Performance: The Output/Input Ratio

Thijs ten Raa

1 Introduction

A production unit, organization, firm, industry, or economy performs well if it produces much output per unit of input, in other words, when the output/input ratio is high. The main performance measure is productivity. There are subtle connections between performance, productivity, efficiency, and profitability. Analysis of their interrelations will take us through many issues and concepts of measurement and will connect different bodies of literature, namely in economics and operations research.

The measurement of performance using an output/input ratio presumes that output components and input components can each be aggregated. This is particularly true for inputs. Production requires multiple inputs, typically labor and capital services. On the output side, the aggregation issue is often circumvented. One way is to break down production in micro-activities, one for each type of output. This approach moves the aggregation issue away from commodities toward the micro-performance measures (Blackorby and Russell 1999). An obvious alternative way to circumvent output aggregation is to assume that there is a single performance criterion, such as profit, but this approach raises the question if profit is a better measure of performance than, say, real profit (profit divided by a price index). A windfall profit due to a price shock, without any change in the input-output

T. ten Raa (✉)
Utrecht School of Economics, Utrecht University, Utrecht, The Netherlands
e-mail: tenraa@uvt.nl

T. ten Raa and W. H. Greene (eds.), *The Palgrave Handbook of Economic Performance Analysis*, https://doi.org/10.1007/978-3-030-23727-1_3

structure of production, does not reflect an improvement in management performance. In other words, we better disentangle profit in a real performance component and a nominal price effect. This issue is related to the design of bonus schedules for managers, where profit is shown to be a good proxy for effort only if the distribution of the windfall component fulfills a certain property (the likelihood ratio monotonicity of Milgrom 1981).

Throughout this chapter, I assume constant returns to scale, unless explicitly stated otherwise. With increasing complexity, I will discuss, first, single input-single output production; second, multiple input-single output production; third, single input-multiple output production; and, fourth, multiple input-multiple output production. The simplest of these cases, single input-single output production, suffices to discuss the relationship between performance, productivity, efficiency, and profitability.

Consider a single input-single output industry with two firms, a duopoly. Denote the input quantities by x and the output quantities by y. Use superscripts to indicate to which firm a symbol pertains: firm 1 or firm 2. Let the first firm be the more productive than the second: $y^1/x^1 > y^2/x^2$. (This is an innocent assumption, because we are free to relabel the firms.) Then firm 1 can produce no more than it produces, at least under the assumptions that the data represent all conceivable practices of production and that the firm's input is fixed. Firm 2, however, could perform better by adopting the production technique of firm 1. That way it would produce y^1/x^1 units per unit of input and since it commands x^2 inputs, its potential output is $(y^1/x^1)\ x^2$. By the presumed productivity inequality, this exceeds the actually produced quantity, y^2.

In our discussion, we must distinguish observed market prices and competitive shadow prices. Market prices are observed and may vary. Some firms negotiate tighter labor conditions than others, and some firms may have shrewder salesmen, extracting higher prices. Someone who "could sell sand to the Arabs" exercises market power but is not productive; the market price exceeds the production price. Production prices are shadow prices which in turn are associated with the constraints of a program that determines the optimum allocation of resources. Later on, optimality will be linked to the consumer's preferences, but in the introductory Mickey Mouse duopoly, it reduces to the maximization of output subject to an input constraint. The maximization program can be applied to a firm (1 or 2) and to the industry (the duopoly), to determine firm and industry efficiencies. The simplest program features constant returns to scale and is applied to a firm, say firm 1:

$$\max_{\theta_1,\theta_2,c\geq 0} y^1 c : x^1\theta_1 + x^2\theta_2 \leq x^1, y^1\theta_1 + y^2\theta_2 \geq y^1 c. \tag{1}$$

In program (1), firm 1 runs activities 1 (input x^1, output y^1) and 2 (input x^2, output y^2) with intensities θ_1 and θ_2, respectively, and c is the expansion factor for output. The first constraint binds the required input by the available input. Denote the Lagrange multiplier or shadow price of this constraint by w (the labor wage). The second constraint binds the expanded output by the sum of the activity outputs. Denote the Lagrange multiplier or shadow price of this constraint by p (the product price). The shadow prices are relevant for performance measurement and are the variables of the dual program associated with the primal program, (1) in this case.

