Can Specific Policy Indicators Identify Reform Priorities?

Aart Kraay (The World Bank)

Norikazu Tawara (Akita International University)

First Draft: March 2010

This Draft: October 2011

Abstract: Several detailed cross-country datasets measuring very 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 policy reforms that policymakers and aid donors can target in their efforts to improve regulatory and institutional outcomes. Doing so, however, requires evidence on the partial effects of the many very specific policy choices reflected in such datasets on intermediate outcomes of interest, such as investor perceptions of the quality of the regulatory environment, or perceptions of corruption. In this paper we use Bayesian Model Averaging (BMA) to systematically document the partial correlations between detailed policy indicators and a range of outcome variables of interest. We find major instability across outcomes in the set of policy indicators identified by BMA as important partial correlates of 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 conclusively identify reform priorities using cross-country data.

1818 H St. NW, Washington, DC, akraay@worldbank.org, and Okutsubakidai, Yuwa-Tsubakigawa, Akita, 010-1292 Japan, nori.tawara@gmail.com, respectively. We would like to thank Nathaniel Heller, Daniel Kaufmann, Eduardo Ley, Chris Papageorgiou, Luis Serven, Stefan Zeugner, and two 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.

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 development 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 perceptions of overall corruption from commercial risk rating agencies to growth and investment across countries.

Turning such important and influential findings into concrete policy advice for countries seeking to improve their 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 very detailed indicators of specific laws, regulations, and policies that plausibly influence broad institutional outcomes such as the business regulatory environment or public sector integrity. In this paper we consider two such datasets, covering large cross-sections of countries: (i) the Doing Business project of the World Bank, which reports 41 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 outcomes.[1] However, realizing this promise requires an understanding of the relative magnitude of the partial effects of each of the specific policy indicators on the corresponding outcomes

[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).

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 might 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 very 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. We do this in three steps. First, we identify a set of potential intermediate outcome variables which we think a policymaker might reasonably want to influence through reforms to specific policies measured by our two datasets. In the case of Doing Business, which focuses on business regulatory policies, 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 focus on 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.

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

[2] More sophisticated approaches to aggregation, such as principal components, will have the same difficulty. For example, the factor loadings on the first principal component of a set of policy indicators will reflect only the pairwise correlations among the policy indicators, which need not correspond to the partial effects of the indicators on outcomes. See Lubotsky and Wittenberg (2006) for a more extensive discussion of this problem.

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 ask how these sets differ 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 application 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 20th most important for the EIU outcome variable.

More systematically, for each outcome variable, we identify a set of 10 policy indicators in the Doing Business dataset that are the most important partial correlates of the outcome. Comparing these sets of important policy indicators across outcomes, we find very little overlap: for example, despite the high correlation between the EIU and CPIA outcome variables, 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 41 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. We also cannot reject the null hypothesis that membership in the set of important correlates of one outcome variable is independent of membership in the set of important correlates of other indicators. This finding suggests that policymakers may find it very difficult to use these kinds of datasets to isolate a small number of specific policy reforms that are likely to matter for the set of outcomes they are interested in influencing.

Our main finding of instability across 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 Tables dataset. We share with that paper an emphasis on the instability of regression results across different closely-related outcome variables. Our work differs however in that Ciccone and Jarocinski (2010) 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 very specific policy levers and intermediate outcomes that policymakers might want to influence using these levers.

The rest of this paper proceeds as follows. In Section 2 we describe the Doing Business dataset and the corresponding set of outcome variables. In Section 3 we explain the Bayesian Model Averaging methodology. Section 4 contains our main results, and Section 5 discusses the robustness of the results. Section 6 offers conclusions. To conserve space, our findings based on the Global Integrity Index are confined to an online appendix.

2. A First Look at the Data

We work with data from the 2009 edition of the Doing Business report (World Bank (2009)), which reports on 41 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 41 indicators are also aggregated into an overall "Ease of Doing Business" indicator which averages together each country's rank on the individual indicators.

The DB 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.[3]

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, policy reforms reflect policymakers' ambition to influence a potentially very large set of ultimate outcomes, ranging from their own political interests to broad considerations of social welfare. In our empirical exercise we take the more limited view that one relevant intermediate outcome policymakers are likely to care about when formulating business regulatory reforms is the extent to which these 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 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.

Of course, there is no one unique measure of investor perceptions of the business regulatory environment, and policymakers might very well want to inform their decisions by looking at the relationship between specific policy reforms and a variety of measures of perceptions of regulatory quality. We use seven such closely-related measures as outcome variables for the Doing Business dataset of policy indicators. 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

[3] While for terminological convenience we refer to the 41 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 involve changing or eliminating individual steps such as these, which would in turn have implications for time, cost and number of procedures.

from a large cross-country survey, the Global Competitiveness Report (GCS) survey of firms in 134 countries, that asks firm managers a variety of questions about the quality of the business environment.

