The weight of weighting – an empirical study based on the OECD better life index

Galambosné Tiszberger Mónika
University of Pécs, Faculty of Business and Economics, HUNGARY
Institute of Economics and Econometrics

Key words
well-being measure, weighting, composite indicator, ranking

Abstract

The measurement of the well-being of a country has important roles, such as supporting policy makers and researchers, making comparisons between countries or regions, or informing citizens. The Stiglitz-report in 2009 declared that well-being is a multidimensional phenomenon and any properly selected dimensions should be considered simultaneously. Therefore, most well-known well-being indicators are so called composite indices, where the dimensions are aggregated with a specific weighting scheme.

This paper focuses on the potential weighting schemes and their effect on the final value of indicators together with the rankings they produce. After the description of the main features and elements of well-being measures and aggregating procedures in general, the author introduces the results of an empirical study that is based on the Better Life Index (BLI). The Organization for Economic Co-operation and Development (OECD) proposed the Better Life Initiative in 2011. Since then, the data of 36-38 countries for five years have been made available. BLI assesses well-being of countries on the basis of two domains (material living conditions and quality of life) and eleven dimensions (income and wealth, jobs and earnings, housing, health status, education and skills, work and life balance, civic engagement and governance, social connections, personal security, environmental quality and life satisfaction). Equal weights are applied to each of these dimensions, however, users can select different (0-5) weights based on their personal preferences, which changes the ranking of the countries.

The main findings of this research are the results that indicate the impact of the different combinations of weights. Descriptive statistics and correlation analysis have been applied in order to illustrate the findings of the varying weight simulations. The paper concludes with additional questions that outline possible further research directions.

The ideas and results of this research are the original work of the author and have not been published elsewhere.

Acknowledgement
Supported by the ÚNKP-17-4-III New National Excellence Program of the Ministry of Human Capacities

Introduction

The purpose of creating composite indicators (CI) is to measure some complex, directly unmeasurable phenomenon (such as competitiveness, sustainable development, innovation or well-being), so that it may become possible to support policy related decisions, government programs, regional development, etc. Usually, composite indicators serve the basis of ranking countries or regions or of comparing their performance.

There is no "single-best practice" for creating such indices, as their complexity, their extent and their objectives may differ in many ways. However, the main steps and decision points of this process follow a certain logic, which forms the first part of this paper, with special emphasis on the potential ways of aggregation (weighting). In the second part, the Better Life Index (BLI) of the Organization for Economic Co-operation and Development (OECD) – as an example of CI within

well-being measures – is used to illustrate the importance of weight selection. Results of earlier studies that tackle the methodology of BLI are discussed along with the inferences of weighting simulations in the third section.

The Construction of Composite Indicators

Mostly, it is easier for the general public or policy makers to interpret and understand CIs than it would be to make sense of individual trends and details of numerous indicators. Rankings and the benchmarking of countries are also easier and more transparent through CIs (Saltelli, 2007). They can also assess progress of countries over time. However, CIs may communicate misleading messages through oversimplification or if they are ill-structured. This is what makes CI development a great responsibility.

The main steps of creating a well-structured CI are (OECD, 2008):

Developing the theoretical framework: A sound theoretical framework should clearly define the concept behind the phenomenon that it intends to measure, including its purpose. Based on the multidimensional concept, it is also necessary to determine the sub-groups and the precise selection criteria of potential components.

Selecting the relevant and available indicators: The overall strength of a CI largely derives from the quality of the underlying indicators. There are several criteria that can be used to evaluate data quality (accessibility, timeliness, relevance, soundness, etc.) and ideally all of them should be taken into account. In most cases, official statistics are the best to apply, as they comply with several regulations, and are harmonized in space and time. In case of cross-country comparisons, harmonized data content is especially important. Handling of missing data: There are several statistical methods for the imputation of missing data. However, imputation can only be a solution if the ratio of missing information is reasonably low. And even in these cases the methodology and the extent of the method must be described in a transparent way. Researchers should aim to work with indicators that are available for every necessary observation unit on each level.