The dual program minimizes the value of the bounds subject to the dual constraint. The bounds are x^1 and 0, so the objective of the dual program is wx^1 or, equivalently, w. The dual constraint is $(w\ p)\begin{pmatrix} x^1 & x^2 & 0 \\ -y^1 & -y^2 & y^1 \end{pmatrix} \geq (0\ 0\ y^1)$, featuring the row vector of shadow prices, the matrix of coefficient rows, and the objective coefficients. The first two components of the dual constraint, $wx^1 \geq py^1$ and $wx^2 \geq py^2$, state that the prices render the two activities unprofitable. Rewriting, $w/p \geq y^1/x^1$ and $w/p \geq y^2/x^2$. By assumption that the first firm is more productive and because w is minimized, the first dual constraint is binding, $w/p = y^1/x^1$. In other words, *the real wage rate equals the highest productivity* and, therefore, this activity would break even.

There is an interesting connection between shadow prices and competitive markets. Without loss of generality, program (1) has been set up (by inclusion of coefficient y^1 in the objective function) such that the third component of the dual constraint, $py^1 = y^1$, normalizes the price system $(w\ p)$ such that $p = 1$. Hence $w = y^1/x^1$. The second, less productive, activity would be unprofitable under these prices. In other words, if shadow prices prevail and entrepreneurs are profit maximizers, they would select the optimal activity to produce output. The solutions to the profit maximization problems are not unique, but there exists a combination of solutions which is consistent with equilibrium; see ten Raa and Mohnen (2002).

In the primal program (1), the less productive activity is suppressed by setting $\theta_2 = 0$ and, therefore, $\theta_1 \leq 1$ and the maximum value is $c = 1$. Firm 1 cannot expand its output. Next consider firm 2. In program (1), in the right sides of the constraints, superscripts 1 are replaced by 2. The first two dual constraints, $wx^1 \geq py^1$ and $wx^2 \geq py^2$, remain the same, as does the conclusion that the first activity would be adopted to produce output. The maximum expansion factor equals the ratio of the highest productivity to the actual productivity, $c = (y^1/x^1)/(y^2/x^2)$. For example, if this number is 1.25,

potential output of firm 2 exceeds actual output by 25%. Conversely, actual output is only 80% of potential output. Firm 1, however, produces 100% of its potential output. The efficiency of firm 1 is 100% and the efficiency of firm 2 is 80%. Here efficiency is defined as the inverse expansion factor, $(y^2/x^2)/(y^1/x^1)$ for firm 2 and $(y^1/x^1)/(y^1/x^1)=1$ for firm 1. Efficiency is the performance measure. *Efficiency is equal to the ratio of actual to optimal productivity.* In the single input-single output case with constant returns to scale, introduced in this section, efficiency must be technical efficiency. However, in more general settings, the inverse expansion factor of efficiency will also encompass allocative efficiency, as we will see in Sects. 2 and 4.

2 Multiple Input-Single Output Production

In the bulk of the economic literature, including macro-economics, there are multiple inputs, such as labor and capital, but a single output. The inputs, x^1 for firm 1 and x^2 for firm 2, turn vectors and the input price will be represented by row vector w. The previous set-up is maintained and the extension to more than two firms is straightforward. However, because of the multiplicity of inputs, several activities may now be activated when a firm maximizes output given its input vector, x. The potential output given an input vector is a scalar, the value of a function, $y=F(x)$. This is the reduced form of program (1) with y equal to scalar y^1c and x equal to vector x^1. Mind that potential output y is the product of actual output and the expansion factor. F is called the *production function*. To define productivity as an output/input ratio, we must aggregate the input components, if only because division by a vector is impossible. The way to do this is intuitive, making use of a well-known property of Lagrange multipliers, namely that they measure the gain in output per unit of input. The rate of potential output with respect to input k is given by w_k, the shadow price of the k^{th} component of the constraint $\sum_i x^i\theta_i \leq x$. The productivity of input k is shadow price w_k. This is output per unit of input. Now the problem is that a unit of input is arbitrary. For example, sugar can be measured in kilograms or in metric pounds. Using the latter, a unit has half the size of the former, the number of units is doubled, and the shadow price is halved. We must aggregate across inputs in a way that is not sensitive with respect to the units of measurement. The way to do this is to first aggregate productivity over the units of the same input, k. The contribution to output of input k is w_kx_k and in this product, the two effects of taking metric pounds instead of kilograms cancel. Summing over inputs k, the value of the dual program is obtained. However, by the main theorem

of linear programming, the value of the dual program is equal to the value of the primal program, potential output y. The aggregate output/input ratio, $y/\sum_i wx^i$, is thus unity. The reason for this peculiar limitation is that output and input are different commodities; there is no common denominator. This is the economic problem of value and the classical solution is to express output in terms of resource contents, like labor values. Then, indeed, the output/input ratio is bound to be one.

