Forthcoming, Journal of Economic Growth

Can Specific Policy Indicators Identify Reform Priorities?

Aart Kraay (The World Bank)

Norikazu Tawara (Akita International University)

First Draft: March 2010

This Draft: February 2013

Abstract: Several detailed cross-country datasets measuring specific policy indicators relevant to business regulation and government integrity have been developed in recent years. The promise of these indicators is that they can be used to identify specific reforms that policymakers and aid donors can target in their efforts to improve the regulatory and institutional environment. Doing so, however, requires evidence on the partial effects of the many specific policy choices reflected in such datasets. In this paper we use Bayesian Model Averaging (BMA) to document the cross-country partial correlations between detailed policy indicators and several measures of regulatory and institutional outcomes. We find major instability in the set of policy indicators identified by BMA as important partial correlates of similar outcomes: specific policy indicators that matter for one outcome are, on average, not important correlates of other closely-related outcomes. This finding illustrates the difficulties in using highly-specific policy indicators to identify reform priorities using cross-country data.

1818 H St. NW, Washington, DC, akraay@worldbank.org, and Okutsubakidai, Yuwa-Tsubakigawa, Akita, 010-1292 Japan, ntawara@aiu.ac.jp, respectively. We would like to thank Nathaniel Heller, Daniel Kaufmann, Eduardo Ley, Chris Papageorgiou, Luis Serven, Stefan Zeugner, and several anonymous referees for helpful feedback, and especially Martin Feldkircher and Stefan Zeugner for providing their R-code for implementing Bayesian Model Averaging. Financial support from the Japan Consultant Trust Fund and the Knowledge for Change Program of the World Bank is gratefully acknowledged. The views expressed here are the authors' and do not reflect those of the World Bank, its Executive Directors, or the countries they represent.

1. Introduction

Strong institutions, including a sound regulatory environment for private sector economic activity, are widely considered to be crucial to successful economic development. This consensus has been informed by a large empirical literature linking various measures of regulatory and institutional quality to differences in economic growth performance across countries. Out of necessity, much of this literature has relied on fairly broad summary measures of institutional and regulatory outcomes. For example, in one of the most widely-known papers in the institutions and growth literature, Acemoglu, Johnson and Robinson (2001) proxy for institutional quality using the risk of expropriation, as perceived by analysts at a commercial risk rating agency. In another pioneering paper, Mauro (1995) relates cross-country differences in growth and investment to perceptions of overall corruption from commercial risk rating agencies.

Turning such influential findings into concrete policy advice for countries seeking to improve the regulatory and institutional environment has, however, been more difficult. One reason for this has been the shortage of systematic cross-country data on specific policies that governments might implement in order to influence the institutional outcomes that have been identified as important correlates of growth and development. Recognizing this gap, a number of organizations have in recent years developed highly detailed indicators of specific laws, regulations, and policies that plausibly influence broad institutional outcomes such as perceptions of the business regulatory environment or public sector integrity. In this paper, we consider two such datasets, both covering large cross-sections of countries: (i) the Doing Business project of the World Bank, which reports 38 indicators of specific rules and regulations relevant to the business environment, and (ii) the Global Integrity Index, which reports over 300 specific policy indicators relevant to public sector accountability mechanisms.

The promise of such detailed indicators is to provide guidance on reform priorities, by pinpointing specific policy levers under the control of policymakers that can be changed in order to improve institutional quality, ultimately leading to better growth performance.[1] However, realizing this promise requires an understanding of the relative magnitude of the partial effects of each of the specific

[1] For example, the World Bank has supported the development of "actionable" indicators under the control of policymakers, which are intended to provide *"...convenient and replicable guidance on the features (rules of the game, organizational capabilities) for which reform interventions are likely to prove most helpful for improving the performance of particular governance elements"* (see www.agidata.org). See also Trapnell (2011) for a discussion of the promise of such "actionable" policy indicators.

policy indicators on the corresponding outcomes that policymakers might want to improve. For example, a policymaker considering business regulatory reform would want to know whether investor perceptions of the quality of the business environment respond more to streamlined procedures for specific regulatory processes, or to reduced fees for these same processes. The same policymaker would likely also want to know whether the answer depends on which particular measure of investor perceptions of the business environment is used.

On this crucial question of partial effects of specific policy indicators on outcomes of interest, empirical evidence has not kept pace with the proliferation of detailed indicators relevant to the regulatory environment and institutional quality. Confronted with such large numbers of potential policy indicators, one common empirical approach has been to simply average them together into a composite measure that can be related to outcomes of interest. This, however, embodies the unappealing assumption that the partial effects on outcomes of all of the components of the composite policy indicator are equal.[2] Another approach is to pick a subset of specific policy indicators to relate to outcomes, which may be misleading if the chosen policy indicators are correlated with other policy measures that also matter for the outcome of interest.

In this paper, we illustrate the challenges of linking specific policy indicators to regulatory and institutional outcomes, without imposing unappealing prior restrictions about the relative importance of individual indicators. We do this in three steps. First, we identify a set of intermediate outcome variables which we think a policymaker might reasonably want to influence through reforms to specific policies captured in our two datasets. In the case of Doing Business, which focuses on business regulation, we choose as outcome variables seven closely-related subjective measures of the quality of the regulatory environment, as perceived by a variety of firm survey respondents and expert assessments. In the case of Global Integrity, which focuses on public sector accountability mechanisms, we choose seven subjective measures of perceptions of corruption. A key feature of these outcome measures is that they tend to be quite highly correlated across countries. For example, the median pairwise correlation among the seven measures of perceptions of the regulatory environment is 0.68.

[2] More sophisticated approaches to aggregation, such as principal components, will have similar difficulties, since they implicitly impose the hard-to-justify assumption that the effects of the different policy indicators on outcomes are proportional to the intercorrelations of the policy indicators themselves. See Lubotsky and Wittenberg (2006) for a more extensive discussion of this problem in a different context.

Second, we use Bayesian Model Averaging (BMA) to systematically document the partial correlations between the many detailed policy indicators and the seven outcome variables of interest. As discussed in more detail below, BMA is a powerful tool for systematically identifying robust partial correlates of outcomes when there are many potential explanatory variables and the precise empirical specification is unknown. Third, having identified a set of important partial correlates for each outcome variable, we investigate how similar these sets are across different outcome variables.

For any given outcome variable, we find that the BMA procedure readily identifies a small number of specific policy indicators that are strongly partially correlated with the outcome of interest. However, this good news is tempered by an important negative finding: there is a great deal of instability across similar outcomes in the set of policy indicators which the BMA procedure identifies as important for each outcome. To take a specific example, two of our outcome variables for the Doing Business indicators are assessments of the quality of the business environment produced by the Economist Intelligence Unit (EIU) and the World Bank's Country Policy and Institutional Assessments (CPIA). Both measures are conceptually similar, summarizing respondents' views of the quality of business regulation and the restrictiveness of international trade. In the set of 110 countries for which data on both outcomes are available, these two measures are correlated at 0.78. Yet despite the similarity of these two outcome variables, we find little correspondence in the specific policy indicators that are identified as their important partial correlates. For example, the BMA procedure identifies legal protections for creditors as the most important partial correlate of the CPIA outcome variable. However, it ranks only 19th most important for the EIU outcome variable.

More systematically, for each outcome variable, we identify a set of 10 policy indicators that are the most important partial correlates of the outcome. Comparing these sets of important policy indicators across outcomes, we find little overlap: for example, despite the high correlation between the EIU and CPIA outcome variables noted above, only three indicators fall into the set of top ten important correlates for both of these outcome variables. Moving beyond these two specific outcome measures, we find that not one of the 38 specific policy indicators in the Doing Business dataset turns up as an important correlate of all seven outcome variables, and just one turns out to be important for six out of seven outcome variables. Conversely, we find that 29 out of the 38 policy indicators turn up in the set of important partial correlates for at least one outcome. This suggests a great deal of instability across outcomes in the set of important policy indicators. We find a similar degree of instability in the

set of important correlates across alternative measures of corruption when we perform the same exercise using the Global Integrity dataset.

This key instability finding can be interpreted in two alternative ways. One possibility is that there is a common model linking the same small set of policy indicators to each of the outcomes, but the available data and empirical techniques that we deploy are not able to conclusively identify the set of policy indicators included in this common model. Another possibility is that the true mapping from policy indicators to outcomes in reality is different across the different outcome variables, and this is reflected in our finding of instability in the set of important indicators across outcomes. We discuss these two alternative interpretations in more detail later in Section 5.6 of the paper. Importantly, however, we note that both interpretations share the same implication: policymakers will find it difficult to use these kinds of datasets of highly-detailed policy indicators to isolate a small number of specific reforms that are likely to matter for the set of outcomes they are interested in influencing.

Our results should not be interpreted as a criticism of the specific policy indicator datasets we use in the paper -- in our view both Doing Business and Global Integrity are credible and careful data gathering exercises. Nor do we view our findings as a critique of Bayesian Model Averaging as a technique for isolating robust partial correlations when there is little prior information about the true empirical specification -- as we discuss in more detail below, our findings are not driven by key assumptions required to implement the BMA methodology. Rather we view our empirical results as a cautionary tale for creators and users of cross-country datasets of highly specific policy indicators. While such indicators can serve a useful role in documenting differences in regulatory and institutional practices across countries, it may be difficult to use them as a roadmap to reforms in the real world, where policymakers must choose to spend their political capital on a limited set of high-impact reforms.

Our main finding of instability of explanatory variables across similar outcomes is most closely related to Ciccone and Jarocinski (2010), who document a high degree of instability in the set of important growth determinants, when the dependent variable of economic growth is calculated using alternative revisions and updates of the widely-used Penn World Table dataset. We share with that paper an emphasis on the instability of BMA results across different closely-related outcome variables. Our work differs however in that they focus on a set of quite broad growth determinants, including many historical and geographical features of countries as well as initial conditions that, while relevant for growth, are not amenable to specific policy interventions. In contrast, our emphasis is on

understanding the links between specific policy levers and intermediate outcomes that policymakers might want to influence using these levers.[3]

The rest of this paper proceeds as follows. In Section 2, we describe the Doing Business data and the corresponding set of outcome variables. In Section 3, we explain the Bayesian Model Averaging methodology, and Section 4 contains our main results using the Doing Business dataset. In Section 5, we explore a range of potential explanations for our instability finding, and Section 6 offers concluding remarks. To conserve space, our findings based on the Global Integrity Index are confined to an online Appendix A, and the detailed results of a number of robustness checks discussed in Section 5 are available in an online Appendix B.

2. A First Look at the Data

We work with data from the 2009 edition of Doing Business (World Bank (2009)), which reports on 38 specific policy indicators covering ten dimensions of the business regulatory environment (Starting a Business, Dealing with Construction Permits, Employing Workers, Registering Property, Getting Credit, Protecting Investors, Paying Taxes, Trading Across Borders, Enforcing Contracts, and Closing a Business). For example, "Starting a Business" is based on four indicators measuring the number of procedures, the number of days, the cost of associated fees, and the minimum capital requirement, that are required to start a new business. The 38 indicators are also aggregated into an overall "Ease of Doing Business" indicator which averages together each country's rank on the individual indicators. The Doing Business data is scenario-based. Doing Business respondents, typically business law practitioners in the country in question, are given a detailed scenario about a hypothetical transaction, for example, registering a firm with particular characteristics in the capital city of the country. The data collected by Doing Business then correspond to the specific regulatory procedures that the hypothetical firm described in the scenario would have to comply with.[4]

[3] Our work is also related to Pritchett (1996), who observes that various indicators of trade policies tend to be uncorrelated with each other, and moreover are not highly correlated with trade outcomes.

