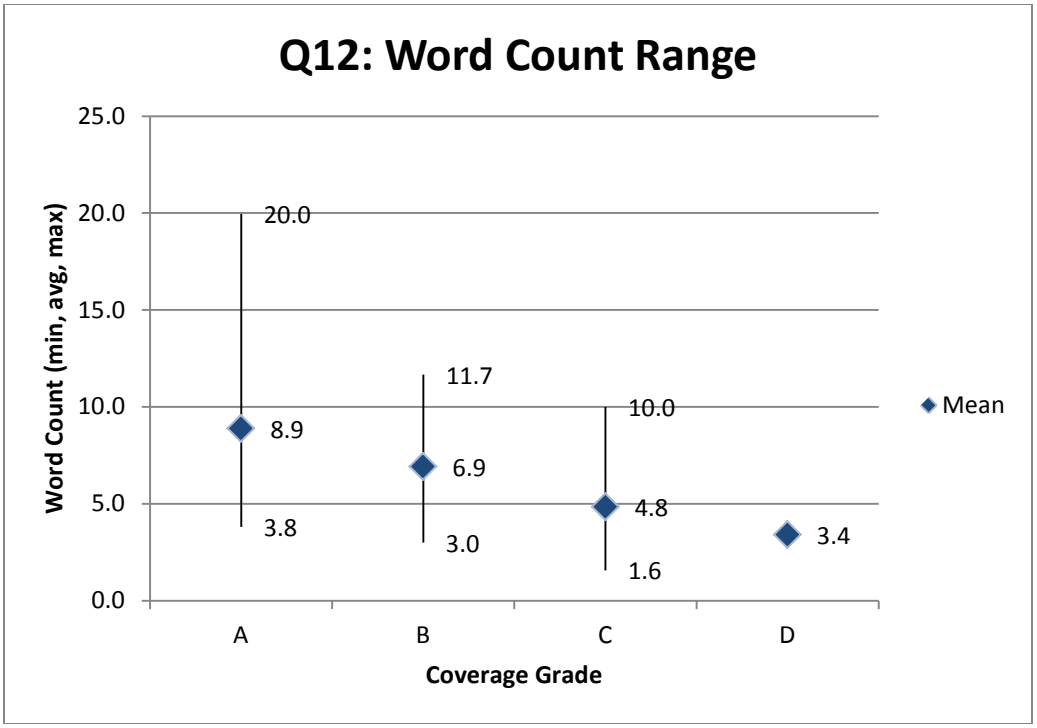

Figure 2
Mean, maximum and minimum word count (in hundreds) for Question 12

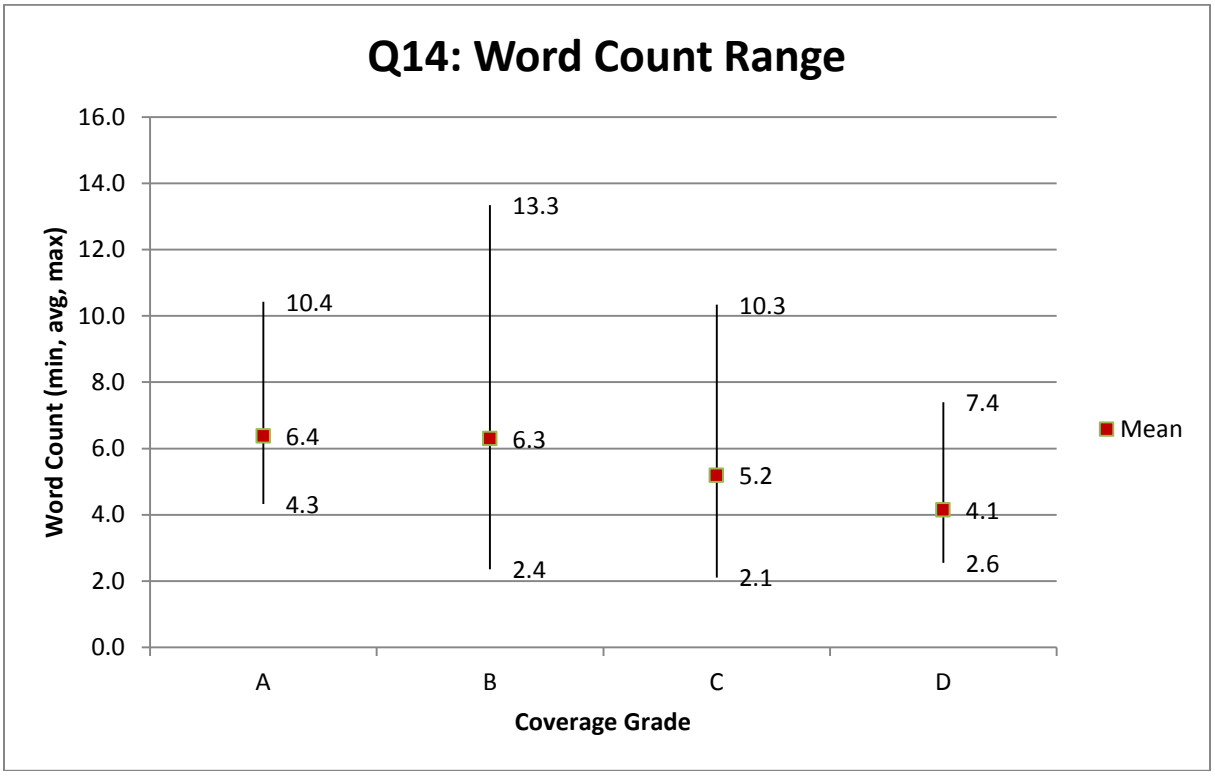

Figure 3
Mean, maximum and minimum word count (in hundreds) for Question 14

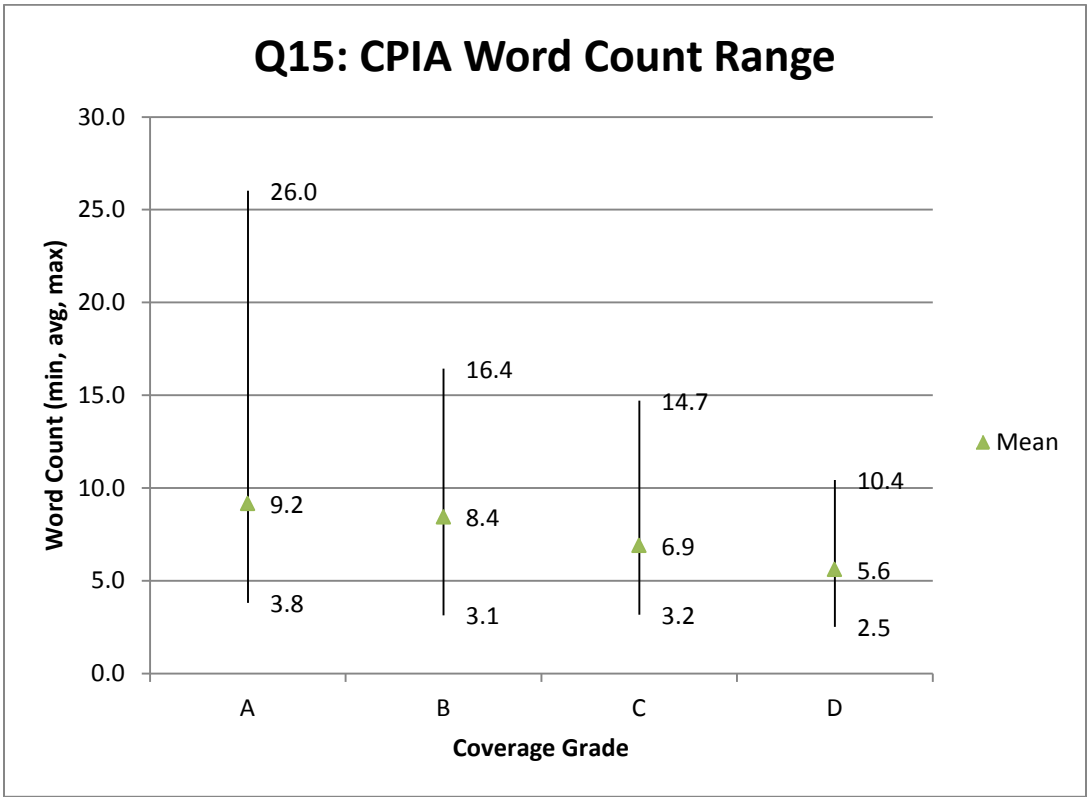

Figure 4
Mean, maximum and minimum word count (in hundreds) for Question 15

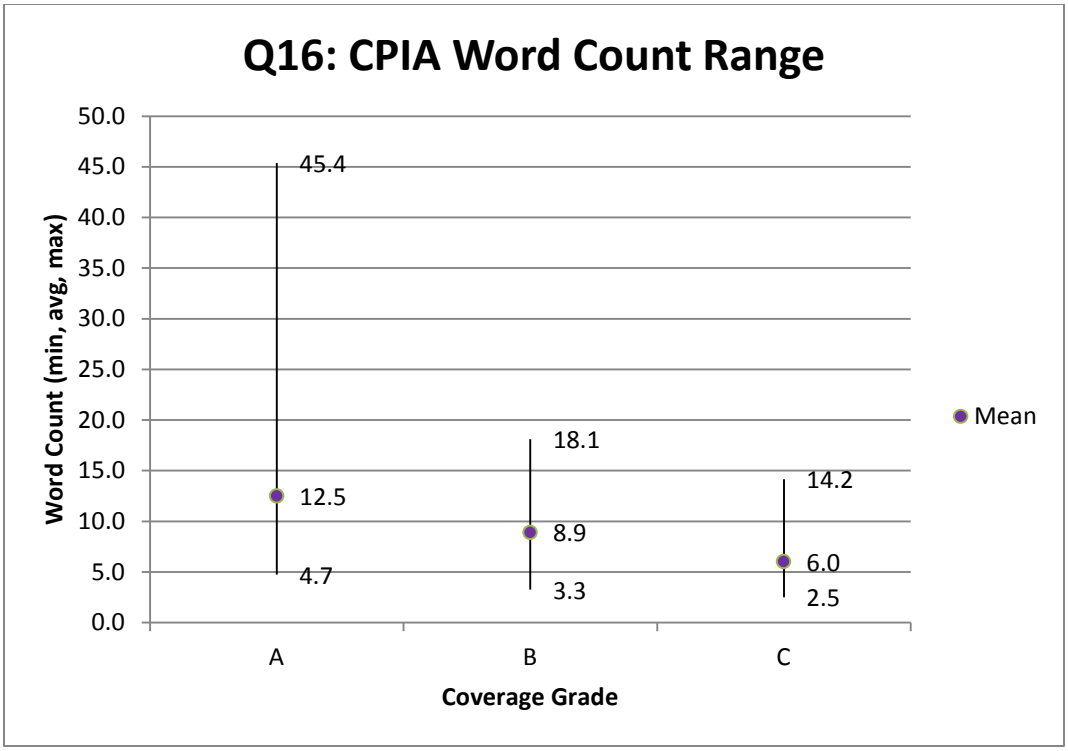

Figure 5
Mean, maximum and minimum word count (in hundreds) for Question 16

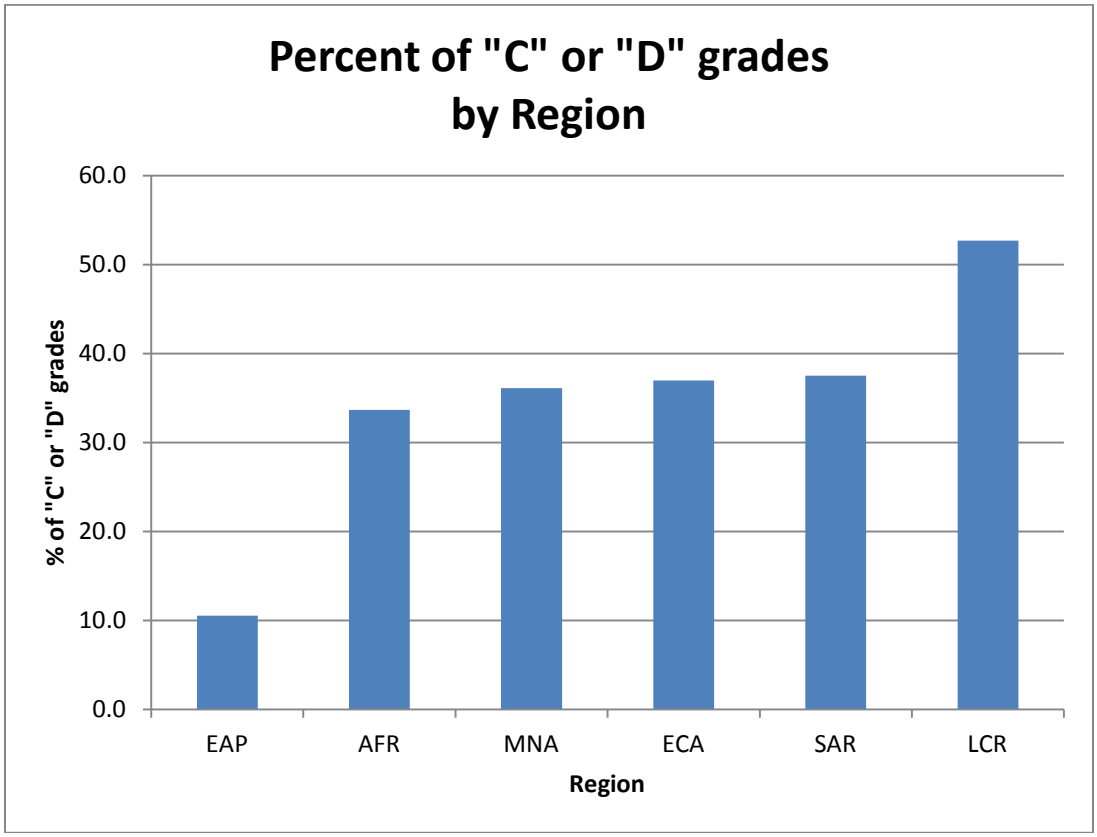

Figure 6