Yet this framework is useful, because productivity levels are determined relative to a base economy, a base year. We do observe changes in the output/input ratio over time. For example, if the productive firm in Sect. 1, firm 1, increases output in the next period, then $w=y^1/x^1$ remains valid, hence productivity w increases. This argument is extendable to the multi-input case. Dropping firm indices, productivity growth of input k is $\dot{w}_k$, where the dot stands for the derivative with respect to time. This, again, is sensitive with respect to the unit of measurement. However, aggregating across inputs, weighing by the units of inputs, $\sum_k \dot{w}_k x_k$, the sensitivity gets lost, because $\dot{w}_k x_k = \frac{\dot{w}_k}{w_k} w_k x_k$, in which the ratio is a growth rate while the subsequent product was already seen to be insensitive with respect to the unit of measurement. It is also customary to express the change in the output/input ratio as a growth rate, by dividing by the level, $\sum_k w_k x_k$. In short, the output/input ratio grows at the rate

$$TFP = \sum\nolimits_k \dot{w}_k x_k / \sum\nolimits_k w_k x_k. \tag{2}$$

Expression (2) is called *total factor productivity growth*. *TFP* is the most prominent performance measure. The expression can be rewritten as a weighted average of the factor productivity growth rates, $\dot{w}_k/w_k$, with weights $w_k x_k / \sum w_k x_k$. These weights sum to one.

This direct approach from Lagrange multiplier-based input productivities to total factor productivity growth can be related to the Solow residual approach. Recall that the values of the primal and dual programs match, $py = wx$, where the right-hand side is the product of row vector w and column vector x, and that we normalized $p=1$. Differentiating totally, $\dot{w}\,x = \dot{y} - w\,\dot{x}$ and, therefore, expression (2) equals

$$TFP = (p\,\dot{y} - w\,\dot{x})/py. \tag{3}$$

Expression (3) is called the Solow residual; see ten Raa (2008), Sect. 7, and the references given there. Solow (1957) modeled technical change by letting the production function depend on time,

$$y = F(x,t). \tag{4}$$

Differentiating production function (4) with respect to time, indicating partial derivatives by subscripts, $\dot{y} = F'_x \dot{x} + F'_t$ or

$$(\dot{y} - F'_x \dot{x})/y = F'_t/F, \tag{5}$$

Now, if inputs are rewarded according to their marginal products, $w = pF'_x$, then the left-hand sides of Eqs. (3) and (5) match, and, therefore, the Solow residual (3) reduces to F'_t/F, i.e., *technical change*. This condition is fulfilled if the input prices are the shadow prices of the program that maximizes output subject to technical feasibility. The production possibility set, $\{(x, y): y \leq F(x, t)\}$, is the set which is either spanned by the observed input-output pairs or postulated by some functional form of function F. This distinction corresponds with nonparametric and parametric performance measurement. The first underpinning, by observed input-output pairs, is more fundamental, as the second underpinning, by a production function, can be shown to be generated by a distribution of input-output pairs, where the distribution represents the capacities of the activities. Houthakker (1955) demonstrated this for the Cobb–Douglas function, $Y = AK^{\alpha}L^{\beta}$, where K and L are inputs, Y is output, and A, α, and β are parameters with $\alpha + \beta < 1$, meaning there are decreasing returns to scale. The returns to scale decrease because of constraining third input, as will be explained next. Output notation Y is customary in the Cobb–Douglas literature. Moreover, we may now reserve y for full capacity output.

