

Corruption Perceptions Index 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 have been adjusted and refined. In 2012, important changes were made to the methodology to allow for score comparison across time¹, which was not possible prior to 2012.

Methodology

The methodology follows four basic steps: selection of source data, rescaling source data, aggregating the rescaled data and then reporting a measure for uncertainty. The calculation process also incorporates a strict quality control mechanism which consists of parallel independent calculations conducted by two in-house researchers and two academic advisors with no affiliation to Transparency International.

1. Selection of data sources

The CPI draws upon 13 data sources which capture the assessment of experts and business executives on a number of corrupt behaviours in the public sector, including:

- Bribery
- Diversion of public funds
- Use of public office for private gain
- Nepotism in the civil service
- State capture

Some of the sources also look at the mechanisms available to prevent corruption in a country, such as:

- The government's ability to enforce integrity mechanisms
- The effective prosecution of corrupt officials
- Red tape and excessive bureaucratic burden
- The existence of adequate laws on financial disclosure, conflict of interest prevention and access to information
- Legal protection for whistleblowers, journalists and investigators

¹ The methodology used to calculate the CPI 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 cpi@transparency.org.

Each of the data sources used to calculate the CPI is evaluated against the following criteria:

- A) Methodological reliability and institutional reputation:** For a source to be included in the CPI, it is necessary to ensure the quality and adequacy of its methodological approach. For that reason, each source must originate from a professional institution that clearly documents its data collection methods and measurement approach. Transparency International then evaluates the soundness of the methodology.
- B) Conceptual alignment of the data:** As it is a measurement of corruption in the public sector, all data sources used to construct the CPI must be explicitly linked to the levels of corruption or corruption risks in the public sector. The questions can relate to a defined 'type' of corruption (for example, petty corruption) or to the effectiveness of corruption prevention mechanisms, which can also be used as a proxy for the perceived level of corruption in a country.
- C) Quantitative granularity:** The scales used by the data sources must allow for sufficient differentiation in the data (that is, 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. Moreover, the source must measure the same concept across countries and with the same scale. While there is currently no criteria regarding the minimum coverage a source must have to be part of the CPI, the data source with the lowest coverage provides scores for a total of 16 countries.
- E) Multi year data availability:** Since the CPI measures corruption across countries and time, sources that capture corruption perceptions for a single point in time, but that are not designed to be repeated over time, are excluded.

In order to carry out this quality assurance process, Transparency International reaches out to each one of the institutions providing data in order to verify the methodology used to generate their scores. Since some of the sources are not publicly available, Transparency International also requests permission to publish the rescaled scores from each source alongside the composite CPI score. Transparency International is, however, not permitted to share the original scores given by private sources with the general public.

2. Standardise data sources

Each of the sources included in the CPI is standardised to allow for the aggregation into the CPI score. The standardisation converts all the data points to a scale of 0-100 where a 0 represents the highest level of perceived corruption, and 100 the lowest level of perceived corruption. While most of the underlying CPI sources are also coded in the same direction (with lower scores indicating higher levels of perceived corruption), three sources are scaled the opposite way, i.e. with lower scores denoting lower levels of corruption. For comparability purposes, these four sources are reversed by multiplying every score by -1.

The sources that need to be reversed are:

- Economist Intelligence Unit's corruption indicator
- Political and Economic Risk Consultancy Asian Intelligence's perceptions of corruption score
- Varieties of Democracy Project's political corruption index

Since many of the sources used for the CPI do not have a global coverage, the missing values for these sources are imputed for the baseline year.² This process is conducted using the statistical software package STATA and, more specifically, the programme's "impute" command. The "impute" command in STATA estimates a value for each missing data point using only those data sources with at least 50% of the total number of countries covered by the CPI in any given year.³ 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.

After the imputation process, the mean and standard deviation for each data source is calculated and used as parameters to standardise the original data into z scores. The standardised z scores are calculated by subtracting the mean of each source from each country score and dividing by the standard deviation of the respective source. This results in a data set centred around 0 and with a standard deviation of 1. Critically, the z scores are calculated using the mean and standard deviation parameters from the imputed baseline year sources. This ensures that CPI scores are comparable year-on-year since the baseline year.

