Business Research
https://doi.org/10.1007/s40685-018-0077-z

ORIGINAL RESEARCH

Is data envelopment analysis a suitable tool for performance measurement and benchmarking in non-production contexts?

Victoria Wojcik[1] · **Harald Dyckhoff**[1] · **Marcel Clermont**[2]

Received: 29 November 2017 / Accepted: 11 September 2018

Abstract After 40 years of research with thousands of application-oriented scientific papers, empirical evidence that data envelopment analysis (DEA) has really improved the practice of performance measurement and benchmarking in real-life non-production contexts is rare. The main reason for this deficit may be that DEA is founded on the concepts of production theory such as production possibility set or returns to scale. These concepts can hardly be applied to pure multiple-criteria evaluation problems, which are often attempted to be solved using DEA. This paper systematically investigates strengths and weaknesses of DEA in the exemplary case of welfare evaluation using real data on 27 countries of the European Union. We analyze and explain the differences in the results of various frequently used DEA models for two different, but strongly connected sets of welfare indicators, thereby demonstrating the pitfalls, which often arise in the application of DEA, as well as some approaches for avoiding them.

Keywords Benchmarking · Performance measurement · Data envelopment analysis application · Model choice and building · Indicator selection · Welfare evaluation

✉ Harald Dyckhoff
dyckhoff@lut.rwth-aachen.de

Victoria Wojcik
victoria.wojcik@rwth-aachen.de

Marcel Clermont
m.clermont@tu-braunschweig.de

[1] School of Business and Economics, RWTH Aachen University, Templergraben 64, 52062 Aachen, Germany

[2] Institute of Management Control and Business Accounting, TU Braunschweig, Fallersleber-Tor-Wall 23, 38102 Braunschweig, Germany

Published online: 22 September 2018

JEL Classification C61 · C67 · I31

1 Introduction

Having been researched for 40 years, data envelopment analysis (DEA) still appears to be an ever-growing field. Up to 2015, the Web of Science exhibits 10,720 DEA publications for this topic, with 1020 entries for the year 2015 alone (Wojcik et al. 2017). The majority considers applications of DEA to performance evaluation and benchmarking of various areas. However, even though DEA is essentially build on the foundations of production theory (cf., e.g., Charnes et al. 1985 or Färe et al. 1994), many of these applications do not take place in a context of real production. This involves several difficulties, e.g., the question of how to select and define inputs and outputs which are relevant for the performance analysis at hand. With respect to this particular crucial open question (Cook et al. 2014, p. 2) state in a recent methodological review on 'DEA: Prior to choosing a model':

> "In summary, if the underlying DEA problem represents a form of 'production process', then 'inputs' and 'outputs' can often be more clearly identified. The resources used or required are usually the inputs and the outcomes are the outputs.
>
> If, however, the DEA problem is a general benchmarking problem, then the inputs are usually the 'less-the-better' type of performance measures and the outputs (…) the 'more the-better' type (…). DEA then can be viewed as a multiple-criteria evaluation methodology where DMUs are alternatives, and the DEA inputs and outputs are two sets of performance criteria where one set (inputs) is to be minimized and the other (outputs) to be maximized."

Indeed, DEA is a powerful performance measurement and benchmarking tool for applications where the evaluated 'decision-making units' (DMUs) are described by activities representing real processes which generate products or services and are based on a convex (or even linear) technology. Otherwise, however, serious doubts can be cast on the proposition that DEA can really "be viewed as a multiple-criteria evaluation methodology" which is, in general, appropriate to derive resilient information for benchmarking the performance of DMUs. At least, some further reflections and a solid reasoning for applying DEA are necessary in these cases, if the results of DEA should be accepted as valid at all. To the best of our knowledge, such reflections are rare in the DEA literature (cf. Dyson et al. 2001 for some general remarks on pitfalls and protocols).

This paper demonstrates the problems and difficulties that occur with such DEA applications to explain why the application in non-production contexts might often not lead to the derivation of empirically valid results in practice. To provide concrete illustrations of an awkward use of DEA in the creation of performance indexes, we use the example of welfare evaluation of 27 countries of the European Union. Despite the exemplary nature of our investigation, the conclusions for this specific application field can be viewed as characteristic for other non-production

contexts, too. Our goal is to raise more awareness for the distinct performance and benchmarking results that may be obtained when applying different variants of DEA models and when modifying the selected inputs and outputs. This is of particular importance when the model assumptions cannot be easily verified by the nature of the production system considered.

The paper is structured as follows: Illustrated by numerical examples, Sect. 2 gives a short overview of standard DEA models and procedures, explains pitfalls of their application, and presents some useful extensions. While these considerations are of general interest for the application of DEA, the following sections concentrate on specific difficulties usually appearing in non-production contexts, here exemplarily demonstrated with case studies on welfare evaluation. To begin with, Sect. 3 summarizes typical procedures and some pitfalls of applying DEA to this area of performance measurement. The case study of Sect. 4 then identifies concrete problems involved in the application of DEA standard models to the 'Prosperity Quintet', consisting of five relative welfare key indicators of European countries. To solve these problems, in a first step, the same standard models are applied to a modified data set in Sect. 5, whereby the five key indicators are converted into six more basic ones. In a second step, in Sect. 6, we use the DEA model extensions of Sect. 2 for this basic set of welfare indicators to analyze and explain the differences in the results of all DEA models used. While the studied cases demonstrate the pitfalls which typically arise during the application of DEA to non-production contexts, as well as approaches for avoiding them, Sect. 7 summarizes our key findings on the advantages and disadvantages of applying DEA in practice to such areas of performance measurement and benchmarking. Section 8 concludes with an outlook on possibilities for and requirements of future research.

2 Standards, pitfalls, and extensions of DEA

We start with an introduction of the well-known radial CCR and BCC models and their main features.[1] Then, we discuss a specific additive model which avoids some pitfalls of the prior models, however, at the expense of other disadvantages. Third, we present a recent approach which allows to measure the balance or specialization of a DMU's performance within the usual DEA framework.

2.1 Oriented radial measurement of efficiency

Most DEA applications refer to the radially oriented models as a standard (cf. Kerpen 2016). The pioneer model by Charnes et al. (1978), the so-called *CCR model*, reads as follows in its output-oriented variant (CCR-O):

[1] For more comprehensive introductions to DEA, see, e.g., Färe et al. (1994), Coelli et al. (2005) or Cooper et al. (2007). In this paper, we concentrate on the envelopment form of the DEA models and do not discuss their multiplier form which is received by building the dual Linear Programming model.

$$\begin{aligned} &\text{Maximize } \eta_{\mathrm{o}} \\ &\text{such that } \sum_{j=1}^{n} \lambda_j \cdot x_{ij} \leq x_{\mathrm{io}} \quad \forall i \\ &\sum_{j=1}^{n} \lambda_j \cdot y_{rj} \geq \eta_{\mathrm{o}} \cdot y_{\mathrm{ro}} \quad \forall r \\ &\lambda_j \geq 0 \quad \forall j \end{aligned} \tag{1}$$

and in its input-oriented variant (CCR-I):

$$\begin{aligned} &\text{Minimize } \theta_{\mathrm{o}} \\ &\text{such that } \sum_{j=1}^{n} \lambda_j \cdot x_{ij} \leq \theta_{\mathrm{o}} \cdot x_{\mathrm{io}} \quad \forall i \\ &\sum_{j=1}^{n} \lambda_j \cdot y_{rj} \geq y_{\mathrm{ro}} \quad \forall r \\ &\lambda_j \geq 0 \quad \forall j \end{aligned} \tag{2}$$

Here, x_{ij} denotes the quantity of the inputs $i = 1,\ldots,m$, y_{rj} that of the outputs $r = 1,\ldots,s$ and λ_j denotes the activity level of the DMUs $j = 1,\ldots,n$.

In the CCR-O case, the value of the objective function of the linear program (LP) as optimization problem indicates the productivity factor η_{o} (greater than one) by which all outputs of the currently considered DMU o can be proportionately (i.e., radially) increased without decreasing the input. Here, we will use its reciprocal value $\theta_{\mathrm{o}} = 1/\eta_{\mathrm{o}}$ instead, namely as an efficiency score between zero and one. In the CCR-I case, the value of the objective function indicates the rationalization factor θ_{o} by which all inputs can be proportionately decreased without reduction of the outputs. For the two CCR model variants (1) and (2), both efficiency scores are identical due to the intercept theorem of geometry. Then, the reciprocal value of the maximum possible productivity increase is equal to the minimum rationalization factor: $\theta_{\text{CCR-O}} = \theta_{\text{CCR-I}} = :\theta_{\text{CCR}}$ (Thanassoulis et al. 2008: 263). However, the specific solutions of both LPs for the dominant efficient combination of other DMUs—as the so-called reference units or benchmarking partners—can differ and, thus, suggest different benchmarks as target values.

Both variants of the CCR model exhibit constant returns-to-scale (cRTS), whereas the *BCC model* developed by Banker et al. (1984) is defined for variable returns-to-scale (vRTS) in its output- and input-oriented variants. This second standard model differs from the LPs (1) and (2) only by an additional restriction on the activity levels λ_j of the following form:

$$\tau_{\min} \leq \sum_{j=1}^{n} \lambda_j \leq \tau_{\max} \tag{3}$$

For the BCC model variants, the following applies: $\tau_{\min} = \tau_{\max} = 1$. Radial-oriented models with non-increasing (niRTS) or non-decreasing (ndRTS) returns-to-scale are seldom used. For niRTS, restriction (3) applies in addition to (1) and (2) with $0 \leq \sum_{j=1}^{n} \lambda_j \leq 1$, i.e., $\tau_{\min} = 0$ and $\tau_{\max} = 1$; for ndRTS with $1 \leq \sum_{j=1}^{n} \lambda_j < \infty$, i.e., $\tau_{\min} = 1$ and $\tau_{\max} = 1/\varepsilon$ (with $0 < \varepsilon \ll 1$, i.e., ε is infinitesimally small). Due to the varying strengths of restrictions (3) in comparison of the various RTS properties,

the following inequalities generally apply for the (optimal) efficiency scores of the relevant LPs:

$$\theta_{\text{BCC}-\text{O}} \geq \theta_{\text{niRTS}-\text{O}} \geq \theta_{\text{CCR}} \leq \theta_{\text{niRTS}-\text{I}} \leq \theta_{\text{BCC}-\text{I}} \tag{4}$$

$$\theta_{\text{BCC}-\text{O}} \geq \theta_{\text{ndRTS}-\text{O}} \geq \theta_{\text{CCR}} \leq \theta_{\text{ndRTS}-\text{I}} \leq \theta_{\text{BCC}-\text{I}} \tag{5}$$

Therefore, the CCR models are the most rigorous ones, because they compare the respective DMU with all linear combinations of the other DMUs, and not only with their convex combinations as is the case with the BCC models. In the literature of DEA, the ratios SE_O: = $\theta_{\text{CCR}}/\theta_{\text{BCC-O}}$ and SE_I: = $\theta_{\text{CCR}}/\theta_{\text{BCC-I}}$ are called *scale efficiency* regarding the associated output or input orientation (cf. Banker et al. 1984). Because of (4) and (5), $0 \leq \text{SE} \leq 1$ holds, attaining the maximum of 100% if the considered DMU is both CCR- as well as BCC-efficient.

The eight mentioned model variants and their relations are now demonstrated with a simple numerical example of six DMUs A,…,F which produce one output y with exactly one input x. This example is displayed in Fig. 1. The corresponding data as well as the DEA results are shown in Table 1.

Columns 2 and 3 of Table 1 contain the corresponding input and output quantities. The ray in Fig. 1 spanning from the origin through DMUs B and C marks the efficient frontier of the linear envelopment of the six DMUs in the CCR model. Therefore, the four remaining DMUs are (CCR) inefficient. Their (in)efficiency scores θ_{CCR} are given in the fourth column of Table 1 as percentages, rounded to integer values (as for all such scores in the following tables, too). Though, Fig. 1 also indicates that DMUs A and D are in fact (BCC) efficient if different properties are assumed for the underlying production technology concerning the data envelopment of the six DMUs, in this case convexity and strong disposability according to the BCC model.

Columns 5 and 6 in Table 1 contain the input- and output-oriented efficiency scores $\theta_{\text{BCC-O}}$ and $\theta_{\text{BCC-I}}$ as well as the scale efficiencies SE_O and SE_I of the six DMUs. However, it has to be noted that, for the four BCC-efficient DMUs in column 5, the efficiency scores of 100% are replaced by the corresponding super-efficiency score (> 100%), e.g., 120% for DMU B in case of an output-oriented optimization. The super-efficiency score of a DMU can be obtained by excluding the respective DMU from the data envelopment.[2] As Fig. 1 indicates for DMU A in the case of output orientation as well as for DMU D in the case of input orientation, such a super-efficiency score does not always exist. In Table 1 and all the following tables, non-existing super-efficiency scores are marked as "inf" (= infeasible). Because DMUs B and C lie on the same ray, their CCR super-efficiency scores in column 4 equal 100%.