An *activity* is a pair of proportionate inputs and an output. The assumption of input proportionality facilitates normalization of the activity by the output to $(k, l; 1)$, with $k = K/Y$ and $l = L/Y$ fulfilling $Ak^{\alpha}l^{\beta} = 1$. The activities can be parameterized by one input, e.g., k. Then $l = (Ak^{\alpha})^{-1/\beta}$ and, therefore, the technology set of activities is $\{(k, (Ak^{\alpha})^{-1/\beta}; 1): k > 0\}$. Each activity can be run with intensity s_k. Total output will be $\int s_k \mathrm{d}k$, where the integral is taken over the positive numbers. The constraints are $\int s_k k \mathrm{d}k \leq K$ and $\int s_k l \mathrm{d}l \leq L$, where K and L are the factor endowments. However, Houthakker (1955) assumes there is a capacity constraint for each activity. A fixed input causes the capacity constraint. The fixed input is different than the variable inputs, capital, and labor. Houthakker (1955) suggests entrepreneurial resources. The distribution of entrepreneurial resources (i.e., of the capacity constraint) across activities $(k, l; 1)$ is considered to be given and denoted by $y(k, l)$. This distribution need not be concentrated on a frontier-like $\{(k, l): Ak^{\alpha}l^{\beta} = 1\}$. Some activities may dominate others, with

both components of (k, l) smaller. Yet a dominated activity may be run, because the superior activity, like all activities, has a capacity constraint. Activities can be run with intensities $0 \leq s(k, l) \leq y(k, l)$. Subject to the factor constraints $\iint s(k,l)k\mathrm{d}k\mathrm{d}l \leq K$ and $\iint s(k,l)l\mathrm{d}k\mathrm{d}l \leq L$, we maximize output $\iint s(k,l)\mathrm{d}k\mathrm{d}l$. This is an infinite-dimensional linear program, with a continuum of variables $s(k, l)$. Denote the shadow prices of the two factor constraints by r and w, respectively. By the phenomenon of complementary slackness, unprofitable activities, with unit cost $rk + wl > 1$, are not run, $s(k, l) = 0$. By the same argument, profitable activities, with unit cost $rk + wl < 1$, are run at full capacity, $s(k, l) \leq = y(k, l)$. Activities which break even, $rk + wl = 1$, have activity $0 \leq s(k, l) \leq y(k, l)$, but since the set of such activities has measure zero, we may set $s(k, l) = y(k, l)$. It follows that inputs and output are $K = \iint\limits_{rk+wl\leq 1} y(k,l)k\mathrm{d}k\mathrm{d}l$, $L = \iint\limits_{rk+wl\leq 1} y(k,l)l\mathrm{d}k\mathrm{d}l$, and $Y = \iint\limits_{rk+wl\leq 1} y(k,l)k\mathrm{d}k\mathrm{d}l$, respectively. The implicit assumption is that all factor input can be fully employed. There must be activities with factor intensity k/l below endowment ratio K/L and activities with factor intensity above the endowment ratio.

The three expressions, for inputs K and L and output Y, are interrelated by the two shadow prices r and w. The idea of Houthakker (1955) is to use the first two expressions to solve for r and w in terms of K and L. Substitution of the results in the third expression yields output as function of the inputs. Houthakker (1955) carries out this calculation for the capacity distribution with Pareto density function, $y(k, l) = \mu k^{\kappa-1} l^{\lambda-1}$, where μ, κ, and λ are positive constants. The result is $Y = AK^{\alpha}L^{\beta}$ with $\alpha = \kappa(\kappa + \lambda + 1)$, $\beta = \lambda(\kappa + \lambda + 1)$ and A a positive constant depending on μ, κ, and λ. In other words, a Pareto capacity distribution yields a Cobb–Douglas production function. This is Houthakker's Theorem. At the micro-level, activities have fixed input-output ratios—it takes given amounts of labor to operate given machinery and equipment—but a change in resources, such as the inclusion of the East German labor force in the year 1989, is accommodated by the activation of new activities and the deactivation of some incumbent activities. Reallocations of resources across activities manifest as substitutions.

The capacity distribution is not concentrated on a single isoquant in input space. Both k and l can be bigger, less efficient. In solving the output maximization, smaller input combinations are activated, but only to full capacity. Residual inputs are employed by more input intensive activities. The capacity constraints thus yield decreasing returns to scale. Indeed, the Cobb–Douglas function has exponents summing to a number less than one.

Houthakker's activity foundation of neoclassical production functions works only if returns to scale are decreasing.