Conceptually, these data sources are closely related in the sense that they broadly 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 that are 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 a 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.[4] 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 7 outcome variables accounts for 73 percent of the total variation in these variables.[5]

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 very 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

[4] 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 (2010) for more detailed descriptions).

[5] The median pairwise Kendall rank correlation coefficient, which compares the ranking of all pairs of observations for two variables, is slightly lower, at 0.54.

significantly correlated with the overall Ease of Doing Business indicator, which averages together countries' ranks on all of the specific policy indicators. 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 unconditional 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.

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 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 41 policy indicators in the Doing Business dataset.

[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 (Durlauf, Kourtellos, and Tan (2008)), panel data applications (Moral-Benito (2009)), and instrumental variables estimation (Eicher, Lenkoski, and Raftery (2009). Finally, several papers including Fernandez, Ley and Steel (2001b), Ley and Steel (2009), Eicher, Papageorgiou and Raftery (2009) and Feldkircher and Zeugner (2009) all discuss the consequences of alternative prior assumptions for the outcome of BMA.

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.:

(1) $$y = \iota\alpha_j + X_j\beta_j + \varepsilon_j$$

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.

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

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 for variables inside the parenthesis. The prior distribution for the slope coefficients, conditional on σ, α_j, and model j, is 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.

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

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 higher 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:

(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 effect 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 improving one of the outcome variables, who 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 very close to one. [7]

We implement the BMA procedure seven times, corresponding to the seven different choices of outcome variable y, using standard computational tools from this literature.[8] Before turning to the 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 very simple linear OLS regressions. As such, all of our conclusions are subject to the usual limitations of such a model. 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 very 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 multiple outcomes of interest.

[7] 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.

[8] Implementing BMA in principle poses major computational problems, as the number of models to be estimated and averaged increases in the number of explanatory variables at the rate 2^K. Fortunately, fast and accurate algorithms for identifying and sampling only those models with 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 MC3 sampler (Madigan and York 1995, 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 100,000 discarded drawings and 300,000 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 very 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.

4. Results

Our main findings are reported in Table 3. The rows of this table correspond to the 41 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 0.83, indicating that the joint posterior probability of the set of models in which this variable appears is 83 percent. The same is true for the "Time" component of "Enforcing Contracts", which has a PIP of 0.83 as well. A few other variables also have fairly high PIPs, including "Firing Costs" under "Employing Workers (at 0.79) and the "Profit Tax" variable under "Paying Taxes" (at 0.74).

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 1.3 percent, 1.0 percent, and 0.9 percent respectively. These low probabilities for even the most likely individual models highlight the importance of averaging inferences across 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 237 models are required to cover 50 percent of the posterior probability. In contrast, more than three times as many (923 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 0.97 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 41 policy indicators, as ranked by their PIPs, for each outcome variable. This allows us to identify at a glance the relatively 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 variables with the highest PIPs as the most "important" for a given outcome variable.

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 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 0.09. 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 0.09, or about one-tenth of its potential range.[9] We note also that the ranking of variables by their PIPs is very 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,

[9] 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. This is the expected impact of this particular policy indicator on this particular outcome variable, averaging across all models. Considering only models in which this policy indicator appears, the expected impact is slightly larger at 0.09/0.828.

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 outcome variables in the set of policy indicators that BMA identifies as important. To take a specific example, consider the EIU and CPIA outcome variables, which as noted earlier have a pairwise correlation of 0.78. Yet despite this very 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 20th 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.

More systematically, looking through Table 3, not one of the 41 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 two other policy indicators are in the "top ten" set for five out of seven outcomes (the "Import Time" component of "Trading Across Borders" and the “Legal Rights” component of “Getting Credit”). In contrast, 31 out of 41 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.

We perform a simple non-parametric test in order to provide a more formal method of documenting this key instability finding. The null hypothesis is that the event that a policy indicator is included in the "top ten" list for a given outcome variable is independent of the event that it is in the "top ten" list for any other outcome variable. This would correspond to the extreme case of no stability at all across outcome variables in terms of which policy indicators are identified as important by the BMA procedure. Since by construction there is a 10/41 chance of a given policy indicator being included in the "top ten" set, 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 10/41 under the null hypothesis. We can then compare the predicted proportions from this distribution with the observed proportions, using a standard chi-squared test.[10] The p-value for the null hypothesis of independence across outcomes is 0.99, suggesting we cannot reject even this extreme version of instability across outcome variables.[11]

In the online appendix 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 very unstable across outcome variables. Specifically, we again cannot reject the null hypothesis that the event of inclusion in the set of "important" policy indicators for one outcome variable is statistically independent of the event of inclusion in the set of "important" indicators for other outcomes. This is true despite the fact that our outcome measures are all conceptually similar and fairly strongly-correlated measures of perceptions of corruption.

