

The Tyranny of International Index Rankings*

Bjørn Høyland

Karl Moene

Fredrik Willumsen[†]

First version: April, 2009

This version: July, 2009

Abstract

International index rankings are popular, but perhaps too persuasive. They emphasize country differences where similarity is the dominant feature. The rankings of the Human Development Index, Freedom House, and Doing Business can be misleading, not because of wrong indicators, but because the estimation of the scores ignores inherent uncertainty. Re-estimated with a method that captures this uncertainty, it becomes clear that the practice of ranking every adjacent country is a rather courageous activity.

Keywords: International indexes, ranking, latent variables, Bayesian estimation

JEL Codes: C43, O15, H77

“Looking-glass upon the wall, Who is fairest of us all?”

1 Introduction

One can hardly open a newspaper without finding a reference to an international country ranking. The appeal of the reported rankings lies in their simplicity. They provide an instant idea of the success of countries relative to others. Their users need no more statistical knowledge than readers of sports news. Just as football teams are ranked according to their performance, countries are ranked according to their ability to provide a high standard of living, democratic rights, and an appealing business environment. It

*We thank Ottar Hellevik, Marie Ingvaldsen, Carl Henrik Knutsen, Oddbjørn Knutsen, Jo Thori Lind, Halvor Mehlum, and seminar participants at PRIO and Department of Political Science, University of Oslo, for useful comments. While carrying out this research, we have been associated with the ESOP centre at the Department of Economics, University of Oslo. ESOP is supported by The Research Council of Norway.

[†]Høyland: Department of Political Science, University of Oslo. Moene and Willumsen: Department of Economics, University of Oslo. E-mail addresses: `bjorn.hoyland@stv.uio.no`, `k.o.moene@econ.uio.no`, `f.h.willumsen@econ.uio.no`.

seems like we are blessed with an excellent tool that everybody can understand and that everybody can use for a wide range of purposes.

But are we? Each ranking appears precise and it often ends up emphasizing differences where similarity is the dominant feature. To demonstrate the magnitudes of the inherent uncertainty in the rankings we single out three indexes for scrutiny: the United Nations' *Human Development Index*, the Freedom House's *Freedom in the World*, and the World Bank's *Doing Business Index*. The three are chosen not because they are constructed in a particularly bad manner, but because they cover the most important aspects of social, political, and economic life, and because they are popular and widely noticed. The Human Development Index provides the most famous ranking; it ranks countries according to health, knowledge, and material resources. Doing Business is the most read ranking; it ranks countries on the basis of how 'business-friendly' the regulatory environment is. Freedom House produces the most cited international regime categorization; it creates an annual index of the political rights and civil liberties enjoyed in different countries.

There are good reasons to applaud the data-collecting efforts behind the production of these indexes. The data are highly valuable both to policy makers and researchers alike. Our criticism is only related to how the data are summarized in a one-number-per-country fashion, as this practice can be highly misleading when the inherent uncertainty in this one number is not reported. We substantiate this criticism by estimating the uncertainty contained in the rankings using a Bayesian latent variable approach, where each of the indicators are considered as signals of the underlying performance level.

Why is it so important to incorporate uncertainty? First, by appearing precise and certain, the rankings seem so persuasive that they cannot be ignored. A particularly favorable or unfavorable position is likely to be widely noticed, and governments stand to lose by not commenting upon them. To attack an index is never appropriate for politicians. When the ranking is unfavorable, an attack would just make things worse; when the ranking is favorable, the praise is too tempting.

In Norway, for instance, leading politicians regularly insist that the United Nations has chosen Norway as the best country to live in—based on its position in the Human Development Index (at least prior to 2007). To give an indication of the hidden uncertainty, let us consider Iceland, Norway, Canada, and Australia, denoted INCA. The four countries were the top 4 countries in 2007. The observed performance indicators, we suggest, are uncertain signals of the underlying quality. Given the results in section 5, it is easy to calculate that the four are more likely to belong to the top 10 human development performers than not. The index, however, ranks them as top 4 in the order INCA in 2007, conveying the picture that Iceland (prior to the financial crisis) is best, and thus clearly better than Norway, which is clearly better than Canada, and so on.

How persuasive is this ranking compared to the information, neither conveyed nor discussed, that the four countries have about the same chances of being among the top

10 and that none of them have a higher chance than 3/4 of being in this group? The probability that INCA is the “true” top rank must then be extremely low. Just the probability that all four are among the top 10 is less than 1/3, and if all sequences of the four countries are almost equally likely, the probability that INCA is the true top rank must be below one percent.¹ As we show below, this imprecision is shared with the rankings of Doing Business and Freedom House.

Second, winners of a close race often attract the most attention and the rankings that in reality are the least certain are therefore considered the most exciting. Rankings may provide clear-cut orderings of countries that in fact are difficult to tell apart. By incorporating uncertainty, however, it becomes clear that each of the rankings draws attention to one specific, but rather arbitrary, ordering. The popularity of the indexes may simply reflect this weakness as people seem obsessed by differences when everything is more or less on an even level. The most exciting index rankings may therefore be the most uncertain, and the attention that such rankings receive in the international community is often inversely related to their accuracy. Media, policy makers and researchers often end up discussing the deep causes of a slight alteration in the internal rankings when there really is no significant change at all.

Third, precisely reported but inherently uncertain rankings may invite what we denote rank-seeking behavior. As governments may design policies more to improve their rank on the index than to improve their real performance, this is a highly relevant problem. Aspirations for rank-seeking are announced as when the director-general of the Malaysian Industrial Development Authority insists that “Malaysia aims to move from the 24th to a top 10 position in the World Bank’s ‘Doing Business’ ranking list. We continue to ask ourselves what it will take to reach the top 10, and are we willing to do what it takes to get there.” (Asia in Focus, Jan. 8 2007). Similarly, Akylbek Zhaparov, Minister of the Economic Development in Kyrgyzstan, expresses a hope (in 2008) that his country “shall rank among top twenty countries in the Doing Business rating in three years”. Achievements from rank-seeking are also celebrated as when Macedonia placed a one-page advert in the Economist’s annual forecasting report “The World in 2008”, where the key message was that Macedonia had improved their position on the Doing Business index.² Such rank-seeking behavior is magnified by the gap between reported preciseness and the underlying uncertainty, as rank-seeking behavior is crucially dependent on a precise link between indicators and rankings and a fuzzy link between indicators and real performance.

¹Using the posterior distribution from Section 5, we find that the probability that INCA is the exact top ranking is 0, while the probability that $I > N > C > A$, but not necessarily occupying the four top places, is .037.

²Using the posterior distribution from Section 5, the 95 per cent credibility intervals for rankings is from 14th place to 60th place for Malaysia, from 67th to 119th place for Macedonia, and from 117th to 167th place for Kyrgyzstan (of 181 countries).

To provide the intuition of how we incorporate the uncertainty in the rankings, consider first how an index is normally constructed. The general idea is to start out from a set of observable country outcomes that one thinks are partial signals of the relevant performance of the countries vis-à-vis each other. Next, one can aggregate them into a number or an index for each country. Finally, one ranks countries according to this index. This is the approach that the three indexes follow. They do not, however, consider the important question as to how much uncertainty is carried over from *the noisy signals* of performance to the aggregate index score and the rankings based on that index score.^{3,4,5}

How can we observe the uncertainty in the signals? Imagine that we ranked every country on each of the available indicators for a particular index. If it were the case that we obtained the same ranking in all of these sub-rankings, then the indicators would simply be deterministic representations of the same underlying performance variable, and a ranking based on these would therefore be exact. If the rankings differed a lot across the sub-rankings, however, it would make more sense to think of the indicators as stochastic representations of the same underlying performance variable, and the level of variability across the rankings based on the sub-indicators would be a potential measure of the degree of the uncertainty contained in the signals of the underlying performance variable. This is the approach we pursue in this paper.

Our paper contributes to a small literature on estimating index scores. In a pioneering contribution Kaufmann, Kraay and Zoido-Lobaton (1999) estimate “Governance” scores from overlapping and non-overlapping sub-indexes using an unobserved components model. They identify two problems with their methodology: correlated disturbances and the assumed linearity of the unobserved components model. We address both of these problems by introducing group-specific shocks and allowing for more flexible measurements. Since we have a large number of indicators, we can allow for correlation between “similar” indicators that are due to correlation in error terms and not in the underlying latent variable, and since we use a very general Bayesian approach, we can also allow for different types of variables, continuous, ordinal, count, and censored outcomes, within the same estimation framework. Both group-specific shocks and flexible measurements are important, as the main point of this paper is to estimate the uncertainty bounds in

³For other weaknesses of relying on rankings see the overview in Søreide (2006).

⁴The indexes may or may not incorporate the most relevant information, but we will not enter into the quarrel over which variables should be included, since our critique of index rankings is orthogonal to the critique concerning which variables to include. There has been a great deal of discussions about this for all the indexes used in this paper; for the Human Development Index see for example Sagar and Najam (1998), for Freedom House see among others Munck and Verkuilen (2002), and for Doing Business see for example Lee, McCann and Torm (2008) and Hampel-Malagrosa and Frickenstein (2008). There are also problems with comparing these indexes over time, see for example Morse (2003).

⁵Some scholars distinguish between reflective and formative measurements (e.g. Bollen and Lennox (1991) and for an overview, Jarvis, Mackenzie and Podsakoff (2003)). While the direction of the link goes from the latent construct to the observable indicators in a reflective measure, the direction of the link goes the other way in a formative measure. We take the view that the indexes we consider here are of the reflective sort.

the index scores, which will be affected by either correlation in the error terms or model misspecification. Our paper also builds on Jackman (2004), who estimates the quality of graduate students based on different types of assessments with flexible measures, and especially to Treier and Jackman (2008), who find that the uncertainty in the Polity scores is both notable and may be detrimental for researchers using indexes as explanatory variables in regressions.

2 The problem: measuring unobserved performance

To avoid unnecessary abstractions we illustrate how to interpret the indexes by focusing on one of them, the Human Development Index (HDI). It is the simplest and most transparent index of the three, and, for good and bad, therefore gives a stark representation of the approach that we use.

Historically the Human Development Index was constructed as a response to the common practice of ranking countries solely on the basis of their GDP per capita. What matters, many thought, is not only national income, but also health, an ability to read and write, and longevity. The index thus emphasizes three different “aspects” of human development: a decent standard of living, knowledge, and a long and healthy life. How should we interpret the index?

2.1 Interpretation

In contrast to our approach, one may insist that the HDI actually *is* human development—or that human development *is* the HDI. The index makers may have approached the experts, one might imagine, who have told them that HDI is it. But the most prominent scholars on the topic, Anand and Sen (1994, p. 1), who also are leading proponents of human development, motivate the index quite differently when they claim that “[h]uman beings are the real end of all activities, and development must be centered on enhancing their achievements, freedom, and capabilities. It is the lives that they lead that is of intrinsic importance, not the commodities or incomes they happen to possess.” Accordingly, human development must be about how people can lead good lives and develop to their full potential. It is therefore related to self-worth, dignity, and creativity—none of which can be completely defined by the rise or fall of national incomes, life expectancy, and schooling. Clearly, the assertion ‘HDI equals human development’ cannot be a reasonable interpretation. The observed factors that enter the index, however, may of course be correlated with human development in one way or another.