Clearly, different capacity distributions for the activity levels will generate different production functions. Houthakker (1955) has generated a stream of theoretical and applied research. The bulk of this literature features a lower dimension, with only one variable input, namely labor, and again one fixed output, which is now capital. In this one fixed-one variable input framework, Levhari (1968) found the capital distribution for which total output is a CES function of the total fixed input (capital) and the total variable input (labor) and showed it encompasses the Cobb–Douglas function. Muysken (1983) has consolidated the Cobb–Douglas, CES, and VES functions by showing they are all generated by beta distributions, with alternative parametrizations. Two books on the distribution approach to production are Johansen (1972) and Sato (1975). In this literature, activities have fixed input-output proportions and capacity constraints explain the existence of inefficient activities. Increases in levels of inputs prompt the activation of less efficient activities, in Ricardian style. The law of one price yields rents to the more efficient activities. The activation of different activities prompts different proportions between the input totals and the output. Substitution is considered a symptom of the change in the range of active activities (run with positive intensity).

3 Single Input-Multiple Output Production

In classical economics, labor is the only factor input. All other inputs are produced commodities, also called intermediate inputs. Production output is used to fulfill intermediate demand and final demand, where the latter is defined residually, as the difference between output and intermediate input. Production output is also called gross output; similarly, final demand is also called net output. In standard input-output analysis, each output has a single technique to produce it. Assuming constant returns to scale, the input of commodity j, $j = 1, \ldots, n$, per unit of output, is denoted by the input vector $(a_{1j}, \ldots, a_{nj}, l_j)$, of which the components represent the n intermediate inputs and the factor input (labor), respectively. If these unit input requirements are constant and fixed, they cannot be reduced and, therefore, are necessarily efficient (actual and optimal production coincide). If, however, there is a set of input vectors for each product j, there is room to reallocate labor between alternative techniques, which may save labor or, alternatively, increase output. This would increase the output/input ratio from

actual productivity to optimal productivity. The ratio of the two would be efficiency. A deep result states that the optimal input vectors, one for each product, are independent of the composition of final demand. This is the substitution theorem, but for an obvious reason also called non-substitution theorem, which goes back to Samuelson (1951). The proof of the theorem has a long history, in which details have been worked out and minor flaws eliminated. This culminated in a proof based on the efficiency program of maximizing the expansion factor of a some net output vector, determining the optimal input vectors, one for each product, and showing that this combination of input vectors remains optimal when the net output vector is replaced by another one (ten Raa 1995).

The substitution theorem yields an all-purpose optimal technology, featuring one technique for each product. Given any net output vector, one can compare the optimal labor input to the actual labor input. The ratio is the efficiency of the economy.

4 Multiple Input-Multiple Output Production

The determination of efficiency is simple in the single output case: One maximizes output given the inputs and in the single input case, one can minimize the input given the output. A mechanical extension to the multiple input-multiple output world would be to expand the output vector while preserving its component proportions. This procedure, however, presumes that the mix of outputs should not be changed and is optimal. Yet it is a useful procedure and I will detail it and discuss its merits. The fundamental paper of this approach is Debreu's (1951) now classic "The Coefficient of Resource Utilization," which will be discussed first.

The economy comprises m consumers with preference relationships $\succsim_i$ and observed l-dimensional consumption vectors $y^i (i = 1, \ldots, m)$, where l is the number of commodities.[1] Z is the set of possible l-dimensional input vectors (*net* quantities of commodities consumed by the whole production sector during the period considered), including the observed one, z. A combination of consumption vectors and an input vector is *feasible* if the total sum—the economy-wide *net* consumption—does not exceed the vector of

[1] I stick to the performance literature notation of (factor) inputs x, (consumed) outputs y, and intermediate inputs z. In the general equilibrium literature, including Debreu (1951), the notation is (factor) inputs z, (consumed) outputs x, and intermediate inputs y.

utilizable physical resources, l-dimensional vector x.[2] Vector x is assumed to be at least equal to the sum of the observed consumption and input vectors, ensuring the feasibility of the latter.