[4] While for terminological convenience we refer to the 38 Doing Business variables as "policy indicators", many of them are themselves amalgams of even more specific rules and regulations. For example, the number of days, number of procedures, and costs to start a business reflect the particular combination of steps required for this formality in each country, which may include items as diverse as complying with requirements to (i) obtain a company seal, (ii) register with various government agencies, (iii) document the uniqueness of the company name, (iv) obtain the criminal record of the company manager, and many more. Specific policy reforms would then

Our next step is to identify outcomes that policymakers might reasonably want to influence through reforms that would be captured by changes in the individual policy indicators. In principle, reforms could reflect policymakers' ambition to influence a potentially large set of ultimate outcomes, ranging from their own political interests to broad considerations of social welfare. In our empirical exercise we take a more limited view -- that policymakers are likely to care about the extent to which regulatory reforms influence perceptions of the quality of the regulatory environment. For example, policymakers presumably care whether the reforms they implement will lead to their country or city being perceived by the business community as being a better or worse place to do business. Of course, we do not claim that this is the *only* outcome policymakers might want to influence by changing the specific policy indicators that are captured by Doing Business. Rather, we think these perceptions of the quality of the regulatory environment might plausibly be among the many considered important by policymakers, and so can provide a good illustration of the challenges of identifying the partial effects of the many specific policy indicators on outcomes of interest. Moreover, these variables are also of interest given that many papers in the empirical growth literature have related growth and other development outcomes to such perceptions-based measures of the regulatory and institutional environment.

Of course, there is no single unique measure of investor perceptions of the business regulatory environment, and policymakers might well want to inform their decisions by considering a variety of such measures of perceptions of regulatory quality. In the case of Doing Business, we use seven such closely-related measures as outcome variables. Five of these are expert assessments of business environment quality taken from commercial business information providers (Economist Intelligence Unit (EIU), Political Risk Services (PRS), Global Insight Global Risk Service (DRI), Global Insight Business Risk Conditions (WMO), and Cerebus Corporate Intelligence Gray Area Dynamics (GAD)). One additional expert assessment is the World Bank's Country Policy and Institutional Assessment (CPIA). Finally, we draw on responses from a large cross-country survey, the Global Competitiveness Report (GCS) survey of firms in 130 countries, that asks firm managers a variety of questions about the quality of the business environment.

---

involve changing or eliminating individual steps such as these, which would in turn have implications for time, cost and number of procedures.

Conceptually, these data sources are closely related in the sense that they reflect the perceptions of members of the business community regarding aspects of the overall business regulatory environment. This is most clearly the case for the GCS survey of firms, where the respondents are a random sample of managers of firms in the country. However, this is also true for the five other data sources from commercial business information providers, who market their assessments primarily to the local and international business communities. The one exception to this are the CPIA assessments of the World Bank, where the respondents are World Bank country economists rather than members of the business community. Nevertheless, a look at the published scoring criteria for the CPIA questions also reveals an emphasis on a regulatory environment that facilitates trade and private sector business activity, which presumably would also be valued by private sector respondents.

Table 1 lists the specific dimensions of the quality of the business environment that are assessed by each of these data sources, while the full list of countries included in the Doing Business dataset and also covered by each source is available in online Appendix B.[5] While there are of course some differences across these outcome variables in terms of the specific questions being addressed, all of them share a common emphasis on business regulation, with several also explicitly including international trade regulation. This common emphasis on views of the business regulatory environment contributes to the quite high observed correlation across these data sources. The median pairwise correlation between the seven outcome indicators is 0.68, and the first principal component of the seven outcome variables accounts for 73 percent of the total variation in these variables.

Our interpretation of these data sources is that each one is a reasonable proxy for perceptions of the quality of the business regulatory environment. In light of this, policymakers interested in influencing perceptions of the business environment would like to know how each of these measures respond to the various specific aspects of the business regulatory environment captured by the Doing Business indicators. Before delving into the relationship between specific policy indicators and the seven outcome variables, it is useful to first document that all seven outcome variables are in fact significantly correlated with the overall Ease of Doing Business indicator, which averages together countries' ranks on all of the specific policy indicators.

[5] The specific measures from each of these seven data sources are constructed in the same way as they are used in the Worldwide Governance Indicators project (see www.govindicators.org, and Kaufmann, Kraay and Mastruzzi (2009) for more detailed descriptions).

We show this in Table 2, which summarizes the results of regressing each of the outcome variables on the aggregate Ease of Doing Business measure. The bivariate regressions deliver t-statistics ranging from 7 to 16. Conditioning on GDP per capita weakens the correlations of Ease of Doing Business with the outcomes, but they remain strongly significant. Of course, we cannot interpret these correlations in Table 2 as purely reflecting a causal effect from the Doing Business indicators to the outcomes of interest – there are many potentially-confounding omitted variables. However, it seems reasonable to think that they at least in part reflect an effect running from the specific regulations and institutions measured by Doing Business to the relevant outcomes. To the extent that this is the case, our goal in this paper is to try to unbundle these aggregate correlations into differential impacts of the many detailed subcomponents of the overall Ease of Doing Business indicator on perceptions of the business regulatory environment.

Finally, it is worth noting that each of these intermediate outcome measures based on perceptions of the business regulatory environment is also strongly correlated with the level of development. This can be seen in the last row in Table 2, which reports correlation coefficients ranging between 0.47 and 0.84. While for the same reasons as given in the previous paragraph we cannot interpret this correlation as reflecting a causal impact of the business environment on the overall level of development, it nevertheless is reassuring that these intermediate outcomes are related to broader development outcomes that policymakers might ultimately want to influence.

3. Bayesian Model Averaging

We now briefly describe the Bayesian Model Averaging (BMA) procedure used in the remainder of the paper to document the partial correlations between the many specific policy indicators in the Doing Business dataset and the seven corresponding outcome variables capturing perceptions of the business environment.[6] The basic idea of BMA is simple. Rather than base inferences about parameters

[6] Over the past several years, BMA has become a widely-used tool for assessing the robustness of regression results to variations in the set of included control variables. The seminal application to cross-country growth empirics is Fernandez, Ley and Steel (2001a), followed by Sala-i-Martin, Doppelhofer and Miller (2004), and then many others. Brock, Durlauf and West (2003) particularly emphasize the decision-theoretic aspects of BMA as a useful tool for guiding policy choices. Recently Ciccone and Jarocinski (2010) have used BMA to document the non-robustness of growth empirics to minor data revisions in the dependent variable. There is also an active literature extending the BMA methodology in various dimensions, including groups of regressors as proxies for various growth theories (Brock and Durlauf (2001) and Durlauf, Kourtellos, and Tan (2008)), panel data applications (Moral-Benito (2012)), and instrumental variables estimation (Eicher, Lenkoski, and Raftery (forthcoming)). Finally, several papers including Fernandez, Ley and Steel (2001b), Ley and Steel (2009, 2012), Eicher, Papageorgiou and

of interest on just one preferred model consisting of one particular set of explanatory variables, BMA combines inferences about parameters of interest across many candidate models corresponding to different sets of explanatory variables. To be more precise, let y denote an $N \times 1$ vector of observations on the dependent variable of interest, and let X denote an $N \times K$ matrix of potential explanatory variables for y. In our case, y is one of the seven outcome variables, and X is the set of 38 policy indicators in the Doing Business dataset.

Let $j \in \{1,2, \dots, 2^K\}$ index models, distinguished by their included set of regressors. In particular let X_j denote an $N \times K_j$ matrix containing a subset of $K_j \leq K$ regressors from X. A model j consists of a linear regression of y on the variables in X_j, i.e.:

$$y = \iota\alpha_j + X_j\beta_j + \varepsilon_j \tag{1}$$

where ι is an $N \times 1$ vector of ones and ε_j is an $N \times 1$ vector of i.i.d. normal disturbances with zero mean and variance σ^2. The scalars σ and α_j, and the $K_j \times 1$ vector β_j, are the parameters of model j, and following the bulk of the literature on BMA we use Zellner's g-prior for them, i.e.

$$f(\alpha_j, \beta_j, \sigma) \propto \sigma^{-1}\phi\left(\beta_j; 0, \frac{\sigma^2}{g}\left(X_j'X_j\right)^{-1}\right) \tag{2}$$

where $\phi(x; a, b)$ denotes a normal density function for x with mean a and variance b, and $f(.)$ denotes a joint density function. The prior distribution for the slope coefficients in the vector β_j, conditional on σ, α_j, and model j, is multivariate normal and centered on zero, with a variance equal to that of the OLS estimator, but scaled by g. As the prior parameter g becomes small, the prior variance expands, so that the prior for the slopes becomes more diffuse.[7]

Raftery (2011) and Feldkircher and Zeugner (2009, 2012) all discuss the consequences of alternative prior assumptions for the outcome of BMA.

[7] We implement BMA separately for each of the seven different outcome variables. An alternative approach would be to define the model space in terms of a sequence of seemingly-unrelated regression (SUR) models, in which the seven equations corresponding to the seven outcomes are estimated simultaneously. In the presence of cross-equation correlations in the residuals, SUR estimation which exploits this information would be more efficient, but at the same time this approach would be more prone to misspecification problems. As a rough check of the importance of the efficiency gains from SUR estimation, we chose the "top" model for each outcome variable, and re-estimated this set of seven equations using SUR. In this case we found little evidence of efficiency gains, in the sense that t-statistics were higher with the SUR estimates for only about half of the parameter

The key ingredient in BMA is the assignment of probabilities to different models. Let $p[M_j|y,X]$ denote the posterior probability of model j. These are computed using Bayes' Rule, i.e.

(3) $$p[M_j|y,X] \propto \mathcal{L}[M_j|y,X]p[M_j]$$

where $\mathcal{L}[M_j|y,X]$ is the marginal likelihood of model j, and $p[M_j]$ is the prior probability assigned by the researcher to model j. Fernandez, Ley and Steel (2001a) show that, given the g-prior and the assumption of homoskedastic normal disturbances, the marginal likelihood is given by:

(4) $$\mathcal{L}[M_j|y,X] \propto \left(\frac{g}{1+g}\right)^{\frac{K_j}{2}} \left(1-\frac{R_j^2}{1+g}\right)^{-\frac{N-1}{2}}$$

where R_j^2 is the R-squared associated with model j. This expression tells us that models with better fit, as measured by a higher R-squared, have a higher marginal likelihood. However the marginal likelihood trades off improvements in fit against increases in model size, with the model size penalty captured by the first term. The prior parameter g plays two roles here: the smaller is g, the greater is the model size penalty, but at the same time the more responsive is the likelihood to improvements in R-squared.

We will use a standard prior for model j that reflects the assumption that there is a fixed probability θ that any one of the variables in X is included in model M_j. Assuming independence of inclusion across the variables in X, this prior implies a mean prior model size of $\mu \equiv \theta K$, and a prior probability for model j given by:

(5) $$P[M_j] \propto \left(\frac{\mu}{K-\mu}\right)^{K_j}$$

As long as prior model size $\mu < K/2$ then the prior favours more parsimonious models with fewer regressors.

Putting these ingredients together delivers the following expression for the posterior probability of model j:

---

estimates, and lower for the other half. It is however unclear a priori how a full BMA-SUR approach would differ from the results presented here.

(6) $$p[M_j|y,X] \propto \left(\frac{\mu}{K-\mu}\right)^{K_j} \left(\frac{g}{1+g}\right)^{\frac{K_j}{2}} \left(1-\frac{R_j^2}{1+g}\right)^{-\frac{N-1}{2}}$$

This expression summarizes how BMA assigns probabilities to models with different sets of regressors, with higher probabilities assigned to models with better fit, subject to a model size penalty. These posterior model probabilities can then be used to average inferences across different models. A key quantity we will use is the Posterior Inclusion Probability (PIP) of a particular explanatory variable k. This is defined as the sum of the posterior probabilities of all models including variable k, and is a useful summary of how "important" a variable is in the sense of being included in models that are more likely. Similarly, a useful summary of the magnitude of the effect of a particular regressor on the dependent variable is its posterior probability-weighted average slope across all models.