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 very 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 very difficult to use this type of data and analysis to narrow down the set of potential reforms to a small set of policy indicators that matter systematically across outcomes.

---

[10] In particular the sum of the squared deviations between expected and observed proportions, normalized by expected proportions, will be a chi-squared random variable with 6 degrees of freedom.

[11] It is interesting to note that, by the standard of this test, the results in Ciccone and Jarocinski (2010) show a similar degree of instability across alternative choices of dependent variable. Their Table 2 reports PIPs for 67 candidate growth determinants, using economic growth data based on three revisions of the Penn World Tables as dependent variables. The pairwise correlations of growth from these three sources range from 0.93 to 0.97 (their Appendix Table B1). Performing the same chi-squared test in their application, for membership in the set of top-10 growth determinants, delivers a p-value of 0.72.

5. Robustness

Our key finding of interest is that there is a surprising degree 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 a range of potential explanations for this instability finding, including the extent to which our results are driven by differences in sample size across outcome indicators; the choice of the PIP cutoff used to identify the set of important correlates of outcomes; the specification of the prior distribution in the BMA analysis; the role of near-collinearity among the policy indicators; whether our results simply reflect low correlations across outcome variables; and the possibility that the true models linking policy indicators to outcomes are in reality very different across outcome variables.

5.1 Differences in Country Samples

A first very pedestrian potential explanation for our instability result across outcomes 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. In this dataset, we find the same pattern of strong instability across outcomes. The p-value for the null hypothesis that the event of inclusion in the set of "important" regressors is independent across outcomes is 0.96 in the common set of countries. This suggests to us that our instability across outcomes finding is not primarily due differences in country samples.

5.2 Choice of Cutoff PIP for Defining "Important" Policy Indicators

A second straightforward potential concern is that our finding on instability across outcomes may be driven by 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 very low PIPs, it would not be very surprising to find a lot of instability across outcome variables. We explore this possibility by simply taking alternative thresholds for defining the set of "important" policy indicators. If, for example, we focus on the "top five" rather than "top ten" policy indicators, the p-value for the null hypothesis of independence across outcome variables falls slightly to 0.94. For the set of "top two" policy indicators, the p-value for the null of independence is still 0.72. We conclude from this that, short of focusing on a very small set of indicators for each outcome, membership in the set of "important" indicators is quite unstable across outcome variables.[12]

5.3 Sensitivity to Choice of Prior

We next consider the extent to which our conclusions may be driven by our choice of the 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”. [13] 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 very large changes in posterior model probabilities and PIPs as we move from one dependent variable to another, if the "supermodel" is different for the different outcomes.

There are, however, two reasons why the concentration of posterior probabilities is not driving our instability finding. First, recall that we have focused throughout on the *ordering* of right-hand-side variables by their PIPs, rather than the magnitudes of the PIPs themselves. While the latter may be very sensitive to the choice of g, the former are less so. Second, we have also reproduced the BMA analysis, but using a hyperprior distribution for the prior parameter g, as advocated by Liang et. al. (2008) and

---

[12] We reject the null hypothesis of independence across outcomes only if we focus only on the extreme case of considering just the single best-performing policy indicator for each outcome. This reflects the fact that one variable (the "Legal Rights" component of "Getting Credit") appears as the highest-PIP variable for four of the seven outcome variables. Under the null hypothesis of independence, getting four "successes" with a success probability of only 1/41 is extremely unlikely, and this drives the rejection of the null hypothesis.

[13] 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).

Feldkircher and Zeugner (2009). This effectively averages all results across a range of alternative values of g, and so smoothes out any effects of g on inferences. When we do this, we find that the rank ordering of variables by their PIPs is nearly identical. This in turn means that our finding of instability across outcome variables is also preserved when we use the hyperprior for g. Based on this evidence, we conclude that our instability finding is probably not driven by the particular choice of the prior parameter g.

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 very 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 very surprising that the data are not very informative about the relative importance of policy indicators across various outcome variables.

There are, however, two main reasons to discount this objection. The first is conceptual, and reflects the criteria used by BMA to assign posterior probabilities to models. Recall from Equation (6) that, for our choice of a small value for the prior parameter g, there is a strong tradeoff between model size and R-squared when calculating posterior probabilities for individual models. This helps to ensure that models with near-collinear regressors are assigned low posterior probabilities, since adding a near-collinear regressor to a model will result in only a very small improvement in R-squared.