The efficient frontier of the niRTS model consists of the line segments spanning from the origin through DMUs B and C up to DMU D. In turn, the efficient frontier of the ndRTS model can be described by the line segments from DMU A to B and C and further infinitely along the ray. Therefore, the envelopment of these two models

[2] Super-efficiency always takes a value above 100% and indicates how much an efficient DMU could increase its inputs proportionally, or how much it could decrease its outputs proportionally, without becoming inefficient (cf. Andersen and Petersen 1993).

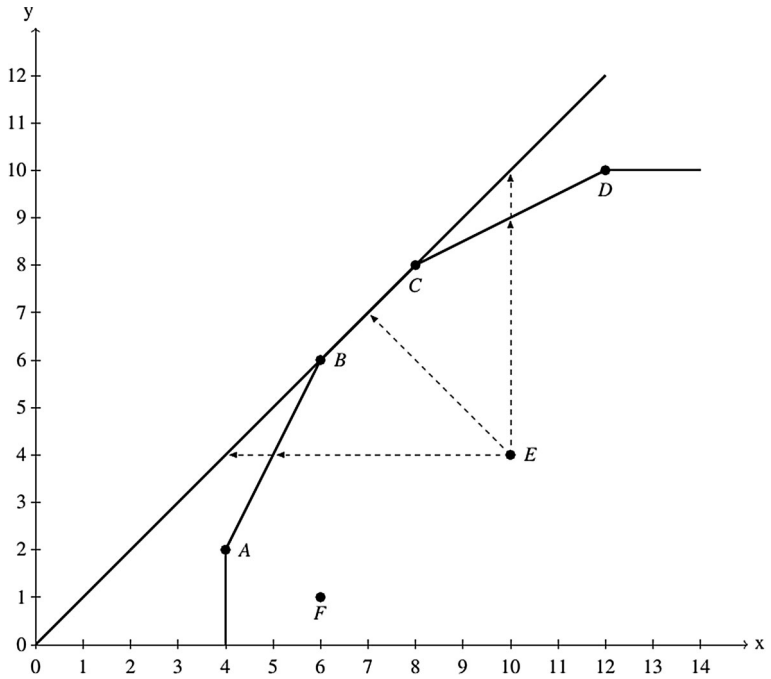

Fig. 1 Two-dimensional example with six DMUs

consists of the respective parts of the CCR and the BCC envelopments, which entails similar results for the corresponding models. Hence, niRTS and ndRTS models are rarely considered explicitly in the literature. Instead, the corresponding RTS properties of the various DMUs are mostly analyzed ($\sum_{j=1}^{n} \lambda_j$ from the CCR model variants in column 7 of Table 1).

Considering inequalities (4) and (5), one might think at first that up to seven different efficiency scores can occur for a DMU. In that case, the question would arise as to which of the seven scores is the "right" one. In fact, however, only three distinct efficiency scores are possible for the eight model variants. However, these three efficiency scores can differ substantially from each other (as shown in Table 1). Emrouznejad and De Witte (2010:1580) also report that there can be "significant differences" between the results of the CCR and BCC model variants. This emphasizes the importance of a systematic and justified selection of the RTS assumption as well as the choice between output and input orientation.

The realization that, for each DMU, the eight DEA models attain at most only three different (super) efficiency scores is not a coincidence but rather a regularity. It is based on certain characteristics of Linear Programming. Accordingly, the efficiency score of a niRTS or ndRTS model variant—as optimal value of the objective functions (1) or (2), taking account of the relevant restriction (3)—must be identical to that of either the CCR model or the BCC model with the same orientation, or of both models if they are equal. That is, in each of the four different inequality chains of (4) and (5), one of the two inequality signs has to be an equality.

Table 1 Results of standard DEA models for the two-dimensional example with six DMUs (all scores in %)

DMU	x	y	CCR O&I	BCC O/I	SE O/I	$\Sigma\lambda$ O/I
A	4	2	50	inf/150	50	67/33
B	6	6	100	120/111	100	100
C	8	8	100	109/113	100	100
D	12	10	83	125/inf	83	150/125
E	10	4	40	44/50	90/80	125/67
F	6	1	17	17/67	100/25	100/17

In the case of an optimum with $\sum_{j=1}^{n} \lambda_j = 1$, the niRTS and the ndRTS efficiency score are identical with the score of the BCC model and, otherwise, that of the CCR model. Nevertheless, it is a priori unpredictable as to which of the two possible scores (CCR or BCC score) for the respective orientation will be attained by a niRTS or ndRTS model for the DMU under consideration. Therefore, column 7 of Table 1 displays the sum $\sum_{j=1}^{n} \lambda_j$ of the activity levels for the CCR model variants (1) and (2) as further information on RTS, characterizing the respective DEA results of the considered DMU. Since there are two CCR-efficient DMUs on the same ray, the value of the sum $\sum_{j=1}^{n} \lambda_j$ depends on the choice of DMU B or C as benchmarking partner; in Table 1, it is always the DMU nearest to the reference point, i.e., DMU C above and DMU B below.

For example, let us consider the inefficient DMU E. Its three efficiency scores in columns 4 and 5 differ only slightly between 40, 44, and 50% (resulting after rounding from 4/10, 4/9, and 5/10). These scores can readily be derived from the lengths of the four vertical and horizontal dashed arrows, as shown in Fig. 1. However, as already stressed before, the four so-called *reference units* of DMU E on the efficient frontier (shown by the arrowheads in Fig. 1) as well as the corresponding *benchmarking partners* (DMU C for CCR, respectively, DMUs C and D for BCC in the two vertical cases as well as DMU B for CCR, respectively, DMUs A and B for BCC in the two horizontal cases) can differ and thus suggest distinct benchmarks as *target values*.

The two scale efficiencies connected with DMU E are vertically $SE_O = 9/10 = 90\%$ and horizontally $SE_I = 4/5 = 80\%$. In Fig. 1, they equal the ratio of the smaller and the larger output or input of the corresponding two reference points onto which DMU E is projected. It must be noted that the actual input or output quantities of DMU E do not play an immediate role in the calculation of the scale efficiency SE. Only the output or input quantity of their two reference points determines the ratio. Thus, scale efficiencies describe the distance between the BCC and the CCR-efficient frontiers regarding those parts of the frontier onto which the considered DMU is projected subject to the chosen orientation. For example, the scale

efficiency of DMU E is 100% (instead of 90% or 80%) if it is not projected vertically or horizontally, but simultaneously in both directions onto the line segment between DMUs B and C, as shown by the third, bisecting dashed arrow to the upper left.[3] As another example, DMU F leads to $SE_O = 6/6 = 100\%$, but $SE_I = ¼ = 25\%$. Therefore, it is important to note that it makes no sense to talk of the scale efficiency of a single DMU *without disclosing its supposed projection onto the efficient frontier*. In fact, scale efficiency is a property which characterizes the distance of the CCR and the BCC-efficient frontier, and in no way, it is a property of any inefficient DMU itself.[4]

The same crucial reservation has to be made regarding the returns-to-scale (RTS) property of a DMU. In production theory, RTS are originally defined as a property of the whole technology, only.[5] This can be generalized as a *local property*—for the neighborhood *of certain parts of the efficient frontier*—which characterizes the extent of the total trade-off between inputs and outputs (cf. Cooper et al. 2007: Ch. 5). If it is proportional, we have constant RTS, otherwise variable, in particular increasing or decreasing RTS if the outputs change disproportionately with the inputs. Hence, the RTS are increasing along the line segment between DMUs A and B, constant between B and C, as well as decreasing between C and D. In this sense, the efficient frontier of the BCC model in the neighborhood of the respective reference point of DMU E shows decreasing RTS if E is projected vertically ($\Sigma\lambda_j = 125\% > 100\%$), increasing RTS if projected horizontally ($\Sigma\lambda_j = 67\% < 100\%$) as well as constant RTS if projected to the upper left by the bisecting arrow ($\Sigma\lambda_j = 100\%$). However, RTS are not defined (or infinite) for DMU E itself, because, for any (strongly) inefficient point, it is even possible to increase (all) outputs without increasing any input or to decrease (all) inputs without decreasing any output.[6]

2.2 Non-oriented additive measurement of efficiency

The significant difference in the BCC-efficiency scores of DMU F between 17 and 67% (more precisely: 1/6 and 2/3) originates from the fact that F is projected onto a weakly efficient part of the data envelopment in case of an input orientation. That is why, LP models (1) and (2), including restriction (3), are usually complemented by an infinitesimally small summand in their objective function or by a second optimization step, which identifies possible slacks for individual inputs and outputs. In this way, the originally purely radial projections of inefficient DMUs are modified, so that (strongly) efficient points of the envelopment will be identified as

[3] E.g., by a non-oriented additive model with equal weights for the input and output slacks (as will be further discussed in the next subsection).

[4] In this sense, Dyckhoff et al. (2009) have used scale efficiencies to empirically characterize the RTS of the best practice research production function of German business schools.

[5] Cf. Dyckhoff and Spengler (2010: 63). As already mentioned in our introduction, DEA is essentially being built on production theory as economic scientific foundation (cf. also—more generally—Dyckhoff 2018). .

[6] The reservations stated before regarding SE and RTS for *single* (inefficient) DMUs seem to be not clearly recognized in much of the DEA applications in the literature.

benchmarks and targets, e.g., data point A for DMU F in the case of an input orientation. However, the radial efficiency score itself remains the same, which is why it only indicates weak efficiency in general.

To avoid this deficit of weak efficiency inherent to all radial DEA models, additive *slack-based models* (SBM) can be used instead. These models take all slacks s_i^- and s_r^+ in the definition of the efficiency measure into account. Thus, these models directly identify strongly efficient solutions, without the additional calculations which are necessary for radial models. Because of its compatibility with the radial models, the model by Tone (2001) is particularly suitable for enabling comparisons to the results of the standard DEA models. Moreover, since it is often hardly possible to justify the orientation of a model meaningfully in DEA applications, the non-orientation of additive models represents yet another benefit. Instead of (1) and (2), the non-oriented SBM model by Tone (2001) takes the following non-linear form:

$$\begin{aligned} \text{Minimize } \rho_{\text{o}} \;=\; & \frac{1-\frac{1}{m}\left(\sum_{i=1}^{m}\frac{s_i^-}{x_{i0}}\right)}{1+\frac{1}{s}\left(\sum_{r=1}^{s}\frac{s_r^+}{y_{r0}}\right)} \\ \text{such that } & \sum_{j=1}^{n}\lambda_j\cdot x_{ij}+s_i^- = x_{\text{io}} && \forall i \\ & \sum_{j=1}^{n}\lambda_j\cdot y_{rj}-s_r^+ = y_{\text{ro}} && \forall r \\ & \lambda_j, s_i^-, s_r^+ \geq 0 && \forall j, \forall i, \forall r \end{aligned} \tag{6}$$

It implies constant RTS. If we add—analogously to (1) or (2)—the restriction (3) in an appropriate form for (6), we obtain corresponding versions of the model with variable, non-increasing or non-decreasing RTS. With the same reasons as for (4) and (5), the following applies:

$$\theta_{\text{Tone,vRTS}} \geq \theta_{\text{Tone,niRTS}} \geq \theta_{\text{Tone,cRTS}} \leq \theta_{\text{Tone,ndRTS}} \leq \theta_{\text{Tone,vRTS}} \tag{7}$$

To facilitate the calculation, model (6) can be linearized (Tone 2001, Cooper et al. 2007: Ch. 4.4.3). In addition, a super-efficiency score can be determined (Tone 2002). However, because its definition differs from that one of the efficiency scores, the super-efficiency score cannot be easily compared with that of the radial models.

By the same reasoning as in Sect. 2.1, also in the case of the SBM model, its niRTS and ndRTS versions must each attain one of the efficiency scores under cRTS and vRTS, so that each DMU can have a maximum of two efficiency scores due to the absence of input or output orientation.

