Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

OECD's 'Better Life Index': can any country be well ranked?

Jérôme Kasparian & Antoine Rolland

To cite this article: Jérôme Kasparian & Antoine Rolland (2012) OECD's 'Better Life Index': can any country be well ranked?, Journal of Applied Statistics, 39:10, 2223-2230, DOI: 10.1080/02664763.2012.706265

To link to this article: http://dx.doi.org/10.1080/02664763.2012.706265

Published online: 13 Jul 2012.

Journal of Applied Statistics
Vol. 39, No. 10, October 2012, 2223–2230

OECD's 'Better Life Index': can any country be well ranked?

Jérôme Kasparian[a]* and Antoine Rolland[b]

[a]*Université de Genève, GAP-Biophotonics, Chemin de Pinchat 22, CH-1211 Geneva 4, Switzerland;*
[b]*Laboratoire ERIC, Université Lumière Lyon 2, 69676 BRON cedex, France*

(*Received 25 July 2011; final version received 21 June 2012*)

We critically review the Better Life Index (BLI) recently introduced by the Organization for Economic Co-operation and Development (OECD). We discuss methodological issues in the definition of the criteria used to rank the countries, as well as in their aggregation method. Moreover, we explore the unique option offered by the BLI to apply one's own weight set to 11 criteria. Although 16 countries can be ranked first by choosing *ad hoc* weightings, only Canada, Australia and Sweden do so over a substantial fraction of the parameter space defined by all possible weight sets. Furthermore, most pairwise comparisons between countries are insensitive to the choice of the weights. Therefore, the BLI establishes a hierarchy among the evaluated countries, independent of the chosen set of weights.

Keywords: well-being; aggregation; ranking; scoring; parameter space

1. Introduction

In May 2011, the Organization for Economic Co-operation and Development (OECD) proposed a new well-being index named 'Better Life Index' (BLI) [9]. Following a tradition of research in multi-crietria evaluation in the economic and social fields [8], this index aims at offering an alternative to the gross domestic product (GDP) to compare countries, taking into account not only the global amount of their wealth, but also well-being indicators.

The BLI evaluates the 34 member states of OECD on 11 criteria, such as housing, income, education, governance, etc. Each criterion is evaluated on a scale ranging between 0 and 10. A global country score is obtained by a weighted mean of the criteria. As emphasized by OECD, the innovative aspect of the BLI is the possibility offered to anyone to choose her/his own weights (as integer values between 0 and 5) in order to represent her/his own preferences on well-being indicators: 'The OECD is NOT deciding what makes for better lives. YOU decide for yourself' [9].

In this paper, we present a critical review of this statement. In Section 2, we discuss biases due to the way how the criteria are built, as has been done for the Shanghai index of universities ranking [1] or various other composite indicators [2]. In Section 3, we investigate whether and

*Corresponding author. Email: jerome.kasparian@unige.ch

ISSN 0266-4763 print/ISSN 1360-0532 online

http://dx.doi.org/10.1080/02664763.2012.706265
http://www.tandfonline.com

how an appropriate set of weights allows a given country to be ranked first and analyze the most frequent rankings in the space of all the possible weights. By doing so, we exhibit an implicit and quite rigid country ranking established by the BLI, merely independent of the chosen set of weights.

2. Critical discussion of the criteria

It is well known that an indicator is always a partial view of the reality: its choice is then a matter of personal preference, i.e. of political approach. Therefore, it is always criticizable. In this regard, the option offered by the BLI to arbitrarily chose the weightings and/or select specific criteria among the 11 proposed ones constitutes a progress. For this reason, we will not discuss here the choice of the 11 criteria included by OECD in the BLI. However, even within this framework, technical criticism should be made about the construction of the indicators and associated scores.

