INDEX FOR EASTERN PARTNERSHIP COUNTRIES: CODING

The Eastern Partnership Index relies on two types of data: expert assessments commissioned by us and numerical data from publicly available sources. This general design is intended to use the best existing knowledge and to improve this body of knowledge by focused, systematic data collection that benefits from OSI's unique embeddedness and access to local knowledge in EaP countries. However, expert surveys are prone to subjectivity. Many existing expert surveys are characterized by a mismatch between "soft", potentially biased expert opinions and "hard" coding and aggregation practices that suggest a degree of precision not matched by the more complex underlying reality and their verbal representation in country reports.

The expert survey underlying the Index therefore avoids broad opinion questions, and instead tries to verify precise and detailed facts. Complex issues are disaggregated into detailed questions that enable experts to provide more specific responses. Guided by a detailed questionnaire, experts are less often forced to assign subjective weights to different aspects of reality in their evaluation. Most survey questions asked for a "Yes" or "No" response to induce experts to take a clear position and to minimize misclassification errors. Experts were requested to explain and document their responses.

As a rule, all questions to be answered with yes or no by the country experts were coded 1 = yes or positive with regard to EU integration and 0 = negative with regard to EU integration (labeled "1-0"). If the expert comments and the correspondence with experts suggested intermediate scores, such assessments were coded as 0.5 (labeled "calibration").

Expert assessments: coding examples

Question	Assessment	Score
Is the electoral management commission perceived as impartial, transparent and legitimate by parties and voters? Yes/No	No. According to national opinion polls conducted in 2007, before the pre-term parliamentary elections only 22.3% of the citizens felt confidence in the Central Election Commission, while 43.8% of the citizens felt complete distrust with the CEC. The CEC is generally perceived as transparent. Meanwhile, impartiality of the CEC raise serious doubts, since its members are political appointees.	0
	According to OSCE/ODIHR, "The election administration performed in an overall transparent and professional manner and was perceived as impartial by the majority of stakeholders" during the June 5, 2011 elections, the Central Electoral Commission operated in a transparent and impartial manner and generally enjoyed the confidence of political parties. The level of confidence in electoral bodies at the regional and local level is lower.	1
Are there systems in place to preclude vote buying? Yes/No	Yes, but the system is ineffective. Though under the Electoral Code, a political party or candidate could be de-registered if fact of vote-buying is proved by the court, the system is totally ineffective. In practice, none of the cases of vote buying identified by the political parties and non-governmental organizations and brought to the attention of the election administration and courts have been effectively examined or followed up. In 2011, the Parliament adopted more strict regulations on vote buying in connection with the announcement of political plans by the biollionaire Bidzina	0.5

	Ivanishvili. It is expected that the government will actively apply the new legal provisions against the opposition during future elections	

For items requiring numerical data (quantitative indicators), the figures were coded through a linear transformation, using the information they contain about distances between country scores. The transformation used the following formula:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x refers to the value of the raw data; y is the corresponding score on the 0-1 scale; x_{max} and x_{min} are the endpoints of the original scale, also called "benchmarks". We preferred this linear transformation over other possible standardization techniques (eg., z-transformation) since it is the most simple procedure.

Benchmarks may be based on the empirical distribution or on theoretical considerations, on the country cases examined or on external standards. In the case of the Eastern Partnership Index, this problem is intertwined with the question of the *finalité* of the Eastern Partnership. Whereas the EU refuses to consider accession an option, but tends to expect standards similar to the standards of the accession process, some EaP countries aspire for EU membership. In addition to this uncertain finalité, many items entail a problem of determining unambiguous best or worst practice benchmarks, both in terms of theory and empirical identification. Given these difficulties, we have opted for a mix of empirical and theoretical benchmarks.

For items scored with 0-1 or the intermediate 0.5, benchmarks are defined theoretically by assigning 1 and 0 to the best and worst possible performance. In contrast, benchmarks for quantitative indicators were defined empirically: in the *Linkage* dimension, we assigned 1 and 0 to the best and worst performing EaP country to emphasize the relative positioning of a country vis-à-vis its peers. In the *Approximation* and *Management* dimensions, we defined benchmarks either on the basis of theoretical considerations or based on the performance of other East European countries (including new EU member states) in order to focus on gaps or catching-up relative to this group.