One way that the observed factors can be correlated with human development is via their links to the immediate *causes* of good or bad performance. To visualize the underlying performance we might attempt to model the economic, social, and political

mechanisms that determine income, longevity, and schooling, and to base the comparisons across countries on a ranking of how human-development-friendly the underlying mechanisms are. In doing this one would of course have to incorporate mutual interdependencies of the three dimensions: Health and knowledge are inputs to the production of the material standard of living, a high material standard of living can help finance health and education, there is a political economy and a governance structure that affect all three dimensions, and so on.

A more feasible approach, the one that we follow in this paper, is to exploit that the indicators are *consequences* of good or bad policies, and that the basic point of the ranking based on the Human Development Index is to identify to what extent countries have, say, a policy orientation in favor human development. In fact, for the estimation we perform it is only important to agree that the HDI is a measure of some underlying quality, called ‘human development’. The precise interpretation of this quality may be controversial, but it is enough that it is a quality reflected primarily in income, longevity and schooling.

Taking this view on what the indexes measure amounts to viewing each indicator as an uncertain signal of the underlying quality that they all are partial observations of. Instead of just taking the average of the performance on the different indicators, as the Human Development Index does, we try to tap the information that the different indicators contain on the underlying variable in an optimal way.

In doing this we do not violate that human development has several dimensions. Just as the index collapses the three dimensions into one aggregate number for each country, with an obvious loss of information, we assert that the three dimensions in part are determined by the country’s ‘human development orientation’—and that this orientation is correlated with the actual human development of the country. The aggregation of the index can be justified by similar arguments as we apply when asserting that it in part is caused by the orientation and capacity of the country. This justification is even more accurate when we come to the Doing Business Index and Freedom House.

3 The approach: measuring latent variables

In this section we more technically demonstrate how we use the different indicators to trace out the underlying performance. We continue to use the Human Development Index as an example when discussing the model and the estimation. What we show holds true for the other indexes as well.

3.1 Model and identification

Let the true, unobserved performance of country i be denoted by f_i , and let the observable outcomes or indicators be denoted by x_{ij} , where j denotes one of several outcomes. It is reasonable to think of x_{ij} as a *signal* of the unobserved f_i , a signal which hopefully can be used to make inferences on the unobserved performance of country i (relative to country k).

Consider the following specification of how the true, unobserved performance is related to the observable indicators:

$$x_{ij} = \alpha_j + \beta_j f_i + \epsilon_{ij}, \quad \text{var}(\epsilon_{ij}) = \sigma_{\epsilon_j}^2, \quad (1)$$

Clearly, in this specification β_j is important; it measures index j 's ability to discriminate between countries with respect to the latent variable f_i . As we have specified it, the indicators observed for each country are related via the common underlying level of human development in the country, f_i . The goal is to use this relationship to try to tap the different observed indicators for the information they contain on f_i .

By looking at equation (1), we realize that everything that is on the right-hand side of the equation is unknown; although this looks like an equation readily set out for estimation via Ordinary Least Squares it is not, as the “explanatory variable” f_i is not observed. So how can we estimate α, β, σ^2 , and f using only information contained in x ?⁶ Intuitively, we are going to exploit the correlation between the different indicators across countries to estimate α, β and σ^2 , and use the within-country variation to estimate the different f_i 's. For the model to be identified, however, we first need to assume a structure of the underlying distribution of performance (the f_i 's) across countries. As only relative performance matters and this performance does not have any natural scale, we are free to normalize it as we wish. Thus, we assume that the performance is distributed with mean 0 and variance 1.

Since we exploit the correlation between the different indicators to estimate the weights that should be put on them in estimating the index, the β 's, the error terms in equation (1) must not be correlated. Correlation in the error terms on two or more indicators will make us put more weight on these indicators relative to the others, as the correlation will make us think that these indicators tap heavily from the underlying variable f_i , while in fact they only tap information from the correlated error term. This implies that it is important to have indicators that are unrelated to each other in other ways than through the underlying quality variable, f_i . By looking at the data we have at hand, this conditional independence assumption, that the observed indicators are unrelated once we

⁶ α, β, σ^2 denote the vectors of respectively $\alpha_j, \beta_j, \sigma_j^2$ for all j , f denotes the vector of f_i 's, x_i denotes the vector of the observable indicators for country i , and finally x denotes the matrix of the observed indicators over all countries.

condition on f_i , may fail to hold. One example is the Human Development Index that has three different classes of measures: schooling, life expectancy, and GDP per capita. While schooling is represented by two indicators (enrollment and literacy rates), the two other groups are represented by only one indicator each. Hence, by just using the four raw variables directly, we will not take into account that two of the variables may be more closely related than the others for reasons that are not contained in f_i . Fortunately, it is straightforward to take such grouping effects into account, by augmenting equation (1) with a group-specific error term $\delta_{i,k(j)}$:

$$x_{ij} = \alpha_j + \beta_j f_i + \delta_{i,k(j)} + \epsilon_{ij}, \quad \text{var}(\epsilon_{ij}) = \sigma_{\epsilon_j}^2, \quad \text{var}(\delta_{i,k(j)}) = \sigma_{\delta,k(j)}^2, \quad (2)$$

where δ is a group k -specific shock that creates correlation between the indicators j in group k for reasons other than f_i . As we estimate the variance of the group-specific shock, $\sigma_{\delta,k(j)}^2$, we of course still allow for the possibility that the variance of the group-specific shock is estimated to be 0. A variance of 0 implies no correlation in the error terms between the indicators within a cluster.

Is it possible to identify all the parameters in (2) with the data we have? If we have identification for the Human Development Index, we will also have identification for the other models. The reason for this is that the extra moments we get by including additional indicators more than make up for the extra parameters that need to be estimated when we add more indicators.

Again, the Human Development Index contains four measures of performance: school enrollment, literacy rate, GDP per capita, and life expectancy. Assume now that school enrollment and literacy rate are contained in the variables x_{i1} and x_{i2} , GDP per capita in x_{i3} , and life expectancy in x_{i4} . Exploiting the normalization $\text{var}(f) = 1$, the covariance matrix is

$$\text{cov}(x_i) = \begin{pmatrix} \beta_1^2 + \sigma_{\epsilon_1}^2 + \sigma_{\delta}^2 & & & & \\ \beta_2\beta_1 + \sigma_{\delta}^2 & \beta_2^2 + \sigma_{\epsilon_2}^2 + \sigma_{\delta}^2 & & & \\ \beta_3\beta_1 & \beta_3\beta_2 & \beta_3^2 + \sigma_{\epsilon_3}^2 & & \\ \beta_4\beta_1 & \beta_4\beta_2 & \beta_4\beta_3 & \beta_4^2 + \sigma_{\epsilon_4}^2 & \end{pmatrix}$$

where x_i is the vector consisting of the four variables in the Human Development Index and σ_{δ}^2 is the covariance among the error terms for the two schooling measures. We immediately see that by using the empirical covariance matrix, we are able to identify β , σ_{ϵ}^2 , and σ_{δ}^2 .⁷ From the mean of the variables we then identify α . Finally, using the identified parameters and x_i , we can identify f_i .⁸

⁷We have ten unique elements in the covariance matrix, and nine parameters ($\beta_1, \beta_2, \beta_3, \beta_4, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_{\delta}^2$). The Jacobian (the matrix of derivatives of all the unique elements of $\text{cov}(x)$ with respect to the parameter vector) is nonsingular, so the system is identified.

⁸It should be noted that we only have *local* identification, meaning that if both β and f are multiplied

The statistical approach just described can be contrasted to the one applied by the index makers. By using similar notation, one should ask how the index makers infer the unobserved performance from the observed indicators, i.e. how do they choose α_j and β_j ? The Human Development Index is constructed using the average of three development outcomes for each country: health, as measured by life expectancy; knowledge and skills, as measured by a weighted average of functional literacy and combined elementary and secondary net school-enrollment rates; and resources, as measured by the level of real per capita income. The formula for calculating the index is

$$f_i = \frac{1}{3} \sum_{j=1}^3 \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})}$$

To cast this formula in terms of the model in equation (1), the HDI parameters β_j and α_j for $j = 1, 2, 3$ are set as follows

$$\beta_j = \max x_j - \min x_j, \quad \alpha_j = \min x_j, \quad \sum_{j=1}^3 \frac{\epsilon_{ij}}{\max x_j - \min x_j} = 0,$$

where the max and the min in principle are the highest and lowest observed outcomes across all countries.⁹

First note that the three sub-indexes, and hence also the aggregate index, are bounded between 0 and 1. By normalizing the observed variables x_{ij} such that maximum value of each of the sub-indexes is 1 and the minimum value 0, it readily follows that the approach behind the Human Development Index is basically to set α_j equal to 0 and $\beta_j = 1$, i.e. what the Human Development Index implicitly assumes is that all the sub-indicators are equally good at discriminating between countries. Again, the approach we use is fundamentally different, as we try to estimate the discrimination parameters from the available data.

3.2 Flexible measurements

In the presentation of the model in the previous section, we assumed that the indicators were continuous variables. By looking at the indicators for the different indexes (the full list of indicators for all the indexes is given in Appendix B), we immediately realize that

by -1 the results will be the same. This is entirely unproblematic, as previously stated, the scale of the performance is arbitrary. We can therefore normalize it such that the best performers have positive values and the poorest performers have negative values, remembering that we have normalized f such that $\text{var}(f) = 1$ and $E(f) = 0$.

⁹The Human Development Index does not always use the observed max and min from the data, but rather relies on so-called “goalposts”, a theoretical max/min for the different indicators. It also uses the natural logarithm rather than the level of GDP per capita. For simplicity, we still refer to this variable as GDP per capita. For more information, see Technical Note 1 in HDR (2007).

this is not a valid assumption. Several of the indicators are defined as if they are censored. For example the literacy rate in the Human Development Index cannot exceed hundred percent. While some of the indicators are of an ordinal nature, such as the indicator for “Rule of law” in Freedom House, other indicators are counts, such as the number of documents needed for importing in Doing Business. It is important to take these properties of the indicators into account when we estimate the model, as a misspecification such as fitting a model suitable for continuous variables on an ordinal indicator may give rise to greater estimated uncertainty in the index scores and the rankings.

So how do we take this into account in the estimation? For the ordinal variables, we follow the approach in Jackman (2004), Treier and Jackman (2008), and Clinton and Lewis (2008), and estimate an ordered choice model. We still take into account within-group correlation for the different variables as described in section 3.1. For the variables that are censored, we modify the likelihood function to take this censoring into account, following Gelman and Hill (2006, pp. 404-5). Finally, the count data is modeled using a Poisson regression type model with overdispersion, following Gelman and Hill (2006, p. 382).