Percentage of write-ups graded “C” or “D”, by region

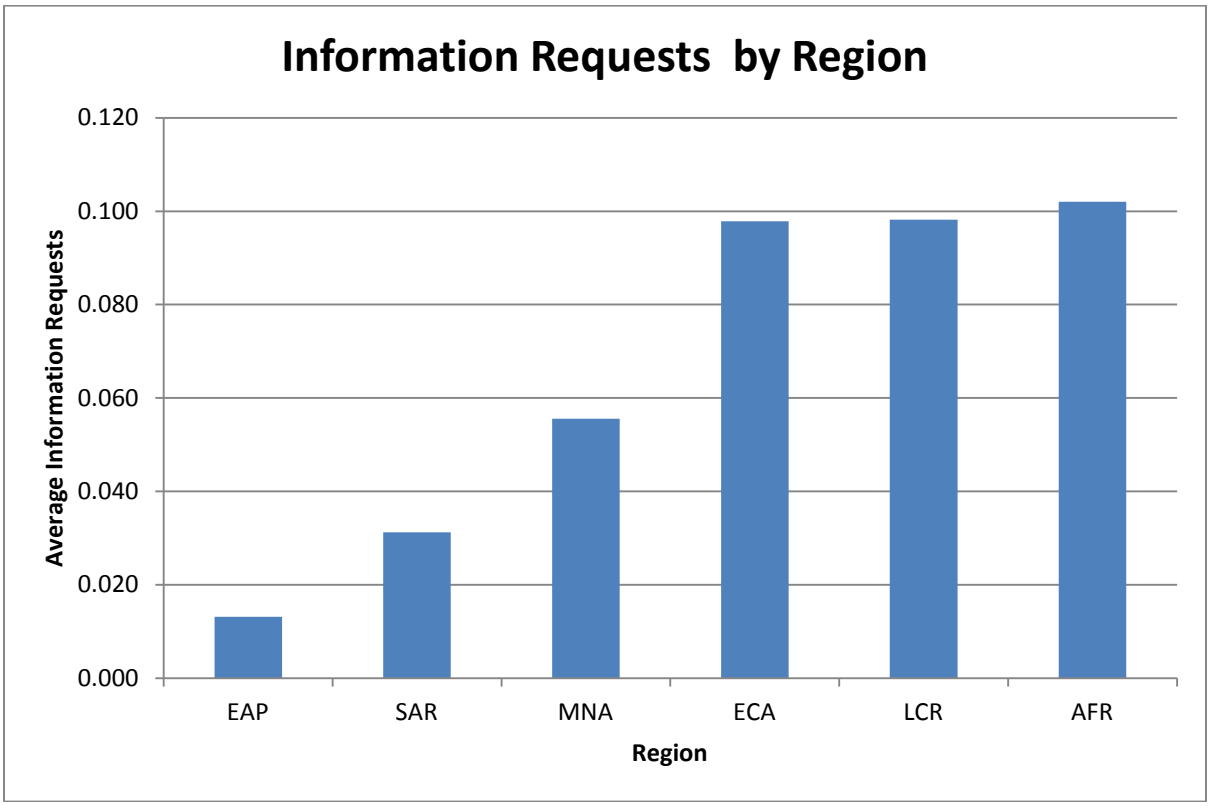

Figure 7
Network reviewer requests for additional information (%), by region

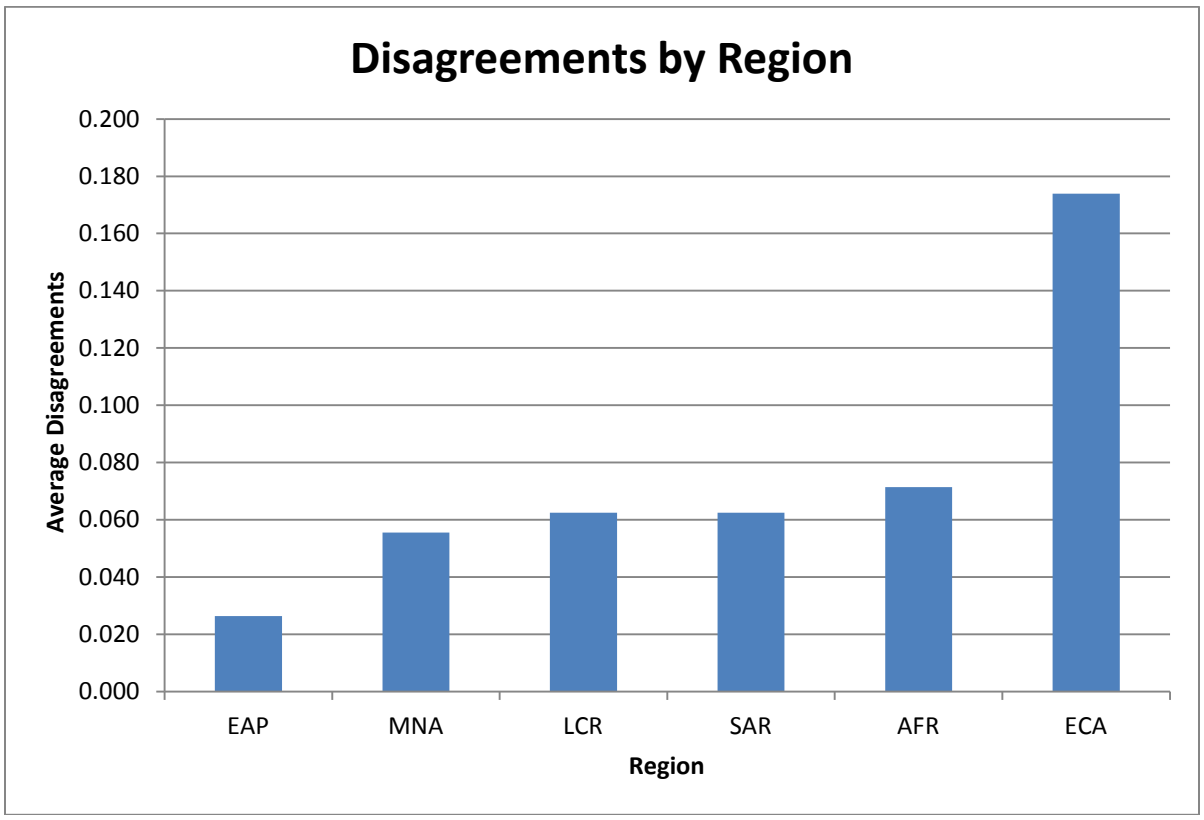

Figure 8
Network reviewer disagreements with proposed ratings (%), by region

Appendix

CPIA Cluster D

Short Definitions

12. Property Rights and Rule-based Governance

This criterion assesses the extent to which economic activity is facilitated by an effective legal system and rule-based governance structure in which property and contract rights are reliably respected and enforced. Each of three dimensions should be rated separately: (a) legal framework for secure property and contract rights, including predictability and impartiality of laws and regulations; (b) quality of the legal and judicial system, as measured by independence, accessibility, legitimacy, efficiency, transparency, and integrity of the courts and other relevant dispute resolution mechanisms; and (c) crime and violence as an impediment to economic activity and citizen security.

14. Efficiency of Revenue Mobilization

This criterion assesses the overall pattern of revenue mobilization, not only the tax structure as it exists on paper, but revenue from all sources as they are actually collected. Separate sub-ratings should be provided for: (a) tax policy; and (b) tax administration. For the overall rating, these two dimensions should receive equal weighting.

15. Quality of Public Administration

This criterion covers the core administration defined as the civilian central government (and sub-national governments, to the extent that their size or policy responsibilities are significant) excluding health and education personnel, and police. The criterion assesses the functioning of the core administration in three areas: (a) managing its own operations; (b) ensuring quality in policy implementation and regulatory management; and (c) coordinating the larger public sector Human Resources Management regime outside the core administration (de-concentrated and arms-length bodies and subsidiary governments).