Multivariate analysis: The underlying nature and the structure of the selected indicators need to be analyzed carefully. This step helps with identifying redundant variables and it also supports later methodological choices regarding aggregation or weighting. Factor analysis can explore the underlying dimensions so the experts can compare these empirical results with the theoretical concept of the subgroups. Cronbach alpha estimates the internal consistency of the items. Cluster analysis groups the observed elements (e.g. countries), not the indicators, based on the selected set of variables. It tells us which variables are indeed different and helps researchers find countries that are similar to each other. The results of multivariate analyses can verify the theoretical framework and the selected set of variables, or they can provide guidance on how these two basic steps may be reconsidered.

Data normalization: The selected indicators are most likely to have different measurement units or they may be measured at different orders of magnitude. It is not an objective to assign a dominant role to a variable that has the highest values, therefore, a common measurement or scale needs to be created. Certain transformations may also be necessary if the distribution of the indicators is, for instance, highly skewed. There are several normalization techniques (simple ranking, categorical scales, min-max transformation, z-scores, above or below the mean indicators, etc.) and each of these techniques has pros and cons. There are no clear rules to assist the choice – the suitable normalization method has to rely on the properties of the data set and the objectives of the CI.

Deciding about the weighting and aggregation scheme: The simplest weighting technique is the use of equal weights (EW) for each indicator or dimension. Sometimes it is said that there are "no weights" in this case, however, equal weights mean that every dimension is equally important. Budget allocation process builds on the opinion of a specific group (experts, citizens, etc.). They have

a limited budget which they have to spend to cover all given dimensions. The weights are calculated as the average budget. Factor analysis can also be applied to create one factor of the dimensions (like the CI itself) that captures the most of the information content (variance) of the dimensions. It can be declared as an objective tool, however it is only based on the "natural" correlation among the dimensions and it is rarely in line with the theoretical framework. Another disadvantage of this technique is the uniqueness of the weights: if we run factor analysis in a different year or for a different group of countries, the weights will always be different. There are more complicated methods, which may appear less transparent to some (like analytic hierarchy process, unobserved components model, benefit of the doubt approach, data envelopment analysis).

The two main types of aggregation methods are additive and geometric aggregations. Additive methods can be based on ranks or on the distance from a benchmark. Linear aggregation or arithmetic averaging (which may also consider the weights) is also an additive tool. Geometric mean is an option as a geometric aggregation. Its specialty is the handling of compensability. Geometric mean rewards the similar values of the different dimensions compared to an unequally distributed pattern. So it says that being good in one dimension does not compensate for the poor performance of the other aspects.

The main guidance can be the demand to maintain interpretability and to stay as close to the original meaning and nature of the variables as possible.

Decisions on weighting or aggregation schemes usually have a notable impact on the final ranking of the observed units. This fact puts even more responsibility in the hand of the experts. The selection of the weighting and aggregation method has to be justified in an objective manner with clear facts and logic, which can be rather challenging.

Uncertainty and sensitivity analyses: Uncertainty analyses (UA) aim to quantify how uncertainty within the input factors (e.g. selection of subindicators, data transformation, data normalization, weighting scheme, aggregation formula) disseminates through the structure of the CI and affects its value. Sensitivity analysis (SA) is performed to check the robustness of the results taking into account each individual source of uncertainty separately, i.e. how much the final CI values depend on the assumptions of the theoretical model and the choices during the process of CI development. If the dependence is not substantial, then the basic decisions behind the CI are verified. If the dependence is too heavy, then the theory has to be revisited in an even more rigorous way or it may have to be modified. UA is more often adopted than SA and synergistic use is rather rare. (Saisana et al. 2005) There are also different aspects behind the results of sensitivity analyses: "While the robustness of results with respect to changes in weights may be a good sign, it can also signal potential problems with the specification of the components: if components are highly correlated, a composite will always have similar indices, no matter what the weights." (Freudenberg, 2003 p24)

All these steps are important and have to be worked out precisely, however, there is no "best approach" in any case, just a CI, which is good enough in the light of its aims and the possibilities. This is why the choices are usually under heavy criticism.