The set of net consumption vectors that are at least as good as the observed ones is

$$B = \left\{ \sum y^{i'} : y^{i'} \succsim_i y^i, \quad i = 1, \ldots, m \right\} + Z. \tag{6}$$

The symbol B stands for "better" set. The minimal resources required to attain the same levels of satisfaction that come with x^i belong to $B^{\min}$, the south-western edge or subset of elements z' that are minimal with respect to $\geq$.[3] Assume that preferences $\succsim_i$ are convex and continuous, and that production possibilities form a convex and closed set, then the separating hyperplane theorem yields a supporting price row vector $p(x') > 0$ (all components positive) such that $x'' \in B$ implies $p(x')x'' \geq p(x')x'$. The *Debreu coefficient of resource utilization* is defined by

$$\rho = \max_{x'} \{ p(x')x'/p(x')x : x' \in B^{\min} \}. \tag{7}$$

Coefficient ρ measures the distance from the set of minimally required physical resources, $x' \in B^{\min}$, to the utilizable physical resources, x, in the metric of the supporting prices (which indicate welfare indeed). Debreu (1951, p. 284) shows that the distance or the max in (7) is attained by

$$x' = \rho x \in B^{\min}. \tag{8}$$

In other words, the Debreu coefficient of resource utilization is the smallest fraction of the actually available resources that would permit the achievement of the levels of satisfaction that come with x^i. Coefficient ρ is a number between zero and one, the latter indicating full efficiency. In modern terminology, this result means that ρ is the *input-distance function*, determined by the program

$$\rho = \min_r \left\{ r : \sum y^{i\prime} + z' \leq rx, y^{i\prime} \succsim_i y^i, z' \in Z \right\}. \tag{9}$$

[2]For example, if the last commodity, l, represents labor, and this is the only nonproduced commodity, then $x = Ne_l$, where N is the labor force and e_l the l-th unit vector.

[3]By convention, this vector inequality holds if it holds for all components.

5 The Farrell Efficiency Measure

Another classic paper is Farrell (1957), which decomposes efficiency in technical efficiency and price efficiency. Here, technical efficiency is measured by the reduced level of proportionate inputs (as a percentage between 0 and 100) such that output is still producible. Price efficiency is the fraction of the value of an input vector with possibly different proportions (but the same output) to the value of the efficient input vector with the given proportions. Farrell (1957) notes the similarity between his technical efficiency and the Debreu coefficient of resource utilization: Both the Farrell technical efficiency measure and the Debreu coefficient of resource utilization are defined through proportionate input contractions, but the analogy is sheer formality and even confusing at a conceptual level. The analogy suggests that Farrell takes the Debreu coefficient to measure technical efficiency and augments it with a reallocative efficiency term, thus constructing a more encompassing overall measure. However, it is the other way round; the sway of the Debreu coefficient is far greater than that of Farrell's efficiency measure. Farrell's price efficiency measure is a partial (dis)equilibrium concept, conditioned on prices. It takes into account the cost reduction attainable by changing the mix of the inputs, given the prices of the latter.

The Debreu coefficient, however, is a general (dis)equilibrium concept. It measures the technical and allocative inefficiency in the economy given only its fundamentals: resources, technology, and preferences. Prices are derived and enter the definition of the Debreu coefficient, see (2). Debreu (1951) then *proves* that the coefficient can be freed from these prices, by Eq. (8) or non-linear program (9). The prices are implicit, supporting the better set in the point of minimally required physical resources. The Debreu coefficient measures technical *and* allocative inefficiency, both in production and in consumption, solving the formidable difficulty involved in assessing prices, referred to by Charnes et al. (1978, p. 438). Farrell refrains from this, restricting himself to technical efficiency and price-conditioned allocative efficiency, which he calls price efficiency.

The formal analogy between the Debreu coefficient and the Farrell measure of technical efficiency prompted Zieschang (1984) to coin the phrase "Debreu-Farrell measure of efficiency," a term picked up by Chakravarty (1992) and Grifell-Tatjé et al. (1998), but this practice is confusing. Debreu's coefficient of resource allocation encompasses both Farrell's technical efficiency and his price efficiency measures and frees the latter from prices. On top of this, Debreu's coefficient captures consumers' inefficiencies. The confusion persists. Färe et al. (2002) speak of the "Debreu-Farrell

measure of technical efficiency." A recent review of Farrell's contribution states

> (Debreu) worked only from the resource cost side, defining his coefficient as the ratio between minimised resource costs of obtaining a given consumption bundle and actual costs, for given prices and a proportional contraction of resources. Førsund and Sarafoglou (2002, footnote 4)

However, Debreu (1951) calculates the resource costs *not* of a given consumption bundle, but of an (intelligently chosen) Pareto equivalent allocation. (And the prices are not given, but support the allocation.) It is true, however, that the Debreu measure would become applicable if the aggregated consumption bundle can be considered given. Ten Raa (2008) demonstrates that this approach is doable and that it is exact if the preferences are Leontief.