For his efficiency scores, Tone (2001: 502) proved an important property which, besides the appropriate definition, is essentially based on the consideration of all slacks:

$$\theta_{\text{Tone,cRTS}} \leq \theta_{\text{CCR}} \tag{8}$$

An analysis of his proof given for the input-oriented CCR model shows that it is equally valid when the set of feasible solutions for LP (6) is further restricted by an additional restriction of the type (3). Accordingly, inequality (8) is applicable to all

Table 2 Results of the Tone models for the two-dimensional example with six DMUs (all scores in %)

DMU	x	y	cRTS	vRTS	SE	$\Sigma\lambda$
A	4	2	50	100	50	33
B	6	6	100	100	100	100
C	8	8	100	100	100	100
D	12	10	83	100	83	150
E	10	4	40	40	100	100
F	6	1	17	17	100	100

(input-oriented standard radial) models with other forms of RTS. The same proposition can be proved in a similar way for the output-oriented models. The respective radial efficiency score tallies exactly with that of Tone if the DMU considered is strongly efficient in terms of the radial DEA model. In that case, no slacks occur, so that efficiency scores under the respective RTS are both 100%. Therefore, in cases of inefficient DMUs, the Tone efficiency score generally attains a value which is genuinely smaller than the radial efficiency score with the same RTS. Thus, an advantage of the SBM model is its much better discrimination between the inefficient DMUs, so that the differences between them are elucidated more clearly.

For the purpose of illustration, Table 2 contains the results of the Tone model for the six DMUs of the prior example in Fig. 1. Because of non-comparability of the super-efficiency scores, columns 4 and 5 only show the usual score 100% in case of (here always strong) efficiency.

For constant RTS, the efficiency scores of all DMUs of the SBM model (column 4) are identical to those of the radial (CCR) model (column 4 of Table 1), also for variable RTS (columns 5) necessarily for the efficient DMUs, but not for the inefficient ones where they are smaller (as stated in general before). Since their vRTS score equals its cRTS score, the scale efficiencies SE of both inefficient DMUs E and F are 100%, i.e., that they are projected to the line between DMUs B and C with cRTS in Fig. 1. For DMUs A–D, the SE scores equal those of the radial models. In the particular case of our numerical example, all those points on the ray (as efficient frontier for cRTS) build reference points which dominate the respective DMU under consideration. Benchmarking partners are the DMUs B or C.

2.3 Measurement of balance or specialization

The dominance of radial models in the DEA literature seems to have several reasons. For example, Ahn et al. (2018) state:

> "(T)he predominantly application of Farrell efficiency measures in DEA surely has an historical background. However, there are also practical reasons to justify their consideration. Compared to other approaches like, e.g., the slack-based measure of Tone (2001), it is easier for practitioners to interpreting them, i.e., to understanding the concept of radially measured

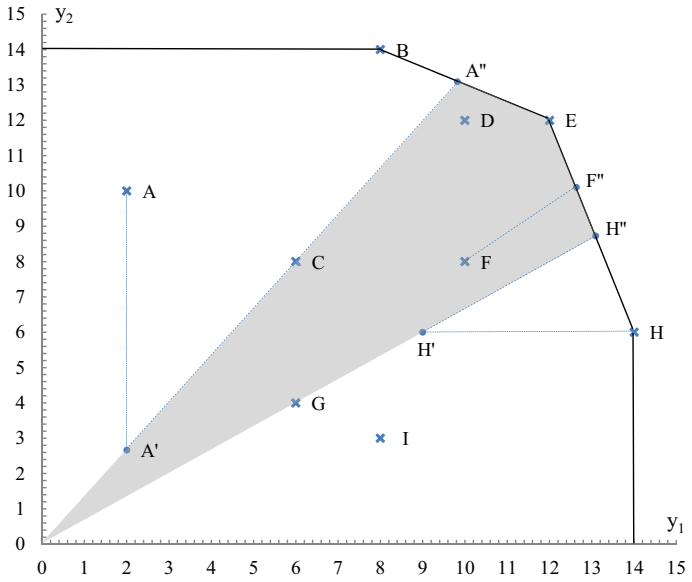

Fig. 2 Three-dimensional example with nine DMUs (Dyckhoff et al. 2013, Fig. 1)

efficiency scores. Furthermore, Farrell efficiency corresponds particularly well with findings in incentive theory. Imposing proportional changes, which is the very essence of Farrell, can under some circumstances be shown to be the optimal response of a principal who lacks information about the relative costs of different activities of his agents …"

If one accepts these reasons one also has to accept that radial efficiency scores themselves in general incentivize weak efficiency, only. Then, it may be interesting how strong the strongly efficient solution deviates from the proportional expansion of the outputs in the output-oriented case or from the proportional reduction of the inputs in the input-oriented case. This may be a cause to define a new, further type of performance indicators which measure the deviation of a DMU from a specific output or input mix relation.

Balanced DEA is a method presented by Dyckhoff et al. (2015) which measures such a kind of deviation, in this case from predetermined relations of the output quantities. Their DEA-integrated balance (or specialization) measure has been developed for output-oriented radial DEA models. As illustrating example, Fig. 2 presents the output diagram of nine DMUs A, …, I which produce two outputs y_1 and y_2 with identical quantities $x = 1$ of one input. It also shows the data envelopment of these DMUs resulting from a BCC model. Table 3 contains the data (columns 2–4) as well as the relevant results. Because of the identical input quantity of all DMUs, the CCR-O and BCC-O efficiency scores θ are identical, too (column 5).

Table 3 Results of the balance measurement for the three-dimensional example with nine DMUs (all scores in %; cf. Dyckhoff et al. 2013, Table 1)

DMU	x_1	y_1	y_2	θ	θ_b	β
A	1	2	10	71	20	29
B	1	8	14	100	81	81
C	1	6	8	61	61	100
D	1	10	12	94	94	100
E	1	12	12	100	100	100
F	1	10	8	79	79	100
G	1	6	4	46	46	100
H	1	14	6	100	69	69
I	1	8	3	57	34	60

The shaded subset **B** inside the data envelope marks the intersection of it with a pointed cone starting from the origin. It is assumed that this cone represents all combinations of both outputs which are considered as a "entirely balanced" output mix (in view of exogenously given information), defined by a (maximal) balance score of $\beta = 100\%$ (or a minimal specialization degree of $\sigma := 1 - \beta = 0\%$). All points of the data envelope outside this cone feature a balance score between 0 and 100%, namely DMUs A, B, H, and I. The balance score of a point $(\boldsymbol{x}_o;\boldsymbol{y}_o)$ outside **B**, e.g., DMU A or H, can be determined by projecting this point appropriately onto a corresponding point $(\boldsymbol{x}_o;\boldsymbol{y}_b)$ inside **B**, here A' and H'. This corresponding point has a lower efficiency score, so that the balance is determined by the ratio of the two points' efficiency scores: $\beta := \theta_b/\theta$. The scores for balanced efficiency θ_b and balance β for the example are given by columns 6 and 7 of Table 3. Points belonging to **B** are projected onto themselves, thus fulfilling the condition $\beta = 100\%$.

The approach of Dyckhoff et al. (2015) deliberately places high requirements on the balance score. That is achieved by two conditions: $(\boldsymbol{x}_o;\boldsymbol{y}_b)$ is dominated by $(\boldsymbol{x}_o;\boldsymbol{y}_o)$ and has the highest possible efficiency score. These conditions result in a non-linear optimization model, which is not easy to be solved in general (cf. Dyckhoff and Gutgesell 2015). There are two exceptions: first, the case of two outputs, as in the example of Fig. 2, where the problems can be solved graphically; second, the case of more than two outputs with an extremely reduced balance cone which is identical to a ray, i.e., the output ratios of balanced output mixes are fixed. The approach of Dyckhoff et al. (2015) can be transferred from output-oriented radial DEA models to non-oriented SBM models if one accepts that the efficient target point itself is not necessarily balanced.

Our explanations on standards, pitfalls, and extensions of DEA in this section refer to general aspects which are valid for all imaginable application fields. In contrast, in the subsequent sections, we primarily analyze such aspects that are specifically characteristic for applications in a non-production context. In this context, we focus on the selection of relevant inputs and outputs in conjunction with

the choice of an appropriate DEA model. As an exemplary example for a non-production context, we examine the field of welfare evaluation.

3 DEA applications to welfare evaluation and some common pitfalls

In a narrower, material sense, welfare is understood to be the standard of living reflected in the level of provision of goods and services to individuals, private households, or to an entire society. Based on this understanding, the traditional measurement approach involves the gross domestic product (GDP) or its growth (per capita). More recent considerations extend that notion by taking account of entirely different—or at least additional—factors which influence the welfare of a country. In the same way as sustainability or sustainable development is understood, these factors can be generally assigned to the dimensions of social and environmental quality of life (e.g., Böhringer and Jochem 2007; Singh et al. 2012).

When designing appropriate welfare measures, two approaches can be differentiated with respect to the indicators, namely either establishing an integrated indicator (*composite or aggregated indicator*) or designing a set of key indicators (*dashboard*). A composite indicator condenses the welfare construct into a single value. This can be monetary, as particularly in the case of the so-called GDP revisions [e.g., Genuine Progress Indicator (GPI) or the Index of Sustainable Economic Welfare (ISEW)], as well as non-monetary, as in the case of multi-component indicators [e.g., Human Development Index (HDI) or Gross National Happiness Index (GNH)]. By contrast, in the cases where a set of key indicators is used, welfare can be described by means of a selection and combination of various indicators, which are themselves mostly the result of several aggregation stages: from raw data via partial indicators to key indicators. The key indicators represent the situation or development of a subdomain of welfare and are not finally condensed (e.g., such as those stated in the Eurostat Monitoring Report 2013).

DEA is a methodology which is used to compare several countries based on such multi-dimensional sets of indicators. The performance of the countries is determined by the respective data without a specification of concrete weights or aggregation rules for the indicators (as is explained in Sect. 2 before). On the contrary, the weights are endogenously calculated, so that the respective country is "depicted in the best possible light"; that is to say, they are selected optimally in favor of the country under consideration. In that way, a welfare profile of a country is determined relatively to the other countries being studied. If a country, despite the most favorable choice of its weights, is dominated by others in the overall assessment, and is thus inefficient with respect to the underlying indicators, role models and benchmarks are derived to provide guidance for the improvement of that country's welfare. However, this presupposes that the methodological assumptions of DEA are compatible with reality.

Next, we illustrate the extent to which the apparent potential of DEA has already been utilized in the context of welfare evaluation. Our assessment is primarily based

on a prior analysis of the literature, presented in depth in Wojcik (2018: Ch. 3.3). In addition, we consider a similar literature review by Mariano et al. (2015).[7] In the following, we refer to typical procedures and pitfalls in the relevant articles, with respect to their *content-related focus* and *methodological aspects* as well as to the approach chosen for the *selection and classification of the performance indicators*. We focus on the findings from Wojcik (2018), as certain aspects of DEA model choice are more relevant to our purpose, and complement them with some more methodology-related findings from Mariano et al. (2015).

The content-related focus of many articles is the human development index (HDI), which is published annually by the United Nations. A large portion of those articles uses the sub-indices of the HDI or the underlying metrics as 'outputs' of DEA. In terms of structural observations, a single, uniform dummy input is often used for all countries, so that, ultimately, they conducted effectivity measurements (e.g., Despotis 2005; Lee et al. 2006; Lozano and Gutiérrez 2008; Bougnol et al. 2010; Blancard and Hoarau 2011). Apparently, these authors often understand DEA as being only a means of determining optimal weights for the HDI sub-indices and subsequently use these weights in further aggregation procedures which build upon the basic ideas of DEA.

Regarding *methodological aspects*, the predominant majority of the literature uses one, and *only one*, model mostly of standard CCR and BCC type, either exclusively or as a basis for other calculations (e.g., Martić and Savić 2001; Jurado and Perez-Mayo 2012; Sabermahani et al. 2013; Wu et al. 2014). Some of the rare exceptions in the selection of models have been, up to now, SBM models (Murphy et al. 2013; Reig-Martínez 2013), Directional Distance Function models (Shetty and Pakkala 2010), and Range Adjusted Measure (RAM) models (Lozano and Gutiérrez 2008). Using standard radial models can be problematic, for example, if possibilities for improvement in the form of slacks are not incorporated in the calculated efficiency score (cf. Section 2.2), or the potentially generated zero weights cause the exclusion of entire criteria (cf. Coelli et al. 2005: 198ff. and Sect. 2.3). Regarding the returns-to-scale (RTS), no specific choice predominates, while respective assumptions are rarely discussed and often only implicitly given by the chosen model formulation. A similar observation applies to the orientation of DEA models. In total, far more models are input-oriented, however (cf. Mariano et al. 2015: 36f.). Summarizing the methodological aspects, Mariano et al. (2015: 40) state that "little attention is paid to important modeling issues, such as the choice of the model and the orientation".