It is important to note that the complete data set with imputed values is used only to generate the baseline global parameters and the imputed values themselves are not used in the final aggregation that produces the CPI scores.

The z scores are then transformed to fit the CPI scale of 0-100. This uses a rescaling formula, which aims to set the mean value of the standardised dataset to 45, and the standard deviation to 20. The following formula is used for the calculation of the re-scaled indicators:

$$\frac{\text{Original Indicator } X_{\text{Country}} - \text{mean imputed Indicator } X^{t=\text{baseline}}}{\text{standard deviation imputed Indicator } X^{t=\text{baseline}}} * 20 + 45$$

Any rescaled scores which take values of less than 0 are made equal to 0 and any

² Until 2016, the global parameters from 2012 were used as a baseline. With the introduction of the VDEM data into the CPI in 2016, the missing values for 2016 were imputed again. A comparison of the global parameters from 2012 and 2016, however, proved to be statistically insignificant, which means that the change in baseline did not affect comparability across time. In 2017, VDEM expanded its country coverage to over 50 per cent of the CPI countries. For this reason, the global parameters had to be re-calculated once again. The changes in the global mean and standard deviation were again not statistically significant. As a result, the effective reference year remains 2012.

³ As per the 50 per cent country coverage rule mentioned above, the following seven data sources were used for the imputation process for the CPI 2017:

- Varieties of Democracy Project 2017 (83.58 per cent sample coverage)
- Bertelsmann Foundation's Transformation Index 2018 (62.32 per cent sample coverage)
- Economist Intelligence Unit Country Risk Ratings 2017 (63.29 per cent sample coverage)
- Global Insights Country Risk Ratings (98.55 per cent sample coverage)
- Political Risk Services International Country Risk Guide (67.63 per cent sample coverage)
- World Economic Forum Executive Opinion Survey 2017 (64.42 per cent sample coverage)
- World Justice Project rule of Law Index 2017-2018 (54.59 per cent sample coverage)

rescaled scores which exceed 100 are capped to 100.

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 per cent confidence interval which reflects the variance in the value of the source data that comprises the CPI score. Furthermore, changes in CPI scores that are statistically significant are identified and reported.

Following the recommendation of the [European Commission Joint Research Centre](#) audit of the conceptual and statistical coherence of the CPI, the formula for computing standard errors was changed starting from CPI2018. The previous formula calculated the standard error terms as the standard deviation of the rescaled source data, divided by the square root of the number of sources. The new formula expressed in the equation below takes into account a small population of sources:

$$\Sigma = \sqrt{\frac{N - n}{N - 1} \frac{\sigma}{\sqrt{n}}}$$

where Σ is the standard error term, N is the total number of sources used for the CPI calculation, n is the number of sources for the CPI score of any given country, and σ is the standard deviation of the CPI score for that given country.⁴

Using this standard error, we can calculate the 90 per cent confidence interval and report the upper and lower bounds of the CPI score for each country, assuming a normal distribution.

After computing the standard errors, we establish whether the change in CPI score for each country is statistically significant or not. For this purpose, we first compute the effect size of the difference between the score of one country in the two years being compared using the below formula:

$$effect\ size = \frac{M_1 - M_2}{\sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2}{N_1 + N_2 - 2}}}$$

where M_1 is the CPI score for a given country in a given year, M_2 is the CPI score for that country in the year of comparison, N_1 is the number of sources in the given year, N_2 is the number of sources in the year of comparison, SD_1^2 is the square of the standard deviation of the CPI score in the given year and SD_2^2 is the square of the standard deviation of the CPI score in the year of comparison.

Having obtained the effect size we then compute the standard deviation (σ) of the effect size using the below formula:

⁴ In 2022 the total population of sources (N) was 13.

$$\sigma(d) = \sqrt{\frac{N1 + N2}{N1 \times N2} + \frac{d^2}{2 \times (N1 + N2)}}$$

where d is the effect size of the differences in the CPI scores, and $N1$ and $N2$ are the number of sources available for each country. Then, the 90% confidence interval is calculated using the below formula:

$$CI_{0.90} = d \pm 1.645 \times \sigma(d)$$

If this confidence interval includes the value of zero, then there are no statistically significant differences between the countries. On the other hand, if zero is outside the range, then the difference is 'statistically significant at the 10% level'.