The second is simply that the policy indicators in the Doing Business dataset are in fact surprisingly uncorrelated with each other. Consider for example the pairwise correlations between the 41 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, 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 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 0.29 to 0.61, 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 very surprising to find that the set of important regressors for each outcome is very 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 1. 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.

To us the more striking feature of Figure 1 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 0.75 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 very 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?

Yet another objection to our finding is that the world is complicated, in the sense that the mapping from policy indicators to outcomes is in reality very different across the different outcome variables we use. If this were the case, it would not be very surprising to find that the sets of policy indicators identified by BMA as important correlates of outcomes are very different across outcomes. For example, while all seven outcome variables broadly reflect the business community's perceptions of the quality of the business environment, we have seen in Table 1 that there are differences across these data sources in terms of the specific dimensions of the business environment that they consider. If these specific dimensions of the business environment happen to respond to different sets of policy indicators, then this could potentially account for our finding that the set of important regressors is very different across outcome variables.

There are however two reasons to discount this objection to our findings. The first is that this interpretation is hard to reconcile with the fact 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 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.[14] 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

[14] 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.

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.

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 very correlated across countries, 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. The difficulty however is that the available data and techniques used in this paper cannot do a very good job of identifying the necessary overlaps in the sets of policy indicators that are relevant for the various outcome variables of interest.

A second reason to discount this objection comes from considering a very specific empirical example. The concern in this subsection is that the different outcome variables might simply reflect very different aspects of the business regulatory environment. If these different aspects of the regulatory environment depend on different policy indicators, then our instability finding would not be very surprising. To address this concern, we would ideally like to control for any definitional differences across outcome variables. 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. [15]

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. This exercise is useful because it controls for the potential problem of differences in evaluation criteria across our different outcome variable, since the two CPIA assessments are based on responses to identical questionnaires, and moreover but quite similar sets of respondents (country economists at multilateral development banks).

The results are reported in Table 4, which has the same structure as Table 3 for our main results. The striking feature of Table 4 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 very high correlation between the two CPIA outcome variables of 0.81, only four 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 third 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 17th most important policy indicator. More formally, the p-value for the null hypothesis that membership in the set of top ten important policy indicators for the World Bank CPIA is independent of membership in the same set for the African Development Bank CPIA is 0.84.[16] This example suggests that, even when the two outcome variables are very 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.

---

[15] The full CPIA for the African Development Bank is available at http://cpia.afdb.org.

[16] This p-value does however have to be taken with a grain of salt. Since there are only two binomial "trials" corresponding to the two outcome variables, the simple chi-squared test we are performing does not have strong power against the alternative hypothesis that the sets of important regressors are correlated across outcomes. For example, even if we were to have observed the BMA procedure identifying exactly the same set of 10 important regressors for the two outcomes, the p-value for the null hypothesis of independence would have been 0.32. In contrast, in our main results in Section 4, the chi-squared test is much more powerful because there are seven "trials" corresponding to the seven outcome variables.

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 41 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 very 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 very unclear how one might find instruments or other sources of exogenous variation in the very many different dimensions of regulatory policy and governance measured by Doing Business. The assumption of linearity is also very 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 more possible combinations of interactions between them. 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. Rather, we simply note that it seems extremely difficult to quantify the partial effects of these many indicators on relevant outcomes. This in turn illustrates why it may be very 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. and Robinson, J.A. (2001). 'The Colonial Origins of Comparative Development: An Empirical Investigation', *American Economic Review*, vol. 91 (5), pp. 1369-1401.

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

Candes, E. and Tao, T. (2005). 'The Dantzig selector: statistical when *p* is much larger than *n*', Manuscript, California Institute of Technology.

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

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

Eicher, T.S., Lenkoski, A. and Raftery, A.E. (2009). 'Bayesian Model Averaging and Endogeneity Under Model Uncertainty: An Application to Development Determinants', Manuscript, University of Washington.

Eicher, T.S., Papageorgiou, C. and Raftery, A.E. (2011). 'Default Priors and Predictive Performance in Bayesian Model Averaging, with Application to Growth Determinants', *Journal of Applied Econometrics*, Vol. 26 (1), pp. 30-55.

Eicher, T.S., Papageorgiou, C. and Roehn, O. (2007). 'Unraveling the Fortunes of the Fortunate: An Iterative Bayesian Model Averaging (IBMA) Approach'. *Journal of Macroeconomics*, vol. 29, pp. 494-514.

Feldkircher, M. and 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. and Zeugner,S. (2010). 'Data Revisions Revisited: A Comment on 'Determinants of Economic Growth: Will the Data Tell?'', University of Salzburg Working Paper No. 2010-16.