3.3 Estimation

The goal of the Bayesian estimation is to characterize the joint posterior probability distribution over all the parameters in the model given the observed indicators, i.e. to estimate $p(\Theta|x)$ where $\Theta = (\alpha, \beta, \sigma, f)$. This posterior probability distribution is, according to Bayes theorem, proportional to the product of the likelihood of the data given the parameters and the prior distribution over the parameters: $p(\Theta|x) \propto p(x|\Theta)p(\Theta)$. We use diffuse priors¹⁰ on all the parameters; the priors we use are given in Table 1 in Appendix A.2. By using diffuse priors, we put no a priori restrictions on the values that the parameters can or are likely to take, and the identification of the parameters therefore relies solely on the data through the likelihood function $p(x|\Theta)$. The likelihood function $p(x|\Theta)$ comes from the model we specify for how the observed indicators are linked to the underlying performance and the assumed distribution of the error terms. The model we estimate is, as previously stated, an augmented version of the model from Section 3.1; all details about the model are given in Appendix A. Again, the reason that we augment the model to take censoring, ordinal outcomes, and count data into account, is that the precision with which we can estimate the underlying quality scores will depend crucially on whether the model is misspecified or not, that is, whether the likelihood function we use is the correct one.

We use a Markov Chain Monte Carlo (MCMC) algorithm implemented in WinBUGS

¹⁰Also known as weakly informative priors, i.e. proper priors (as opposed to non-informative priors which are improper) with much uncertainty. The reason for not choosing fully non-informative priors is that we are using WinBUGS to estimate the model, and WinBUGS require proper priors.

(Lunn, Thomas, Best and Spiegelhalter, 2000) to calculate the posterior from the likelihood function and the priors. The MCMC algorithm, after it has converged, provides us with a random sample from the joint posterior distribution. This sample is what we use for statistical inferences, for example for constructing credibility intervals as in Figures 1, 4, and 6 and for testing hypotheses as in Figures 3, 5, and 7. For brevity, all computational details are relegated to Appendix A.

4 Data and definitions

Here we briefly describe the data; for more details on the variables and classification see Appendix B.

The Human Development Index

The Human Development Index contains four indicators: life expectancy at birth, school enrollment, literacy rates, and GDP per capita. School enrollment and literacy rates are on a 0-100 scale, while life expectancy is in years, and GDP per capita is in 2005 PPP US \$.

Freedom House

Freedom House collects information about political rights and civil liberties. The “Political Rights” pillar has three ordinal scaled indicators: the electoral process, political pluralism and participation, and the functioning of government. The “Civil Liberties” pillar has four ordinal scaled indicators: freedom of expression and belief, associational and organizational rights, rule of law, and personal autonomy and individual rights. All indicators are based on questionnaires.¹¹

Doing Business

The scores on the Doing Business indicators are based on a survey of international companies operating in the major economic center of a country. The indicators include information on a variety of aspects, from the degree of investor protection to the amount protection that workers enjoy.

More precisely, Doing Business collect indicators on 10 different stages of a business’s life: starting a business, dealing with licenses, employing workers, registering property, getting credit, protecting investors, paying taxes, trading across borders, enforcing contracts, and closing a business. Each of these sub-indexes has one or more indicators. The

¹¹Unfortunately, Freedom House does not release the underlying data material used to generate the indicators. This is unfortunate for us, as this information could have helped reduce the estimated uncertainty in our rankings.

sub-index for starting a business, for example, contains a variable counting the number of procedures required for starting a business, the number of days it takes to start a business, cost (as % of income per capita), and minimum capital requirement (as % of income per capita). The other sub-indexes contain similar variables; all the variables in the different groups are given in Appendix B.

5 Results: uncertain rankings

To see how well each of the indexes distinguishes among the countries, we estimate the underlying performance and its corresponding uncertainty, i.e. how much uncertainty is carried over from the noisy signals to index scores and rankings.

5.1 Rankings

The rankings are based on the index scores. As our critique of the indexes is on their focus on rankings, we first present the uncertainty in the rank positions for all countries, before illustrating the uncertainty in the rankings by looking at the probabilities that particular countries are among the top or bottom ten countries in the indexes (exploiting our Bayesian framework).¹² We start with the Human Development Index, followed by Freedom House and Doing Business.

The Human Development Index

The ranking based on the re-estimation of the index score is given in Figure 1. We see that the Human Development Index distinguishes well among the countries at the lower end of the ranking; the Democratic Republic of Congo and the Central African Republic are clearly doing worse than most other countries in terms of Human Development. As a device to distinguish between the 100 least developed countries in the world, the Human Development Index is quite successful.

At the other end of the scale, we find Australia, Canada, Norway, and Iceland. It should be noted that there are tiny differences, if any, between the countries at the top of the ranking.

Figure 2 shows the original Human Development Index ranking versus the re-estimated ranking. The median ranking according to our re-estimated index (the black circles) is highly correlated with the ranking of the original Human Development Index indicating that our assessment of the inherent uncertainty provides particularly relevant information for the interpretation of the original index.

Figure 3 shows that the indicators do a better job in distinguishing less developed countries than the more developed. In the Figure we display the countries for which we

¹²The actual scores are presented in Appendix C.

Uncertainty in HDI ranking

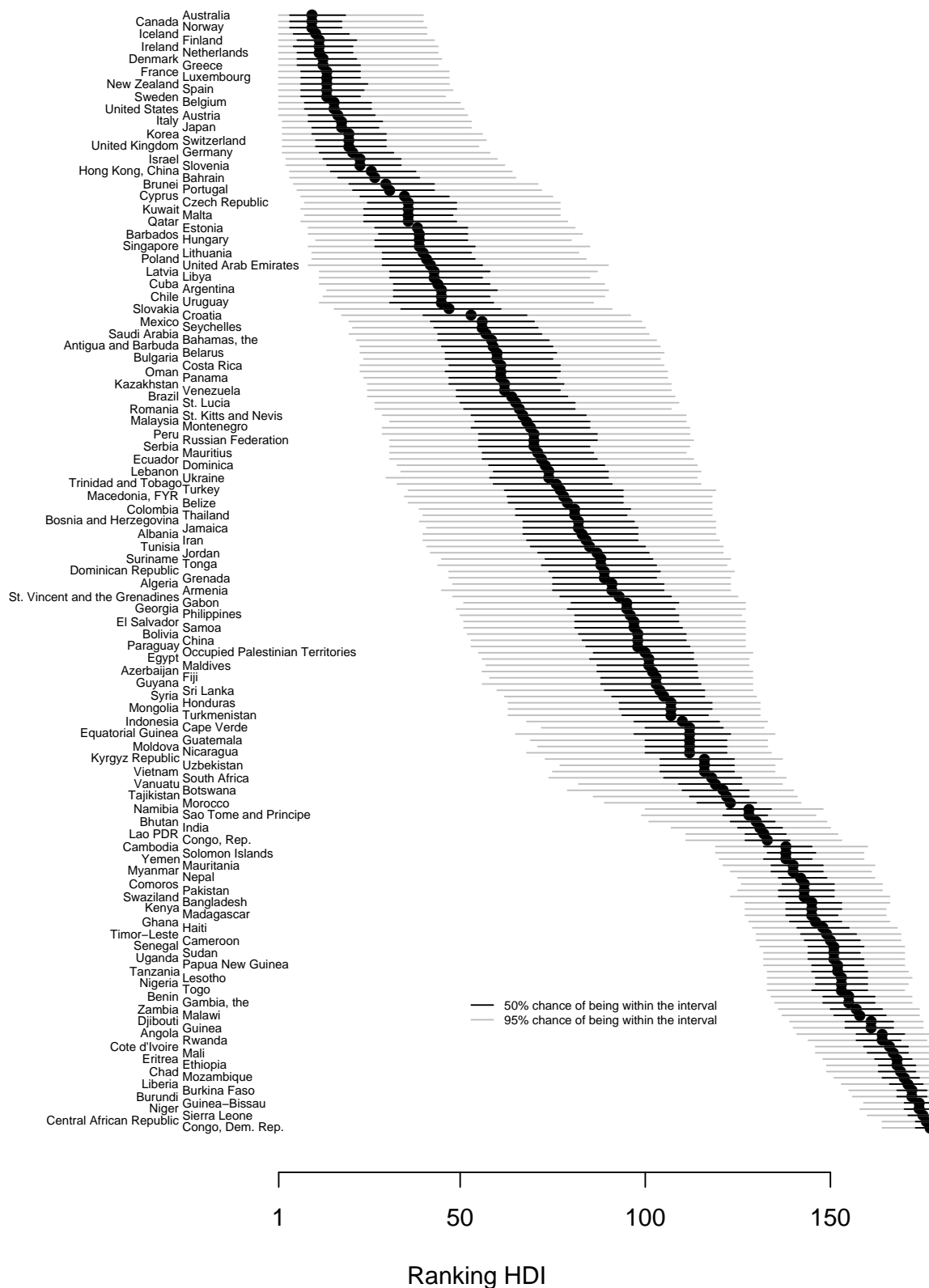


Figure 1: Estimated rankings for the Human Development Index.

Notes: Posterior inferences. The black circle indicates the median ranking, the black line gives the rank position interval for which the country is more likely to fall inside than outside, and the gray line indicates the ranking interval for which the country is more than 95 per cent certain to be located within.

