Corruption Perceptions Index 2012: Technical Methodology Note

Background

The Corruption Perceptions Index (CPI) was established in 1995 as a composite indicator used to measure perceptions of corruption in the public sector in different countries around the world. During the past 17 years, both the sources used to compile the index and the methodology have been adjusted and refined. Following a rigorous review process, some important changes have been made to the methodology in 2012. The method we use to aggregate different data sources has been simplified and also now includes just one year's data from each data source. Crucially, this method will allow us to compare scores over time, which was not methodologically possible previously. Given the changes to the methodology, it must be emphasised that country scores of the CPI 2012 cannot be compared against those of 2011 or previous editions. Year to year comparison will be possible from 2012 onwards.

Methodology

The methodology follows 4 basic steps: selection of source data, rescaling source data, aggregating the rescaled data and then reporting a measure for uncertainty.

1. Selection of data sources

The CPI draws upon a number of available sources which capture perceptions of corruption. Each source is evaluated against the criteria listed below. Contact has been made with each institution providing data in order to verify the methodology used to generate scores and for permission to publish the rescaled scores from each source, alongside the composite index score.

A) **Reliable data collection and methodology from a credible institution:** It is necessary that we trust the validity of the data we are using. As such, each source should originate from a professional institution that clearly documents its methods for data collection. These methods should be methodologically sound, for example, where an 'expert opinion' is being provided, we seek assurance on the qualifications of the expert or where a business survey is being conducted, that the survey sample is representative.

B) **Data addresses corruption in the public sector:** The question or analysis should relate to a perception of the level of corruption explicitly in the public sector. The question can relate to a defined 'type' of corruption (e.g. specifically petty corruption), and where appropriate, the effectiveness of corruption prevention as this can be used as a proxy for the perceived level of corruption in the country.

C) **Quantitative granularity:** The scales used by the data sources must allow for sufficient differentiation in the data (i.e., at least a four-point scale) on the

perceived levels of corruption across countries so that it can be rescaled to the CPI's 0-100 scale.

D) **Cross country comparability:** As the CPI ranks countries against each other, the source data must also be legitimately comparable between countries and not be country specific. The source should measure the same thing in each country scored, on the same scale.

E) **Multi year data-set:** We want to be able to compare a country's score, and indeed the index in general, from one year to the next. Sources that capture corruption perceptions for a single point in time, but that are not designed to be repeated over time, are therefore excluded.

2. Standardise data sources

Each source is then standardised to be compatible with other available sources, for aggregation to the CPI scale. The standardisation converts all the data sources to a scale of 0-100 where a 0 = highest level of perceived corruption, and 100 = lowest level of perceived corruption.

Any source that is scaled such that lower scores represent lower levels of corruption must first be reversed. This is done by multiplying every score in the data set by -1.

Every score is then standardised (to a z score) by subtracting the mean of the data and dividing by the standard deviation. This results in a data set centred around zero and with a standard deviation of 0.5.

For these z scores to be comparable between data sets, we must define the mean and standard deviation parameters as global parameters. Therefore where a data set covers a limited range of countries, we impute scores for all those countries that are missing in the respective data set. We impute missing values for missing countries in each data set using the statistical software package STATA and, more specifically, the programme's *ice* command. This command uses multiple regressions with all available data sets to estimate values for each country that is missing data in each individual data set. This command runs the imputation multiple (10) times, generating 10 estimated values for each 'missing' score. The mean and standard deviation for the data set is calculated as an average across all 10 complete data sets and is used as the parameter to standardise the raw data. Importantly, the imputed data is used only to generate these parameters and is not used as source data for CPI country scores.

The z scores are then rescaled to fit the CPI scale between 0-100. This uses a simple rescaling formula, which sets the mean value of the standardised dataset to [TBC], and the standard deviation to [TBC]. Any score which exceeds the 0 to 100 boundaries will be capped.

As this is the first year we will be using this methodology, 2012 will be defined as the baseline year. In subsequent years (post 2012), the mean and standard deviation parameters used for the standardisation calculations will be the same values as generated for the baseline 2012 index. By using the same parameters year to year, we can compare these standardised and rescaled scores over time. When new sources enter the index, in order to appropriately reflect changes over time, the

rescaling calculation will allow for these to be consistent with 2012 baseline parameters. This is done by first estimating if there has been a global change in the mean and standard deviation since 2012, and then using these new values, which may have deviated from [TBC] and [TBC] to rescale the new data set.

3. Aggregate the rescaled data

Each country's CPI score is calculated as a simple average of all the available rescaled scores for that country (note, we do not use any of the imputed values as a score for the aggregated CPI). A country will only be given a score if there are at least three data sources available from which to calculate this average.

4. Report a measure of uncertainty

The CPI score will be reported alongside a standard error and confidence interval which reflects the variance in the value of the source data that comprises the CPI score.

The standard error term is calculated as the standard deviation of the rescaled source data, divided by the square root of the number of sources. Using this standard error, we can calculate the 90% confidence interval, assuming a normal distribution.

- ENDS -