With respect to *indicator selection for inputs and outputs*, the examination of the welfare literature yielded that three different general approaches can be identified:

- *Selection from available welfare studies*: By far, the most frequently used procedure for selecting the relevant DEA indicators is the reference to already established welfare indices and rankings (e.g., Jablonsky 2004; Murias et al. 2006; Lee et al. 2006; Bougnol et al. 2010). Thereby, sub-indicators are often used, or their underlying indices are adopted directly as DEA indicators. In

[7] While Mariano et al. (2015) have a broader view and analyze all 57 papers found in depth, Wojcik (2018) focuses on the more macro-oriented part of the literature, thus analyzing 38 papers in detail.

principle, this approach offers the advantage of data sets not needing to be elaborately collected and of their presumably having been checked for errors during their initial use. However, the use of index numbers can constitute a problem if volume measures are also used at the same time (Dyson et al. 2001).

- *Own selection via normatively selected attribute groups of welfare*: Some authors are inclined to select indicators first by (normatively) forming groups of attributes or partial aspects of welfare. Individual indicators are then assigned to these attribute groups or partial aspects which reflect them or are associated with them; for example, Ramanathan (2006: 158) identifies the groups "economic, educational and health attributes"; Li and Ma (2011) similarly determine four categories of welfare to derive relevant indicators. In this procedure, however, it is questionable and is barely justified as to how precisely the relevant attribute groups should be determined.
- *Own selection without substantiation or reference to the literature*: Often, justifications for selecting indicators are totally absent (e.g., Malul et al. 2009; Jahanshaloo et al. 2011). Some authors argue that there is already general consensus on the relevant indicators in the literature (e.g., Dominguez-Serrano and Blancas 2011: 485) and, therefore, that they do not need to be justified any further. This often leads to difficulties in the interpretability of the models or of the results. Sabermahani et al. (2013), for example, use the employment rate as input, to be minimized in their radial input-oriented model; this is hardly justifiable without due explanation.

It can be concluded from the reviewed literature that the approaches adopted are rarely accurately described nor do they follow any specific process. In particular, the frequent references to the sub-indicators of the HDI suggest that the mere fact of the availability of data ultimately justifies the data being used. This contrasts with the fact that the choice of inputs and outputs is supposed to represent a "key stage in the DEA assessment" (Martić and Savić 2001: 345, and similarly Cook et al. 2014).

Finally, with respect to the *indicator classification* to the categories 'input' and 'output', most literature either does not recognize the production-theoretical foundation of DEA and its resulting classification of indicators as input or output as a challenge or circumvents it by means of simplification. In this way, positively connoted (the more the better) factors are used as outputs and, conversely, negatively connoted (the less-the-better) factors are classified as inputs, as suggested by Cook et al. (2014: 2). However, as already mentioned in our introduction, the original justification of indicator classification is based on production processes. Thus, the fundamental question must be asked as to whether the generation of welfare can even be interpreted as a production process at all and whether the chosen indicators can thus be reasonably assigned to the two groups of input and output. The fact that this question is ignored by the literature on welfare evaluation—and on other non-production contexts, too—leads to problems which are of essential importance for the validity of the DEA results. This will be demonstrated by the case study of the 'Prosperity Quintet'.

4 DEA case study for the Prosperity Quintet

In the last section, some critical points have been raised which illustrate typical pitfalls of the DEA literature on welfare evaluation. In the current section, we intend to clarify essential points with the help of the five welfare indicators of the so-called 'Prosperity Quintet' for an exemplary illustration.

4.1 The Prosperity Quintet for the countries of the European Union

The *Prosperity Quintet* is a set of key indicators developed by the Denkwerk Zukunft Foundation. It was first proposed in 2011 as an alternative welfare evaluation with the purpose of enhancing the informative value. Thus, systematic and transparent assessments of the welfare of early industrialized countries are intended to be obtained in the context of the 27 countries of the EU.[8] Due to the problem of confusion and complexity when too many key indicators are involved, a comparatively small number of only five relative indicators are used (Denkwerk Zukunft 2011: 25ff, 71ff, 2014):

- *Gross Domestic Product per capita* (GDP/POP in €/cap): It stands for the purely economic dimension and the material level of prosperity.
- *80/20 ratio* (HIC/LIC measured in %): It covers the socio-economic dimension and represents the income distribution in a national economy. The incomes of the upper fifth of the population are considered in relation to those of the lower fifth, where the calculation of income refers to the equivalized disposable income per household according to the OECD standard.
- *Social exclusion rate* (SER in %): It shows the social coherence and indicates the proportion of individuals interviewed (15 years or older) who perceive themselves to be excluded from society.
- *Ecological footprint per capita in relation to available global biocapacity of a human being* (EFC/BC in %): It represents the ecological dimension. The ecological footprint contained in the numerator measures the areas of land and water which an inhabitant requires, on average, for the production of goods and services consumed by him or her as well as the absorption of emissions generated in the process, including the areas required by imports. The resulting total area is then considered in relation to the total biocapacity that can be regenerated by the ecosystem and which is available to a person on average (at present, 1.8 global hectares).
- *Public debt rate* (PDR = DEB/GDP in %): It is used as a barometer for the credit financing of a national economy and is calculated as the gross public debt of a country in relation to the GDP. This expresses the extent to which material wealth is generated at the expense of future generations and thus limiting the future viability of a country.

Table 4 represents the respective values of the Prosperity Quintet for all 27 EU countries as they were used by Denkwerk Zukunft (2014) with the data available at

[8] Croatia is not included in this data set, as it only joined the EU in 2013.

Table 4 Prosperity Quintet of 27 EU countries

EU-27	GDP/POP (€/cap)	80/20-ratio (%)	SER (%)	EFC/BC (%)	PDR (%)	Evaluation
A-Austria	32200	420	7.5	289	73	+++ −+
B-Belgium	29630	391	15.2	406	100	++−−−
BG-Bulgaria	3700	611	17.9	161	19	−−−+*
CY-Cyprus	17365	467	20.7	244	85	−+−++
CZ-Czech Republic	11389	350	14.1	261	46	−+−−*
D-Germany	30070	429	9.0	250	82	+++ ++
DK-Denmark	37263	451	6.6	422	46	+++ −*
E-Spain	20300	716	7.9	239	84	−−+++
EST-Estonia	9400	544	12.3	278	10	−−−−*
F-France	27600	454	15.9	272	90	++−−−
FIN-Finland	31100	370	4.3	311	53	+++ −*
GB-Great Britain	30400	533	11.9	250	90	+−−+−
GR-Greece	14900	662	15.6	261	157	−−−−−
H-Hungary	8809	395	9.4	167	79	−+++ +
I-Italy	22800	552	6.5	244	127	−−++−
IRL-Ireland	37100	465	9.8	317	118	+++ −−
L-Luxembourg	62600	415	18.7	833	21	++−−*
LT-Lithuania	8100	534	10.3	228	41	−−++*
LV-Latvia	6800	657	14.0	211	41	−−−+*
M-Malta	13500	395	11.1	244	72	−+−++
NL-Netherlands	32833	360	3.7	328	71	+++ −+
P-Portugal	14200	589	9.9	244	124	−−++−
PL-Poland	8500	494	14.8	228	56	−+−+*
RO-Romania	4400	637	8.2	133	38	−−++*
S-Sweden	35536	374	8.8	300	38	+++−*
SK-Slovakia	9409	372	8.3	194	52	−+++*
SLO-Slovenia	15000	342	4.6	228	54	−+++*
Ø-EU 27	23200	500	10.5	250	85	

Data based on Denkwerk Zukunft (2014); collected subsequently from EuroStat (2013), Eurofound (2012) and Global Footprint Network (2013). The abbreviations of the countries correspond to their official country codes

The data presented in Table 4 could not directly be obtained from Denkwerk Zukunft (2014) and were thus collected subsequently from the sources cited there. For that reason, (marginal) differences in the data can occur, which (nevertheless) can result in a divergent evaluation with "+"or "−". This is evident in the evaluations regarding the debts, because their data are very close to the determined relative limit

that time. In the last column, the evaluation according to Denkwerk Zukunft is shown in the order of the indicators listed. To make concrete statements about the overall welfare of a country and to generate a rating from the set of indicators, Denkwerk Zukunft assesses the countries on the basis of their characteristics in the five key indicators. They are assessed in comparison to endogenously computed

average values (*relative aspiration levels*). The five average values used by Denkwerk Zukunft (2014) are listed in the last row of Table 4. Those countries that are better than this average, and thus reach the relative aspiration level, receive a + for each indicator. A larger GDP per capita is better, while, in the case of the other four criteria, usually smaller indicator values are preferred. With respect to the ecological and the debt criterion, Denkwerk Zukunft even sets an exogenous limit (*absolute aspiration levels*). Thus, no country fulfills the absolute requirements of a relative ecological footprint of not more than 100%, and 13 countries meet the Maastricht criterion of a maximum public debt rate of 60% (* in the last column of Table 4).

4.2 Standard DEA results for the prosperity quintet

The classification by Denkwerk Zukunft into rating classes does not allow or consider any gradual nuances in the evaluation. This can be seen in the last column of numbers of Table 4, for example, in the case of Spain and Cyprus with public debt rates close to the EU average of 85%, which is why minor data changes would lead to a different classification. In contrast, DEA does consider gradual nuances in the form of a quantitative efficiency measure, which can attain any value between zero and one. To this extent, DEA is more informative. In contrast to the rating by Denkwerk Zukunft, DEA compares the 27 EU countries on the basis of their best indicator values (*best practice*). By doing so, a country is compared not only with some of the 26 other countries but also with certain combinations of several of these countries, namely with such combinations which dominate that country, that is to say, with combinations which are not worse than that country in any of the five indicator values. Thus, a country can be 100% efficient if it is particularly good in only a few of the five indicators. DEA permits, therefore, individual emphases to be placed on the different aspects of welfare and so reveals the prevailing profile of a country.

Table 5 contains the results of the standard (radial) DEA models applied to the data of Table 4. They are displayed in the same manner as columns 4–7 of Table 1 in Sect. 2.1, i.e., the CCR- and BCC-efficiency scores, the scale efficiencies (SE) as well as the returns-to-scale (RTS) as rounded integer percentages, divided into their output- (O) and their input-oriented value (I) if both differ. The GDP per capita is treated as the only 'output', and the other four indicators are treated as 'input'. For efficient countries, their super-efficiency scores ($\geq$ 100%) are listed instead if they exist; for variable RTS, there is not always a solution (inf = infeasible). This usually applies to countries, which have indicator characteristics that differ strongly from those of the others. Column 2 of Table 5 shows the rating values of Denkwerk Zukunft according to Table 4 in an abbreviated form. The number is given for the frequency with which the welfare indicators of the country are better than the EU average (= number of + or * in Table 4), while * indicates compliance with the Maastricht limit for debt.

Table 5 clearly shows that, for many countries, the CCR- and BCC-efficiency scores differ substantially from each other. Another observation has the character of a tendency statement, although it is applicable in all the cases of Table 5 if we do

Table 5 Results (in %) of the standard DEA models for the Prosperity Quintet data of Table 4

EU-27	Rating	CCR O&I	BCC O/I	SE O/I	$\Sigma\lambda$ O/I
A-Austria	4	97	98/99	99/98	95/92
B-Belgium	2	71	72/91	99/78	103/73
BG-Bulgaria	2*	20	inf/127	20	51/10
CY-Cyprus	3	59	60/85	98/70	98/57
CZ-Czech Republic	2*	37	inf/102	37	90/33
D-Germany	5	101	115/106	100	100
DK-Denmark	4*	105	109/106	100	100
E-Spain	3	71	81/93	88/77	89/63
EST-Estonia	1*	42	inf/192	42	40/17
F-France	2	84	85/91	100/93	107/90
FIN-Finland	4*	104	inf/112	100	100
GB-Great Britain	2	101	101	100	100
GR-Greece	0	47	47/71	99/66	104/49
H-Hungary	4	43	inf/109	43	67/29
I-Italy	2	81	99	81	80/65
IRL-Ireland	3	98	102/104	98	111/109
L-Luxembourg	3*	318	inf	100	100
LT-Lithuania	3*	30	36/90	82/33	79/24
LV-Latvia	2*	27	35/82	76/33	74/20
M-Malta	3	46	53/89	87/52	94/43
NL-Netherlands	4	123	inf/134	100	100
P-Portugal	2	48	49/79	98/61	98/47
PL-Poland	3*	31	36/83	87/37	84/26
RO-Romania	3*	27	inf/134	27	50/14
S-Sweden	4*	128	inf/128	100	100
SK-Slovakia	4*	40	inf/105	40	73/29
SLO-Slovenia	4*	60	inf/118	60	73/44

Inf—infeasible: country with efficiency score 100%, but without feasible LP solution for super-efficiency

Values x with $0.995 \leq x < 1$ are rounded down to 99% in Tables 5, 8, and 9; therefore, 100% always indicates at least weak efficiency or balance

not take account of any super-efficiency scores. Thus, the output-oriented BCC efficiency is not greater than the input-oriented one: $\theta_{\text{BCC-O}} \leq \theta_{\text{BCC-I}}$. From this, it immediately follows: $\text{SE}_\text{O} \geq \text{SE}_\text{I}$. This inequality is induced by the significantly smaller number of 'outputs' (here, 1) as against the 'inputs' (here, 4). Then, in the case of the radial projection of the inputs in the input-oriented model, more cases of slacks occur in the LP solution than in the corresponding projection of the output-oriented model. The last two columns of Table 5 show that nearly all countries are projected onto parts of the efficient frontier with increasing RTS ($\Sigma\lambda_\text{j} < 100\%$), i.e., that the inputs and outputs of their CCR benchmarking countries are larger as a tendency, most notably with an input-oriented projection. There are only five

exceptions with decreasing RTS, four in the output-oriented case (B, F, GR, and IRE) and one in the input-oriented case (IRE). Seven countries are projected in both orientations to parts of the efficiency frontier with constant RTS and SE = 100% (D, DK, FIN, GB, L, NL, and S).