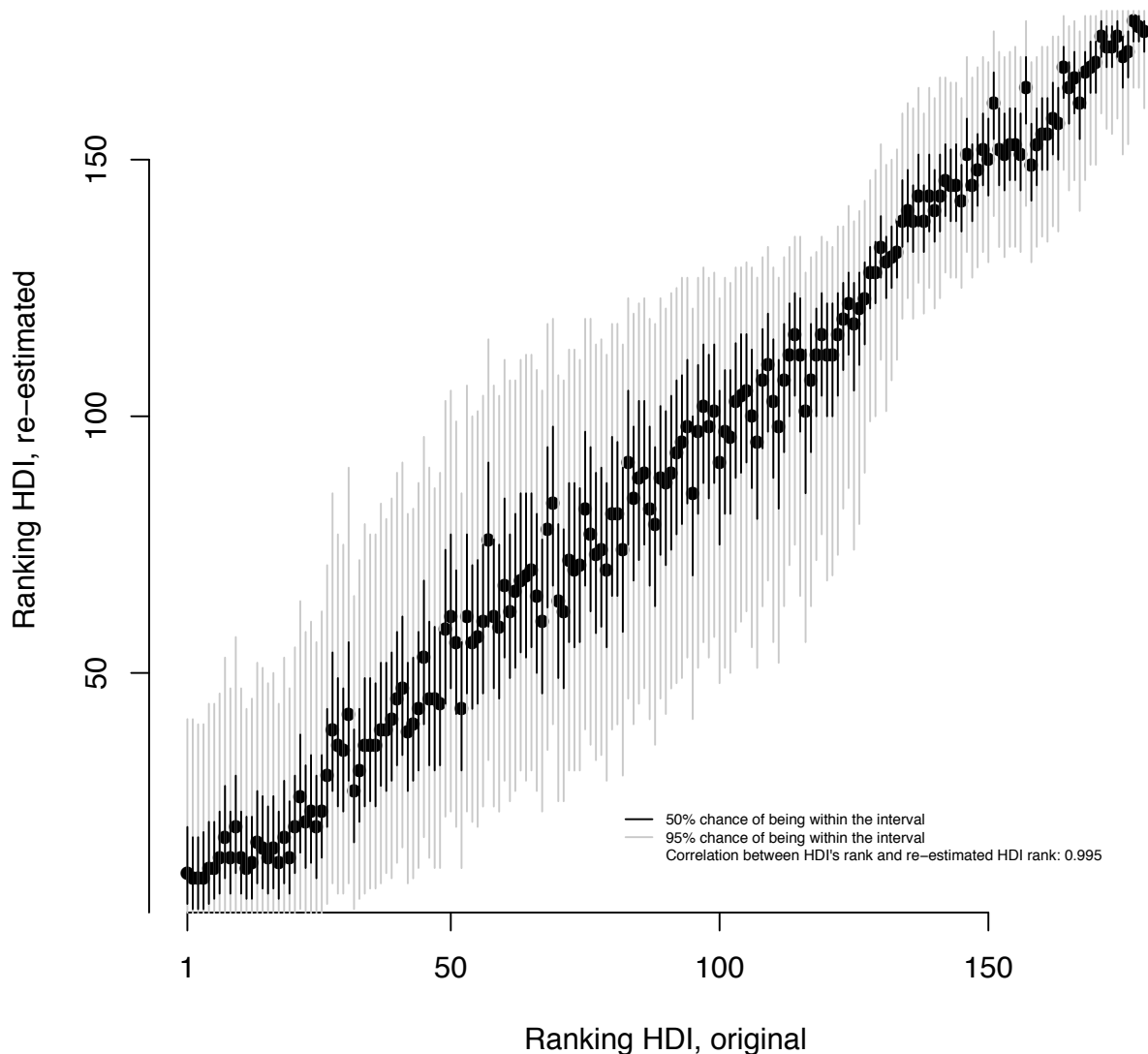


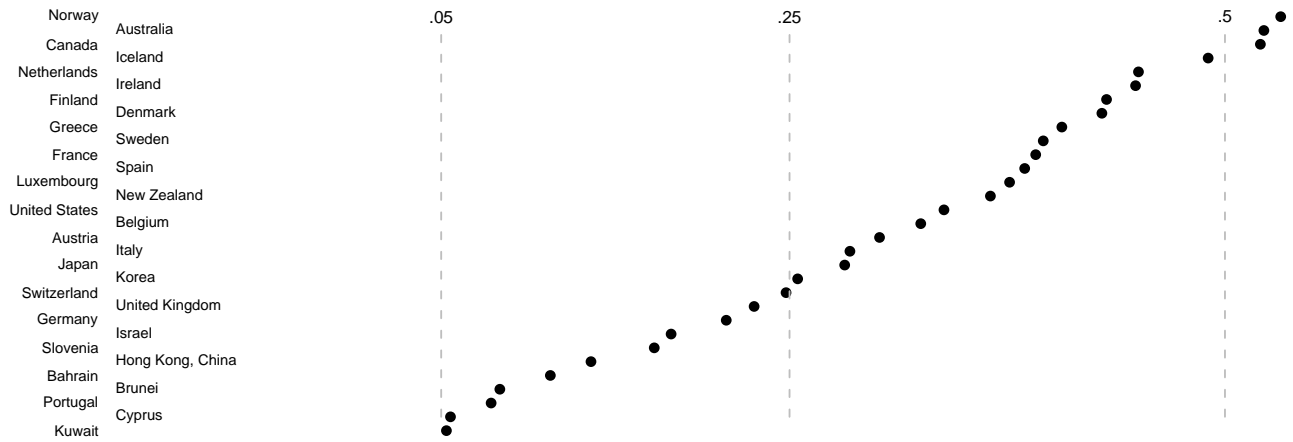
Figure 2: The original HDI ranking versus the re-estimated ranking.

Notes: Posterior inferences. The black circle indicates the median ranking, the black line gives the rank position interval for which the country is more likely to fall inside than outside, and the gray line indicate the ranking interval for which the country is more than 95 per cent certain to be located within. The correlation coefficient is evaluated using the median of the posterior ranking.

cannot reject the hypothesis that they are among the top and bottom 10 countries in the world in terms of Human Development at a 5 per cent level.¹³ While 31 countries have at least a 5 per cent chance of being among the top 10 countries in Human Development, 22 countries have at least a 5 per cent chance of being among the bottom 10 countries on the Human Development Index. No country has more than a 75 per cent chance of being among the top 10, while 5 countries have more than a 75 per cent chance of being among the bottom 10. The Central African Republic and the Democratic Republic of Congo are

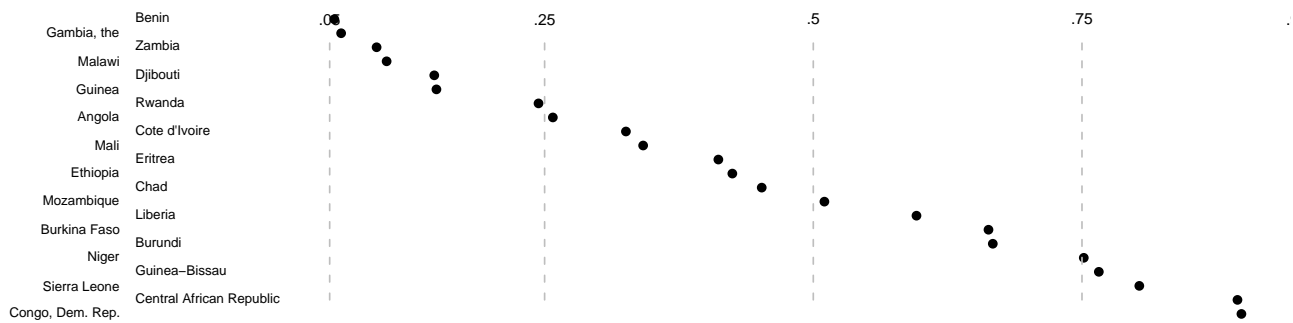
¹³Since we are using diffuse priors, the probabilities can also be seen as the p-values of the classical test H_0 : the country is among the top/bottom ten versus the alternative hypothesis that it is not.

HDI – Probability of being among top 10



No. of countries: 31

HDI – Probability of being among bottom 10



No. of countries: 22

Figure 3: Rank uncertainty for the Human Development Index: countries that have more than a 5 per cent chance of being among the respectively top and bottom 10.

the worst countries out, with an almost 95 per cent chance of being among the 10 least developed countries in the world.

Freedom House

The rankings based on the estimated index scores for Freedom House are given in Figure 4. Freedom House distinguishes better along a wider range of the underlying dimensions than the Human Development Index does. It is only at the very top of the scale that the index clearly fails to distinguish between countries. There is a group of 7 countries that, for all practical purposes are identical in the underlying dimension. The reason for this is that these 7 countries all score the maximum value on all the indicators that go into the index. For this reason it is very hard to know how good these countries really are.

Figure 5 shows that 17 countries have at least a 5 per cent chance of being among

Uncertainty in Freedom House ranking

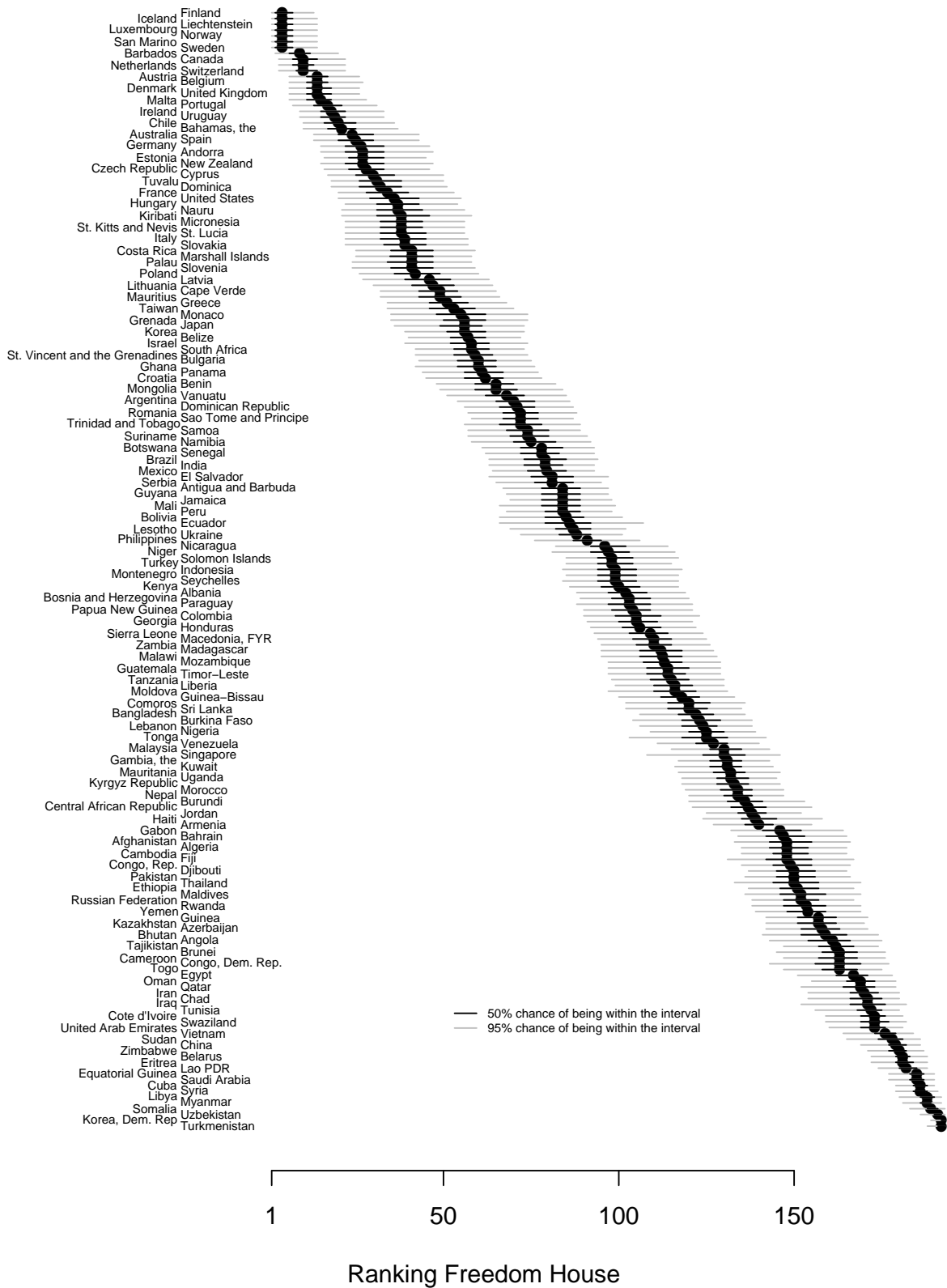
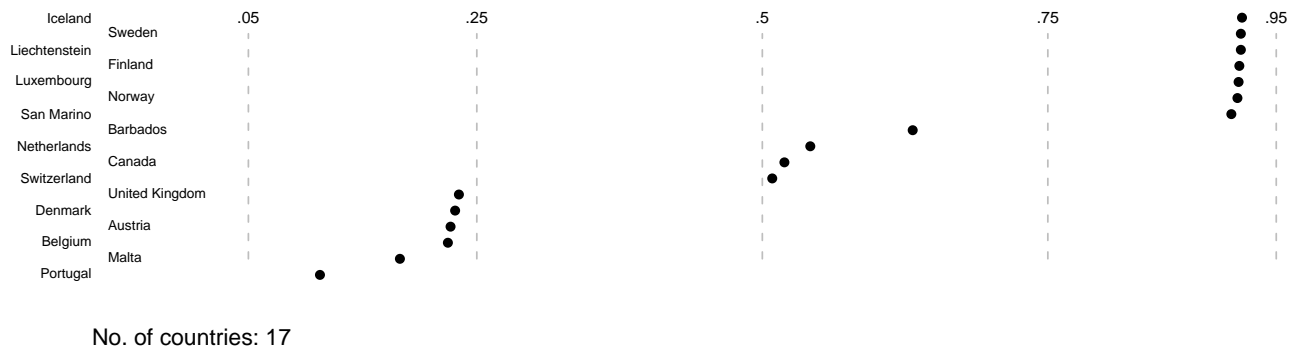


Figure 4: Estimated rankings for Freedom House.