6 The Debreu–Diewert Coefficient of Resource Utilization

Diewert (1983) had the idea that Leontief preferences remove misallocations between consumers as a source of inefficiency. The consequent coefficient of resource utilization yields a more conservative estimate of inefficiency than Debreu's coefficient resource of utilization. Ten Raa (2008) shows that Leontief preferences not only separate production efficiency from consumption efficiency, but also solve an aggregation problem: The Leontief preferences may vary between consumers, with different preferred consumption bundle proportions, but information of this preference variation need not be given. This useful fact is explained now.

Leontief preferences $\succsim_i$ with nonnegative bliss point y^i are defined for nonnegative consumption vectors by $y'' \succsim_i y'$ if $\min y''_k/y_k \geq \min y'_k/y_k$ where the minimum is taken over commodities $k = 1, \ldots, l$. If so, the consumption term in better set (6) fulfills (ten Raa, 2008)

$$\left\{\sum y^{i\prime} : y^{i\prime} \succsim_i y^i,\ i = 1, \ldots, m\right\} = \left\{y' : y' \geq \sum y^i\right\}. \tag{10}$$

Equation (10) shows that "more is better" at the micro-level if and only if "more is better" at the macro-level. Equation (10) is a perfect aggregation result. One might say that if preferences are Leontief with varying bliss points (according to the observed consumption baskets), there is a social

welfare function. The better set is freed from not only preferences, $\succeq_i$, but also individual consumption baskets, y^i. Only *aggregate* consumption is required information.

This result creates the option to determine the degree of efficiency in terms of output. If the production set X features the impossibility to produce something from nothing and constant returns to scale, then $\gamma = 1/\rho$ transforms the input-distance function program (9) into the *output-distance function* program

$$1/\rho = \max\{c : c \sum y^i + z' \leq x, z' \in Z\}. \tag{11}$$

Output-distance program (11) determines the expansion factor and potential consumption, i.e., net output. The ratio of actual output to potential output is equal to efficiency, the Debreu–Diewert coefficient of resource utilization, ρ. This has been applied and analyzed, including decompositions in different inefficiency components, for various economies.

Ten Raa and Mohnen (2001) evaluate the gains from free trade between the European and Canadian economies. The results show that bilateral trade liberalization would multiply the trade volume and let Canada, which is a small economy, to specialize in a few sectors. Perfect competition and free trade together will result in the expansion factors of 1.075 for Europe and 1.4 for Canada, while without free trade the economies expand to 1.073 and 1.18, respectively. The gains of free trade are evaluated at 0.2% for Europe and 22% for Canada. Sikdar et al. (2005) apply a similar model for measuring the effects of freeing bilateral trade between India and Bangladesh. The study was conducted against the background that Bangladesh was about to join the South Asian Association for Regional Cooperation (SAARC, established in 1985), in which India participated from the very beginning. Using the linear program version of the model, the authors locate comparative advantages in both economies and contrast them with the observed trade pattern. While the patterns are generally comparable, there are notable differences for some products. For example, it turns out that although India is an exporter of "Livestock, fishing, forestry" and "Other food products," the free trade model suggests that these should be import products for India. While on its own, each economy's expansion factor equals 1.37, the introduction of free trade would increase it to 1.43 for India and 1.97 for Bangladesh. This means that the potential gains of free trade for these two countries are 6% and 60%. Similarly to the previous paper, a small economy—Bangladesh—has much more to gain by joining the free trade agreement with a large economy. Ten Raa (2005) evaluates the contribution of

international trade, disentangling trade efficiency from domestic efficiency and splits the domestic efficiency of the economy into X-efficiency and allocative efficiency.

Another interesting decomposition of efficiency is provided by Cella and Pica (2001), who use a convex piecewise linear envelopment of the observed data (DEA) to disentangle sectoral inefficiencies in five OECD countries, Canada, France, Denmark, Germany, and the UK, into internal sectoral inefficiencies and inefficiencies imported from other sectors through the price distortion of intermediate product prices. These imported inefficiencies are also called "spillovers" from other sectors. The study shows that inefficiency spillovers are empirically relevant in all sectors of the five considered countries.