Implementing BMA requires choosing the two prior parameters, μ and g. Our choice of these parameters is driven primarily by the logic of the thought experiment we are performing. We have in mind a policymaker interested in influencing one of the outcome variables, and would like to identify a small subset of specific policy indicators that are robustly correlated with outcomes, on which to focus reform efforts. While the threshold determining "small" is of course unclear, we think that a reasonable prior is to set $\mu = 10$. This will lead to posterior mean model sizes in the range of typically 6 to 9 right-hand-side variables, which seems to us a plausibly small set that a policymaker might consider. Turning to g, our objective here is simply to ensure that the inferences from any given model mimic closely traditional frequentist ones, and accordingly we set g to be small, i.e. $g = 0.01$, so that the shrinkage factor is close to one. [8]

We implement the BMA procedure seven times, corresponding to the seven different choices of outcome variable y, using standard computational tools from this literature.[9] Before turning to the

[8] When g is small, Bayesian inference for the parameters of the model mimics frequentist ones. In particular, the posterior distribution of the slope coefficients for a given model is a multivariate-t distribution with mean and variance equal to that of the conventional OLS estimator, but both scaled by a "shrinkage factor" of $\frac{1}{1+g}$ that approaches 1 as the prior becomes more and more diffuse. In contrast, larger values of g reflect a stronger prior belief that the slope coefficients are in fact zero, and so the posterior mean shrinks towards zero and the posterior variance is smaller. In Section 5.3 we will explore the robustness of our main findings to alternative prior specifications.

[9] Implementing BMA in principle poses major computational problems, as the number of models to be estimated and averaged increases rapidly with the number of explanatory variables -- with K regressors there are 2^K possible models to consider. Fortunately, fast and accurate algorithms for identifying and sampling only those models with

results of this exercise in the next section, we acknowledge the important caveat that we are using BMA to combine inferences from a series of simple linear OLS regressions. As such, all of our conclusions are subject to the usual limitations of such models. In particular, a maintained assumption is that the error term is independent of the regressors in all models, an assumption that would clearly be violated if there were reverse causation or omitted variables. We also assume away any plausible nonlinearities such as interactive effects between variables. As we discuss further below, however, addressing these likely important issues we think would only further reinforce our basic point – that it is extremely difficult to identify a small subset of indicators that are robust determinants of closely-related alternative outcomes of interest.

4. Results

Our main findings are reported in Table 3. The rows of this table correspond to the 38 specific policy indicators captured in the Doing Business dataset, while the sets of columns correspond to the seven outcome variables we are considering. For each combination of outcome variable and policy indicator, we first report the PIP, which is simply the sum of the posterior probabilities across all models in which the variable appears. A policy indicator will have a high PIP if the set of models in which it appears jointly has a high posterior probability. Consider, for example, the DRI outcome variable in the first panel of Table 3. The "Legal Rights" component of "Getting Credit" has a high PIP of 0.92, indicating that the joint posterior probability of the set of models in which this variable appears is 92 percent. The same is true for the "Number of Days" component of "Enforcing Contracts", which has a PIP of 0.89. A few other variables also have fairly high PIPs, including "Firing Costs" under "Employing Workers (at 0.83) and the "Number of Days to Import" variable under "Trading Across Borders" (at 0.71).

---

the largest posterior probabilities have been developed, greatly reducing the computational burden, and we rely on them here. Following the BMA literature, the posterior distribution is approximated by simulating a sample from it by applying an MC3 sampler (Madigan and York (1995) and Raftery, Madigan and Hoeting (1997), as described in Fernandez, Ley and Steel (2001a)). We also follow Fernandez, Ley and Steel (2001a) in using the correlation between analytical and empirical posterior model probabilities as a criterion for convergence of the sampling chain. We will report results in the next section from a simulation run with a burn-in of 250,000 discarded drawings and 250,000 recorded drawings. We choose this number so that a high positive correlation between posterior model probabilities based on empirical frequencies and the exact analytical likelihoods is obtained. We also report estimated total posterior model probabilities visited by the chain using a measure of George and McCulloch (1997). We are grateful to Martin Feldkircher and Stefan Zeugner whose R-code (available at http://feldkircher.gzpace.net/links/bma) we used to implement BMA in this paper.

We also report some summary statistics on the distribution of posterior probabilities across models at the bottom of Table 3. We first report the posterior probability of the top three models (ranked by posterior probabilities), and then also the number of models required to cover 50 percent, 75 percent, and 90 percent of the posterior model probabilities. For example, in the case of the DRI outcome variable, the top three models have posterior probabilities of 1.9 percent, 1.5 percent, and 1.3 percent respectively. These low probabilities for even the most likely individual models highlight the importance of considering multiple models. Looking across the columns of Table 3, we see that there are some differences across outcome variables in terms of the concentration of posterior probability across models. For example, for the WMO variable we see that only 214 models are required to cover 50 percent of the posterior probability. In contrast, more than four times as many (912 models) are required to cover half of the posterior probability for the EIU outcome variable, indicating that the posterior probabilities are much more dispersed across models in this case.

These differences across outcome variables in the concentration of posterior probabilities complicate somewhat the interpretation of magnitudes of the PIPs. For example, in the case of EIU, where the posterior probabilities are much more dispersed across models, we also find that the largest PIPs are not very large (the maximum PIP is 0.77 for EIU, while it is 0.98 for WMO). Moreover, as we discuss further below, the concentration of posterior probability mass across models is sensitive to our choice of prior parameter g. In order to compare results across outcome variables, we instead simply emphasize the *ranking* of models, and thus also variables, by their posterior probabilities. In particular, in Table 3 we have highlighted the top 10 out of 38 policy indicators, as ranked by their PIPs, for each outcome variable. This allows us to identify at a glance the most important determinants of each outcome without reference to the precise magnitudes of the associated PIPs, which in some cases are quite small. We also think that this exercise of picking the top few policy indicators as ranked by PIP is analogous to the kind of exercise that a policymaker interested in allocating scarce political capital across a few high-impact reforms might do. In what follows we refer to these policy indicators with the highest PIPs as the most "important" for a given outcome variable.[10]

In the second and third column for each outcome variable, we report the posterior mean and standard deviation of the slope coefficient corresponding to each policy indicator. Returning to the DRI

[10] In Section 5.2 we will show that our main conclusions are robust to a variety of alternative criteria for defining the set of "important" policy indicators.

outcome variable as a specific example, the policy indicator with the highest PIP is the "Legal Rights" component of "Getting Credit", and the corresponding posterior mean for the slope coefficient is 0.10. To interpret the magnitude of these coefficients, note all the policy indicators and the outcome variables have been rescaled to run from 0 to 1. So a change in the value of this policy indicator from its worst possible value of 0 to its best possible value of 1 would lead to an increase in the DRI outcome variable of 0.10, or about one-tenth of its potential range.[11] We note also that the ranking of variables by their PIPs is similar to the ranking of variables by the posterior means of their associated slope coefficients. This tells us that variables that are "important" in the sense of having high PIPs also have high expected impacts on the outcome variable.

Looking at individual outcome variables in isolation, it is clear from Table 3 that the BMA procedure is a powerful tool for isolating a relatively small set of policy indicators that are robustly partially correlated with the outcome variable of interest. One indication of this is the posterior mean model size reported in the bottom of Table 3, which ranges from 6 to 9. This indicates that the posterior probability-weighted average across models of the number of included regressors is reasonably small. Also, looking at the distribution of inclusion probabilities across indicator variables for each outcome, usually it is straightforward to identify a few policy indicators with much larger inclusion probabilities than all the others.

The difficulty, however, is that despite the similarity of the outcome variables, there is a great deal of instability across outcomes in the set of policy indicators that BMA identifies as important partial correlates of these outcomes. To take a specific example, consider the EIU and CPIA outcome variables, which, as noted earlier, have a pairwise correlation of 0.78. Despite this high correlation in outcome variables, there is little overlap in the sets of specific policy indicators that are identified as important partial correlates of outcomes. For example, the BMA procedure identifies legal protections for creditors as the most important correlate of the CPIA outcome variable. However, this policy indicator ranks only 19th most important for the EIU outcome variable. Looking at the set of top ten indicators for both of these outcomes, we find that there are only three policy indicators common to both sets.

[11] Note that these are unconditional means and standard deviations, i.e. averaging across all models including those in which the variable does not appear and for which the slope coefficient is then by definition zero. To obtain the posterior mean conditional on inclusion, we need to scale the reported mean by the inclusion probability.

More systematically, looking through Table 3, not one of the 38 policy indicators scored by Doing Business is in the set of "top ten" important indicators for all seven outcome variables. Only one policy indicator is in the "top ten" set for six outcomes (the "Recovery Rate" component of "Closing a Business"), and only three other policy indicators are in the "top ten" set for five out of seven outcomes (the "Import Time" and "Import Cost" component of "Trading Across Borders" and the "Legal Rights" component of "Getting Credit"). In contrast, 29 out of 38 policy indicators are in the "top ten" set for at least one outcome indicator. This suggests a great deal of instability across outcome variables in terms of the set of policy indicators that are identified by BMA as being important partial correlates of the outcome of interest.

How surprising is this instability across outcomes in the set of important policy indicators? To answer this question, it is useful to consider two polar benchmarks. The first is that there is *perfect stability* in the set of important policy indicators across all seven outcomes. Under this benchmark, we would expect to find that 10 of the policy indicators show up as important for all seven outcome measures, while the remaining 28 are important for none of the seven outcome measures. The second benchmark is that there is *perfect instability* across all seven outcomes in the set of important policy indicators, in the sense that the set of important indicators for each indicator is chosen at random, and this process is repeated independently across the seven outcome variables. Since there is a 10/38 chance of a given policy indicator being included in the "top ten" set, under this benchmark the distribution of the number of outcomes for which each policy indicator is in the top ten list is a binomial random variable with 7 trials and a success probability of 10/38.

We plot these two polar benchmark distributions, together with the actual distribution of the number of outcomes for which each policy indicator is identified as important, in Figure 1. A quick look at Figure 1 suggests that the actual distribution is much more similar to the benchmark of perfect instability than it is to the benchmark of perfect stability. For example, the perfect stability benchmark implies that 10 of the policy indicators should be classified as important for all seven outcomes, while in the data none are. Similarly, the perfect stability benchmark implies that 28 of the policy indicators should be classified as important for none of the seven outcomes, while in fact only 9 policy indicators fall in this category. And finally, the perfect stability benchmark suggests that no policy indicators should be classified as important for between one and six outcome variables, but in fact 29 policy indicators fall in this intermediate category.

In contrast, the actual distribution of the number of outcomes for which each indicator is important has a shape that is similar to that of the binomial distribution implied by the benchmark of perfect instability. Of course, the match with the binomial distribution is not perfect either. For example, the observed distribution has a slightly fatter right tail than the binomial, indicating that at least a few policy indicators show up as important across a larger number of outcome variables than would be the case if the pattern of importance were purely random. And similarly, there are a few more indicators that show up as never being important than would be the case under the benchmark of perfect instability. Overall however, the actual pattern of importance of individual policy indicators appears to be much more similar to the benchmark of perfect instability than to the benchmark of perfect stability. [12]

In Table 4 we report these two benchmark distributions (in Panel A), and then the actual observed distributions for the baseline specification in Table 3 (in Row 1 of Panel B). The remainder of Table 4 reports similar information for several robustness checks described in the following section. In addition, in the final two columns of Table 4, we report a simple summary measure of the difference between the observed distributions and the distributions implied by the two benchmarks of perfect stability and perfect instability. This summary measure is the sum of the absolute deviations between the actual number of policy indicators in each bin and those predicted by each of the benchmarks, normalized to run between zero and 100 percent by scaling by twice the number of policy indicators.[13] Consistent with Figure 1, this dissimilarity index is much larger for the benchmark of perfect stability (at 76 percent) than it is for the benchmark of perfect instability (at 21 percent).