Notes: Posterior inferences. The black circle indicates the median ranking, the black line gives the rank position interval for which the country is more likely to fall inside than outside, and the gray line indicates the ranking interval for which the country is more than 95 per cent certain to be located within.

FH – Probability of being among top 10



FH – Probability of being among bottom 10

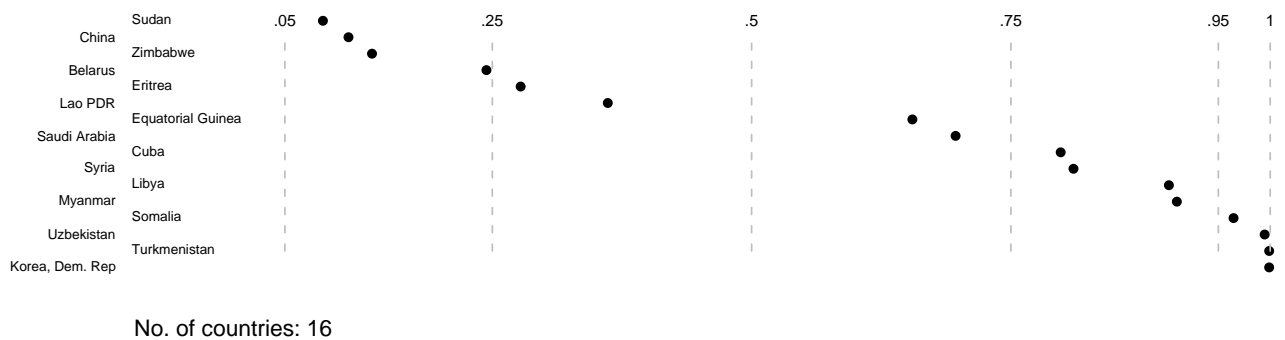


Figure 5: Rank uncertainty for Freedom House: countries that have more than a 5 per cent chance of being among the respectively top and bottom 10.

the 10 most democratic countries in the world, while 16 countries have at least a 5 per cent chance of being among the 10 least democratic countries in the world. Freedom House is hence capable of narrowing down the top and bottom countries fairly accurately. The least free countries in the world are Turkmenistan, North Korea, Uzbekistan, and Somalia. The probability that North Korea is among the 10 least free countries in the world is estimated to be very close to 1.

Doing Business

The results for Doing Business are presented in Figure 6. Our estimates suggest that there is considerable uncertainty around the ranking, in particular for countries in the middle 80 per cent of the scale. In contrast to the key message of the precise ranking published in the Doing Business report, it is clear that the index does not do a very good job in distinguishing between most of the regulatory environments in the world. While the rankings, after taking uncertainty into account, clearly distinguish the best economies from the worst, it does not distinguish particularly well between the economies that are somewhat in between. There is a large group of more than 100 countries, among which

Uncertainty in Doing Business ranking

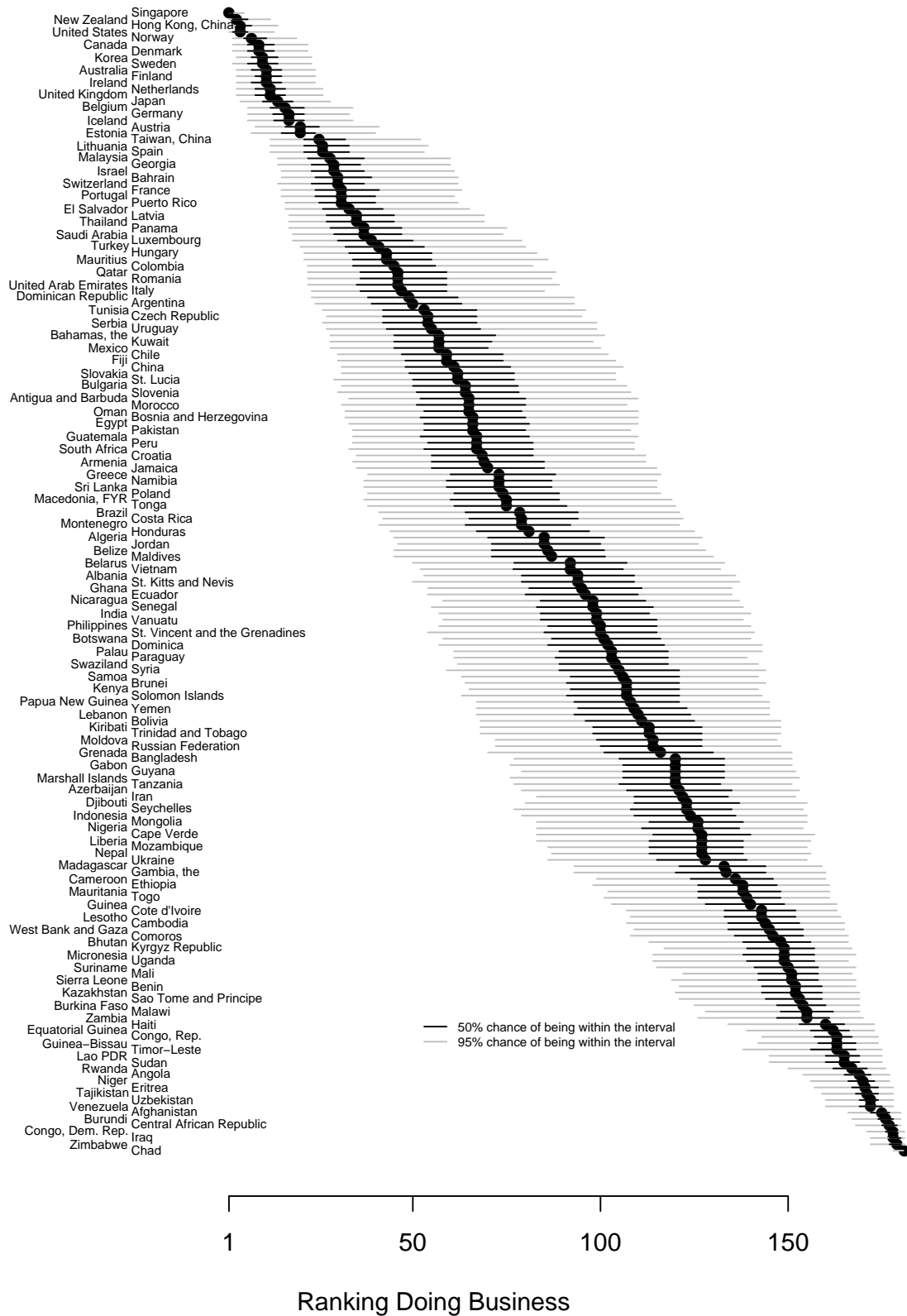


Figure 6: Estimated rankings for Doing Business.

Notes: Posterior inferences. The black circle indicates the median ranking, the black line gives the rank position interval for which the country is more likely to fall inside than outside, and the gray line indicates the ranking interval for which the country is more than 95 per cent certain to be located within.

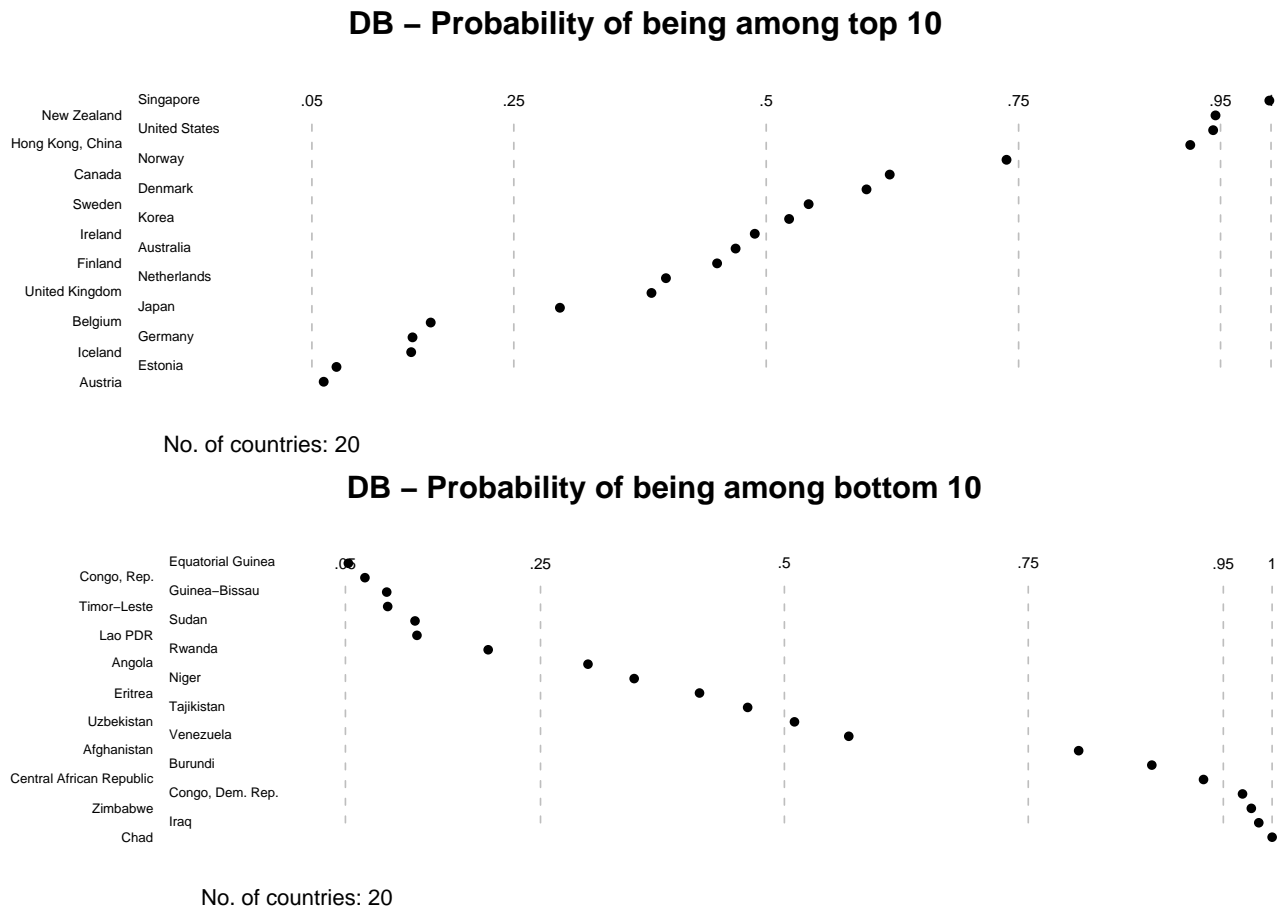


Figure 7: Rank uncertainty for Doing Business: countries that have more than a 5 per cent chance of being among the respectively top and bottom 10.

it is almost impossible to identify any differences.