2.1 *Completeness of the criteria and indicators definition*

The content of the criteria offered for inclusion into the BLI could be enhanced. For example, the 'Environment' criterion includes a single indicator, namely the average number of PM10 (particulate matter of aerodynamic diameter above 10 μm) in cities above 100,000 inhabitants. Other aspects of the environmental quality, such as other air pollutants, either urban or background, biodiversity preservation, water quality, CO_2 emissions and so on, are not considered. Scoring the environmental criterion with reference to a unique indicator, instead of considering the complexity of the matter of interest, could yield results misleading to the non-expert users.

2.2 *Scoring of the criteria*

OEDC has produced and/or compiled the wide set of data needed to evaluate each indicator for each member state. But these data are hard to compare directly since they have heterogeneous scales. Furthermore, some of these indicators have to be maximized, while others should be minimized: this is for example the case for income and atmospheric pollution, respectively. Best practices in the field of multicriteria decision-making use specific methods to fix trade-off values between criteria, taking into account scale differences (see, for example, [4,5]). In the BLI, the score of each criterion is normalized by a ratio method, i.e. it is obtained by applying to the scores of each indicator a linear function that scales to 0 and 10, respectively, the worst and best country scores for the considered indicator.

The scores of a country therefore do not constitute an absolute measurement of its performance. They are rather relative to that of the best and worst countries for this indicator. Consequently, a country can have a bad score on an indicator not because its performance is intrinsically bad, but because one or several other countries have better performances in the considered domain. If the actual performances are almost equivalent, slight differences will result in artificially large contrasts in the scores. This is the case for the indicator 'Time devoted to leisure and personal care' of the 'Work-life balance' criterion, where all countries have very similar performances. The normalization sets the score of Mexico to 0 because of a relatively modest difference of only 2 h weekly with the best countries. This behavior prevents any clear interpretation of the performance of a country or of its temporal evolution, and limits the relevance of the BLI to comparisons between countries at a given time.

As can be seen from this example, due to the use of relative scores, pairwise comparisons between two countries depend not only of their respective performances on each indicator, but also on the performances of all other countries since the latter can reduce (resp. increase) the dynamics of any criterion by compressing (resp. expanding) the corresponding ranking scale through the above-discussed normalization. The relative ranking of two countries can even be influenced by the performances of a third one, as exemplified in Appendix 1.

Finally, the scaling into the 0–10 range is applied to each indicator independently rather than on the composite criteria. When several indicators are aggregated in a criterion, the global score for this criterion is the mean of the individual scores of the indicators. Since this mean is not subsequently renormalized, the scores of multi-indicator criteria do not span over the full 0–10 range, and hence lose dynamics. Its effective weight in the BLI is reduced as compared with mono-indicator criteria, which use the full range of scores from 0 to 10. Although this effect can be compensated by adequately overweighting the multi-indicator criteria, the omission to specify this limitation may be misleading, especially to non-expert users.

2.3 *Global score construction*

The global index score of each country is obtained by a weighted mean of the scores of all criteria. The originality of OECD's BLI is to let the user choose its own weights. While this aggregation technique has the advantage of being easily understandable by all users including non-experts, it is known to strongly constrain the possible rankings, especially because it allows countries with heterogeneous profiles to stay well ranked. Although this discussion lies beyond the scope of our present paper, it would be of great interest to study alternative aggregation procedures [7] relying on the raw data, which are available directly from the OECD web site [10]. Such improved aggregation procedures with more desirable properties [3,6] include the *min* operator, which favors equilibrated countries, the OWA operator for which the weights are associated with the values and not to the criteria, or the Choquet integral which can even model specific interactions between topics, as can, e.g. be expected between health and air quality, and more generally living conditions.

The use of integer weights over a short range (between 0 and 5) is also welcome as a simple approach, but limits the dynamics as well as the possibility of fine tuning of the weights. We shall examine the effect of the discretization of the weights in the next section.