Numerical data sources: coding examples

Item	**Raw data**	**Transformation**	**Score**
Share of commodity trade turnover with the EU, %, last available three-year average	45.5	Linear transformation. Benchmarks defined by best and worst performing EaP countries	$1 = \frac{45.5 - 27}{45.5 - 27}$
Issuing visas by Schengen Area - share in the total number of citizens (%, 2010)	2.34		$0.36 = \frac{2.34 - 0.41}{5.76 - 0.41}$
Vote differential between winner and best-performing rival in most recent presidential elections: Difference between vote shares in percentage points	3.48 percentage points. In the 2010 elections, the winner (Yanukovych) received 48.95% of	Linear transformation, best = 0 (maximum competitiveness), worst = 100 (no competitiveness)	$0.97 = \frac{3.48 - 10_0}{0 - 10_0}$

	the votes, his best performing rival (Tymoshenko): 45.47%.		
Personal autonomy and individual rights (Freedom House, Freedom in the World 2011, subscore) http://www.freedomhouse.org/report/freedom-world-aggregate-and-subcategory-scores	11	Linear transformation, benchmarks defined by best and worst performing EBRD transition countries; best = Estonia (14) worst = Turkmenistan (3)	$0.73 = \frac{11-3}{14-3}$

To construct an Index, that is, a composite indicator, it is necessary to aggregate the individual scores resulting from numerical data and expert assessments. However, aggregation implies decisions about the relative weight of components that need to be explained. The hierarchical structure of the Eastern Partnership Index reflects theoretical assumptions about the components and boundaries of concepts. For example, we define the section *deep and sustainable democracy* as consisting of seven categories: elections; media freedom, association and assembly rights; human rights; independent judiciary; quality of public administration; fighting corruption; accountability. The weights of the seven categories should depend on the importance each category has for deep and sustainable democracy. One could, for example, argue that free and fair elections constitute the core of democracy and should therefore be given a higher weight than the category of association and assembly rights. Conversely, one could also argue that democracy in most EaP countries is mainly impaired by unaccountable governments and lacking media pluralism, while elections are more or less well organized.

Since it is difficult to establish a clear priority of one or several categories over others, we have decided to assign equal weights to all categories. Equal weighting of components is also intuitively plausible since this method corresponds to the conceptual decision of conceiving democracy as composed of five categories placed on the same level. Equal weighting assumes that all components of a concept possess equal conceptual status and that components are partially substitutable by other components.

An arithmetical aggregation of components is, strictly speaking, only possible if components are measured on an interval level, that is, we know that the scores of items, subcategories, categories, sections and dimensions contain information on distances. Most numerical data are measured at interval level: in these cases, we know, for example, that a share of EU exports amounting to 40% of GDP is twice a share of 20% and that this ratio is equal to the ratio between 60% and 30%. For the yes-no questions and items measured with other ordinal scales, we only have information about the ordering of scores, not about the distances between scores.

For example, we do not know the distance between a yes and a no for the question regarding parties' equitable access to state-owned media. Neither do we know whether the difference between yes and no for this question is equivalent with the difference between yes and no for the subsequent question asking whether political parties are provided with public funds to finance campaigns.

In principle, this uncertainty would limit us to determine aggregate scores by selecting the median rank out of the ranks a country has achieved for all components (assuming equal weighting). This would, however, imply omitting the more detailed information contained by the numerical items. To use this information and to put more emphasis on big differences between countries, we have opted to construct quasi-interval level scores by adding the scores of items measured at ordinal level. This has been a standard practice in many indices and could also be justified by the rationale behind equal weighting. Given the frequent uncertainty about the importance of components for aggregate concepts, the safest strategy seems to be assigning equal status to all components. Equal status suggests assuming that a score of 1 used to code a positive response for one question equals a score of 1 for another positive response. Moreover, equal status means that all components constituting a concept are partially substitutable. The most appropriate aggregation technique for partially substitutable components is addition.

Since the number of items differs from subcategory to subcategory and since we want to apply equal weighting, we have standardized the subcategory scores by dividing them through the number of items. Thus, the subcategory score ranges between 1 and 0 and expresses the share of yes-no-questions answered positively in terms of the aggregate concept (and/or the extent to which numerical items or ordinal-level items are evaluated positively).

Quasi-interval level scores allow a range of aggregation techniques at higher levels of aggregation (subcategories, categories, sections and dimensions). The most important methods are multiplication and addition. Multiplication assigns more weight to individual components, emphasizing the necessity of components for a concept; in contrast, addition facilitates the compensation of weaker scores on some components by stronger scores on other components, emphasizing the substitutability of components for a concept.

We apply an additive aggregation of subcategories, categories and sections because this method fits to the method used on the item level, reflects the substitutability of components and is less sensitive with regard to deviating values on individual components. To standardize the aggregate sums and ensure equal weighting, arithmetical means are calculated.