As Figure 7 shows, 20 countries have at least a 5 per cent chance of providing one of the top ten business environments in the world. There are nine countries with more than a 50 per cent chance of belonging to this group, of which four countries—Singapore, New Zealand, United States, and Hong Kong—are almost surely among the top ten. Figure 7 also reveals that the situation at the bottom mirrors the one at the top: 20 countries have at least a 5 per cent chance of having among the 10 least inviting business environments in the world. Nine countries have more than a 50 per cent chance of belonging to this group, of which four countries—Congo, Zimbabwe, Iraq, and Chad—are almost surely among the 10 economies with the worst business environments in the world.

5.2 Do the indexes capture the same thing?

Figure 8 shows the correlation between the rankings in the different indexes. The position of the countries in each of the indexes is evaluated using the median of the estimated rankings for each of the countries. The plots on the diagonal of the matrix show the

density of the median rankings for respectively Doing Business, Freedom House, and the Human Development Index. The off-diagonal elements show scatterplots between the different indexes. The dots are median rankings for each country, the solid line is a non-parametric (lowess) regression, and the dashed line is a linear regression line.

First, we note that although the correlations between all the different indexes are positive, there is substantial spread in the degree of correlation. For all pairs, the correlation is much stronger for higher rankings (closer to 1) than for lower rankings. The extreme case is the correlation between the Human Development Index and Freedom House. These are highly correlated at the upper end of the scale (closer to 1, i.e. in the bottom left corner of the figure). But as demonstrated by the flatness and even turning of the local regression line, there is no or negative correlation between these two measures at the lower half of the scale. The correlations between Doing Business and the other two indexes are positive in both instances, although the positive correlation is much sharper in Doing Business versus the Human Development Index than between Doing Business and Freedom House.

6 Conclusion

We have demonstrated that the Human Development Index, Freedom House, and the Doing Business index all provide country rankings that are highly uncertain. The rankings therefore end up emphasizing imaginary differences between countries as if they were distinct and real. It should be contrasted with the claim of the Economist that “Doing Business [...] put precise numbers on things that people had known about only vaguely, and it allowed citizens and investors to compare their country with 180 others.” (The Economist, 12 March 2009).

Country rankings are used as guides for economic, social, and legal reforms by governments, NGOs and international organizations. Their popularity is not without impact. Even the recognized Forbes Magazine follows suit by emphasizing how one of the indexes can be “a useful starting point in an international stock search.” (September 21, 2005). The rankings also inspire governments to reform. Since its start in October 2003, the *Doing Business* reports, the index has inspired 113 reforms around the world. In particular, it emphasizes how “Georgia targeted the top 25 list and used *Doing Business* indicators as benchmarks of its progress. It now ranks 18 on the ease of doing business, and the government has set an even more ambitious goal. Saudi Arabia and Mauritius have targeted the top 10. Both have made tremendous progress: Saudi Arabia now ranks 23, and Mauritius 27” (World Bank, 2007, p. 7). In the light of our estimates, it is clear that these changes cannot be significant, as our estimated confidence intervals for rankings of the countries are from 14th to 60th place for Georgia, from 18th to 74th place for Saudi Arabia, and from 21st to 86th place for Mauritius.

Whenever the scores of international index rankings are taken literally, the indexes

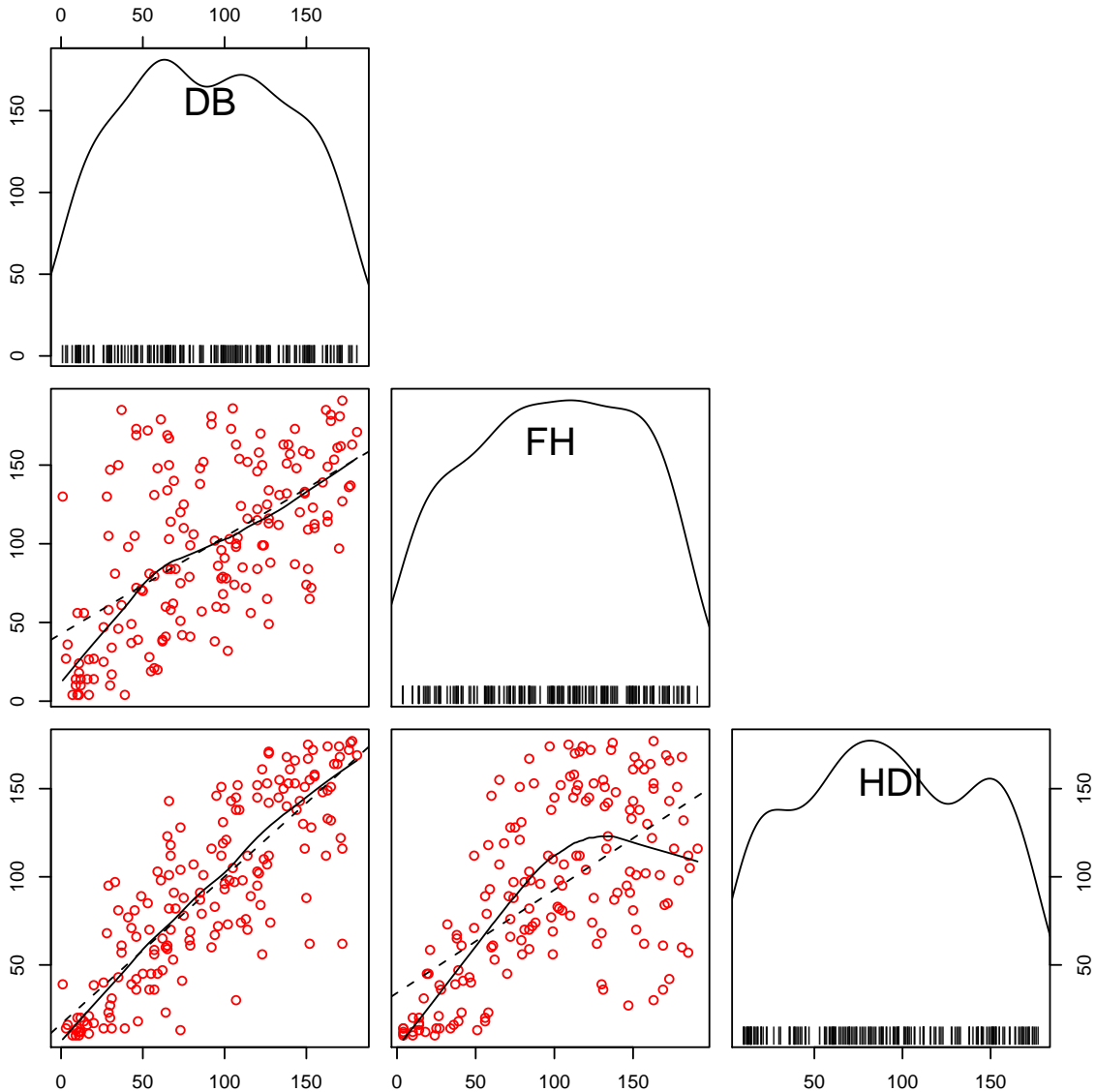


Figure 8: Scattermatrix for the rankings on the different indexes.

Notes: All countries are evaluated at the median ranking on each of the indexes. Best possible rank is 1. The plots on the diagonal of the matrix show the density of the median rankings for respectively Doing Business (top/left), Freedom House (center), and the Human Development Index (bottom/right). The off-diagonal plots show the correlation in rankings between the indexes. The dashed line is a regression line, while the solid line is a non-parametric regression line (lowess).

may be poor guides for policies as each link between indicators and scores is noisy and uncertain, but presented as certain. In applied public economics, performance indexes are rarely used as welfare measures. This is not the case in popular and political debates. To rephrase Keynes, *indexes, both when they are right and when they are wrong, are more powerful than is commonly understood [...] Practical men, who believe themselves to be quite exempt from any influence, are usually the slaves of some defunct conception of an index ranking.*¹⁴ The belief in accuracy in the presence of inaccuracy may lead to a shift in focus among reformers from what really counts to what the makers of these rankings

¹⁴Keynes (1936, p. 383)

count. This is the tyranny of international index rankings.

References

- Anand, Sudhir and Amartya K. Sen**, “Human Development Index: Methodology and Measurement,” Occasional Papers 12, Human Development Report Office, New York 1994.
- Bollen, Kenneth and Richard Lennox**, “Conventional Wisdom on Measurement: A Structural Equation Perspective,” *Psychological Bulletin*, 1991, 110 (2), 305–14.
- Clinton, Joshua D. and David E. Lewis**, “Expert Opinion, Agency Characteristics, and Agency Preferences,” *Political Analysis*, January 2008, 16 (1), 3–20.
- Gelman, Andrew and Jennifer Hill**, *Data Analysis Using Regression and Multi-level/Hierarchical Models*, Cambridge University Press, December 2006.
- Hampel-Malagrosa, Aimee and Judith Frickenstein**, “Taking the woman’s perspective: gender risks of regulatory reforms,” *Enterprise Development & Microfinance*, 2008, 19 (3), 204–19.
- HDR**, “Human Development Report 2007/2008—Fighting climate change: Human solidarity in a divided world,” United Nations Development Programme New York 2007.
- Jackman, Simon**, “What Do We Learn from Graduate Admissions Committees? A Multiple Rater, Latent Variable Model, with Incomplete Discrete and Continuous Indicators,” *Political Analysis*, November 2004, 12 (4), 400–424.
- Jarvis, Cheryl Burke, Scott B. Mackenzie, and Philip M. Podsakoff**, “A Critical Review of Construct Indicators and Measurement Model Misspecification in Marketing and Consumer Research,” *Journal of Consumer Research*, September 2003, 30 (2), 199–218.
- Kaufmann, Daniel, Aart Kraay, and Pablo Zoido-Lobatón**, “Aggregating governance indicators,” Policy Research Working Paper Series 2195, The World Bank October 1999.
- Keynes, John Maynard**, *The General Theory of Employment, Interest and Money*, 2008 paperback ed., BN Publishing, 1936.
- Lee, Sangheon, Deirdre McCann, and Nina Torm**, “The World Bank’s “Employing Workers” index: Findings and critiques—A review of recent evidence,” *International Labour Review*, 2008, 147 (4), 416–32.