3. Which country is whatsoever the most pleasant place to live?

In spite of the limitations discussed in Section 2, we shall discuss here the relevance of OECD's slogan within the framework defined by the specification of the BLI and investigate to which extent the choice of the weights of each criterion influences the ranking of the countries. More precisely, we focus our study on two mains aspects, corresponding to complementary points of view.

First, can the BLI be instrumented by, e.g. a government to exhibit an order in which his/her country is ranked as well as possible? In other words, can one find a set of weights optimizing the ranking of a given country?

Conversely, assuming that one has no preference among the proposed criteria, one could weigh them randomly. What would be the influence on the ranking? We addressed this question by computing the probability for each country to be ranked first, or to be ranked better than a given other one for random weights.

3.1 *Optimizing ranking of a predefined country*

Let us first observe that, among the 34 member states of OECD, 12 are Pareto-dominated by at least another one, i.e. they have lower scores on each of the 11 criteria. Therefore, they cannot be ranked first, regardless of the considered set of weights. This limits *ad hoc* manipulations of the BLI.

Table 1. Optimized score and ranking of each country, as determined by linear programming and considering continuously varying weights.

Country	Best score relative to others	Best possible ranking	Probability of # 1 ranking
Canada	0.97	1	0.47
Australia	1.25	1	0.39
Sweden	0.63	1	0.10
Iceland	1.05	1	1.1×10^{-2}
New Zealand	0.40	1	7.0×10^{-3}
Switzerland	0.81	1	7.0×10^{-3}
United States	2.16	1	6.4×10^{-3}
Denmark	0.83	1	5.4×10^{-3}
Norway	0.44	1	4.1×10^{-3}
Finland	0.63	1	ε
Netherlands	0.14	1	ε
Belgium	0.22	1	ε
Ireland	0.11	1	ε
Japan	0.40	1	ε
Korea	4.5×10^{-3}	1	ε
United Kingdom	0.10	1	ε
Austria	−0.08	2	0
France	−0.18	2	0
Estonia	−0.09	2	0
Germany	−0.06	2	0
Slovak Republic	−0.27	3	0
Poland	−0.28	3	0
Luxembourg	−0.30	3	0
Israel	−0.61	6	0
Spain	−0.69	6	0
Chile	−1.27	8	0
Slovenia	−0.49	9	0
Hungary	−0.58	9	0
Czech Republic	−0.75	9	0
Greece	−0.76	10	0
Portugal	−0.97	10	0
Italy	−0.77	11	0
Mexico	−2.06	14	0
Turkey	−2.34	15	0

Notes: The best score relative to others is the maximum achievable difference between the considered country and the best among the other ones. ε denotes probabilities below 10^{-4}, where the Monte-Carlo algorithm may provide unreliable probabilities.

To go beyond this simple observation, we maximized the score difference with the best of the remaining population (Table 1, second column). For a country A, this score difference is defined as the difference between the score of A, and (i) if A can be ranked first, the country ranked second when using the same set of weights or (ii) if A cannot be ranked first, the country ranked first when using the same set of weights. In this optimization, which we performed using linear programming, the free parameters are the weights of the 11 criteria. We made the problem continuous by releasing the constraint of integer weights, allowing them to vary continuously between 0 and 1. To avoid convergence to ‘pathological’ solutions (e.g. reducing all weights toward zero, in which case all countries would be ranked first *ex aequo*), as well as to allow comparing the score differences between countries (see Column 2 of Tables 1 and 2), the sum of the weights was normalized to 1.

Optimizing the score difference allows a robust optimization since the optimizing observable is continuous. But its interest is mostly restricted to countries that can be ranked first or second. To gain information about the other countries, we minimized the ranking. As displayed in the third

Table 2. Systematic test of all possible set of integer scores between 0 and 5 for each country.