Better Life Index

CIs are commonly applied in the field of well-being measures. Well-being itself has a number of complex definitions, in which the extent of the necessary variables (both objective and subjective) indicates a wide range. The construction of well-being indicators is usually motivated by the interest in exploring differences between countries, or by the need to support policy related decisions.

The well-known Stiglitz-report (Stiglitz et al., 2009), based on academic research and initiatives developed around the world, suggests a multidimensional definition that describes well-being in their Recommendation 5. It proposes a set of dimensions that should be considered simultaneously. OECD's Better Life Initiative was proposed in 2011 and it leans on the recommendations of the Stiglitz-report in many ways. A thorough literature review and analysis of

the determinants of subjective well-being that is based on the Gallup World Poll also supports and verifies the selected dimensions and variables. (Boarini et al., 2012)

The Better Life Index goes beyond GDP (Gross Domestic Product) and relies on the OECD Measuring Progress Framework which assesses the current well-being of countries on the basis of two domains: material living conditions and quality of life. These two domains are represented by eleven dimensions: income and wealth, jobs and earnings, housing, health status, education and skills, work and life balance, civic engagement and governance, social connections, personal security, environmental quality and life satisfaction. Most of these dimensions are also measured by multiple direct indicators (altogether, 24 variables are included). BLI is complemented by the publication series "How's Life?" which offers a detailed and comprehensive picture of the included countries based on the assessment of the eleven dimensions, together with even more aspects.

On the indicator level, the min-max normalization technique is applied, based on the real range of the different indicators in the given time period. The resulting normalized values vary between zero (the bottom performer) and one (the top performer). It is not clear how they deal with normalization where low indicator values are better. In these cases the ranking is based on a descending order instead of the ascending one. These indicators are: housing expenditures, labor market insecurity, long-term unemployment rate, air pollution, homicide rate and employees working very long hours. (In my research, I have applied these values as 1 – normalized value to be in line with others.)

After normalization, the dimensions can be aggregated. In case of "Your" BLI, where users can set their own weights and preferences, this is implemented in two stages. First, within each dimension, indicators are summed up with equal weights (additive aggregation as the indicators can completely compensate each other). Second, the aggregation (which can be called as weighted linear aggregation or weighted arithmetic mean) across dimensions is done automatically and interactively as the individual users rate the topics. On their interactive homepage, anyone can create their own preferences by assigning arbitrary weights to each of the eleven dimensions. The status of the person's home country and the best country based on the personal preferences update real-time. When rating the topics, individuals select weights as integer values between zero and five. Zero or no rate means that the topic is not important at all, and from one to five, higher values indicate higher importance. This idea is similar to the budget allocation process, however the sum of the given rates has no upper limit, therefore, individuals are not forced to work out their preference list seriously enough. The selected rates are automatically converted into weights that vary from zero to 100%, with the constraint that the sum of weights must be 100%. (Boarini, 2011) Uncertainty and sensitivity analysis results about BLI are not available.

Evaluation of the BLI methodology

In this part the first five general steps of CI construction are considered to be acceptable and reasonable, despite of several critiques and supplementary ideas, especially regarding the question of inequality (Decancq, 2017). The author focuses more on the construction of BLI and the associated scores and ranks. Before the background and the results of the simulations are introduced, conclusions of two other critical studies are summarized, as they complete or support the basic ideas of this paper, however the findings are different from those of this paper.