[12] In fact, a natural question one might ask is whether the difference between the observed distribution and the distribution implied by the benchmark of perfect instability is statistically significant. In principle this question can be answered using a chi-squared test to assess the goodness-of-fit of the observed distribution compared with the relevant binomial distribution. Doing so is complicated however by the fact that the expected bin count is quite small in several of the bins in Table 4. For example, with just 38 policy indicators, the predicted bin count is above five for only three of the eight bins in Table 4. A common rule of thumb for chi-squared tests of goodness of fit is that the quality of the asymptotic chi-squared approximation is poor when there are bin counts lower than five. One remedy would be to group the 0 and 1 outcome bins into a single bin, and the 3-7 outcome bins into another single bin, and then perform a chi-squared test based on the resulting three bins. The chi-squared statistic, i.e. the sum of observed minus expected counts squared and normalized by the expected counts in each bin is 2.8. Comparing this with a chi-squared distribution with 2 degrees of freedom suggests a p-value of 0.25, i.e. we would not reject the null hypothesis of perfect instability.

[13] To understand this scaling, consider the case where the observed distribution placed all 38 indicators in a category in which the benchmark distribution places zero indicators. The sum across categories of the absolute deviation between actual and predicted counts would be 2x38. Rescaling by this amount puts an upper bound on the dissimilarity index of 100 percent. In contrast, if the actual distribution were identical to the benchmark distribution, the dissimilarity index would be zero.

In the online Appendix A accompanying this paper we report a version of Table 3 for the much larger set of 303 policy indicators included in the Global Integrity database. Our findings in this dataset are broadly similar. As with Doing Business, we find that the BMA procedure does a good job of identifying a relatively small set of policy indicators that are robustly partially correlated with each outcome variable, and that posterior probabilities are concentrated on a reasonably small set of parsimonious models. However, we again find the same problem that the set of "important" policy indicators with the highest PIPs is unstable across outcome variables, even though our outcome measures are all conceptually similar and fairly strongly-correlated measures of corruption perceptions.[14] For example, we find that only one of 303 indicators is included in the set of "important" policy indicators for all seven outcome variables. The dissimilarity index is 78 percent for the benchmark of perfect stability, and just 15 percent for the benchmark of perfect instability.

We began this paper with the observation that realizing the promise of detailed datasets of specific policy indicators to identify reform priorities requires knowledge of the partial effects of these potentially many indicators on outcomes that policymakers might reasonably care about. We have seen that, for a given outcome, the BMA methodology used here yields useful results by identifying a fairly small subset of the Doing Business indicators that are robustly partially correlated with the outcome of interest. The key challenge, however, is that across quite similar outcomes, we find very different sets of important partial correlates of outcomes. This suggests that a policymaker would find it difficult to use this type of data and analysis to narrow down the set of potential reforms to a small set of indicators that matter systematically across a range of relevant outcomes.

5. Robustness

Our key finding of interest is that there is a great deal of instability across related outcome variables in terms of the set of policy indicators that are important partial correlates of these outcomes. In this section we consider several potential explanations for this instability finding, including (1) the extent to which our results are driven by differences in sample size across outcome indicators; (2) alternative criteria for identifying the set of important correlates of outcomes; (3) the specification of the prior distribution in the BMA analysis; (4) the potential role of near-collinearity among the policy

[14] In order to keep our results on instability comparable across these two datasets, we define the set of "important" policy indicators as the top 25 percent of the 303 policy indicators in the Global Integrity dataset, as ranked by their posterior inclusion probabilities.

indicators; (5) whether our results simply reflect low correlations across outcome variables; (6) the possibility that the true models linking policy indicators to outcomes are in reality different across outcome variables; and (7) possible nonlinearities and omitted variables in the relationship between policy indicators and outcomes.

5.1 Differences in Country Samples

A first pedestrian potential explanation for the observed instability across outcomes in the set of important policy indicators is that it might simply reflect differences across outcomes in the set of countries included in the analysis. Not all of the outcome variables are available for all countries, and in order to use as much information as possible, we have performed our analysis using the largest available set of countries for each choice of outcome variable. However, if the true model relating policy indicators to outcomes were different across different sets of countries, this might contribute to our instability finding. To investigate this possibility, we repeat the analysis in Table 3, but restricting attention to the (much smaller) set of 70 countries for which all seven outcome variables are available. The results are reported in the online Appendix Table B3.1. Moving to this smaller sample consisting only of developing countries (since the CPIA outcome covers only developing countries) naturally leads to differences in the sets of variables identified by BMA as being important for each outcome, in the sense of having a high PIP. Most relevant for our results, however, is that we continue to find the same pattern of strong instability across outcomes, as summarized in Table 4. The dissimilarity index for the benchmark of perfect stability remains much higher than for the benchmark of perfect instability (68 percent versus 29 percent). This suggests that our instability across outcomes finding is not primarily due to differences in country samples.

5.2 Alternative Criteria for Identifying "Important" Policy Indicators

A second straightforward concern is that our finding on instability across outcomes may be driven by the particular criterion we have used to identify important policy indicators (i.e. that they are among the top-ten ranked variables by PIP). In this subsection, we consider a variety of alternative criteria in order to verify that our conclusions about instability are not sensitive to this choice. A first natural variant is to consider alternatives for the unavoidably-arbitrary cut-off value that we have been using to identify the set of important policy indicators. After all, in Table 3, many of the explanatory variables have quite low PIPs, that are practically indistinguishable from zero or from each other. To the

extent that the "top ten" list of important policy indicators includes some of these variables with low PIPs, it might not be surprising to find a lot of instability across outcome variables in the set of important policy indicators.

We explore this possibility by simply taking alternative thresholds for defining the set of "important" policy indicators, looking instead at the "top five" and "top two" indicators. The second and third rows of Panel C in Table 4 summarize the degree of stability across outcomes in the set of important regressors for these two alternative criteria. Consider the case of the "top five" criterion. If the set of important indicators were perfectly stable across outcomes we should expect to see 5 indicators classified as important for all seven outcomes, and the remaining 33 indicators to be important for none of the outcomes. In the data, the corresponding figures are zero and 19, respectively. We also see that the dissimilarity index is once again much higher for the benchmark of perfect stability than it is for the benchmark of perfect instability (at 50 percent versus 24 percent, respectively).[15]

Another natural criterion for identifying important policy indicators would be to focus simply on the level of the PIP, rather than the rank. As we noted earlier, and discuss in more detail in the following subsection, a drawback of this approach is that levels of PIPs are sensitive to the choice of prior in the BMA analysis. However, it has the advantage of not including as "important" some variables that turn out to have quite low PIPs in our baseline results. Accordingly, we classify policy indicators as "important" if their PIPs are greater than 50 percent. Averaging across all seven outcome variables, this results in roughly 5/38=13 percent of policy indicators being classified as "important", as opposed to 10/38=26 percent in our benchmark results. Doing so does not affect our conclusion that the set of important indicators is quite unstable across outcomes. The dissimilarity index for the benchmark of perfect stability is again much higher than for the benchmark of perfect instability (50 versus 24 percent).

[15] In doing this calculation, we appropriately revise the two benchmark distributions. The distribution in the case of perfect stability would have either five or two policy indicators classified as important for all seven outcomes, with the remainder classified as important for none of the outcomes. The distribution in the case of perfect instability would be binomial with success probabilities of either 5/38 or 2/38. We similarly adjust the benchmark distributions as required in the remaining rows of this table, for the additional robustness checks discussed below, when these checks result in changes in either (a) the proportion of policy indicators classified as important, and/or (b) the number of policy indicators being analyzed.

A final interesting exercise is to contrast this finding of instability across outcomes in the set of important policy indicators as identified by BMA with simpler approaches of opportunistically "picking and choosing" among policy indicators according to some ad hoc criterion. To take a specific example, suppose instead we had simply identified the ten policy indicators that have the highest simple correlation with each of the seven outcome indicators. Given that the outcome indicators are quite highly correlated among themselves, it is not surprising that this alternative set of "top ten " indicators is stable across outcomes. For example, four indicators (Number of Days to Export, Number of Days to Import, Recovery Costs under Closing a Business, and Cost of Starting a Business), appear in the top-ten set for all seven outcome variables. However, this rather naive approach does not identify the same set of "important" policy indicators as the BMA procedure, which in effect averages partial correlations (not simple correlations) across all possible combinations of indicator variables. For example, Number of Days to Export is in the top-ten (and even top-five) set of indicators with the highest simple correlation with outcomes for all seven outcomes. But as shown in Table 3 it ranks in the top-10 set of policy indicators by PIP for only two out of seven outcomes.

5.3 Sensitivity to Specification of Prior

We next consider the extent to which our conclusions may be driven by our choice of prior. We first consider the choice of prior parameter g. Recall from Section 3 that the parameter g plays two roles in the assignment of posterior probabilities across models: lower values of g increase the model size penalty for including additional regressors, and lower values of g also increase the sensitivity of the posterior probability to improvements in R-squared. Together, these two forces imply that when g is small, the posterior probability will tend to concentrate on models with few regressors, and among these, on models with high R-squareds. This concentration of posterior model probabilities on a few models can be extreme, a phenomenon which Feldkircher and Zeugner (2009) label the “supermodel effect”. [16] And this in turn can lead to a strong concentration of high PIPs on just a few variables. Potentially, this “supermodel effect” can lead to large changes in posterior model probabilities and PIPs as we move from one dependent variable to another, if the "supermodel" happens to be different for the different outcomes. This is why, in our benchmark results, we measure the importance of policy

[16] Note that Feldkircher and Zeugner (2009) define g as the inverse of how g is defined in this paper, i.e. their prior variance is $\sigma^2 g\left(X_j{}'X_j\right)^{-1}$, and so given their notation they emphasize the adverse consequences of choosing a *large* value of g.

indicators by their *ordering* by PIPs, rather than the magnitudes of the PIPs themselves. While the latter can be sensitive to the choice of g, the former are less so.[17]

Standard choices in the literature include setting $g = 1/N$ or $g = 1/K^2$ (see for example Fernandez, Ley and Steel (2001b) and Ley and Steel (2009)), as well as allowing a hyperprior distribution over this prior parameter g (see for example Feldkircher and Zeugner (2009) and Ley and Steel (2012)). We next consider how our benchmark results (which set $g = 0.01$) change under these alternative assumptions. A full set of results corresponding to these three alternatives can be found in online Appendix Tables B3.2-4. Not surprisingly, our findings change only minimally if we use the first alternative of setting $g = 1/N$, since the sample size across our different outcome variables ranges from 130 to 178, and so g ranges from 0.006 to 0.008, which is not very different from our baseline results which set $g = 0.01$. The ranking of variables according to their PIPs is nearly identical, and as a result, our finding of instability across outcomes is identical to that in the baseline results.

The option of choosing $g = 1/K^2$ implies a much smaller value of $g = 0.0007$ in our case where $K = 38$ regressors. Consistent with the intuitions described above, we find much more concentration of posterior probability mass on a smaller number of top models. Comparing Appendix Table B3.3 with the results in Table 3, we find that now only 51 models are required to cover 50 percent of the posterior probability on average across outcomes (as compared with 526 in our benchmark results). Moreover, we find that the ordering of policy indicators according to their PIPs is somewhat different. However, neither of these differences materially affect our conclusion about instability in the set of important policy indicators across outcomes -- in Row 6 of Panel C of Table 4, we find a nearly identical pattern of instability in the set of important policy indicators across the seven outcome variables.