Country	Best score relative to others	Best possible ranking	Probability of # 1 ranking
Canada	0.95	1	0.52
Australia	0.84	1	0.34
Sweden	0.34	1	8.9×10^{-2}
Iceland	0.35	1	1.4×10^{-3}
New Zealand	0.044	1	8.1×10^{-6}
Switzerland	0.54	1	1.0×10^{-2}
United States	0.66	1	3.1×10^{-3}
Denmark	0.80	1	2.8×10^{-2}
Norway	0.26	1	8.7×10^{-3}
Finland	0.14	1	9.3×10^{-6}
Netherlands	−0.014	2	0
Belgium	−0.51	7	0
Ireland	−0.070	2	0
Japan	−1.2	12	0
Korea	−1.3	13	0
United Kingdom	−0.44	6	0
Austria	−0.33	5	0
France	−0.68	10	0
Estonia	−2.2	22	0
Germany	−0.65	10	0
Slovak Republic	−1.4	17	0
Poland	−1.7	18	0
Luxembourg	−0.63	8	0
Israel	−0.73	6	0
Spain	−1.4	16	0
Chile	−2.3	17	0
Slovenia	−1.5	17	0
Hungary	−2.4	26	0
Czech Republic	−1.5	17	0
Greece	−2.0	21	0
Portugal	−2.6	24	0
Italy	−1.5	17	0
Mexico	−2.3	14	0
Turkey	−3.9	28	0

Notes: The best score relative to others is the maximum achievable difference between the considered country and the best among the other ones. Countries are sorted in the same order as in Table 1.

column of Table 1, 16 countries, i.e. almost half of the OECD, can be ranked first if adequate weight sets are chosen. However, strong discrepancies are observed in the differences of scores with the best of all other countries. Two-third of the OECD countries can be ranked in the first three positions. Conversely, the poor optimum ranking of several countries evidences that some countries are intrinsically better ranked in the BLI than others and illustrates the limitations to the possible manipulations.

We then investigated the impact of discrete weights on this result, by comparing it with the systematic exploration of the 6^{11} possible sets of integer weights between 0 and 5. The sum of weights of each set is then normalized to 1 in order to allow comparison with Table 1. Comparing the results displayed in Tables 1 and 2 shows that the discrete weights strongly constrain the ranking, and to a lesser extent the score differences. They prevent six countries from being ranked first. Among them, Japan and Korea can be only be ranked 12th and 13th at best with discrete weights. This discrepancy illustrates that imposing discrete weights substantially influences the freedom of the user to fine tune her/his preferences or to manipulate the ranking.

3.2 *Robustness analysis*

These results clearly show that, provided one accepts the criteria defining the BLI, the OECD member states do not all provide the same well-being, regardless of the weights being assigned to the BLI criteria. In other words, the BLI defines a hierarchy of the countries, which is only marginally affected by the choice of the weights. To get more insight into this hierarchy, we explored the $[0;5]^{11}$ parameter space defined by the 11 weights, by performing Monte-Carlo calculations. Such an approach is classical for exploring parameter spaces too large to allow a systematic investigation. A textbook example of their use is the estimation of a sub-volume of a space, e.g. to estimate the volume of a sphere and estimate the value of π [11]. In our work, over 430 million sets of weights (i.e. of coordinates in the parameter space) were randomly chosen in the $[0;5]^{11}$ parameter space. For each set, each weight was randomly chosen in the [0;5] interval and the corresponding BLI ranking was calculated for each country. Among the resulting sample of 430×10^6 rankings, we computed the probability of each country to be ranked first and collected its best ranking and relative score relative to other countries. These Monte-Carlo computed values provide approximations of the real ones. Owing to the extremely large number of trials, the 95% confidence intervals are minimal, as the precision for each proportion is $\pm 4.7 \times 10^{-5}$ in the worst case.