4.3 Pitfalls of DEA application to the Prosperity Quintet

The comparison of DEA efficiency scores with the rating of Denkwerk Zukunft (2014) shows not only their different capability of depicting nuances in the evaluation, but also that there is no strong correlation between the two concepts of assessment. Thus, Table 5 also highlights their diversity in the approach: namely, the comparison between the best practice and the averages, on one hand, and between the profile formation and equal weighting, on the other hand, for the five welfare indicators. Great Britain (GB with a rating of only 2) and the extremely super-efficient Luxembourg (L with a rating of only 3*) are two examples of cases where the DEA evaluation is noticeably better than the rating of Denkwerk Zukunft, whereby both countries are identified as efficient in all the standard DEA models. Vice versa, strongly CCR inefficient Slovakia (SK 0.40) and Hungary (H 0.43) have a rather high rating of 4 and 4*. Reasons for this are not always as obvious as in the case of the very high GDP/cap of Luxembourg (L 1.68 times that of DK as second best). The detailed results of DEA can provide valuable information for an in-depth analysis; for example, the endogenously generated weights for each welfare indicator.[9] Consequently, both evaluation concepts complement each other in their aggregation and analysis of data.

When applying DEA, the question arises as to which of their mathematical variants is appropriate for the considered issue or situation. As illustrated in Table 5, the quality of the results largely depends on the model choice. Thus, the inequalities (4) and (5) imply that usually significantly more countries are indicated as (super-) efficient for the BCC model than for the CCR model: 15 countries are in this sense BCC-efficient and seven are CCR-efficient. Therefore, CCR models are frequently chosen in DEA applications to achieve a more pronounced discrimination. However, this is mostly done without further explanation or reasoning regarding the RTS property. The assumption of constant RTS is, however, especially problematic if some of the welfare indicators cannot be arbitrarily increased or reduced, as is partly the case with the Prosperity Quintet. Thus, the exclusion rate (SER) and 80/20 ratio are bounded to 100% by definition: upwards for the exclusion rate and downwards for the 80/20 ratio. In fact, the input-oriented CCR model

[9] Due to space constraints, we avoid an elaborate presentation of such detailed analyses common to DEA and discuss the corresponding results explicitly only to the extent where it is helpful in assessing the basic suitability of the DEA methodology for its application in welfare evaluation. In particular, the reasons for the efficiency or inefficiency of individual countries can be explored through the usual detailed analysis of the DEA results concerning the weights, the targets, and the benchmarking partners. The weights can be interpreted as shadow prices of the relevant restrictions of the six basic indicators in the optimum of the linear programs (1), (2), and (6). Alternatively, the weights are variables of the respective dual linear program (LP) to (1), (2), and (6), the so-called *multiplier model.*

assigns a target value of less than one to five of the 27 countries for the 80/20 ratio (BG 0.38, EST 0.68, LT 0.98, LV 0.86, and RO 0.68).

Since the exclusion rate is defined as an input here, its theoretical maximum of 100%, however, practically fails to play a role, even in the output-oriented CCR (or ndRTS) model. This is because the solution of each DEA model calculates (weakly) dominant indicator scores as target values, i.e., the inputs cannot become greater nor the outputs smaller, as implied by the model inequalities of (1) and (2). On the other hand, to circumvent the problem of the lower limit of 100% for the 80/20 ratio, a scale transformation can be undertaken for this indicator by subtracting 100% from all the values of column 3 in Table 4. Then, the desired ratio scale is given, in which a value of 400% indicates that the richest income quintile has got four times more than the poorest quintile; in other words, five times as much as the latter. Thus, the value of zero would indicate the same level of income for all. However, scale transformations generally affect the results of radial DEA models (Lovell and Pastor 1995; Pastor 1996).

In the instances where constant RTS present a problem due to the definition of the indicators, the BCC model at least ensures that the calculated benchmarks for inefficient countries do not lie outside a realistic bandwidth and, therefore, cannot attain any nonsensical targets (Hollingsworth and Smith 2003). For variable RTS, however, the orientation can significantly influence the model results (e.g., for GR, LT, LV, and PL). In general, countries can actually be efficient in one orientation, but inefficient in the other. This is due to the fact that radial models can generate an efficiency score of 100%, although, in actuality, the DMU is only weakly but not strongly efficient.

A problem for the application of DEA is also the fact that the 80/20 ratio and the public debt rate are defined as the quotient of two quantities that both themselves represent original welfare objectives.[10] This is still acceptable for the public debt rate, because the debt should be minimized in the numerator and the GDP should be maximized in the denominator. In contrast, the incomes of both the richest and the poorest quintile represent quantities which *ceteris paribus* are supposed to be as large as possible. Therefore, the minimization of the 80/20 ratio must be generally questioned, because the minimum of 100% with the same amount of income for all inhabitants of a country is hardly the optimum of income distribution due to the absence of incentives for high performers in the society and thus for the achievable total income.

Our considerations show that interdependencies have to be taken into account during the selection process for the characteristics of the DEA model to be applied—from the definition and classification of indicators up to the choices of RTS, orientation, and efficiency measure. Thus, such aspects cannot be considered independently. It is all the more surprising that the choice of the individual model characteristics is rarely justified or questioned in depth—at least according to the literature on welfare evaluation with DEA reviewed in Sect. 3. In the following, we

[10] The inappropriate use of ratios in DEA has been criticized by Hollingsworth and Smith (2003) as well as by Emrouznejad and Amin (2009).

present the extent to which changes in the specification of indicators can overcome some of the critical points mentioned above.

5 DEA case study for a basic welfare set

In a first step, we calculate the standard DEA models for a second, modified set of welfare indicators. To ensure comparability with the previous results, the following six indicators represent not only the same welfare aspects as the Prosperity Quintet, but form more or less the basis of it, as the Prosperity Quintet can be calculated from them.

5.1 A set of six basic welfare indicators

To avoid the problem of forming quotients from several variables which all themselves represent objectives of welfare and to circumvent the limitations of certain indicators previously identified as problematic, we now (re-)define the original Prosperity Quintet as six basic indicators.[11] The size of the country and, in particular, the number of its inhabitants have a crucial influence on that country's welfare evaluation. Therefore, in principle, the data on the 27 EU countries in absolute terms cannot directly be compared with each other, since, for example, Luxembourg is significantly less populous than Germany. However, the population itself is not a measure of welfare and is, therefore, considered to be exogenously given. To relativize the sizes of countries, the data on the various aspects of welfare are related to the individual (average) inhabitant by weighting every absolute indicator with the inverse of its population size for each country. We use the following six basic welfare indicators:

- *High income per capita* y_1 (HIC): average income of the population's quintile with the highest incomes.
- *Middle income per capita* y_2 (MIC): average income of the three quintiles with middle incomes.
- *Low income per capita* y_3 (LIC): average income of the population's quintile with the lowest incomes.
- *Social exclusion rate* x_1 (SER): proportion of individuals interviewed (15 years or older) who perceive themselves to be excluded from society.
- *Ecological footprint per capita* x_2 (EFC): areas of land and water which an inhabitant requires on average for the production of goods and services consumed by that person as well as the absorption of emissions generated in the process.
- *Debt level per capita* x_3 (DBC = DEB/POP): average gross debt of a country per inhabitant.

[11] Cf. Emrouznejad and Amin (2009) for a general approach. Olesen et al. (2015, 2017) discuss efficiency measures and computational approaches for DEA models when ratio inputs and outputs cannot be avoided.

Table 6 Relationship between basic welfare indicators and the Prosperity Quintet

Prosperity Quintet	Backward calculation using basic welfare indicators
GDP/POP (in €/cap)	$0.2y_1 + 0.6y_2 + 0.2y_3$
80/20 ratio (in %)	$\frac{y_1}{y_3}$
SER (in %)	x_1
EFC/BC (in %)	$\frac{x_2}{1.8}$
PDR (in %)	$\frac{x_3}{0.2y_1+0.6y_2+0.2y_3}$

The welfare of a country increases when the last three indicators decrease, and the first three increase, possibly in a certain proportion to each other. According to the commonly used terminology in the DEA literature, they are classified as 'inputs' and 'outputs'. If we make certain simplifying assumptions with respect to the measurement of income, all key indicators of the Prosperity Quintet can be calculated from the six basic indicators using simple mathematical operations (see Table 6).

In Table 7, the data for the six basic indicators are given. In addition (in column 2), the population size of each of the 27 EU countries is also specified. Based on the relationships in Table 6, the data are consistent with those of Table 4. Thus, the values of the exclusion rate (SER) are identical in the two sets of welfare indicators.[12] The ecological footprint of an inhabitant (EFC) is now measured in absolute terms, whereas before it was specified relatively to the average footprint of a person (1.8 gha/cap). The sixth column of numbers in Table 7 is thus identical to the fourth column of Table 4, the seventh column correspondingly proportional to the fifth of Table 4. For the other indicators, the relationships are more complex and sometimes even non-linear. Therefore, when applying DEA, we can principally assume that the results for the two sets of indicators will differ from each other, although both sets are based on the same basic data.

5.2 Standard DEA results for the basic welfare set

Table 8 includes the results of the standard radial DEA models in the same manner as Table 5. What is striking is that the range of the efficiency scores is smaller than in the case of the Prosperity Quintet. The smallest score is now 49% (in case of H and PL for CCR) as against the previously calculated 20% (in case of BG for CCR in Table 5); the greatest super-efficiency score is now 209% (EST for BCC-I) as opposed to 318% (in case of L for CCR) for the Prosperity Quintet.

In Sect. 4.2, we have discussed that the efficiency score of the output-oriented version of the BCC model does not tend to be smaller than the input-oriented score,

[12] We do not change this welfare indicator, although it is bounded upwards. In view of the methodical goals of this paper, our main reason is that we want to remain comparability with the former results of Table 5 by applying the same standard DEA models to two sets of welfare indicators containing at most the same information. However, for a valid application of DEA to the area of welfare evaluation, such an indicator is problematic (cf. footnotes 10 and 11).

Table 7 Population and set of six basic welfare indicators for EU-27

EU-27	POP (10^3cap)	y_1 HIC (€/cap)	y_2 MIC (€/cap)	y_3 LIC (€/cap)	x_1 SER (%)	x_2 EFC (gha/cap)	x_3 DBC (€/cap)
A-Austria	8443	58765	29409	14007	7.5	5.2	23635
B-Belgium	11095	52096	27577	13320	15.2	7.3	29482
BG-Bulgaria	7327	7456	3275	1221	17.9	2.9	685
CY-Cyprus	862	34539	14964	7395	20.7	4.4	14758
CZ-Czech Republic	10505	20121	10355	5757	14.1	4.7	5221
D-Germany	81844	55535	27291	12943	9.0	4.5	24652
DK-Denmark	5574	67327	34689	14920	6.6	7.6	17083
E-Spain	46196	41412	18101	5786	7.9	4.3	17093
EST-Estonia	1294	18659	8303	3431	12.3	5.0	949
F-France	65328	54510	23828	12006	15.9	4.9	24895
FIN-Finland	5401	55203	28457	14928	4.3	5.6	16483
GB-Great Britain	63256	61560	26296	11552	11.9	4.5	27360
GR-Greece	11290	30098	13286	4545	15.6	4.7	23378
H-Hungary	9932	15972	8008	4048	9.4	3.0	6970
I-Italy	59394	44688	20406	8094	6.5	4.4	28956
IRL-Ireland	4583	70676	33205	15211	9.8	5.7	43630
L-Luxembourg	525	115497	56549	27857	18.7	15.0	13250
LT-Lithuania	3004	15998	7169	2997	10.3	4.1	3297
LV-Latvia	2042	14518	5757	2210	14.0	3.8	2768
M-Malta	418	24503	12263	6210	11.1	4.4	9733
NL-Netherlands	16730	57892	30067	16072	3.7	5.9	23354
P-Portugal	10542	30104	11928	5112	9.9	4.4	17551
PL-Poland	38538	16575	7523	3358	14.8	4.1	4726
RO-Romania	21356	8690	3982	1364	8.2	2.4	1663
S-Sweden	9483	60350	33725	16153	8.8	5.4	13561
SK-Slovakia	5404	16262	8805	4371	8.3	3.5	4897
SLO-Slovenia	2055	25125	14175	7350	4.6	4.1	8115

From EuroStat (2013), Eurofound (2012) and Global Footprint Network (2013)

considering the ratio of one output to four inputs. Here, for the set with six basic indicators, such a regularity cannot be observed. This is due to the fact that the number of outputs equals the number of inputs. However, the efficiency scores of the input- and output-oriented variants of the BCC model do not differ quite as much from each other as they do in case of the Prosperity Quintet.