We also implement the BMA analysis, but using a hyperprior distribution for the prior parameter g, as advocated by Liang et. al. (2008) and Feldkircher and Zeugner (2009). This effectively averages all

[17] Ciccone and Jarocinski (2010) also emphasize that standard choices of g lead to a strong sensitivity of posterior model probabilities to small differences in the goodness of fit of individual models. Their interpretation of this finding is that measurement error in cross-country income and growth determinants data is likely to lead to spurious differences in model fit across specifications, and this in turn makes agnostic Bayesian analysis of growth determinants uninformative. While measurement error surely is present in the data we use in this paper, we think their conclusion about Bayesian model averaging may be overly pessimistic. As noted in the main text, the *ordering* of variables by their posterior inclusion probabilities is much less sensitive to differences in model fit. And moreover, as we show below, our conclusions about instability persist even when we rely on recent advances such as hyperpriors for g that smooth out the influence of this aspect of the prior specification.

results across a range of alternative values of g, and so smooths out any effects of g on inferences. The results for this are in Appendix Table B3.4.[18] We find that the rank ordering of variables by their PIPs is nearly identical to that in our baseline results case. This in turn means that our finding of instability across outcome variables is also preserved when we use the hyperprior for g (Row 7 of Panel C in Table 4). Based on this evidence, we conclude that our instability finding is probably not driven by the particular choice of the prior parameter g.

A final noteworthy feature of our prior is that we are imposing the assumption that all policy indicators are equally likely to be included in a given model. As argued by Brock and Durlauf (2001) and Durlauf, Kourtellos and Tan (2008, 2012), this assumption might not be appropriate when the various right-hand-side variables can naturally be grouped into alternative "theories", for which multiple empirical proxies are available. The concern in a nutshell is that a researcher might have a prior belief that alternative theories are equally likely, but if some theories have a large number of potential corresponding right-hand-side variables as proxies, then assigning equal prior weight to individual variables will implicitly overweight those theories for which many empirical proxies happen to be available. They then propose implementing a hierarchical prior in which (a) theories are equi-probable, but (b) within theories, models with multiple correlated proxy variables are downweighted by a factor proportional to the determinant of the correlation matrix of those proxies.

In the context of the Doing Business dataset, it is natural to think of the 10 different Doing Business "topics" as corresponding to "theories". Under this interpretation, a potential concern is that the number of indicators is different across topics, ranging from three (for five of the 10 topics) to four (for Starting a Business, Employing Workers, and Getting Credit) to five (for Paying Taxes) and six (for Trading Across Borders), and so our assignment of equal prior probabilities to individual policy indicators implicitly overweights these topics with more indicators. We address this potential concern in a somewhat ad hoc, but straightforward, way. For each Doing Business topic with more than 3 corresponding policy indicators, we average some of the policy indicators together to reduce the overall number to 3 indicators for all topics.[19] We then repeat the BMA analysis using this "balanced" dataset

[18] Specifically, we assume that $\frac{1}{1+g}$ follows a $Beta(1, \frac{a}{2} - 1)$, where $2 < a \leq 4$. We choose a =2.02 so that $E\left[\frac{1}{1+g}\right] = \frac{2}{a} = \frac{1}{1.01}$ so that the hyperprior is centered on our default choice for this parameter in the benchmark specification.

[19] Specifically, for Trading Across Borders, we average together the time, cost, and number of documents indicators for exports and imports; for Paying Taxes we average together the three individual tax rates into one

of 30 indicators together with our default prior with $g = 0.01$, and report the results in Appendix Table B3.5. Once again, the key feature of these results for our purposes is the continued instability across outcomes in terms of the set of "important" policy indicators. As shown in Panel C, Row 8 of Table 4, the pattern of importance of individual indicators is much closer to the benchmark of perfect instability than it is to the benchmark of perfect stability (with a dissimilarity index of 33 percent versus 83 percent).

5.4 Collinearity Among Policy Indicators?

Another possible objection to the instability finding is that it reflects multicollinearity problems in the individual models considered by BMA. As is well-known, one consequence of having nearly collinear regressors in finite samples is that parameter estimates are highly sensitive to small changes in model specification. In light of this, if the policy indicators we consider are correlated across countries (i.e. if countries with strong policy performance in one area of Doing Business also tend to have good policy performance in other areas), then it might not be surprising that the data are not informative about the relative importance of policy indicators across various outcome variables.

The most straightforward response to this concern is simply that the individual policy indicators in the Doing Business dataset are in fact surprisingly uncorrelated with each other. Consider for example the pairwise correlations between the 38 policy indicators in Doing Business. The median pairwise correlation is only 0.18, and 90 percent of the pairwise correlations are smaller than 0.39. For models with only two regressors, this suggests that collinearity problems are not very prevalent in our application. More formally, for each outcome variable, and using our baseline specification in Table 3, we retrieve the top 300 models that have at least two explanatory variables. Then for each model, we compute the R-squared of a regression of each right-hand-side variable on the remaining right-hand-side variables. For a model with K_j regressors, there will be K_j such R-squareds, each one summarizing how collinear a given regressor is with the remaining regressors in the model. A common rule of thumb is that an R-squared greater than 0.9 is a signal of potential finite-sample collinearity problems (i.e. it

---

overall rate, and for Employing Workers we average together the procedures and costs of firing a worker. For the remaining three topics, there is no obvious pairing of indicators to average together based on their stated definitions. Instead, we average together the two indicators with the highest pairwise correlation. Specifically, for Starting a Business, we average together the time and number of procedures indicators, while for Getting Credit we average together the credit information and private credit bureau coverage indicators.

corresponds to a variance inflation factor of 10 or more). For each model we retrieve the maximal R-squared as an indicator of the "worst" possible collinearity problem for that model.

We report the median across models, as well as the 90th percentile and maximum value of these R-squareds, in the bottom of Table 3. Typically these maximal R-squareds are small – the median ranges from 0.26 to 0.60, depending on the choice of outcome variable. Even the 90th percentile of the distribution of these maximal R-squareds is well below the rule-of-thumb value of 0.9, indicating that strong multicollinearity problems are not a feature of the vast majority of models that are assigned high posterior probabilities, and on which our findings are based.

5.5 Insufficiently Correlated Outcome Variables?

Another possible explanation for the instability we observe across outcomes in the set of important regressors is that it simply reflects the fact that the outcome variables themselves are insufficiently strongly correlated with each other. To take an extreme case, if the seven outcome variables we consider are measuring different outcomes that happen to be completely uncorrelated across countries, then it would not be surprising to find that the set of important regressors for each outcome is different.

We assess this interpretation by examining pairs of outcome variables. For each pair of outcomes, we first calculate the simple correlation between the two. Next, for each pair, we summarize the extent to which there is agreement on the set of "important" policy indicators for the two outcomes, by calculating the fraction of all policy indicators that fall in the "top ten" list for both outcomes. We plot this measure of pairwise agreement against the correlation between each pair of outcome variables in Figure 2. We find that there is a positive relationship: when considering pairs of outcomes that are more highly correlated with each other, there is also greater agreement across the two outcomes on the set of important regressors. This is not unexpected – after all, if two outcome variables were perfectly correlated, then necessarily the set of right-hand-side variables identified as important by BMA would have to be the same.

The more striking feature of Figure 2 is how low the agreement is even for those pairs of outcomes that are quite highly correlated. For example, four pairs of outcomes are correlated at 0.74 or higher. Yet for these pairs of outcomes, the number of variables included in the set of top 10 by their PIPs for both outcomes ranges from a low of three to a high of only six. This suggests to us that even

when outcome variables are strongly correlated, there still is a surprising extent of disagreement across outcomes as to which right-hand-side variables are important.

5.6 The World Is Complicated?

As discussed in the introduction, our instability finding can be interpreted in two possible ways: (1) there is a common model linking outcomes to policy indicators, that is the same across all outcomes, but is difficult to pin down empirically; and (2) the true mapping from policy indicators to outcomes in reality is different across outcomes, and this is reflected in our instability findings. While both interpretations pose a challenge for policymakers interested in influencing all of the outcomes with a limited number of policy interventions, it is nevertheless interesting to make some effort to try to distinguish between them.

In this subsection we present two arguments in support of the first interpretation. The first is based on the observation that the outcome variables are themselves quite highly correlated with each other, while the disaggregated policy indicators are not. To see why this limits the extent to which one can plausibly allow for differences in the sets of underlying determinants of outcomes, a stylized example is helpful. Suppose there are two outcome variables, y_1 and y_2, and the true data generating process is that each one depends on a different policy indicator, i.e. $y_i = x_i + \varepsilon_i$ for $i = 1,2$, where x_1 and x_2 are the two different policy indicators, and ε_1 and ε_2 are the error terms which we assume to be uncorrelated with each other and with the policy indicators.[20] For the purposes of this example, the slope of the relationship is not important and so we have normalized it to one for both indicators. Suppose further that the explanatory power of the policy indicators is the same for the two outcomes, i.e. $R^2 \equiv \frac{V(x_1)}{V(y_1)} = \frac{V(x_2)}{V(y_2)}$. Then the correlation of the two outcome indicators is $CORR(y_1, y_2) = R^2 CORR(x_1, x_2)$. In the Doing Business database, the median pairwise correlation between policy indicators is 0.18, and so the *maximum* pairwise correlation between the two outcome indicators in this example can be only 0.18 (since the R-squared is bounded above by one). Yet in reality we observe that the median pairwise correlation in outcome indicators is much higher than this, at 0.68.

[20] In particular, for the purposes of this example, we are assuming that the only source of comovement between the two outcome measures is the extent to which they reflect policy indicators that also are correlated with each other. If, in addition, we allowed the error terms to be correlated across countries, then the correlation of the outcome indicators would simply reflect our assumptions about this unobserved correlation, obscuring the content of this example.

A slightly less stark illustration of the same point would be to define x_1 and x_2 as the sum of non-overlapping sets of N policy indicators, i.e. the two outcomes depend on non-overlapping sets of explanatory variables, and again normalizing the slopes on all the explanatory variables to one. It is straightforward to show that in this case $CORR(x_1, x_2) = N\rho/(N\rho + 1 - \rho)$ where ρ is the correlation between the individual policy indicators (which we assume to be equal for all pairs of indicators). Setting $\rho = 0.18$ to mimic the median observed pairwise correlation between indicators in the Doing Business database, and setting N=10 to correspond to our exercise of focusing on "top-ten" policy indicators, implies that $CORR(x_1, x_2) = 0.69$. This in turn means that models in which outcome variables depend on non-overlapping sets of outcomes can deliver the observed correlation across outcome variables only if the explanatory power of these models is perfect, i.e. $R^2 = 1$. In fact, however, the R-squared of a typical model in the BMA analysis is much lower, at 0.47. As a result, in this simple example, models with non-overlapping sets of indicators can only account for correlations among outcome variables of only 0.47x0.69=0.32, or less than half the typical observed pairwise correlations between outcomes.

If we accept the premise that the various policy indicators contained in the Doing Business dataset do matter non-trivially for the outcomes we have considered, and given that the policy indicators themselves are not correlated across countries, this argument suggests that there must be significant overlaps between the sets of policy indicators that matter for the various outcomes in order to account for the high observed correlations among outcome variables.

A second argument in support of this interpretation comes from considering a specific empirical example, where it is possible to consider two outcome indicators that employ exactly the same scoring criteria. While this is unfortunately not feasible in general, it can be done in the context of one of our outcome variables, the World Bank's CPIA ratings. This is because the African Development Bank also produces CPIA ratings of its member countries, using the same questionnaire and scoring criteria as the World Bank. We take the World Bank and African Development Bank CPIA ratings for the much smaller set of 51 African countries where both are available, and we use the same BMA procedure described above to identify the set of important policy indicators for these two outcome variables.[21] This exercise provides a useful benchmark, because we have similar evaluators and evaluation criteria, and so it

[21] The full CPIA for the African Development Bank is available at http://cpia.afdb.org. We use this CPIA data only in this robustness check and not in the main analysis of the paper only because its geographical coverage is limited to countries in Africa.

seems less plausible that the true model linking policy indicators to outcomes will be different across these two outcomes.