- Lunn, D.J., A. Thomas, N. Best, and D. Spiegelhalter**, “WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility,” *Statistics and Computing*, 2000, 10, 325–37.
- Morse, Stephen**, “For better or for worse, till the human development index do us part?,” *Ecological Economics*, 2003, 45, 281–96.
- Munck, Gerardo L. and Jay Verkuilen**, “Conceptualizing and Measuring Democracy,” *Comparative Political Studies*, February 2002, 35 (1), 5–34.
- Sagar, Ambuj D. and Adil Najam**, “The human development index: a critical review,” *Ecological Economics*, June 1998, 25 (3), 249–264.
- Søreide, Tina**, “Is it wrong to rank? A critical assessment of corruption indices,” CMI Working Papers WP 2006: 1, CMI (Chr. Michelsen Institute), Bergen, Norway 2006.
- Treier, Shawn and Simon Jackman**, “Democracy as a Latent Variable,” *American Journal of Political Science*, January 2008, 52 (1), 201–217.
- World Bank**, “Doing Business 2008: Comparing Regulation in 178 Economies,” The World Bank Washington, D.C. 2007.

A Model and estimation

A.1 Model—Continuous, ordinal, count, and censored outcomes

All the observable indicators we have in our data are one of the following kinds: i) continuous with a clearly defined scale (GDP per capita, life expectancy, etc.), ii) continuous but as if censored (literacy rates, enrollment rates, etc.), iii) discrete count data (the number of documents needed to fill out to export or import a good etc.), iv) discrete data of an ordinal nature (scores on scales that do not have a clear-cut meaning). The point of our estimation is to estimate the parameters of the model given in equation (1).

For the ordinal variables, we follow the approach in Treier and Jackman (2008) and Clinton and Lewis (2008), and estimate an ordered choice model. The idea is to still think of f_i as the (continuous) underlying quality of the countries, but that the observable indicator now is observed in steps only. The steps are assumed to be ordered, in the sense that x_{ij} is non-decreasing in f_i , holding the error term constant and assuming that $\beta_j > 0$.

More precisely,

$$x_{ij} = \begin{cases} 1 & \text{if } \alpha_j + \beta_j f_i + \varepsilon_{ij} \in (-\infty, \tau_{j1}^*] \\ 2 & \text{if } \alpha_j + \beta_j f_i + \varepsilon_{ij} \in (\tau_{j1}^*, \tau_{j2}^*] \\ \vdots & \\ K_j & \text{if } \alpha_j + \beta_j f_i + \varepsilon_{ij} \in (\tau_{j,K_j-1}^*, \infty) \end{cases}$$

where K_j denotes the maximum number of values observed for the ordinal variable. Assuming the error terms are distributed according to the cdf F , it is immediate that

$$\begin{aligned} \Pr [x_{ij} = 1] &= F(\tau_{j1} - \beta_j f_i) \\ \Pr [x_{ij} = 2] &= F(\tau_{j2} - \beta_j f_i) - F(\tau_{j1} - \beta_j f_i) \\ &\vdots \\ \Pr [x_{ij} = K_j] &= 1 - F(\tau_{j,K_j-1} - \beta_j f_i), \end{aligned}$$

where $\tau_{jk} = \tau_{jk}^* - \alpha_j$. We assume that ε_{ij} is logistically distributed, which means that the model for the ordered outcomes is a standard ordered logit, but with an unobserved right-hand side. We also here control for correlation among the indicators within the same group, see section 3.1.

We also take censoring into account, by modifying the likelihood function in the appropriate way. If x_{ij} is top-censored at 100, say, the likelihood function for an observation is (assuming that the error terms are normally distributed)

$$p(x_{ij} | \alpha, \beta, \sigma^2) = \begin{cases} \phi((x_{ij} - \alpha_j - \beta_j f_i) / \sigma) & \text{if } x_{ij} < 100 \\ \Phi((\alpha_j + \beta_j f_i - 100) / \sigma) & \text{if } x_{ij} = 100, \end{cases}$$

where ϕ is the density function and Φ is the cumulative density function of the normal distribution, see for example Gelman and Hill (2006, p. 404). We take this censoring into account using the $I()$ -construct in WinBUGS, following Gelman and Hill (2006, p. 404-5).

For the count data, we estimate a overdispersed Poisson regression-type model. Assuming that the left-hand side variable x_{ij} only takes positive integer values, the regression model we use has the form

$$\begin{aligned} x_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \log(\lambda_{ij}) &= \alpha_j + \beta_j f_i + \epsilon_{ij}. \end{aligned}$$

A.2 Estimation

The estimation is performed using WinBUGS.¹⁵ All computer codes used in the paper are available at request. Our approach is a latent variable approach, where we estimate the underlying dimension given the observed indicators; see the previous section and section 3. The model we fit to different indexes is the model given in equation (1), with care taken to what kinds of variables are observed on the left-hand side of the equation.

Priors

We use diffuse priors on all the structural parameters; the priors are given in Table 1. By this we mean that we put no a priori restrictions on the values that the parameters can or are likely to take, the identification of the parameters relies only on the data via the likelihood function.

Computational Details

We run 3 chains each consisting of 300,000 iterations for all of the indexes. The first 100,000 iterations of each of the chains are discarded as burn-in, in order to ensure that our results are not driven by the starting values. The starting values for f_i were obtained by running a standard factor analysis and calculating the Bartlett scores. Starting values for the β_j 's were set to 1, which ensured that all the chains were on the same scale—remember that we only have local identification, so if both f and β are multiplied by -1 the results will be exactly the same. For the same reason we set the starting values for the δ_{ik} -shocks to 0 for the HDI. The remaining starting values were generated at random in WinBUGS.

Each 200th subsequent iteration is recorded. This is in order to ensure low autocorrelation between each of the sampled values. The traceplots show no sign of non-convergence, and formal tests of convergence support the assessment that the models have converged: \hat{R} is below 1.05 for all parameters. There is no indication of serious autocorrelation problems.

B Classification of variables

The Human Development Index

Classification of the variables in the Human Development Index can be found in Table 2.

Freedom House

Classification of the variables in Freedom House can be found in Table 3.

¹⁵WinBUGS is available at <http://www.mrc-bsu.cam.ac.uk/bugs/>.

Index	Variable	Prior
Human Development Index	f	$N(0, 1)$
	(α_j, β_j)	$N(\mathbf{0}, 10^7 \times I_2)$
	$1/\sigma_j^2$	$\Gamma(.05, .05)^a$
	δ_{ik}	$N(0, \sigma_{\delta,k}^2)$
	$\sigma_{\delta,k}^2$	Uniform on $[0, 100]$
	Freedom House	f
β_j		$N(0, 1/.15)^b$
τ_{j1}		$N(0, 1/.15)^c$
τ_{jk}		Exponential w/ rate 2
δ_{ik}		$N(0, \sigma_{\delta,k}^2)$
$\sigma_{\delta,k}^2$		Uniform on $[0, 100]$
Doing Business	f	$N(0, 1)$
	(α_j, β_j)	$N(\mathbf{0}, 10^7 \times I_2)$
	$1/\sigma_j^2$	$\Gamma(.05, .05)$
	δ_{ik}	$N(0, \sigma_{\delta,k}^2)$
	$\sigma_{\delta,k}^2$	Uniform on $[0, 3]^d$
	τ_{j1}	$N(0, 1/.15)$
	τ_{jk}	Exponential w/ rate 2
	ϵ_{ij}	$N(0, \sigma_{\epsilon_{ij}}^2)$
	$\sigma_{\epsilon_{ij}}^2$	Uniform $[0, 5]$

^a $\Gamma(\text{shape}, \text{rate})$. We have also fitted the model using $\Gamma(1, .01)$ and $\Gamma(.01, .01)$. The results were the same.

^b We have also fitted the model using a variance of $1/0.001$. The results were the same.

^c We have also fitted the model using a variance of $1/0.001$. The results were the same.

^d We have also fitted the model using $[0, 100]$. The results were the same, but it converged slower. None of the estimated parameters are close to the boundaries of the distribution.

Table 1: Priors for the different parameters

Group	Variable	Type
Schooling	School enrollment	Continuous, censored at 100
	Literacy	Continuous, censored at 100
GDP	Log GDP per capita	Continuous
Life expectancy	Life expectancy	Continuous

Table 2: Classification of variables, the Human Development Index

Group	Variable	Type
Political Rights	A: Electoral Process	Ordered
	B: Political Pluralism and Participation	Ordered
	C: Functioning of Government	Ordered
Civil Liberties	D: Freedom of Expression and Belief	Ordered
	E: Associational and Organizational Rights	Ordered
	F: Rule of Law	Ordered
	G: Personal Autonomy and Individual Rights	Ordered

Table 3: Classification of variables, Freedom House

Doing Business

Classification of the variables in Doing Business can be found in Table 4.

Group	Variable	Type
Starting a Business	Procedures (number)	Count
	Time (days)	Count
	Cost (% of income per capita)	Continuous
	Min. capital (% of income per capita)	Continuous, censored at 0
Dealing with Licenses	Procedures (number)	Count
	Time (days)	Count
	Cost (% of income per capita)	Continuous
Employing Workers	Difficulty of Hiring Index	Continuous, censored at 0 and 100
	Rigidity of Hours Index	Continuous, censored at 0 and 100
	Difficulty of Firing Index	Continuous, censored at 0 and 100
	Firing costs (weeks of wages)	Continuous, censored at 0
Registering Property	Procedures (number)	Count
	Time (days)	Count
	Cost (% of property value)	Continuous
Getting Credit	Credit Information Index	Ordered
	Private bureau coverage (% of adults)	Continuous, censored at 0 and 100
	Public bureau coverage (% of adults)	Continuous, censored at 0 and 100
	Legal Rights Index	Ordered
Protecting Investors	Disclosure Index	Ordered
	Director Liability Index	Ordered
	Shareholder Suits Index	Ordered
Paying Taxes	Payments (number)	Count

Group	Variable	Type
	Time (hours)	Count
	Profit tax rate	Continuous, censored at 0
	Labor tax rate	Continuous, censored at 0
	Other taxes	Continuous, censored at 0
Trading Across Borders	Documents for export (number)	Count
	Time for export (days)	Count
	Cost to export (US\$ per container)	Continuous
	Documents for import (number)	Count
	Time for import (days)	Count
	Cost to import (US\$ per container)	Continuous
Enforcing Contracts	Procedures (number)	Count
	Time (days)	Count
	Cost (% of debt)	Continuous
Closing a Business	Time (years)	Continuous
	Cost (% of estate)	Continuous
	Recovery rate (cents on the dollar)	Continuous