Table 8 Results (in %) of standard radial DEA models for the set of six basic welfare indicators

EU-27	CCR O&I	BCC O/I	SE O/I	$\Sigma\lambda$ O/I
A-Austria	99	100	99	95
B-Belgium	62	72/64	85/96	144/89
BG-Bulgaria	64	inf/157	64	55/36
CY-Cyprus	67	72/82	93/81	85/57
CZ-Czech Republic	61	76/83	80/73	39/24
D-Germany	102	114/106	100	100
DK-Denmark	101	104/106	100	100
E-Spain	81	92/96	89/84	85/69
EST-Estonia	202	inf/209	100	100
F-France	88	89/91	99/97	105/93
FIN-Finland	113	123/116	100	100
GB-Great Britain	111	111/112	100	100
GR-Greece	50	50/70	100/71	99/49
H-Hungary	49	89/96	56/51	51/25
I-Italy	86	108/103	86	77/67
IRL-Ireland	104	110/118	100	100
L-Luxembourg	162	195/inf	100	100
LT-Lithuania	54	69/83	78/65	34/18
LV-Latvia	56	73/82	77/69	37/21
M-Malta	53	57/76	93/69	72/38
NL-Netherlands	125	inf/151	100	100
P-Portugal	56	60/79	94/71	88/49
PL-Poland	49	60/71	81/69	35/17
RO-Romania	55	inf/152	55	25/14
S-Sweden	130	135/133	100	100
SK-Slovakia	56	79/90	72/63	36/21
SLO-Slovenia	76	inf/128	76	54/41

Inf—infeasible: country with efficiency score 100%, but without feasible LP solution for super-efficiency

Comparing DEA results between Tables 5 and 8, there are now 9 CCR- and 14 BCC-efficient countries as opposed to 7 CCR- and 15 BCC-efficient countries before.[13] The seven countries which are CCR-efficient regarding the Prosperity Quintet (D, DK, FIN, GB, L, NL, and S) are also CCR-efficient in terms of the six basic indicators. Hence, in view of (4) and (5), they have to be BCC, niRTS, and ndRTS efficient, too. One of the additionally CCR-efficient countries in Table 8 is nearly efficient for the Prosperity Quintet (IRL 0.98). Extreme deviations exist for

[13] By speaking of CCR- or BCC-efficient DMUs in this paper, we mean that the efficiency score of the respective DEA model is 100%; the DMU may, however, be weakly efficient only because of some slacks (cf. Section 2.2).

Estonia, only, with 42% for the CCR score as against a super-efficiency of now 202% for the basic welfare set.

More often, the efficiency scores do not differ significantly when we compare both sets of welfare indicators. Furthermore, three of the four countries which are BCC-efficient but not CCR-efficient for the six basic indicators (BG, RO, and SLO) have the same property for the Prosperity Quintet. Vice versa, three of the countries which are BCC-efficient regarding the Prosperity Quintet are inefficient in terms of the basic indicators (CZ 0.76/0.83, H 0.89/0.96, and SK 0.79/0.90).[14]

The DEA results for both sets of welfare indicators are similar not only for efficient countries but also for many inefficient ones. Considering the same models, the deviations for both sets of data are often less than 10%. Nonetheless, sometimes, greater deviations can be observed for inefficient countries, and they mostly pertain to countries which have very low efficiency scores for the Prosperity Quintet. In these cases, the calculations regularly lead to implausible targets, something that we have already criticized (see BG, CZ, EST, LT, LV, and RO regarding CCR, and LT, LV, and PL regarding BCC-O).

Due to the mentioned similarities between the efficiency scores of many countries, their scale efficiencies (SE) are mostly similar, too, as can be seen comparing the corresponding columns of Tables 5 and 8. The scale efficiencies of the basic welfare set have a tendency to be larger than those ones of the Prosperity Quintet, especially regarding the input-oriented projection onto the efficient frontier. Only for a few countries (BG, H, and RO), the scale efficiency of an output-oriented projection is smaller than 70%, which is why the distance between the associated parts of the BCC envelopment and the CCR envelopment is not so high. For several (often larger and higher developed) countries, the scale efficiencies for the basic welfare set are 100% or nearly 100%. Thus, the associated parts of the efficient frontier display (almost) constant RTS regarding the chosen welfare indicators.

Column 5 of Table 8 shows that the input-oriented projection of all inefficient countries refers to parts of the CCR-efficient frontier where the sum of the activity levels of their benchmarking partners is smaller than one, i.e., that their reference points on the BCC-efficient frontier are characterized by (increasing) economies of scale. Except Ireland, this is also true for the welfare evaluation with respect to the Prosperity Quintet in column 6 of Table 5. Regarding the output-oriented projection, similar assertions can be stated. Exceptions are B, F, GR, and IRL in Table 5 as well as B and F in Table 8, which are projected to reference points with decreasing RTS.

The comparison of Tables 5 and 8 has shown that, for most countries, each of the standard radial DEA models leads to similar (super-) efficiency scores and in part to almost identical values (such as for A). Despite the close relationship between the underlying data in Tables 4 and 7, this broad consistency in the welfare measurement could not necessarily be expected, taking into account the different definition and number of the other inputs and outputs. Nevertheless, the exceptions that we have specified demonstrate that the solutions to the individual DEA models

[14] Corresponding statements inevitably apply to niRTS- and ndRTS-efficient countries, too, because, as explained earlier, these must attain the CCR- or the BCC-efficiency score of the same orientation.

in fact differ significantly from each other for some countries (BG, CZ, EST, LT, LV, PL, and RO) when the results of the two sets of welfare indicators are compared.

6 Extended DEA case study for the basic welfare set

As concluded in Sects. 4.3 and 5.2, the set of six basic indicators is superior to the Prosperity Quintet, especially since nonsensical targets for individual countries can be avoided more easily.[15] However, even if the superiority of the set of basic indicators has been demonstrated and discussed, there are still open questions from a methodological point of view. Thus, for example, the 80/20 ratio has not been directly addressed so far, and the problem of weakly efficient solutions has also not yet been considered. We examine these two aspects in the following subsections.

6.1 Measurement of income balance

Table 6 clearly reveals that, on the basis of the average incomes of the three population classes (as 'outputs'), the average income of all inhabitants is maximized simultaneously (as via the GDP earlier in the case of the Prosperity Quintet). However, the question arises as to how the income distribution is taken into account when applying DEA to the six basic indicators. As already mentioned in Sect. 4.3, minimizing the 80/20 ratio does not appear to be meaningful, but conversely, nor does its maximization either. Excessive inequalities in income are perceived as unjust and have a dysfunctional effect on an economy and the social interaction of a society (Wilkinson and Pickett 2009).

The output-oriented radial DEA models are constructed in such a way that the existing ratio of the average income of the three population classes remains constant, because they attempt to increase the income proportionately. Accordingly, the objective function of (1) indicates the factor by which all incomes can be uniformly increased until the boundary of the data envelopment is reached; or the factor by which they can be reduced in the case of super-efficiency without compromising the efficiency of a country. Thus, for example, Belgium should increase all incomes to 100/62 = 161% under constant RTS and to 100/72 = 139% under variable RTS (see columns 2 and 3 of Table 8).

In such an application of DEA, the income distribution of a country is not scrutinized. While we can regard Belgium's 80/20 ratio, which attains a value of 391%, as well balanced, it appears to be not well balanced for Bulgaria, which has an 80/20 ratio of 611%. The almost identical CCR efficiency scores of both countries (B 0.62 and BG 0.64) provide no information about the inequality of the incomes at all. Since neither minimizing nor maximizing the 80/20 ratio is reasonable, it would seem obvious to determine a specific value or a range for this

[15] As already stated before (footnote 12), this superiority does not mean that the basic set of welfare indicators is not to criticize anymore. However, the formulation of a 'right' set of such indicators is beyond the intention of our paper. Moreover, this always depends on the specific evaluation goals as well as other aspects of the intended analysis.

ratio and to integrate it directly as a constraint within the selected DEA model (cf., e.g., Thompson et al. 1990 regarding the so-called assurance regions). However, this approach often leads to infeasible solutions of the models.

Balanced DEA (as presented in Sect. 2.3) appears to be a fruitful approach, since it allows the degree of imbalance of a country's income distribution to be measured within the DEA methodology itself.[16] This provides a second performance measure in addition to that of the efficiency score. A prerequisite for using balanced DEA is the ability to exogenously specify in this case what income distribution may be regarded as 'balanced'. A complete fulfillment of this specification is then indicated by a balance score of 100%. For demonstration purposes, only, we assume in the following that ratios of 4:2:1 of the per capita incomes of the richest to the middle and to the poorest income class are balanced. Columns 3 and 5 of Table 9 show the balance scores for the output-oriented CCR and BCC models adjacent to the efficiency scores already known from Table 8 (now without super-efficiency).

As explained in Sect. 2.3, the balance score β indicates the proportion of the efficiency score θ to which θ relatively decreases if a country is forced to lower the average income of one or two income classes, such that the targeted ratio of 4:2:1 is achieved. Considering columns 2 and 3 as examples for the CCR model, Belgium has, for instance, an income balance of 98% as against that of Bulgaria at only 86%. The product $\beta \cdot \theta$ of the efficiency score and the related balance score can be interpreted as the efficiency score of the country at a balanced distribution of income (e.g., $0.64 \times 0.86 = 0.55$ for BG or $0.88 \times 0.94 = 0.83$ for F).

According to Table 9, even countries without a balanced income distribution can still be efficient (e.g., DK and GB), or vice versa, inefficient countries can have balanced incomes (such as H and M). Similarly, there are countries which are efficient as well as balanced (EST, FIN, L, NL, and S) and, on the other hand, those that have a high inefficiency with very unequal incomes (e.g., GR, LV, and RO). When comparing column 3 of Table 9 with column 3 of Table 4, we find that countries with a (CCR) income balance of below 70% (E 0.56, GR 0.66, LV 0.65, and RO 0.68) have an 80/20 ratio of more than 600% (E 7.16, GR 6.62, LV 6.57, and RO 6.37). Thus, the balance score generally has the advantage that the previously calculated efficiency scores are presented in a different light: It enables a more differentiated view of additional aspects which cannot be usefully modeled through the indicators themselves, but which are still interesting in terms of the interpretation of results.

However, a major deficiency of this type of DEA-integrated balance measurement is based on the general problem of all the standard DEA models that a radial efficiency score of 100% does not necessarily mean that the country is also (strongly) efficient. Countries may be only weakly efficient, with the result that they can no longer improve themselves in all outputs or all inputs, but probably in some inputs or outputs. Consequently, the balance scores of 100% only reflect a *weak*

[16] Less rigid balance measures can be defined, for example, in such a way that a redistribution of rich to poor is undertaken between the income classes without reducing the total income. However, there is still no literature on this.

balance. In fact, no country of the EU-27 displays precisely the (here exemplarily) predefined distribution of income in a ratio of 4:2:1.

6.2 Slack-based welfare measurement

To avoid the deficit of weak efficiency, the application of DEA to welfare measurement can be further improved using additive models (as shown in Sect. 2.2). Because of its compatibility with the radial models, the SBM model by Tone (2001) is particularly suitable for enabling comparisons with our previous results. Since it is hardly possible to justify the orientation of a model meaningfully when measuring welfare, the non-orientation of additive models represents yet another benefit. However, because Tone's (2002) definition of super-efficiency differs from that of his corresponding efficiency score, his super-efficiency scores cannot be easily compared with that of the radial models. This is why, we no longer take super-efficiency into account in Table 9. Columns 6–9 accordingly show the Tone efficiency scores and the related balance factors.