The results are reported in Table 5, which has the same structure as Table 3 for our baseline results. The striking feature of Table 5 is that there is a similar degree of instability in the set of important policy indicators, even when comparing these two outcome variables that assess countries based on exactly the same criteria. Despite the high correlation between the two CPIA outcome variables of 0.81, only five policy indicators fall in the set of top ten policy indicators as ranked by PIP for both outcomes. To take a specific example, the time required to obtain a construction permit is the second most important policy indicator when the World Bank CPIA measure is used as the outcome, but for the African Development Bank CPIA, this variable is only the 16th most important policy indicator. This example suggests that, even when the two outcome variables are similar in terms of criteria and respondent base, there still is considerable instability across the two outcomes in the set of policy indicators identified as important by the BMA procedure.

In summary, in this subsection we have provided two arguments in support of the idea that the true model linking indicators to outcomes is the same across outcomes, but is difficult to identify empirically. The skeptical reader may however remain unpersuaded, and prefer the alternative interpretation that the true mapping from indicators to outcomes in fact is different across the different outcomes. Under either interpretation, though, we emphasize that it is difficult to conclusively identify a small set of policy indicators that matter for all of the various outcome measures of the quality of the business environment. This in turn poses a challenge for policymakers who would like to improve perceptions of the business environment with a limited set of policy interventions.

5.7 Potential Nonlinearities or Omitted Variables in the Mapping from Policy Indicators to Outcomes?

Thus far we have limited the model space to various combinations of the policy indicators themselves. Yet it is plausible on prior grounds that there are a variety of potential nonlinearities in the mapping from policy indicators to outcomes, including interactions of the policy indicators with each other and with additional variables. And if the true common model linking policy indicators to outcomes features such nonlinearities, it could be that our instability result is in part driven by the fact that the true model is not included in our model space. Pursuing this approach exhaustively is an impossibly

open-ended task, given the many potential additional regressors and interactions that one could plausibly consider. For example, simply allowing for pairwise interactions among policy indicators would increase the size of the model space from 2^K to $2^{K+K(K+1)/2}$ models.[22] To keep the proliferation of possibilities manageable, we restrict attention to one plausible set of interactions: of each policy indicator with log GDP per capita, in order to allow for the possibility that the effect of policy indicators on outcomes is dependent on the level of development. In this case, our model space consists of 2^{2K+1} models, corresponding to various combinations of the $K = 38$ policy indicators, the $K = 38$ interactions with log GDP per capita, and log GDP per capita itself.

The results of this robustness check are presented in Appendix Table B3.6.[23] A first observation is that log GDP per capita itself does not appear to be a particularly robust partial correlate of the various outcomes, as it falls in the set of top-10 right-hand-side variables in only two of seven outcomes. There is some evidence that the interaction terms are important partial correlates of the outcome variables. Of the 68 variables that appear in the set of top-10 right-hand-side variables for at least one of the seven outcome variables, 29 are direct effects of policy indicators while 39 are interactions of policy indicators with log GDP per capita. However, the most important observation for our purposes is that there is little improvement in terms of stability across outcomes in the set of important policy indicators. As can be seen in the second-last row of Table 4, the distribution of policy indicators in terms of the number of outcomes for which they are classified as "important" is much closer to the benchmark of perfect instability than it is to the benchmark of perfect stability. While these results are certainly not comprehensive, they do suggest that our instability finding is not due to a failure to allow for one particular natural set of nonlinearities, i.e. interactions with the level of development.

Yet another possible explanation for our instability finding is that there are some omitted variables that (a) are correlated with the policy indicators of interest, and (b) matter for outcomes in different ways for the different outcome variables. If this is the case, then part of our instability finding may be driven simply by the fact that omitted variable bias is different across the different outcome

[22] This would allow the marginal effect of each variable on the outcome of interest to be a linear function of all of the variables in the model, including the variable in question, i.e. such a specification would include quadratic terms in each variable, which is a popular way of capturing nonlinear effects of a single variable in cross-country growth empirics.

[23] We use a burn-in of 300,000 discarded drawings and 700,000 recorded drawings to obtain a satisfactory positive correlation between posterior model probabilities based on empirical frequencies and the exact analytical likelihoods. The greater number of drawings is necessary since the model space is much more expanded here.

variables. As is the case with nonlinearities, addressing the possibility exhaustively is impossible given the large number of potential omitted variables that one might consider. Instead, as a small step in this direction, we repeat our baseline results in Table 3, but controlling for log GDP per capita and a set of regional dummies as crude proxies for potential omitted variables.[24] The full results are reported in Appendix Table B3.7. Once again, we find that this does not materially affect our main conclusion about instability. As shown in the last row of Table 4, the distribution of policy indicators in terms of the number of outcomes for which they are classified as "important" is much closer to the benchmark of perfect instability than it is to the benchmark of perfect stability.

6. Conclusions

Policymakers and aid donors interested in using highly-detailed indicators of specific policies relevant to governance and the business environment to identify reform priorities would like to know the impacts of reforms in specific areas on outcomes that they care about. The results of this paper suggest that the data on policy indicators and outcomes we have considered do not permit sharp discrimination between those specific policies that matter systematically for outcomes of interest, and those that do not. This should be worrisome for a policymaker interested in using these specific policy indicators to identify reform priorities. To give a very stark example, the results here suggest that roughly three-quarters of the 38 detailed policy indicators in the Doing Business dataset are important partial correlates of at least one of the seven closely-related outcome measures of perceptions of the quality of the business environment that we have considered. While this is a tribute to the relevance of the overall Doing Business indicators, it does little to narrow down the set of measures a policymaker might want to target for reforms.

Beyond this, we note that there are likely to be even bigger obstacles to using such specific indicators to identify reform priorities than the ones we have seen here. In this paper we have relied on the simple tool of linear OLS regressions as a means of identifying the effects of specific indicators on outcomes. As we have noted, the standard exogeneity assumptions required to justify such an empirical approach are unlikely to hold in reality. However it is unclear how one might find instruments or other sources of plausibly exogenous variation in the many different dimensions of regulatory policy and governance measured by Doing Business in order to strengthen identification. The assumption of

[24] Specifically, we project each of the policy indicators and each of the outcome variables on these control variables, and then repeat the BMA analysis using the residuals from these regressions.

linearity is also quite restrictive: one could easily imagine that improvements in various combinations of the individual indicators are required to improve outcomes, implying that we need to consider not only the 2^K potential combinations of regressors, but the vastly larger number of possible combinations of interactions between them as well. And many other potential nonlinearities might be present. For example one hypothesis might be that of "weakest links" whereby a country's performance depends primarily on the areas in which it has the lowest scores, regardless of which indicators these might be.

In concluding, we want to be clear that we do not think that the information painstakingly gathered in the many individual policy indicators that comprise the Doing Business or Global Integrity datasets is irrelevant. To the contrary, most if not all of them measure things that plausibly are intrinsically good on their own (it is hard to imagine why it might be a *bad* idea to simplify business entry regulation from current levels in most countries, for example), and it also seems intuitive that they matter for outcomes. We also do not view our results as a critique of the methodology of Bayesian Model Averaging, which we find to be a useful tool for identifying a parsimonious set of partial correlates of each outcome variable in our application with many potential explanatory variables. Rather, we simply note that, despite the combination of careful data and powerful empirical tools, it is extremely difficult to robustly quantify the partial effects of these many indicators on relevant outcomes. This in turn illustrates why it may be difficult to use these indicators as a recipe or a roadmap to reforms in the real world, where policymakers must choose to spend their political capital on a limited number of reform priorities.

References

Acemoglu, D., Johnson, S., & Robinson, J.A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, 91(5), 1369-1401.

Brock, W.A., & Durlauf, S.N. (2001). Growth Empirics and Reality. *World Bank Economic Review*, 15(2), 229-272.

Brock, W.A., Durlauf, S.N., & West, K.D. (2003). Policy Evaluation in Uncertain Economic Environments. *Brookings Papers of Economic Activity*, 34(1), 235-301.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when *p* is much larger than *n*. Annals of Statistics, 35(6), 2313-2351.

Ciccone, A., & Jarocinski, M. (2010). Determinants of Economic Growth: Will Data Tell?. *American Economic Journal: Macroeconomics*, 2(4), 222-46.

Durlauf, S. N., Kourtellos, A., & Tan, C.M. (2008). Are Any Growth Theories Robust?. *The Economic Journal*, 118 (527), 329-346.

Durlauf, S. N., Kourtellos, A., & Tan, C.M. (2012). Is God in the Details? A Reexamination of the Role of Religion in Economic Growth. *Journal of Applied Econometrics*, 27 (7), 1059-1075.

Eicher, T.S., Lenkoski, A., & Raftery, A.E. Bayesian Model Averaging and Endogeneity Under Model Uncertainty: An Application to Development Determinants. forthcoming in Econometric Reviews..

Eicher, T.S., Papageorgiou, C., & Raftery, A.E. (2011). Default Priors and Predictive Performance in Bayesian Model Averaging, with Application to Growth Determinants. *Journal of Applied Econometrics*, 26 (1), 30-55.

Eicher, T.S., Papageorgiou, C., & Roehn, O. (2007). Unraveling the Fortunes of the Fortunate: An Iterative Bayesian Model Averaging (IBMA) Approach. *Journal of Macroeconomics*, 29(3), 494-514.

Feldkircher, M., & Zeugner, S. (2009). Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging. IMF Working Paper No. 09/202, International Monetary Fund.

Feldkircher, M., & Zeugner,S. (2012). The impact of data revisions on the robustness of growth determinants—a note on 'determinants of economic growth: Will data tell? , Journal of Applied Econonometrics, 27 (4), 686-694.

Fernandez, C., Ley, E., & Steel, M.F.J. (2001a). Model Uncertainty in Cross-Country Growth Regressions. *Journal of Applied Econometrics*, 16(5), 563-576.

Fernandez, C., Ley, E., & Steel, M.F.J. (2001b). Benchmark prior for Bayesian model averaging. *Journal of Econometrics*, 100(2), 381-427.

George, E.I., & McCulloch, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica*, 7(2), 339-373.

Hendry, D. F., & Krolzig, H.-M. (2004). We Ran One Regression. *Oxford Bulletin of Economics and Statistics*, 66(5), 799-810.

Hendry, D. F., & Krolzig, H.-M. (2005). The Properties of Automatic *GETS* Modelling. *The Economic Journal*, 115(502), C32-C61.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). Governance Matters VIII: Aggregate and Individual Governance Indicators, 1996-2008. World Bank Policy Research Working Paper No. 4978.

Ley, E., & Steel, M.F.J. (2009). On the Effect of Prior Assumptions in Bayesian Model Averaging with Applications to Growth Regression. *Journal of Applied Econometrics*, 24(4), 651-674.

Ley, E., & Steel, M.F.J. (2012). Mixtures of g-Priors for Bayesian Model Averaging With Economic Applications. Journal of Econometrics, 171,251-266.

Liang, F., Paulo, R., Molina, G., Clyde, M.A., & Berger, J.O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103(481), 410-423.

Lubotsky, D., & Wittenberg, M. (2006). Interpretation of Regressions with Multiple Proxies. *The Review of Economics and Statistics*, 88(3), 549-62.

Madigan, D., & York, J. (1995). Bayesian Graphical Models for Discrete data. *International Statistical Review*, 63(2), 215-232.

Mauro, P. (1995). Corruption and Growth. *Quarterly Journal of Economics*, 110(3), 681–712.

Moral-Benito, E. (2012). Determinants of Economic Growth: A Bayesian Panel Data Approach. The Review of Economics and Statistics, 94(2), 566-579.

