Corruption Perceptions Index 2016: 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 20 years, both the sources used to compile the index and the methodology has been adjusted and refined. The most recent review process took place in 2012[1], and some important changes were made to the methodology in 2012. The method that was used up until 2012 to aggregate different data sources has been simplified and now includes just one year's data from each data source. Crucially, this method now allows us to compare scores over time, which was not methodologically possible prior to 2012.

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

[1] The methodology used to calculate the CPI 2016 builds on the work examining alternative approaches for constructing the CPI carried out by Prof. Andrew Gelman: Professor, Department of Statistics and Department of Political Science, Columbia University and Dr Piero Stanig: Fellow, Methodology Institute, London School of Economics and Political Science. This work was presented to Transparency International in a report that is available on request. Please email Santhosh Srinivasan at ssrinivasan@transparency.org.

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 0 and with a standard deviation of 1.

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 impute command. This command regresses each data set against the CPI data sources that are at least 50% complete to estimate values for each country that is missing data in each individual data set. This is with the exception of the Bertelsmann Foundation's Transformation Index data, which is not used for the imputation of the Bertelsmann Foundation's Sustainable Governance Indicators because there is no overlap in country coverage of these two data sources. The mean and standard deviation for the data set is calculated as an average of the complete data sets and is used as the parameter to standardise the raw data. Importantly, the complete data set with imputed values is used only to generate these parameters and the imputed values themselves are not used as source data for CPI country scores.

Critically, the z scores are calculated using the mean and standard deviation parameters from the imputed 2012 scores. This is so that 2012 is effectively the baseline year for the data and the rescaled scores can be comparable year on year. When new sources enter the index, in order to appropriately reflect changes over time, the rescaling calculation allows for these to be consistent with 2012 baseline parameters. This is done by first estimating if there was a global change in the mean

and standard deviation since 2012, and then using these new values, which may have deviated from 50 and 20 to rescale the new data set.[2]

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 approximately 45, and the standard deviation of approximately 20. Any score which exceeds the 0 to 100 boundaries will be capped.

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 is reported alongside a standard error and 90% 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.

[2] Since a new data source was added to the CPI, the above procedure was used to check if there was a change in the mean and standard deviation since 2012. We established that the mean and standard deviation had not changed and thereby maintaining year on year comparison of CPI scores.