It becomes obvious that no country is balanced at 100%. For almost all countries, the differences in the balance values are low between the SBM models (exceptions: D 0.78 vs. 0.96; I 0.82 vs. 0.63), as well as between the radial models (except BG, RO, and SLO, probably due to the slacks resulting from their weak efficiency).

As explained in Sect. 2.2, the respective radial efficiency score tallies exactly with that of Tone if the country considered is strongly efficient in terms of the radial DEA model. In that case, no slacks occur, so that efficiency scores under the respective RTS are both 100%. In cases of inefficient countries, however, the Tone efficiency measure generally attains a score which is genuinely smaller than the radial efficiency measure with the same RTS.

In our case, all nine countries having a CCR score of 100% are also SBM-cRTS efficient, as all 14 countries with a BCC score of 100% are also SBM-vRTS efficient. To that extent, the results of the SBM model variants surprisingly fit well to those of the radial models, saying that all countries with a radial DEA score of 100% have already been identified as strongly efficient ones. Hence, there are no weakly efficient countries in the data set. However, for inefficient countries, the SBM efficiency score for all RTS variants is usually significantly smaller than the respective radial one. Thus, an advantage of the SBM model is its much better discrimination between the inefficient countries. The SBM model actually discriminates inefficient countries so strongly that efficiency scores above 60% are rare for constant RTS (except A 0.99 and SLO 0.69). For variable RTS, there is only one inefficient country with an efficiency score of above 70% (SK 0.73). On one hand, this is due to the fact that the SBM models take into account all slacks and, on the other hand, due to the fact that the efficiency measure is defined differently in comparison with the radial models, namely through the arithmetic means of the relative slacks.

Due to space limitations, the results concerning SE and RTS are not listed in Table 9. Fact is that, on one hand, the SBM-niRTS efficiency scores are identical with the SBM-cRTS scores, while, on the other hand, the scores for ndRTS and vRTS are equal, too. As the example of Belgium in Table 10 in Sect. 7 (with

Table 9 Radial- and slack-based efficiency and balance scores (in %) for EU-27 with data of Table 7

Country	CCR		BBC-O		SBM-cRTS		SBM-vRTS	
	θ	β	θ	β	θ	β	θ	β
A-Austria	99	95	100	96	99	81	100	85
B-Belgium	62	98	72	99	49	97	49	97
BG-Bulgaria	64	86	100	100	47	78	100	78
CY-Cyprus	67	86	72	88	42	95	47	95
CZ- Czech Republic	61	87	76	87	44	95	60	95
D-Germany	100	98	100	100	100	78	100	96
DK-Denmark	100	89	100	90	100	83	100	83
E-Spain	81	56	92	58	51	74	66	71
EST-Estonia	100	100	100	100	100	84	100	84
F-France	88	94	89	93	57	96	64	92
FIN-Finland	100	100	100	100	100	96	100	96
GB-Great Britain	100	87	100	89	100	58	100	69
GR-Greece	50	66	50	67	28	77	31	76
H-Hungary	49	100	89	99	39	99	70	99
I-Italy	86	75	100	78	57	82	100	63
IRL-Ireland	100	92	100	94	100	67	100	67
L-Luxembourg	100	100	100	100	100	98	100	98
LT-Lithuania	54	79	69	87	39	86	63	86
LV-Latvia	56	65	73	72	33	79	55	78
M-Malta	53	100	57	100	44	99	50	99
NL-Netherlands	100	100	100	100	100	95	100	95
P-Portugal	56	71	60	72	36	85	43	85
PL-Poland	49	82	60	87	34	90	49	90
RO-Romania	55	68	100	100	34	77	100	74
S-Sweden	100	100	100	100	100	94	100	94
SK-Slovakia	56	94	79	93	46	95	73	95
SLO-Slovenia	76	89	100	100	69	91	100	91

Sweden as unique benchmarking partner) proves, this does not necessarily mean that, for cRTS, the sum of activity levels of the enveloping DMUs must not be larger than one. Our numerical example in Sect. 2.2 has shown for additive DEA models that many different reference points on the efficient frontier with the same efficiency score are possible. Hence, it is problematic to conclude any kind of RTS for an inefficient country under consideration (cf. Section 2.2). Except Belgium, all other SBM-cRTS-inefficient countries, however, have reference points with activity levels sums smaller than one. On the contrary, the SBM scale efficiencies can directly be calculated by dividing the respective SBM-cRTS and SBM-vRTS efficiency scores (columns 6 and 8 of Table 9). With the exception of Belgium again, the SE_{Tone} scores are either equal to the SE_O and SE_I scores in columns 6 and

Table 10 Results of various DEA models for Belgium

DEA model	Prosperity Quintet: targets					Eff.-score θ (%)	Scale $\sum \lambda_j$	Benchmarks
	GDP/POP (€/cap)	80/20-ratio (%)	SER (%)	EFC/BC (%)	PDR (%)			
Original data	29630	391	15.2	406	100	–	–	–
BCC-O (5)	40987	391	10.8	406	43	72	1.00	0.10 IRL; 0.20 L; 0.71 S
BCC-I (5)	29630	357	3.9	310	68	91	1.00	0.82 NL; 0.18 SLO
CCR-O (6)	48279	391	12.8	406	53	62	1.44	0.53 D; 0.91 S
CCR-I (6)	29822	391	7.9	251	53	62	0.89	0.32 D; 0.56 S
SBM-cRTS (6)	48039	374	11.9	406	38	49	1.35	1.35 S
SBM-vRTS (6)	35536	374	8.8	300	38	49	1.00	1.00 S
SBM-niRTS (6)	30675	374	7.6	259	38	49	0.86	0.86 S

7 of Table 8 or are smaller, mostly only marginally smaller. For three countries (BG, I, and RO), however, they are considerably smaller. For nine countries (BG, H, I, LT, LV, PL, RO, SK, and SLO), we have $SE < 70\%$ which implies that the respective parts of the efficient frontiers (of the linear and the convex envelopment to which they are projected) considerably differ.

The reasons for the efficiency or inefficiency of individual countries can be further explored through the usual detailed analysis of the DEA results concerning the weights, the targets, and the benchmarking partners. For variable RTS (and analogously for niRTS or ndRTS), however, the reason for the efficiency of some countries is obvious, since they have a minimum input or a maximum output. In these cases, they cannot be dominated by any point of such a kind of data envelope (L regarding indicators HIC, MIC and LIC; NL for SER; RO for EFC; BG for DBC). This reflects the specific nature of DEA: It enables each DMU to focus on individual welfare indicators whilst disregarding others.

7 Discussion of the case studies' results

In the present paper, we have stated that there are drawbacks to the existing welfare evaluations with DEA, since these evaluations frequently use a widely available data set without challenging it, and most of them adopt a single standard radial DEA models without questioning it. To illustrate the problems resulting from such an approach and how they might be (partly) solved, we have applied such standard DEA models to two formally different sets of welfare indicators which are closely related to each other. We have compared the results of these applications, discussed them critically, as well as gradually extended and improved the models (income balance and SBM models). In this section, we draw conclusions from our findings for the application of DEA in the context of welfare evaluation—as an illustrative and typical example of non-production contexts—regarding three aspects:

Selection and composition of welfare indicators: We have shown that indicators like that of the Prosperity Quintet are not directly suitable as inputs and outputs for a DEA analysis. Some of them consist of quotients of more basic welfare indicators. The values of some quotients are limited upwards or downwards. BCC models can avoid nonsensical values as benchmarks in these cases. However, they may calculate different efficiency scores for one and the same country, depending on their input or output orientation. Basic (volume) indicators may avoid such constellations. Even though it is problematic to take income distributions into account when analyzing the efficiency of incomes themselves, the calculation of the balance score as an additional performance measure ensures a representation of this important aspect of welfare.

Comparability of the selected DMUs: Denkwerk Zukunft considers the Prosperity Quintet to be more suitable for the early industrialized countries. However, this characteristic hardly applies to a number of Eastern European countries. Accordingly, their efficiency score is rarely high in the DEA calculations, at least for constant RTS; on the other hand, it is also not so much lower than the one of many Western European countries. Therefore, a comparison between these EU countries seems to be quite appropriate. It is questionable, however, whether a small state like Luxembourg is comparable with more populous countries, especially due to its 'business model'—special tax avoidance policy for large multinational concerns—which is only limitedly replicable for bigger countries. In such cases, further analysis would be needed for testing the robustness and sensitivity of DEA results. Indeed, eliminating Luxembourg from the data set leads to significant improvements for several countries. For the CCR model, this is the case for the efficiency scores of seven (Eastern European) countries with an improvement of more than 10%.

Choice of the concrete DEA model: On one hand, there are close formal relationships between the efficiency scores of different radial DEA models for the same set of welfare indicators, which often result in similar classifications as efficient or inefficient or in strongly correlated rankings of countries. On the other hand, however, we may not conclude that the selection of the RTS property or the orientation does not play a significant role, because, for a particular country, significant changes in the efficiency evaluation may result in extreme cases even efficiency versus strong inefficiency. This result intensifies if slack-based, non-oriented models are used, as demonstrated in Sect. 6.2 with the Tone model. Therefore, it is decisive for welfare evaluation which DEA model is used for the investigation; especially when only one model is considered without further sensitivity analyses, which is usual practice in the literature on welfare evaluation.

In many cases, the large differences in various model results do not occur for the efficiency scores themselves, but rather for the obtained target values. For example, in the case of Belgium: the SBM models lead to the same efficiency score of 49% under all variations of RTS. However, the desired targets differ significantly, as Table 10 shows in the last three rows. In addition to the original data for the Prosperity Quintet, Table 10 contains detailed results for seven selected models.

These are the results for the input- and output-oriented radial standard models, once with variable RTS for the data of the Prosperity Quintet (BCC-O/I), then with

constant RTS for the data of the six basic indicators (CCR-O/I), as well as the results for three Tone models (SBM-c/v/niRTS). The targets for the set of six basic indicators are recalculated into respective values of the Prosperity Quintet, using the relations of Table 6 to make them comparable. Moreover, the respective efficiency score and the relevant benchmarking partners are listed along with their proportions. The respective sum of these proportions indicates the RTS of those parts of the efficient frontier to which Belgium is projected.

In summary, the 80/20 ratio changes only moderately, while the targets for the other four welfare indicators vary strongly in part. The variation of the orientation in the BCC and CCR models alone has serious consequences for the DEA results. Unlike for the inputs and outputs in the case of real production processes, however, a particular orientation can hardly be meaningfully justified with respect to the 'inputs' to be minimized and 'outputs' to be maximized for the investigated sets of welfare indicators. The same applies to the choice of returns-to-scale and the efficiency measure. However, from the point of view of a decision maker or politician, the question might be essential of which DEA model elicits the best relative evaluation of their own country.

8 Conclusions and outlook

Welfare evaluation of countries is an important topic. Our literature review as well as our detailed case studies on the welfare of 27 countries of the European Union have shown that data envelopment analysis (DEA) seems to be a methodology that is problematic if used in this context and that can lead to findings which are empirically not well founded. In our view, this conclusion also seems to be true in other application fields of DEA where the specific modeling assumptions of DEA cannot be approved easily.

In particular, our paper has outlined, in an iterative process, the problems which can occur if DEA is used without reflections on the context, particularly in non-production cases, and on how these problems might be solved in parts by gradually changing the model specifications. Especially for a rather inexperienced user, DEA poses characteristics and pitfalls, through which methodically related results can be misinterpreted erroneously as being empirical findings. However, since the production-theoretical foundation of DEA (Charnes et al. 1985) is hardly sustaining in the welfare context, DEA itself provides no evidence as to which model is suited best for the specific case of welfare measurement. It is hardly possible to pick out the best model, because several models can lead to plausible results, and the underlying assumptions are partly difficult to justify. Each model has its specific advantages and disadvantages, whether it is the simplicity of calculation and interpretation or the sophistication of the results, which in return involves a more difficult formulation of a model. Our observations make it evident that caution is required when interpreting DEA results. Since DEA is used to gain insights and management recommendations in various application fields, it must always be considered that, besides empirical effects, purely methodological effects can emerge.