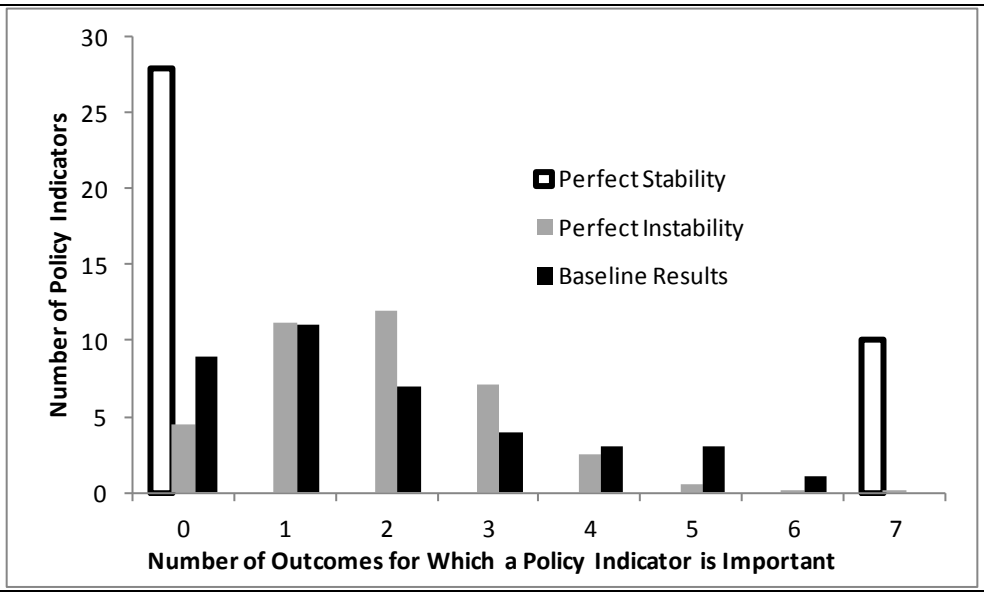
Pritchett, L. (1996). Measuring Outward Orientation in Developing Countries: Can It Be Done? *Journal of Development Economics*, 49(2), 307-35.

Raftery, A., Madigan, D., & Hoeting, J. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92, 179-191.

Sala-i-Martin, X., Doppelhofer, G., & Miller, R. (2004). Determinants of Long-Term Growth: A Bayesian Averaging of Classical Estimates (BACE) Approach. *American Economic Review*, 94(4), 813-835.

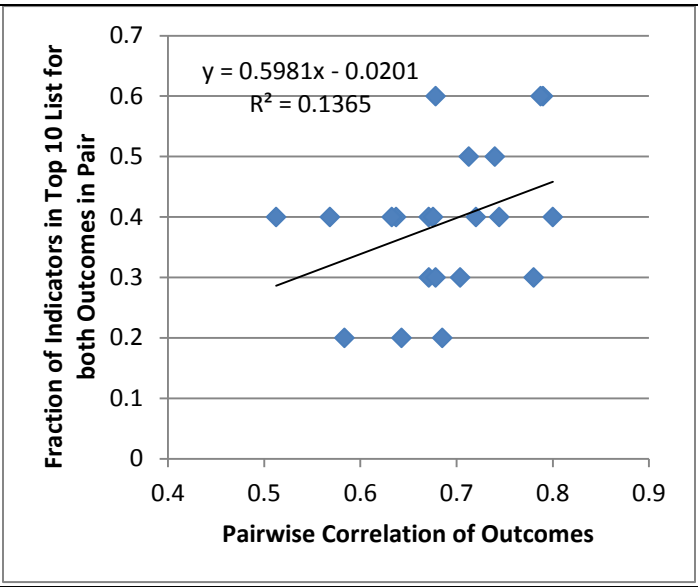
Trapnell, S. (2011). Actionable Governance Indicators: Turning Measurement Into Reform. *Hague Journal on the Rule of Law*, 3(2), 317-348.

Figure 1: Distribution of Policy Indicators According to the Number of Outcomes for Which They Classified as Important



Notes: This figure shows the distribution of the policy indicators according to the number of outcome variables for which they are classified as important. The empty bars indicate the benchmark of perfect stability across outcomes. The light-shaded bars indicate the benchmark of perfect instability across outcomes. The solid bars indicate the actual distribution observed in Table 3.

Figure 2: Pairwise Stability and Pairwise Correlation in Outcomes



Notes: This figure shows the relation across different pairs of the seven outcome variables between the fraction of Doing Business indicators that fall in the top 10 list for both outcome variables and the correlation of the two outcome variables.

Table 1: Outcome Variables for Doing Business

	Concepts Measured	*Summary Statistics*					
		Mean	Std.Dev.	P25	P50	P75	N
Global Insight Global Risk Service (DRI)		**0.84**	**0.12**	**0.80**	**0.89**	**0.92**	**137**
(Average of)	Export Regulation						
	Import Regulation						
	Other Regulatory Burdens						
	Restrictions on Foreign Business Ownership						
	Restrictions on Foreign Equity Ownership						
Economist Intelligence Unit (EIU)		**0.55**	**0.21**	**0.40**	**0.53**	**0.70**	**144**
(Average of)	Unfair competitive practices						
	Price controls						
	Discriminatory tariffs						
	Excessive protections						
	Discriminatory taxes						
Merchant International Group Gray Area Dynamics (GAD)		**0.60**	**0.21**	**0.45**	**0.55**	**0.75**	**158**
(Average of)	Stock Exchange / Capital Markets						
	Foreign Investment						
Global Competitiveness Report Executive Opinion Survey (GCS)		**0.52**	**0.11**	**0.43**	**0.50**	**0.61**	**130**
(Average of)	Administrative regulations are burdensome						
	Tax system is distortionary						
	Import barriers / cost of tariffs as obstacle to growth						
	Competition in local market is limited						
	It is easy to start company						
	Anti monopoly policy is lax and ineffective						
	Environmental regulations hurt competitiveness						
Political Risk Services International Country Risk Guide (PRS)		**0.72**	**0.20**	**0.59**	**0.73**	**0.91**	**133**
	Risk to operations from contract viability, expropriation, repatriation and payment delays.						
Global Insight Business Conditions and Risk Indicators (WMO)		**0.59**	**0.23**	**0.44**	**0.56**	**0.81**	**178**
(Average of)	Efficiency of Tax Collection						
	Business Legislation Complete and Compatible						
World Bank Country Policy and Institutional Assessments (CPIA)		**0.55**	**0.14**	**0.45**	**0.55**	**0.65**	**139**
(Average of)	Business regulatory environment						
	Trade policy						

Table 2: Correlation Between Aggregate Doing Business Indicator and Outcomes

		Perceptions of Regulatory Quality from:						
		DRI	EIU	GAD	GCS	PRS	WMO	CPIA
Unconditionally								
	Slope for Overall DB	0.49	1.11	1.17	0.60	0.98	1.33	0.90
	Standard error	0.06	0.09	0.08	0.05	0.10	0.08	0.08
	t-statistic	7.82	12.54	15.07	12.63	10.28	16.71	11.10
	R-squared	0.31	0.52	0.59	0.55	0.44	0.61	0.47
	N	137	144	158	130	133	178	139
Conditional on Log GDP Per Capita								
	Slope for Overall DB	0.37	0.72	0.79	0.46	0.56	0.57	0.78
	Standard error	0.09	0.13	0.12	0.07	0.14	0.10	0.10
	t-statistic	3.96	5.73	6.85	6.58	3.93	5.79	7.61
	R-squared	0.30	0.57	0.63	0.57	0.50	0.75	0.48
	N	133	141	156	128	131	173	137
Correlation with Log GDP Per Capita		**0.473**	**0.689**	**0.723**	**0.66**	**0.668**	**0.842**	**0.514**

Notes: This table reports the results of OLS regressions of each of the outcome variables on the overall Doing Business Index. The top panel reports results from simple bivariate regressions while the bottom panel conditions on log GDP per capita. The bottom row reports the simple correlation of each outcome indicator with log GDP Per Capita.

Table 3: BMA Results for the Doing Business Dataset

Dependent Variable=		**DRI**			**EIU**			**GAD**			**GCS**			**PRS**			**WMO**			**CPIA**	
	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD
1. Starting a Business																					
Number of Procedures	0.046	-0.001	0.007	0.253	0.026	0.050	0.355	0.033	0.050	0.435	0.024	0.031	0.157	0.015	0.040	0.043	0.001	0.008	0.133	0.009	0.027
Number of Days	0.089	0.004	0.018	0.044	0.001	0.014	0.379	0.039	0.056	0.264	0.015	0.028	0.046	0.002	0.015	0.069	0.003	0.015	0.829	0.094	0.054
Cost	0.098	0.006	0.022	0.117	0.012	0.040	0.113	0.010	0.034	0.051	0.001	0.008	0.490	0.090	0.106	0.834	0.125	0.071	0.230	0.019	0.040
Minimum Capital Requirement	0.114	-0.004	0.014	0.051	-0.002	0.011	0.037	0.000	0.006	0.051	0.001	0.005	0.042	-0.001	0.009	0.048	-0.001	0.007	0.076	-0.002	0.010
2. Construction Permits																					
Number of Procedures	0.243	-0.015	0.030	0.057	0.002	0.014	0.066	0.003	0.015	0.040	0.000	0.005	0.053	0.002	0.015	0.042	0.001	0.008	0.072	0.003	0.012
Number of Days	0.634	0.056	0.050	0.045	0.002	0.013	0.051	-0.002	0.012	0.665	0.044	0.036	0.146	0.014	0.040	0.042	0.000	0.008	0.049	0.001	0.009
Cost	0.037	0.000	0.007	0.338	0.042	0.066	0.148	0.011	0.031	0.213	0.010	0.022	0.221	0.028	0.061	0.968	0.140	0.045	0.241	0.017	0.034
3. Employing Workers																					
Difficulty of Hiring	0.157	-0.007	0.018	0.052	-0.002	0.012	0.049	-0.001	0.009	0.101	0.003	0.010	0.043	0.001	0.011	0.081	-0.003	0.013	0.129	-0.006	0.018
Rigidity of Hours	0.048	0.000	0.008	0.076	-0.004	0.018	0.069	-0.003	0.015	0.191	0.008	0.020	0.042	0.001	0.011	0.099	-0.005	0.018	0.060	-0.002	0.010
Difficult of Firing	0.125	0.005	0.017	0.243	0.021	0.043	0.035	0.000	0.007	0.036	0.000	0.004	0.049	0.002	0.014	0.046	-0.001	0.009	0.116	0.006	0.018
Cost of Firing	0.828	0.071	0.040	0.311	0.031	0.052	0.119	0.007	0.024	0.104	0.003	0.012	0.720	0.104	0.077	0.493	0.041	0.047	0.218	0.014	0.032
4. Registering Property																					
Number of Procedures	0.036	0.000	0.006	0.060	0.003	0.014	0.135	0.008	0.023	0.761	0.046	0.031	0.039	0.001	0.010	0.117	0.006	0.018	0.430	0.030	0.039
Number of Days	0.050	-0.001	0.008	0.041	0.001	0.010	0.116	0.007	0.023	0.070	0.002	0.009	0.054	-0.002	0.016	0.059	0.002	0.012	0.047	0.001	0.009
Cost	0.053	0.001	0.009	0.046	0.002	0.013	0.043	0.000	0.008	0.042	-0.001	0.006	0.041	0.001	0.013	0.048	-0.001	0.010	0.312	0.024	0.040
5. Getting Credit																					
Legal Rights	0.923	0.102	0.042	0.100	0.007	0.027	1.000	0.211	0.041	0.308	0.016	0.027	0.110	0.010	0.035	0.977	0.137	0.043	0.936	0.104	0.041
Credit Information	0.045	0.001	0.007	0.111	0.007	0.025	0.065	0.002	0.014	0.189	-0.007	0.017	0.107	0.008	0.028	0.039	0.000	0.007	0.797	0.077	0.047
Credit Registry Coverage	0.051	-0.002	0.013	0.044	-0.001	0.018	0.047	-0.002	0.019	0.033	0.000	0.007	0.038	-0.001	0.018	0.038	0.000	0.012	0.049	-0.001	0.017
Credit Bureau Coverage	0.052	-0.001	0.014	0.150	0.023	0.065	0.415	0.076	0.103	0.059	-0.002	0.013	0.037	0.001	0.020	0.101	0.011	0.040	0.106	0.013	0.049
6. Protecting Investors																					
Disclosure	0.041	0.000	0.005	0.035	0.000	0.008	0.038	0.000	0.007	0.064	-0.001	0.008	0.034	0.001	0.009	0.038	0.000	0.007	0.043	0.001	0.008
Director Liability	0.053	0.001	0.009	0.036	0.001	0.009	0.040	0.001	0.009	0.051	0.001	0.006	0.073	0.004	0.021	0.036	0.000	0.007	0.066	0.002	0.012
Shareholder Suits	0.054	0.001	0.009	0.047	-0.001	0.012	0.040	0.001	0.009	0.123	-0.004	0.014	0.277	0.033	0.060	0.039	0.001	0.008	0.038	0.000	0.007
7. Paying Taxes																					
Number of Payments	0.038	0.000	0.006	0.615	0.076	0.069	0.963	0.143	0.049	0.075	0.002	0.009	0.548	0.079	0.082	0.951	0.122	0.045	0.044	0.000	0.007
Number of Days	0.043	0.001	0.007	0.193	0.017	0.040	0.046	-0.001	0.010	0.989	0.090	0.024	0.199	0.022	0.051	0.147	0.009	0.027	0.048	0.001	0.009
Profit Tax	0.692	-0.055	0.044	0.066	-0.003	0.016	0.036	0.000	0.008	0.066	0.002	0.009	0.045	-0.001	0.012	0.050	-0.001	0.009	0.040	0.000	0.007
Labour Tax	0.091	-0.004	0.017	0.046	0.000	0.011	0.452	-0.043	0.053	0.060	-0.001	0.008	0.053	0.000	0.015	0.107	-0.006	0.022	0.057	-0.002	0.011
Other Tax	0.536	0.037	0.040	0.089	0.006	0.024	0.039	0.000	0.007	0.035	0.000	0.004	0.278	0.033	0.061	0.050	-0.001	0.010	0.044	0.001	0.008