Von Reumont et al. (2017) discuss the embedding effects in the OECD BLI. They conducted an experiment, where they searched for the difference in the weights because of the different versions of notations. Namely they took the Jobs dimension of BLI (which is built of four indicators), and in the treatment group they replaced it with two dimensions: Job Quality (including personal earnings and job security) and Labor Market (including long-term unemployment rate and employment rate) and created a similar surface for the rating of the dimensions as the OECD BLI website. They kept the original dimension in the control group. They have found that embedding effects truly exist in case of BLI. "Individual preconceived notions of the dimensions seem to

decisively affect subjects' ratings, whereas the underlying indicators are largely ignored." (p16) They also conclude that BLI "does not meet its target of providing societies with a credible order of aspects of quality of life which aggregates citizens' preferences." (p17) However, the main idea was to give citizens a tool to create "their own" BLI, which reflects their personal preferences. This study has also measured and analyzed the amount of time that subjects spent on completing the ratings of BLI dimensions. The final results did not show significant relationship by time, even if they kept those who took less than 45 seconds.

Thinking through the embedding effects I have to add another aspect. Importance mentioned by a citizen is sometimes affected by the actual status of that particular indicator. For instance, if an individual has unlimited access to Internet at home, at work or in school, they might not feel that it is important (as it is available) and mark it lower; or an unemployed person might feel employment is much more important compared to those who have always had a job. Or we might consider Maslow's Hierarchy of Needs (Maslow, 1987). As long as a person has shortage in the physiological or safety stages, they might not evaluate the higher stages important (not because they are not important, but because they seem to be too far or even unattainable). These aspects might also affect the personal ratings and weightings of the dimensions.

Kasparian and Rolland (2012), besides other critical statements about the completeness and scoring of the criteria, systematically tested all possible sets of integer weights from zero to five for each dimension by each country, based on the 2011 dataset (34 countries). They conclude that the innovative framework of Your BLI, as a weighted sum model, defines a quasi-hierarchy among the evaluated countries, as the pairwise comparisons depend little on the selection of the weight set. It is in line with the findings of Von Reumont et al. (2017), as they mentioned that the amount of time spent on the rating of the dimensions did not indicate real effect on the final results, but sheds new light upon them. The reason for the insignificant effect of time spent on rating might simply be the "predefined" quasi-hierarchy of the countries.

Kasparian and Rolland also released the constraint of integer weights of the dimensions. Best possible ranks and the probability of #1 position were calculated. They came to similar conclusions: "The BLI defines a hierarchy of the countries, which is only marginally affected by the choice of the weights" (p4).

My research expands on this idea and evaluates the effect of multiple possible weighting schemes. I work with the most recent data of 38 countries from 2017, available on the OECD website (OECD, 2017). My assumption is that weights can continuously vary and each dimension has a non-zero weight. This latter assumption can be justified by the fact that these dimensions were put together to describe well-being within the theoretical framework of the Stiglitz-report and several other researches, hence, each dimension has at least a low importance, but none of them can be ignored completely. I have generated one set of 11 random values (one for each dimension) using the continuous uniform distribution on [0,1], excluding 0. The relative weight of the random value compared to the sum of the 11 random values formed the final weight in case of each dimension, so that the sum of the final weights is 1. I have repeated this 100 times. Then I used these 100 weight sets for each country, and obtained 100 versions of the BLI for each country. Based on these index values I have computed the ranks for each country in each of the 100 scenarios. The mode, minimum and maximum ranks for each country are shown in Figure 1. The range of the ranks by countries vary from 2 to 16. 15 countries have potential ranks that fluctuate by at least ten positions. As we have altogether 38 countries in our list, the differences can be considered relevant.