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

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

George, E.I. and McCulloch, R.E. (1997). 'Approaches for Bayesian Variable Selection', *Statistica Sinica*, vol. 7, pp. 339-373.

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

Hendry, D. F. and Krolzig, H.-M. (2005). 'The Properties of Automatic *Gets* Modelling', *The Economic Journal*, vol. 115, pp. C32-C61.

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

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

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

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

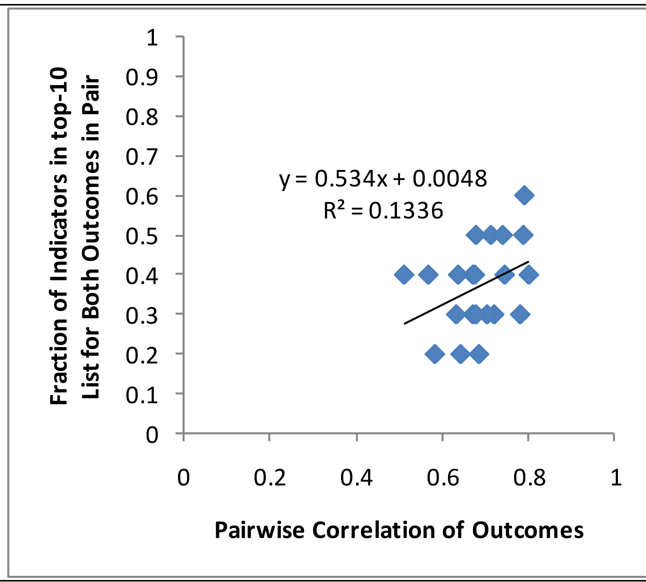
Madigan, D. and York, J. (1995). 'Bayesian Graphical Models for Discrete data', *International Statistical Review*. Vol. 63, pp. 215-232.

Mauro, P. (1995). 'Corruption and Growth', *Quarterly Journal of Economics*, vol. CX (3), pp. 681–712.

Moral-Benito, E. (2009). 'Determinants of Economic Growth: A Bayesian Panel Data Approach', World Bank Policy Research Working Paper, No. 4830.

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

Figure 1: 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 DB-41 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	
Global Insight Global Risk Service (DRI)	
(Average of)	Export Regulation
	Import Regulation
	Other Regulatory Burdens
	Restrictions on Foreign Business Ownership
	Restrictions on Foreign Equity Ownership
Economist Intelligence Unit (EIU)	
(Average of)	Unfair competitive practices
	Price controls
	Discriminatory tariffs
	Excessive protections
	Discriminatory taxes
Merchant International Group Gray Area Dynamics (GAD)	
(Average of)	Stock Exchange / Capital Markets
	Foreign Investment
Global Competitiveness Report Executive Opinion Survey (GCS)	
(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)	
	Risk to operations from contract viability, expropriation, repatriation and payment delays.
Global Insight Business Conditions and Risk Indicators (WMO)	
(Average of)	Efficiency of Tax Collection
	Business Legislation Complete and Compatible
World Bank Country Policy and Institutional Assessments (CPIA)	
(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