Table 3 Continues on Next Page

Table 3, Cont'd: BMA Results for the Doing Business Dataset

Dependent Variable=		**DRI**			**EIU**			**GAD**			**GCS**			**PRS**			**WMO**			**CPIA**	
	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD	PIP	Mean	SD
8. Trading Across Borders																					
Export: Number of Documents	0.138	0.008	0.026	0.444	0.057	0.073	0.154	0.013	0.035	0.036	0.000	0.005	0.092	0.008	0.031	0.918	0.127	0.058	0.512	0.046	0.051
Export: Number of Days	0.169	0.014	0.040	0.085	0.004	0.047	0.471	0.070	0.084	0.108	0.005	0.019	0.120	0.013	0.053	0.934	0.199	0.077	0.131	0.013	0.043
Export: Cost	0.263	0.025	0.049	0.113	-0.004	0.056	0.059	-0.002	0.015	0.737	0.090	0.058	0.100	0.008	0.033	0.197	-0.015	0.034	0.079	0.003	0.019
Import: Number of Documents	0.063	-0.002	0.016	0.063	0.002	0.024	0.332	0.038	0.061	0.038	0.000	0.006	0.038	-0.001	0.017	0.191	-0.020	0.049	0.243	0.023	0.047
Import: Number of Days	0.705	0.096	0.074	0.767	0.193	0.130	0.191	0.022	0.055	0.066	0.002	0.013	0.530	0.108	0.116	0.098	0.010	0.047	0.727	0.113	0.082
Import: Cost	0.663	0.081	0.067	0.476	0.080	0.104	0.052	-0.001	0.013	0.292	0.032	0.055	0.179	0.021	0.052	0.216	-0.017	0.037	0.197	0.016	0.038
9. Enforcing Contracts																					
Number of Procedures	0.095	-0.004	0.017	0.383	0.043	0.062	0.158	0.012	0.031	0.113	0.004	0.014	0.053	-0.002	0.016	0.047	0.001	0.009	0.077	0.004	0.016
Number of Days	0.885	-0.088	0.041	0.045	0.000	0.011	0.067	-0.003	0.014	0.069	0.002	0.009	0.093	-0.007	0.027	0.038	0.000	0.007	0.048	0.001	0.008
Cost	0.040	0.000	0.006	0.136	-0.011	0.033	0.032	0.000	0.007	0.038	0.000	0.005	0.055	-0.003	0.018	0.081	0.004	0.017	0.045	-0.001	0.009
10. Closing a Business																					
Number of Days	0.037	0.000	0.006	0.318	0.042	0.069	0.049	-0.001	0.017	0.042	0.000	0.009	0.141	-0.018	0.056	0.114	-0.008	0.031	0.074	-0.004	0.023
Cost	0.042	0.000	0.006	0.106	0.009	0.032	0.086	-0.005	0.021	0.073	0.002	0.011	0.046	-0.001	0.016	0.062	0.003	0.015	0.267	0.018	0.035
Recovery Rate	0.052	0.001	0.010	0.609	0.113	0.103	0.994	0.226	0.057	0.990	0.124	0.030	0.639	0.143	0.134	0.889	0.125	0.064	0.263	0.025	0.051
Posterior Probability of:																					
First-best model	0.019			0.011			0.014			0.020			0.016			0.040			0.021		
Second-best model	0.015			0.006			0.014			0.018			0.013			0.037			0.019		
Third-best model	0.013			0.006			0.012			0.015			0.011			0.021			0.008		
Posterior Mean Model Size	8.308			6.813			7.543			7.679			6.027			9.346			7.869		
Number of Models Visited	27086			41532			25311			25646			33769			19851			36131		
Number of Models Covering																					
x% of Posterior Probability																					
x=50%	391			912			448			398			714			214			608		
x=75%	1608			3264			1931			1769			2771			1173			2447		
x=90%	4431			7451			4821			4572			6524			3285			6037		
Corr(PMP)	0.985			0.904			0.969			0.974			0.943			0.990			0.951		
G&M Measure of Probability																					
Mass Visited	0.296			0.469			0.626			0.618			0.548			0.678			0.506		
Number of Observations	137			144			158			130			133			178			139		
Maximal Partial R-Squared for top 300 Models																					
Median	0.457			0.459			0.421			0.260			0.416			0.600			0.518		
90th Percentile	0.549			0.570			0.583			0.504			0.526			0.713			0.617		
Maximum	0.895			0.890			0.751			0.881			0.883			0.895			0.886		

Table 4: Distribution of Policy Indicators According to the Number of Outcomes for Which They Are Classified as Important

	Never			*Sometimes Important*				*Always*	% Deviation From:	
Number of Outcome Variables for Which	***Important***							***Important***	***Perfect***	***Perfect***
a Policy Indicator is Classified as Important	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	***Stability***	***Instability***
Panel A: Two Benchmarks										
1. Perfect Stability	28	0	0	0	0	0	0	10		
2. Perfect Instability	4	11	12	7	3	1	0	0		
Panel B: Baseline Specification										
1. Doing Business (38 Policy Indicators)	9	11	7	4	3	3	1	0	76%	21%
2. Global Integrity (303 Policy Indicators)	67	83	79	35	24	8	6	1	78%	15%
Panel C: Robustness Checks										
1. Common Sample of Countries	12	7	7	5	3	2	2	0	68%	29%
2. Top 5 Criterion	19	12	2	1	4	0	0	0	50%	24%
3. Top 2 Criterion	29	7	0	1	1	0	0	0	24%	13%
4. PIP>50%	19	12	3	0	3	1	0	0	50%	24%
5. g=1/N	9	11	7	4	3	3	1	0	76%	21%
6. g=1/K^2	11	9	5	6	3	3	1	0	71%	26%
7. Hyperprior for g	7	11	11	3	2	4	0	0	82%	16%
8. Balanced Across Theories	5	9	3	4	4	3	2	0	83%	33%
9. Log(GDP) plus Interactions	36	23	11	3	4	0	0	0	54%	13%
10. Log(GDP) plus Regional Dummies	11	4	12	5	3	3	0	0	71%	24%

Table 5: BMA Results for the World Bank and African Development Bank CPIA Outcome Variables

Dependent Variable=	World Bank CPIA			African Dev. Bank CPIA		
	PIP	Mean	SD	PIP	Mean	SD
1. Starting a Business						
Number of Procedures	0.192	0.019	0.046	0.067	0.005	0.031
Number of Days	0.350	0.042	0.065	0.947	0.205	0.075
Cost	0.088	0.011	0.046	0.115	0.017	0.057
Minimum Capital Requirement	0.055	0.002	0.011	0.139	0.009	0.025
2. Construction Permits						
Number of Procedures	0.138	0.013	0.039	0.167	0.017	0.045
Number of Days	0.351	0.043	0.066	0.094	0.007	0.030
Cost	0.038	0.001	0.013	0.105	-0.010	0.036
3. Employing Workers						
Difficulty of Hiring	0.066	-0.003	0.016	0.039	0.000	0.009
Rigidity of Hours	0.042	0.001	0.014	0.047	0.002	0.016
Difficult of Firing	0.049	0.003	0.020	0.061	-0.002	0.021
Cost of Firing	0.049	0.001	0.017	0.066	0.005	0.025
4. Registering Property						
Number of Procedures	0.114	0.008	0.028	0.107	0.008	0.029
Number of Days	0.123	-0.011	0.037	0.050	-0.003	0.019
Cost	0.102	0.009	0.034	0.145	0.016	0.046
5. Getting Credit						
Legal Rights	0.052	0.002	0.018	0.127	0.014	0.045
Credit Information	0.041	0.000	0.017	0.143	0.014	0.046
Credit Registry Coverage	0.144	-0.034	0.098	0.324	-0.110	0.182
Credit Bureau Coverage	0.065	0.015	0.095	0.307	0.170	0.294
6. Protecting Investors						
Disclosure	0.040	0.000	0.012	0.052	0.002	0.018
Director Liability	0.048	0.001	0.012	0.039	0.001	0.012
Shareholder Suits	0.091	0.007	0.026	0.100	0.008	0.030
7. Paying Taxes						
Number of Payments	0.051	-0.003	0.021	0.044	-0.002	0.018
Number of Days	0.037	0.000	0.012	0.044	-0.001	0.013
Profit Tax	0.066	-0.003	0.017	0.080	-0.004	0.020
Labour Tax	0.059	-0.002	0.016	0.066	-0.004	0.020
Other Tax	0.249	0.027	0.054	0.148	0.014	0.040

Dependent Variable=	World Bank CPIA			African Dev. Bank CPIA		
	PIP	Mean	SD	PIP	Mean	SD
8. Trading Across Borders						
Export: Number of Documents	0.090	0.006	0.024	0.075	0.004	0.021
Export: Number of Days	0.092	0.008	0.038	0.042	-0.001	0.018
Export: Cost	0.093	0.007	0.029	0.053	0.003	0.018
Import: Number of Documents	0.224	0.025	0.055	0.060	0.003	0.020
Import: Number of Days	0.166	0.021	0.056	0.043	0.001	0.017
Import: Cost	0.111	0.009	0.031	0.061	0.003	0.019
9. Enforcing Contracts						
Number of Procedures	0.115	0.010	0.035	0.182	0.021	0.051
Number of Days	0.040	-0.001	0.012	0.042	0.000	0.012
Cost	0.135	-0.013	0.039	0.093	-0.008	0.032
10. Closing a Business						
Number of Days	0.085	-0.001	0.034	0.082	0.002	0.037
Cost	0.115	0.010	0.038	0.052	0.000	0.021
Recovery Rate	0.742	0.132	0.094	0.830	0.182	0.104
Posterior Probability of:						
First-best model	0.015			0.055		
Second-best model	0.014			0.024		
Third-best model	0.014			0.015		
Posterior Mean Model Size	4.710			5.235		
Number of Models Visited	30920			24007		
Number of Models Covering						
x% of Posterior Probability						
x=50%	513			262		
x=75%	2201			1375		
x=90%	5400			3701		
Corr(PMP)	0.962			0.987		
G&M Measure of Probability						
Mass Visited	0.563			0.620		
Number of Observations	51			51		