Table 4: Classification of variables, Doing Business

C Estimated latent scores

Human Development Index

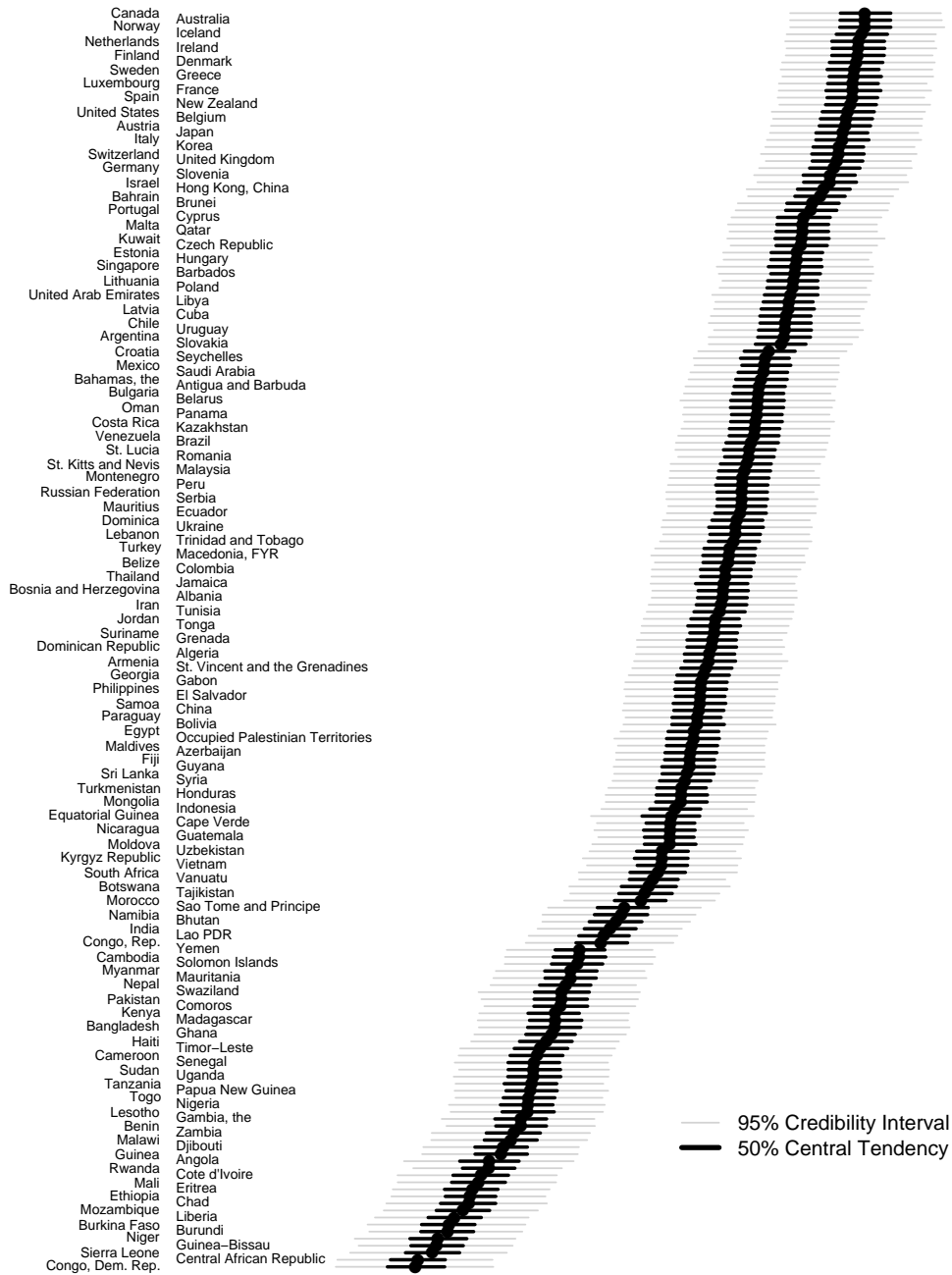


Figure 9: Estimated positions for the Human Development Index.

Notes: Posterior inferences. The black circle indicates the median value, the black line the central tendency, and the gray line the 95 per cent highest posterior density region for f , the position of the countries on the Human Development Index.

Freedom House

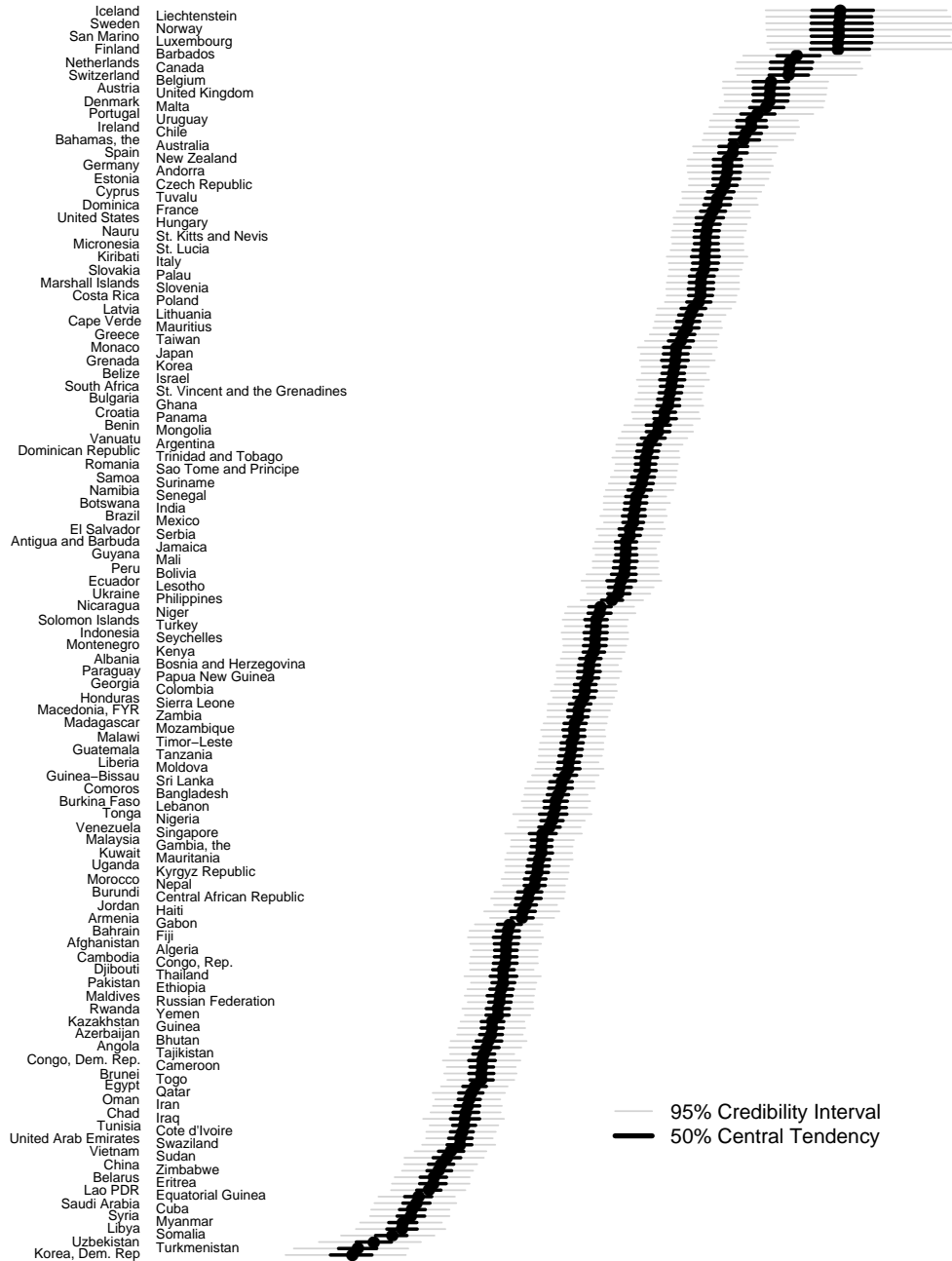


Figure 10: Estimated positions for Freedom House.

Notes: Posterior inferences. The black circle indicates the median value, the black line the central tendency, and the gray line the 95 per cent highest posterior density region for f , the position of the countries on Freedom House.

Doing Business

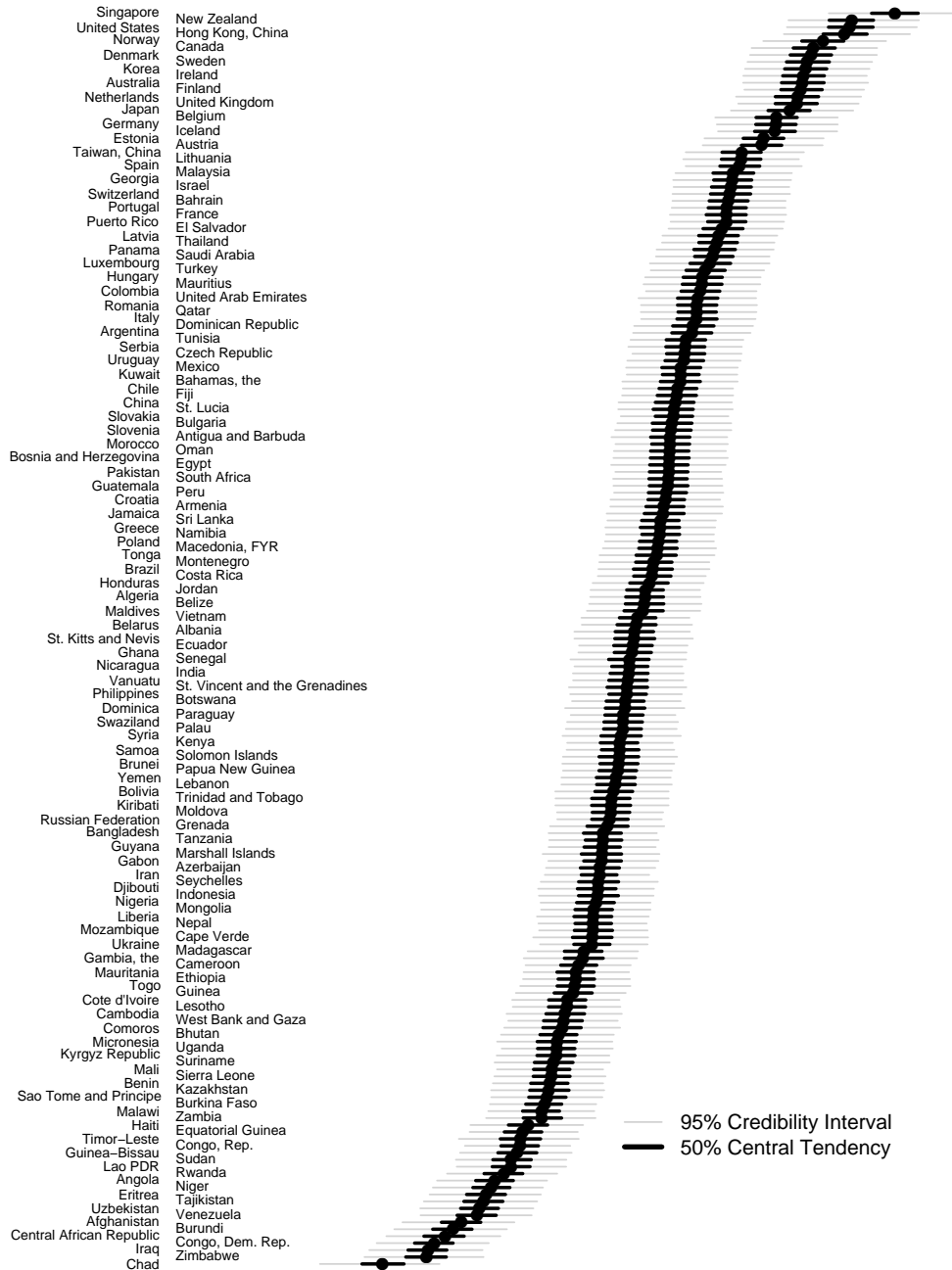


Figure 11: Estimated positions for Doing Business.

Notes: Posterior inferences. The black circle indicates the median value, the black line the central tendency, and the gray line the 95 per cent highest posterior density region for f , the position of the countries on Doing Business.