16. Transparency, Accountability, and Corruption in the Public Sector

This criterion assesses the extent to which the executive, legislators, and other high-level officials can be held accountable for their use of funds, administrative decisions, and results obtained. Accountability is generally enhanced by transparency in decision-making, access to relevant and timely information, public and media scrutiny, and by institutional checks (e.g., inspector general, ombudsman, or independent audit) on the authority of the chief executive. The criterion covers four dimensions: (a) the accountability of the executive and other top officials to effective oversight institutions; (b) access of civil society to timely and reliable information on public affairs and public policies, including fiscal information (on public expenditures, revenues, and large contract awards); (c) state capture by narrow vested interests; and (d) integrity in the management of public resources, including aid and natural resource revenues.

Descriptions for Rating of "3"

Question 12

a. The law protects property rights in theory, but in fact registries and other institutions required to make this protection effective function poorly, making the protection of private property uncertain. Contract enforcement through formal mechanisms is costly and unreliable. Laws and regulations are not changed arbitrarily, but may not be publicly available.
b. Judges and prosecutors are sometimes subject to political interference, and laws are sometimes selectively applied (e.g., against the political opposition). Merit plays some role in judicial appointments. Legal claims against government officials or other elites are commonly prosecuted, but rulings against them are not always enforced. Courts are costly and time-consuming to use, even for small claims. Delays are common. Bribes are known to occur occasionally in the system. Judicial decisions are sometimes publicly available.
c. The state is somewhat effective in limiting violence and crime against citizens and their property. The state actively attempts to combat organized crime, which accounts for a relatively small share of economic activity. A majority of victims report crimes to the police, and citizens generally do not view the police as a source of crime and violence.

Question 14

a. Taxes on trade are a major source of revenue; turnover and other distortionary taxes and levies remain. Consumption-based taxes (e.g., a VAT) are planned, or in limited use. Import tariffs are moderate, but there are too many rates. The income tax base is narrow, and the rate structure is only partly rationalized. Exemptions are moderate.
b. Tax administration is weak, but tax laws are not inordinately complex, and discriminatory enforcement is the exception rather than the rule. Information systems are functioning (e.g., unique taxpayer identification numbers are used). Corruption exists, but there are efforts to improve integrity as well as capacity. Collection and compliance costs are nevertheless somewhat excessive, and collection rates are relatively low.

Question 15

a. The core administration demonstrates modest internal management capacity: major personnel actions, such as recruitment and selection, promotions, and dismissals sometimes reflect merit and performance; terms of employment, and pay are barely sufficiently attractive to ensure that the public administration can compete reasonably effectively for any scarce skill sets it requires; the public sector pay regime is sometimes unable to motivate effort within the public service.
b. The core administration demonstrates modest capacity to ensure quality in policy and regulatory management: Cabinet decisions, presidential or ministerial policy announcements are occasionally dropped or otherwise not implemented; the institutional responsibilities for data collection, analysis and reporting in the sectors are occasionally weak or unclear; and the bodies with responsibility for sector regulation (infrastructure, transport, etc.) are occasionally not regarded as independent in practice and few have adequate regulatory quality management arrangements in place.
c. The core administration demonstrates modest capacity to coordinate the broader public sector HRM regime: (i) merit is the predominant factor in obtaining appointments or promotion in many entities; and (ii) the aggregate public sector wage bill is at some risk of unsustainability.

Question 16

a. Checks and balances on executive power are somewhat effective. External accountability mechanisms may exist, but they have inadequate resources or authority. Regulation of political financing is poorly enforced, usually to the benefit of incumbents. Anticorruption efforts tend to focus on the political opposition. Citizens are sometimes able to bring claims against the state, and legitimate claims are sometimes successful.
b. Decision-making is generally not transparent, and public dissemination of information on government policies and outcomes is a low priority. Some key budget documents are not publicly available. Official restrictions on the media, as well as violence against and harassment of journalists, limit the media's potential for information-gathering and scrutiny.
c. Boundaries between the public and private sectors are moderately well defined, but violations are frequent and often not investigated or sanctioned. Elected and other high-level public officials often have private interests that conflict with their professional duties. Conflict of interest and asset disclosure rules do not apply to high-level officials or are enforced only selectively.
d. Public funds are sometimes diverted to unintended uses by high-level officials, but the prospect of sanctions has some deterrent effect. Bribery and collusion between bidders are common in public contracting, and value for money is often a minor consideration in contract awards.