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

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	0.042	-0.001	0.007	0.255	0.026	0.050	0.346	0.033	0.050	0.431	0.025	0.032	0.113	0.010	0.034	0.035	0.001	0.007	0.111	0.007	0.025
Number of Days	0.088	0.004	0.018	0.044	0.000	0.013	0.360	0.036	0.055	0.286	0.016	0.029	0.040	0.001	0.014	0.063	0.003	0.014	0.840	0.096	0.053
Cost	0.089	0.005	0.020	0.117	0.012	0.040	0.105	0.009	0.033	0.048	0.001	0.008	0.379	0.068	0.099	0.835	0.126	0.072	0.194	0.016	0.038
Minimum Capital Requirement	0.095	-0.003	0.012	0.047	-0.001	0.010	0.040	0.000	0.007	0.045	0.001	0.005	0.040	-0.001	0.009	0.043	-0.001	0.007	0.062	-0.002	0.009
2. Construction Permits																					
Number of Procedures	0.210	-0.012	0.027	0.047	0.002	0.013	0.076	0.003	0.016	0.040	0.000	0.005	0.044	0.002	0.013	0.039	0.001	0.007	0.060	0.002	0.011
Number of Days	0.567	0.051	0.051	0.038	0.001	0.012	0.046	-0.001	0.011	0.597	0.039	0.037	0.131	0.013	0.038	0.033	0.000	0.007	0.041	0.001	0.008
Cost	0.038	0.000	0.007	0.304	0.037	0.064	0.129	0.010	0.030	0.209	0.010	0.022	0.213	0.027	0.060	0.971	0.141	0.045	0.228	0.016	0.034
3. Employing Workers																					
Difficulty of Hiring	0.122	-0.005	0.016	0.041	-0.001	0.011	0.045	-0.001	0.009	0.092	0.002	0.010	0.042	0.001	0.011	0.078	-0.003	0.013	0.109	-0.005	0.017
Rigidity of Hours	0.042	0.000	0.009	0.060	-0.003	0.018	0.069	-0.003	0.016	0.160	0.007	0.019	0.035	0.001	0.010	0.067	-0.003	0.015	0.048	-0.001	0.010
Difficult of Firing	0.142	0.008	0.024	0.235	0.021	0.044	0.040	0.001	0.008	0.042	0.000	0.005	0.042	0.001	0.012	0.039	-0.001	0.008	0.131	0.007	0.023
Rigidity of Employment	0.084	-0.005	0.023	0.045	-0.001	0.018	0.056	-0.002	0.015	0.153	0.007	0.019	0.046	0.002	0.015	0.108	-0.006	0.022	0.063	-0.003	0.017
Cost of Firing	0.794	0.067	0.042	0.306	0.031	0.052	0.120	0.007	0.024	0.095	0.003	0.011	0.768	0.119	0.079	0.475	0.040	0.048	0.219	0.015	0.032
4. Registering Property																					
Number of Procedures	0.026	0.000	0.005	0.060	0.003	0.015	0.125	0.007	0.023	0.716	0.043	0.032	0.036	0.001	0.009	0.112	0.005	0.018	0.400	0.028	0.039
Number of Days	0.050	-0.001	0.008	0.038	0.001	0.010	0.113	0.007	0.023	0.076	0.002	0.010	0.046	-0.002	0.014	0.064	0.002	0.013	0.047	0.001	0.009
Cost	0.045	0.001	0.008	0.045	0.002	0.013	0.030	0.000	0.007	0.041	-0.001	0.005	0.040	0.001	0.012	0.040	-0.001	0.009	0.304	0.023	0.040
5. Getting Credit																					
Legal Rights	0.828	0.090	0.051	0.088	0.006	0.026	1.000	0.211	0.041	0.268	0.014	0.026	0.086	0.007	0.029	0.971	0.138	0.044	0.892	0.098	0.046
Credit Information	0.041	0.001	0.007	0.105	0.007	0.024	0.065	0.003	0.014	0.167	-0.006	0.016	0.091	0.006	0.025	0.032	0.000	0.006	0.798	0.078	0.048
Credit Registry Coverage	0.044	-0.001	0.012	0.038	-0.001	0.017	0.040	-0.002	0.017	0.038	0.000	0.007	0.035	0.000	0.017	0.032	0.000	0.011	0.043	-0.001	0.017
Credit Bureau Coverage	0.037	-0.001	0.012	0.130	0.020	0.061	0.387	0.071	0.101	0.047	-0.002	0.012	0.033	0.001	0.018	0.096	0.010	0.038	0.093	0.012	0.047
6. Protecting Investors																					
Disclosure	0.036	0.000	0.005	0.032	0.000	0.008	0.035	0.000	0.007	0.055	-0.001	0.007	0.046	-0.001	0.016	0.038	0.000	0.007	0.042	0.001	0.008
Director Liability	0.055	0.001	0.009	0.040	0.001	0.010	0.036	0.001	0.008	0.045	0.001	0.006	0.049	0.002	0.017	0.038	0.000	0.007	0.048	0.001	0.010
Shareholder Suits	0.049	0.001	0.009	0.039	-0.001	0.011	0.037	0.001	0.008	0.096	-0.003	0.013	0.189	0.021	0.051	0.038	0.001	0.008	0.035	0.000	0.006
Investor Proection	0.051	0.002	0.010	0.036	0.001	0.012	0.039	0.001	0.010	0.089	-0.003	0.014	0.363	0.053	0.079	0.038	0.001	0.010	0.104	0.006	0.024
7. Paying Taxes																					
Number of Payments	0.041	0.000	0.006	0.582	0.071	0.070	0.963	0.144	0.049	0.071	0.002	0.009	0.473	0.066	0.079	0.950	0.123	0.045	0.039	0.000	0.007
Number of Days	0.043	0.001	0.007	0.191	0.017	0.040	0.043	-0.001	0.009	0.989	0.091	0.024	0.168	0.018	0.046	0.144	0.009	0.027	0.041	0.001	0.008
Profit Tax	0.740	-0.068	0.050	0.058	-0.003	0.015	0.032	0.000	0.007	0.057	0.001	0.008	0.035	-0.001	0.010	0.037	-0.001	0.007	0.037	0.000	0.007
Labour Tax	0.221	-0.020	0.043	0.038	0.000	0.010	0.444	-0.042	0.054	0.047	-0.001	0.007	0.043	0.000	0.014	0.088	-0.005	0.020	0.072	-0.004	0.018
Other Tax	0.372	0.025	0.037	0.079	0.005	0.022	0.035	0.000	0.008	0.037	0.000	0.005	0.283	0.033	0.059	0.047	-0.001	0.009	0.038	0.001	0.008
Total Tax	0.316	0.029	0.048	0.077	0.004	0.020	0.041	0.000	0.011	0.038	0.000	0.005	0.077	0.006	0.025	0.059	-0.002	0.012	0.134	0.008	0.026