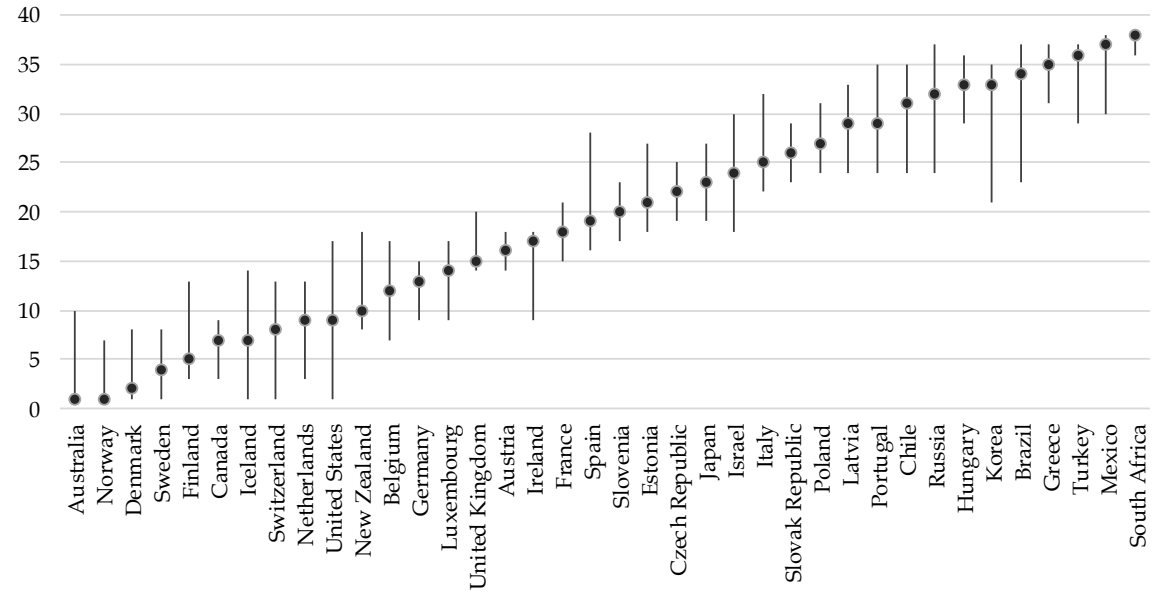

Figure 1: Minimum, maximum and mode of potential ranks by country

On the one hand, only South Africa and Austria have rather fixed ranks, regardless of how the weights are structured. On the other hand, the United States, Brazil and Korea have the chance to gain or lose considerably by a well-established restructuring of the weighting scheme. Kasparian and Rolland (2012) proceeded by looking at the probability of the ranks from a purely mathematical approach. However, subjective opinions or even professional assumptions about the optimal weighting scheme might largely differ from the mathematical chances. So if the question is, say, whether a government can tune BLI to display an order in which the country they represent can get the best possible rank, the answered is "yes" in almost every case (considering the continuous ratio scale of the weights). The answer would only be a real "no" if all dimensions would have completely identical normalized values. If this is not the case, a properly worked out weight set can push the final BLI into the direction of the strongest dimension (or into the direction of the weakest). One might have the feeling, that giving the freedom in the weighting structure creates a similar effect as the geometric mean. Namely a more stable BLI can be achieved if the dimension values are more alike. This statement is supported by the results, as the range of the possible ranks show strong positive correlation ($r=.911$; $p<.001$) with the standard deviation of the dimension values. However a stable BLI does not guarantee a stable position at all, as the same stable score might have higher/lower rank among the countries. According to the analysis, significant negative relationship ($r=-0.482$; $p=.002$) exists between the coefficient of variation (CV) of the dimension values and the CV of the possible ranks.

Based on the results of the empirical analysis, several important inferences can be drawn:

The "freedom" in creating a weighting scheme can carry risk and allows many different orders, which does not reflect or communicate transparency or soundness. It gives the users plenty of rope to manipulate and misinterpret the results. (Mizobuchi, H. (2014) offers the "benefit of the doubt" approach to set final weights to each dimension.)

The selection of the proper weighting scheme in case of this specific indicator puts high responsibility in the hands of the creators. Justified weighting system is to be applied, as the rank of many countries highly depends on it.

As there is no single-best approach, researchers should endeavor to reach consensus by inviting a wide range of potential users of the CI.