As shown in the last column of Table 1, only three countries are ranked first over a substantial fraction of the parameter space: Canada (47%), Australia (39%) and to a lesser extent Sweden (10%), for a total of 96%. Thirteen other countries can only be ranked first at the cost of the

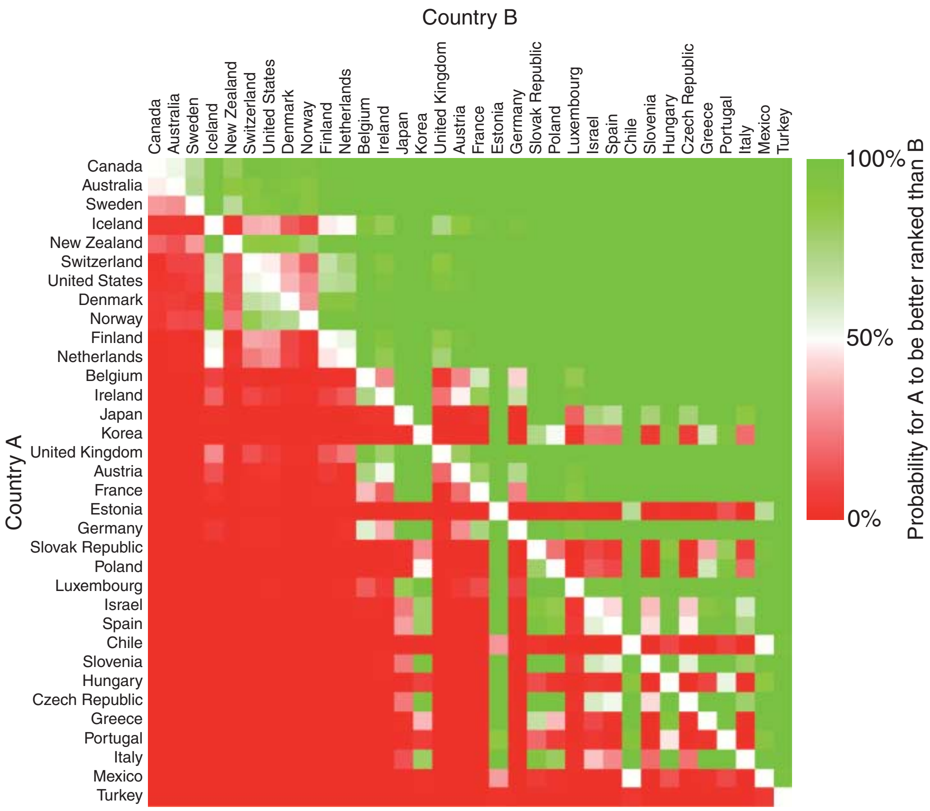

Figure 1. Preference probability for each pair of countries. On the axis, countries are sorted according to their probability within the parameter space to be ranked first, then according to their best ranking (see Table 1).

choice of a very specific set of weights, as evidenced by the small, or even infinitesimal volume of the parameter space in which their scores exceeds that of the other countries. This meta-view of the 'neutral' approach of OECD confirms the hierarchy between the countries set by the BLI and contradicts to a large extent OECD's statement that the BLI does not intrinsically bear a predetermined country ranking.

Another approach in trying to rank the countries is to determine, for a given pair of them, the respective volumes of the weight space where one is preferred to the other, i.e. it obtains a better score. The result obtained for all pairs of countries using 100,000 random weight sets in a Monte-Carlo calculation, offering a 95% confidence interval of ±0.003, is displayed in Figure 1. The most striking feature is that the pairwise comparisons depend little on the choice of the weight set. Over 528 non-trivial pairs, the preference of 262 ones (49.6%) cannot be inverted by the choice of the weight set, while for another 312 of them (59%), the preference in one direction is achieved over less than 0.1% of the parameter space. Conversely, only 33 pairs (6.3%) share the parameter space within 40–60% of preference probability. This confirmation of an implicit underlying hierarchy of the countries is evidenced in Figure 1 by the large areas of the graph with intense red or green colors, as opposed to light values. Furthermore, the hierarchy exhibited from the pairwise preference largely overlaps that issued from overall ranking, as evidenced by the segregation of preferences over the diagonal of Figure 1.