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	0.161	0.011	0.029	0.411	0.054	0.073	0.153	0.013	0.036	0.031	0.000	0.005	0.080	0.007	0.029	0.918	0.125	0.058	0.518	0.047	0.051
Export: Number of Days	0.174	0.015	0.042	0.089	0.006	0.047	0.469	0.071	0.084	0.102	0.005	0.018	0.122	0.014	0.056	0.947	0.199	0.073	0.115	0.012	0.041
Export: Cost	0.252	0.023	0.048	0.101	-0.004	0.052	0.048	-0.002	0.013	0.750	0.092	0.058	0.080	0.006	0.029	0.192	-0.014	0.034	0.067	0.003	0.017
Import: Number of Documents	0.065	-0.003	0.019	0.054	0.002	0.023	0.312	0.037	0.060	0.038	0.000	0.006	0.037	-0.001	0.017	0.180	-0.019	0.047	0.248	0.024	0.048
Import: Number of Days	0.690	0.098	0.078	0.777	0.199	0.130	0.189	0.023	0.056	0.056	0.002	0.013	0.611	0.134	0.122	0.071	0.008	0.040	0.734	0.114	0.081
Import: Cost	0.646	0.078	0.067	0.461	0.077	0.101	0.041	-0.001	0.012	0.279	0.031	0.054	0.138	0.015	0.045	0.189	-0.015	0.035	0.185	0.015	0.036
9. Enforcing Contracts																					
Number of Procedures	0.084	-0.004	0.017	0.376	0.043	0.062	0.159	0.012	0.032	0.103	0.004	0.013	0.043	-0.002	0.014	0.037	0.001	0.008	0.077	0.004	0.016
Number of Days	0.826	-0.080	0.046	0.033	0.000	0.009	0.049	-0.002	0.012	0.063	0.002	0.009	0.084	-0.006	0.025	0.042	0.000	0.008	0.044	0.001	0.008
Cost	0.035	0.000	0.006	0.112	-0.009	0.031	0.031	0.000	0.007	0.031	0.000	0.005	0.047	-0.002	0.016	0.072	0.004	0.017	0.045	-0.001	0.009
10. Closing a Business																					
Number of Days	0.038	0.000	0.006	0.296	0.039	0.068	0.043	-0.001	0.015	0.044	0.000	0.010	0.122	-0.015	0.052	0.108	-0.008	0.030	0.064	-0.003	0.020
Cost	0.035	0.000	0.006	0.094	0.008	0.030	0.074	-0.004	0.019	0.068	0.002	0.011	0.040	-0.001	0.015	0.057	0.002	0.015	0.243	0.017	0.033
Recovery Rate	0.040	0.001	0.009	0.630	0.118	0.103	0.994	0.227	0.056	0.987	0.125	0.031	0.577	0.127	0.132	0.899	0.126	0.063	0.259	0.024	0.049
Posterior Probability of:																					
First-best model	0.013			0.012			0.014			0.018			0.022			0.042			0.022		
Second-best model	0.010			0.006			0.013			0.017			0.012			0.035			0.018		
Third-best model	0.009			0.006			0.013			0.015			0.011			0.020			0.008		
Posterior Mean Model Size	8.352			6.646			7.460			7.625			5.966			9.320			9.320		
Number of Models Visited	33551			44824			28256			30607			36099			22212			42196		
Number of Models Covering																					
x% of Posterior Probability																					
x=50%	554			923			456			452			585			237			656		
x=75%	2130			3407			2004			2017			2476			1319			2655		
x=90%	5438			7892			5171			5175			6123			3643			6617		
Corr(PMP)	0.951			0.910			0.971			0.967			0.961			0.990			0.955		
G&M Measure of Probability																					
Mass Visited	0.540			0.484			0.620			0.585			0.539			0.681			0.481		
Number of Observations	137			144			158			130			133			178			139		
Maximal Partial R-Squared for top 300 Models																					
Median	0.517			0.508			0.476			0.287			0.438			0.610			0.353		
90th Percentile	0.606			0.643			0.613			0.509			0.542			0.712			0.492		
Maximum	0.719			0.903			0.769			0.899			0.878			0.898			0.843		