It is also important in how the weights are chosen to invite subjective opinions (however not limited to six, integer categories), as the weights will determine the hierarchy of the countries. Well-being is a phenomenon which affects the individuals directly, so their preferences are, indeed, relevant. If we consider the individual or household level data, weighting and aggregation at this lower level can form a possible solution as the basic method in the creation of the overall measure. In this case, the chance of using different aggregation levels (e.g.: county, region) is also given. However, the application of this method is only possible if we work with variables that make sense from the perspectives of the individuals. Another obstacle in this individual level evaluation is the potential lack of interest of most citizens to participate in such surveys.

Conclusions

BLI has many pros and cons. It has a well-prepared theoretical background and collects many important and harmonized variables about the OECD countries. The main idea was to give citizens a chance to create "their own" BLI that reflects their personal preferences. However, the limited range of the weights does not allow to see a clear picture of the preferences, added to the fact that the dimension set pre-defines the hierarchy of the countries within this framework. The "How's Life?" report partially compensates for these deficiencies.

If we consider complete freedom in the determination of the weighting system, it generates different orders and creates a risk of misuse and of misinterpretation. The methodology of BLI did not take and simply passed the responsibility to make decisions about the weighting scheme onto the citizens, while the clear aim of the use of BLI is also questionable.

Developing any CI is a very complicated challenge. A well-defined theoretical framework is essential, and so is a transparent and reliable methodology. The description of the clear objectives and of the intended use of the CI is another crucial element. This can guarantee the application of the proper procedure. As CIs are meant to measure unmeasurable phenomena, there will never be a best, widely accepted concept. However, the consensus of a wide range of potential users might result in the overall adoption of the concept.

References

Boarini, R. Comola, M. Smith, C. Manchin, R. de Keulenaer, F. (2012). *What Makes for a Better Life?: The Determinants of Subjective Well-Being in OECD Countries – Evidence from the Gallup World Poll OECD Statistics Working Papers*, No. 2012/03, Paris: OECD Publishing

Boarini, R. (2011). *Designing Your Better Life Index from a methodological perspective* http://wikiprogress.org/designing-your-better-life-index-from-a-methodological-perspective/ Cited 15 Dec 2017

Decancq, K. (2017). *Measuring Multidimensional Inequality in the OECD Member Countries with a Distribution-Sensitive Better Life Index* Social Indicators Research. Volume 131, Issue 3, pp 1057–1086

Freudenberg, M. (2013). *Composite Indicators of Country Performance: A Critical Assessment* OECD Science, Technology and Industry Working Papers. 16. Paris: OECD Publishing

Kasparian, J. and Rolland, A. (2012). *OECD's 'Better Life Index': can any country be well ranked?* Journal of Applied Statistics 39(10), pp. 2223–2230.

Mizobuchi, H. (2014): *Measuring World Better Life Frontier: A Composite Indicator for OECD Better Life Index* Social Indicators Research 118(3), pp. 987–1007.

Maslow, A. H. (1987). *Motivation and personality (3rd ed.)* Delhi, India: Pearson Education.

OECD (2008). *Handbook on Constructing Composite Indicators – Methodology and User Guide* OECD 158 p.

OECD (2017) http://stats.oecd.org/Index.aspx?DataSetCode=BLI Cited 5 Jan 2018

von Reumont, L. Schöb, R. and Hetschko, C. (2017). *Embedding Effects in the OECD Better Life Index*. Beiträge zur Jahrestagung des Vereins für Socialpolitik 2017: Alternative Geld- und Finanzarchitekturen - Session: Political Economy II, No. E20-V1

Saisana M. Saltelli A. and Tarantola S. (2005). *Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators* Journal of the Royal Statistical Society Series A vol. 168, issue 2, 307-323.

Saltelli, A. (2007). *Composite indicators between analysis and advocacy* Social Indicators Research, 81(1), 65-77.

Stiglitz, J. E., Sen, A., Fitoussi, J-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*

https://www.uio.no/studier/emner/sv/oekonomi/ECON4270/h09/Report Cited 15 Oct 2017