4. Conclusion

As a conclusion, we critically reviewed the BLI recently introduced by the OECD. Methodological issues include the use of relative scores rather than absolute ones and aggregation of the several indicators used for each criterion. Some criteria could also be enriched by the addition of complementary indicators. Furthermore, the constraint for integer weights clearly constrains the rankings.

We also explored the unique option offered by the BLI to the users to choose their own weight sets between 11 criteria. Although adequate weightings allow 16 countries to be ranked first, only Canada, Australia and Sweden do so over a substantial fraction of the parameter space defined by all possible weight sets. Furthermore, most pairwise comparisons between countries are either totally or highly insensitive to the choice of the weight set. Therefore, the choice of these weights does not affect much the country ranking. Rather, as a weighted sum model, the BLI defines a quasi-hierarchy among the evaluated countries.

References

[1] J.-C. Billaut, D. Bouyssou, and Ph. Vincke, *Should you believe in the Shanghai ranking? An MCDM view*, Scientometrics 84 (2010), pp. 237–263.

[2] D. Bouyssou, T. Marchant, M. Pirlot, P. Perny, A. Tsoukias, and P. Vincke, *Evaluation and Decision Models: A Critical Perspective*, Springer, Heidelberg, 2001.

[3] M. Grabisch, J.L. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions*, Cambridge University Press, Cambridge, UK, 2009.

[4] E. Jacquet-Lagrèze and J. Siskos, *Assessing a set of additive utility functions for multicriteria decision-making, the UTA method*, Eur. J. Oper. Res. 10 (1982), pp. 151–164.

[5] R.L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*, Wiley, New York, 1976.

[6] J.L. Marichal, *Aggregation functions for decision making*, in *Decision-Making Process – Concepts and Methods*, D. Bouyssou, D. Dubois, M. Pirlot, and H. Prade, eds., ISTE/John Wiley, New York, 2009, pp. 673–721.

[7] P. Meyer and G. Ponthière, *Eliciting preferences on multiattribute societies with a Choquet integral*, Comput. Econom. 37 (2011), pp. 133–168.

[8] G. Munda, *Social Multi-criteria Evaluation for a Sustainable Economy*, Springer, Heidelberg, New York, 2008.

[9] OECD, available at http://www.oecdbetterlifeindex.org/wpsystem/wp-content/uploads/2011/05/YourBetterLifeIndex_ExecutiveSummary2.pdf
[10] OECD, available at http://www.oecdbetterlifeindex.org/wpsystem/wp-content/uploads/2011/06/BetterLifeIndex_Data_2011V6.xls
[11] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 2007.

Appendix 1. Example 1

A sentence like 'life is better in New Zealand than in Finland. But if South Africa and Canada are making efforts to reduce unemployment, life will be better in Finland than in New Zealand' seems dummy. However, this counterintuitive behavior is induced by the way scores on each topic are calculated in the BLI, whatever the weight on each topic. The following example illustrates this possibility.

Let us consider the scores of four fictious countries on two indicators, and the corresponding normalized scores on the topics. In this example, we consider equal weights for both topics:

	Indicator data		Scores		
Country	Topic 1	Topic 2	Topic 1	Topic 2	Global index
Country A	100	100	10	10	10
Country B	55	40	5.5	4	4.75
Country C	45	60	4.5	6	5.25
Country D	0	0	0	0	0

Country C is ranked before country B with regard to the global index. Assume now that the situation has changed only on countries A and D as follows:

	Indicator data		Scores		
Country	Topic 1	Topic 2	Topic 1	Topic 2	Global index
Country A	100	200	10	10	10
Country B	55	40	4.375	2	3.18
Country C	45	60	3.125	3	3.06
Country D	0	20	0	0	0

Absolute performances of countries B and C have not changed, but country B is now better ranked than country C.