Facing the difficulties and advantages as well as disadvantages of certain DEA model choices, extended and more detailed frameworks could be helpful to guide the user through a well-founded selection of indicators as well as through the reasonable choice of proper DEA model characteristics. They should persuade the user to adequately consider all necessary questions, which is why such frameworks should be systematically and iteratively designed. Although there are already some useful and supportive works on frameworks for DEA (see, e.g., Emrouznejad and De Witte 2010), their validity still has to be successfully approved in practice. Moreover, such frameworks need to be further developed continuously to meet the requirements of all different types of DEA applications. Particularly, adopting a generalized perspective on DEA—such as the one developed in Dyckhoff and Allen (2001) which is based on a multi-criteria production theory (Dyckhoff 2018)—may provide some more valuable advice for the further enhancement of those frameworks (see Wojcik 2018 for such an approach). This is especially true regarding more detailed insights for the systematic derivation and justification of performance indicators, a topic which has not been discussed sufficiently in the literature so far. Thus, under certain circumstances, the application of DEA in instances which are not directly based on a classical production process can either be facilitated and enhanced or otherwise be avoided. Until now, however, empirical evidence that DEA has really improved the practice of performance measurement and benchmarking is lacking, even in pure production contexts.

References

Andersen, Per, and Niels Christian Petersen. 1993. A procedure for ranking efficient units in data envelopment analysis. *Management Science* 39 (10): 1261–1264.

Ahn, Heinz, Peter Bogetoft, and Ana Lopez. 2018. Measuring potential sub-unit efficiency to counter the aggregation bias in benchmarking. *Journal of Business Economics: forthcoming*. https://doi.org/10.1007/s11573-018-0901-0.

Ali, Emrouznejad, and Kristof De Witte. 2010. COOPER-framework: A unified process for non-parametric projects. *European Journal of Operational Research* 207 (3): 1573–1586.

Banker, Rajiv D., Abraham Charnes, and William W. Cooper. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30 (9): 1078–1092.

Blancard, Stéphane, and Jean F. Hoarau. 2011. Optimizing the new formulation of the United Nations' human development index: an empirical view from data envelopment analysis. *Economics Bulletin* 31 (1): 989–1003.

Bougnol, Marie-Laure, José H. Dulá, Marcos P. Estellita Lins, and Angela C. Moreira da Silva. 2010. Enhancing standard performance practices with DEA. Omega 38 (1–2): 33–45.

Böhringer, Christoph, and Patrick Jochem. 2007. Measuring the immeasurable—A survey of sustainability indices. *Ecological Economics* 63 (1): 1–8.

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2 (6): 429–444.

Charnes, Abraham, William W. Cooper, Boaz Golany, and Lawrence M. Seiford. 1985. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics* 30: 91–107.

Coelli, Timothy J., D.S. Prasada Rao, Christopher J. O'Donnell, and George E. Battese. 2005. *An introduction to efficiency and productivity analysis*, 2nd ed. New York: Springer.

Cooper, William W., Lawrence M. Seiford, and Kaoru Tone. 2007. *Data envelopment analysis—A comprehensive text with models, applications, references and DEA-solver software*. New York: Springer.

Cook, Wade D., Kaoru Tone, and Joe Zhu. 2014. Data envelopment analysis: Prior to choosing a model. *Omega* 44 (1): 1–4.

Denkwerk Zukunft. 2011. The prosperity quintet 2011—Measuring prosperity in Germany and other early industrialised countries. http://www.wohlstandsquintett.de. Accessed 25 Nov 2016.

Denkwerk Zukunft. 2014. The Prosperity quintet 2014—Measuring prosperity in Germany and other early industrialised countries. http://www.wohlstandsquintett.de. Accessed 25 Nov 2016.

Despotis, D.K. 2005. A reassessment of the human development index via data envelopment analysis. *Journal of the Operational Research Society* 56 (8): 969–980.

Domínguez-Serrano, Mónica, and Francisco J. Blancas. 2011. A gender wellbeing composite indicator: the best-worst global evaluation approach. *Social Indicators Research* 102 (3): 477–496.

Dyckhoff, Harald. 2018. Multi-criteria production theory—Foundation of non-financial and sustainability performance evaluation. *Journal of Business Economics* 88 (7): 851–882. https://doi.org/10.1007/s11573-017-0885-1.

Dyckhoff, Harald, and Katrin Allen. 2001. Measuring ecological efficiency with data envelopment analysis (DEA). *European Journal of Operational Research* 132 (2): 312–325.

Dyckhoff, Harald, Marcel Clermont, Alexander Dirksen, and Eleazar Mbock. 2013. Measuring balanced effectiveness and efficiency of German business schools' research performance. In *Performance Management im Hochschulbereich*, ed. Alexander Dilger, Harald Dyckhoff, and Günter Fandel, 39–60. Wiesbaden: Springer Gabler.

Dyckhoff, Harald, and Sebastian Gutgesell. 2015. Properties of DEA-integrated balance and specialization measures. *OR Spectrum* 37 (2): 503–527.

Dyckhoff, Harald, Eleazar Mbock, and Sebastian Gutgesell. 2015. Distance-based measures of specialization and balance in multi-criteria: A DEA-integrated method. *Journal of Multi-Criteria Decision Analysis* 22 (3–4): 197–212.

Dyckhoff, Harald, Sylvia Rasenhövel, and Kirsten Sandfort. 2009. Empirische Produktionsfunktion betriebswirtschaftlicher Forschung: Eine Analyse der Daten des Centrums für Hochschulentwicklung. *Zeitschrift für betriebswirtschaftliche Forschung* 61 (1): 22–56.

Dyckhoff, Harald, and Thomas Spengler. 2010. *Produktionswirtschaft*, 3rd ed. Berlin: Springer.

Dyson, Robert G., Richard Allen, Ana S. Camanho, Victor V. Podinovski, Cláudia S. Sarrico, and Estelle A. Shale. 2001. Pitfalls and protocols in DEA. *European Journal of Operational Research* 132 (2): 245–259.

Emrouznejad, Ali, and Gholam R. Amin. 2009. DEA models for ratio data: convexity considerations. *Applied Mathematical Modelling* 33 (1): 486–498.

Eurofound. 2012. Third European quality of life survey. Quality of life in Europe: Impacts of the crisis. Brüssel. http://www.eurofound.europa.eu/surveys/eqls/2011/index.htm. Accessed 25 Jun 2016.

Eurostat. 2013. Sustainable development in the European Union—2013 monitoring report of the EU sustainable development strategy. http://ec.europa.eu/eurostat/documents/3217494/5760249/KS-02-13-237-EN.PDF/f652a97e-e646-456a-82fc-34949bbff956. Accessed 25 Jun 2016.

Färe, Rolf, Shawna Grosskopf, and C.A.K. Lovell. 1994. *Production frontiers*. Cambridge: Cambridge University Press.

Global Footprint Network. 2013. National footprint accounts 2013 edition. www.footprintnetwork.org. Accessed 25 May 2016.

Hollingsworth, Bruce, and P. Smith. 2003. Use of ratios in data envelopment analysis. *Applied Economic Letters* 10 (11): 733–735.

Jahanshahloo, Gholam R., Farhad Hosseinzadeh Lofti, Abas A. Noora, and Bahram Rahmani Parchikolaei. 2011. Measuring human development index based on malmquist productivity index. *Applied Mathematical Science* 5 (62): 3057–3064.

Jablonsky, Josef. 2004. Application of alternative methods in recalculation of the human development index. In Proceedings of the 12th international conference quantitative methods in economics, ed. Martin Lukacik, 93–99. Bratislava.

Jurado, Antonio, and Jesus Perez-Mayo. 2012. Construction and evolution of a multidimensional well-being index for the Spanish regions. *Social Indicators Research* 107 (2): 259–279.

Kerpen, Philip. 2016. *Praxisorientierte data envelopment analysis*. Wiesbaden: Springer Gabler.

Lee, Hsuan-Shih, Kuang Lin, and Hsin-Hsiung Fang. 2006. A fuzzy multiple objective DEA for the human development index. In *Proceedings to the 10th international conference on knowledge-based intelligent information and engineering systems*, ed. Bogdan Gabrys, Rorbert J. Howlett, and Lakhmi C. Jain, 922–928. Berlin: Springer.

Li, Peng, and X.Ma. Zhan. 2011. The comprehensive evaluation of regional economic development in Inner Mongolia. *Advanced Materials Research* 230 (1): 44–48.

Lovell, C.A., and J.T. Pastor. 1995. Units invariant and translation invariant DEA models. *Operations Research Letters* 18 (3): 147–151.

Lozano, Sebastian, and Ester Gutiérrez. 2008. Data Envelopment Analysis of the Human Development Index. *Society Systems Science* 1 (2): 132–150.

Malul, Miki, Yossi Hadad, and Avner Ben-Yair. 2009. Measuring and ranking of economic, environmental and social efficiency of countries. *International Journal of Social Economics* 36 (8): 832–843.

Mariano, Enzo Barberio, Vinicius Amorim Sobreiro, Daisy Aparecida, and Daisy Aparecida do Nascimento Rebelatto. 2015. Human development and data envelopment analysis: A structured review. *Omega* 54 (1): 33–49.

Martić, Milan, and Gordana Savić. 2001. An application of DEA for comparative analysis and ranking of regions in Serbia with regards to social-economic development. *European Journal of Operational Research* 132 (2): 343–356.

Murias, Pilar, Fidel Martinez, and Carlos De Miguel. 2006. An economic wellbeing index for the Spanish provinces: a Data Envelopment Analysis approach. *Social Indicators Research* 77 (3): 395–417.

Murphy, Orla A., Ping Wang, Sunny X. Wang, and Greg Tkacz. 2013. An economic efficiency study on different regions of Ghana via slacks-based Data Envelopment Analysis and regression analysis. *Applied Economics* 45 (34): 4773–4780.

Olesen, Ole B., Niels C. Petersen, and Victor V. Podinovski. 2015. Efficiency analysis with ratio measures. *European Journal of Operational Research* 245 (2): 446–462.

Olesen, Ole B., Niels C. Petersen, and Victor V. Podinovski. 2017. Efficiency measures and computational approaches for Data Envelopment Analysis models with ratio inputs and outputs. *European Journal of Operational Research* 261 (2): 640–655.

Pastor, Jeses T. 1996. Translation invariance in Data Envelopment Analysis: a generalization. *Annals of Operations Research* 66 (2): 93–102.

Ramanathan, Ramakrishnan. 2006. Evaluating the comparative performance of countries of the Middle East and North Africa: a DEA application. *Socio-Economic Planning Sciences* 40 (2): 156–167.

Reig-Martínez, Ernest. 2013. Social and economic wellbeing in Europe and the Mediterranean Basin: building an enlarged human development indicator. *Social Indicators Research* 111 (2): 527–547.

Sabermahani, Aasma, Mohsen Barouni, Hesam Seyedin, and Aidin Aryankhesal. 2013. Provincial Human Development Index, a guide for efficiency level analysis: the case of Iran. *Iranian Journal of Public Health* 42 (2): 149–157.

Shetty, Udaya, and T.P.M. Pakkala. 2010. Multistage method of measuring human development through improved directional distance formulation of Data Envelopment Analysis: application to Indian states. *Opsearch* 47 (3): 177–194.

Singh, Rajesh K., H.R. Murty, S.K. Gupta, and Anil K. Dikshit. 2012. An overview of sustainability assessment methodologies. *Ecological Indicators* 15 (1): 281–299.

Thanassoulis, Emmanuel, C.S. Maria Portela, and Ozren Despic. 2008. Data envelopment analysis: The mathematical programming approach to efficiency analysis. In *The measurement of productive efficiency and productivity growth*, ed. Harold O. Fried, C.Knox Lovell, and S.Shelton Schmidt, 251–420. New York: Oxford University.

Thompson, Russell G., Larry N. Langemeier, Chih-Tah Lee, Euntaik Lee, and Robert M. Thrall. 1990. The role of multiplier bounds in efficiency analysis with application to Kansas farming. *Journal of Econometrics* 46 (1–2): 93–108.

Tone, Kaoru. 2001. A slacks-based measure of efficiency in Data Envelopment Analysis. *European Journal of Operational Research* 130 (3): 498–509.

Tone, Kaoru. 2002. A slacks-based measure of super-efficiency in Data Envelopment Analysis. *European Journal of Operational Research* 143 (1): 32–34.

Wilkinson, Richard G., and Kate Pickett. 2009. *The spirit level: Why more equal societies almost always do better*. London: Allen Lane.

Wojcik, Victoria. 2018. *Performanceanalyse mittels Verallgemeinerter Data Envelopment Analysis: Vorgehensmodell und Evaluation*. Hamburg: Dr. Kovač.

Wojcik, Victoria, Harald Dyckhoff, and Sebastian Gutgesell. 2017. The desirable input of undesirable factors in data envelopment analysis. *Annals of Operations Research* 259 (1–2): 461–484.

Wu, Po-Chin, Chiung-Wen Fan, and Sheng-Chieh Pan. 2014. Does Human Development Index provide rational development rankings? Evidence from efficiency rankings in super efficiency model. *Social Indicators Research* 116 (2): 647–658.