Table 4: BMA Results for the World Bank and African Development Bank CPIA Ratings

Dependent Variable=	World Bank CPIA			African Dev Bank CPIA		
	PIP	Mean	SD	PIP	Mean	SD
1. Starting a Business						
Number of Procedures	0.108	0.009	0.032	0.064	0.005	0.030
Number of Days	0.293	0.033	0.058	0.945	0.203	0.075
Cost	0.086	0.011	0.046	0.110	0.016	0.058
Minimum Capital Requirement	0.043	0.001	0.009	0.123	0.008	0.024
2. Construction Permits						
Number of Procedures	0.081	0.006	0.027	0.153	0.016	0.043
Number of Days	0.441	0.056	0.071	0.092	0.008	0.030
Cost	0.039	0.001	0.014	0.096	-0.008	0.033
3. Employing Workers						
Difficulty of Hiring	0.065	-0.003	0.018	0.034	0.000	0.010
Rigidity of Hours	0.034	0.000	0.012	0.041	0.002	0.016
Difficult of Firing	0.047	0.001	0.017	0.058	-0.003	0.022
Rigidity of Employment	0.047	-0.001	0.019	0.043	0.000	0.015
Cost of Firing	0.084	0.007	0.029	0.074	0.007	0.031
4. Registering Property						
Number of Procedures	0.142	0.011	0.033	0.100	0.008	0.028
Number of Days	0.223	-0.026	0.056	0.049	-0.003	0.020
Cost	0.077	0.006	0.027	0.115	0.012	0.041
5. Getting Credit						
Legal Rights	0.045	0.001	0.016	0.101	0.011	0.041
Credit Information	0.044	0.001	0.017	0.136	0.015	0.047
Credit Registry Coverage	0.240	-0.066	0.135	0.310	-0.104	0.177
Credit Bureau Coverage	0.051	0.009	0.071	0.272	0.149	0.278
6. Protecting Investors						
Disclosure	0.032	0.000	0.011	0.041	0.001	0.015
Director Liability	0.036	0.000	0.010	0.037	0.001	0.011
Shareholder Suits	0.109	0.009	0.031	0.096	0.008	0.030
Investor Proection	0.047	0.002	0.016	0.055	0.003	0.020
7. Paying Taxes						
Number of Payments	0.083	-0.008	0.033	0.043	-0.002	0.019
Number of Days	0.063	-0.004	0.022	0.039	-0.001	0.014
Profit Tax	0.052	-0.002	0.014	0.065	-0.004	0.018
Labour Tax	0.151	-0.015	0.041	0.072	-0.006	0.026
Other Tax	0.094	0.008	0.033	0.111	0.010	0.035
Total Tax	0.678	0.107	0.086	0.171	0.019	0.050

Dependent Variable=	World Bank CPIA			African Dev Bank CPIA		
	PIP	Mean	SD	PIP	Mean	SD
8. Trading Across Borders						
Export: Number of Documents	0.067	0.004	0.019	0.060	0.004	0.019
Export: Number of Days	0.117	0.013	0.046	0.037	0.000	0.016
Export: Cost	0.152	0.014	0.040	0.043	0.002	0.017
Import: Number of Documents	0.138	0.014	0.042	0.050	0.003	0.018
Import: Number of Days	0.141	0.017	0.051	0.042	0.001	0.016
Import: Cost	0.133	0.012	0.036	0.055	0.003	0.018
9. Enforcing Contracts						
Number of Procedures	0.064	0.004	0.023	0.157	0.018	0.048
Number of Days	0.036	0.000	0.010	0.041	0.000	0.012
Cost	0.141	-0.013	0.039	0.090	-0.007	0.031
10. Closing a Business						
Number of Days	0.067	-0.001	0.028	0.079	0.003	0.036
Cost	0.122	0.010	0.035	0.052	0.000	0.022
Recovery Rate	0.462	0.073	0.091	0.791	0.172	0.107
Posterior Probability of:						
First-best model	0.019			0.114		
Second-best model	0.019			0.044		
Third-best model	0.018			0.028		
Posterior Mean Model Size	5.177			5.142		
Number of Models Visited	78581			63241		
Number of Models Covering						
x% of Posterior Probability						
x=50%	65			35		
x=75%	150			110		
x=90%	232			203		
Corr(PMP)	0.957			0.995		
G&M Measure of Probability						
Mass Visited	0.152			0.290		
Number of Observations	51			51		