Amores and ten Raa (2014) distinguish three levels of production efficiency of the Andalusian economy: a firm level, an industry level, and the economy level. *Firm level* efficiency measures the potential productivity gains (i.e., output/input ratios) that arise if the firm could choose to use production techniques of other firms from the same industry. (However, intellectual property rights may prevent this.) *Industry efficiency* measures the gains that can be achieved by pooling all the vectors of inputs and outputs of the firms that belong to this industry and reallocating production within the industry to maximize the total output value of the industry. Finally, the total *efficiency of the economy* measures the gains that can be achieved by the economy if there were no barriers to reallocation of inputs and outputs across firms and industries. Based on the results from these three problems, one can distinguish *industrial organization efficiency* and *industrial specialization efficiency*. The former captures the efficiency gains achieved by reorganization within industries, if each industry starts to produce a more valuable (i.e., efficient) output mix. The latter captures the additional efficiency that can be achieved by re-specialization of the output mix of the economy.

7 Interrelation Between the Productivity and Efficiency Measures

Productivity growth, measured by the Solow residual (3), and efficiency, measured by the Debreu–Diewert coefficient of resource utilization (11), can be interrelated.

Productivity is output per input. For an economy, input are the resources and output is the final consumption. Input x and output y are multi-dimensional. Denote the production possibility set at time t, the set of all pairs (x, y)

such that x can produce y at time t by P^t, the so-called *production possibility set*. Following Eq. (9) the input-distance function is

$$D(x, y, t) = \min\{r : (rx, y) \in P^t\}. \tag{12}$$

Input distance r is a number between zero and one. If $r = 1$, input cannot be contracted, is on the frontier of the production possibility set, and is efficient. If $r < 1$, input can be contracted, is not on the frontier, and is inefficient. An increase in the input distance signals an increase in efficiency. *Efficiency change* is the relative change in input-distance function (12) with a dot representing time derivative:

$$EC = \dot{D}/D. \tag{13}$$

The distance to the frontier may grow without any change in input x or output y, simply because the frontier shifts out. This shows a decrease in the input distance. *Technical change* is minus the relative partial derivative of the input-distance function with respect to time, i.e., keeping input x and output y fixed:

$$TC = -D'_t/D. \tag{14}$$

To relate these efficiency change and technical change to the single output Solow residual analysis, we must replace Solow's implicit assumption that output is related to input by the production function, (4), by the more general relationship

$$y = D(x, y, t)F(x, t), \tag{15}$$

where potential output is reduced to actual output. Differentiating Eq. (15) with respect to time, $\dot{y} = \dot{D}F + D(F'_x \dot{x} + F'_t)$ or, dividing by expression (15),

$$(\dot{y}/y - F'_x \dot{x}/F) = \dot{D}/D + F'_t/F, \tag{16}$$

The left-hand side of formula (16) features total factor productivity, see Equation with y 1-dimensional and p canceling out, and the right-hand side features efficiency change (13) plus technical change. The last term is indeed consistent with Eq. (14), as output $y = D(x, y, t)F(x, t)$ and partial differentiation with respect to time yield $D'_t F + DF'_t = 0$. Summarizing,

$$TFP = EC + TC = \dot{D}/D - D'_t/D, \tag{17}$$

where the second equality holds term by term. Expression (17) holds for multi-output production with, however, constant returns to scale. Ten Raa (2008) proves that the efficiency change term is measured by the growth rate of the Debreu–Diewert coefficient of resource utilization and the technical change term by a generalized Solow residual of net frontier output growth evaluated at the supporting price vector.

In applied work, time is in discrete periods and the main performance measure that accommodates this is the Malmquist productivity index (Caves et al. 1982). Its derivation is as follows. The first term on the right-hand side of Eq. (17) is the total derivative of input distance $D(x, y, t)$ and the last term subtracts the third partial derivative. What remains are the first two partial derivatives,

$$TFP = \frac{\partial \ln D(x,y,t)}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{\partial \ln D(x,y,t)}{\partial y}\frac{\mathrm{d}y}{\mathrm{d}t}. \tag{18}$$

In discrete time expression (18) is a local approximation to

$$\ln D(x^{t+1}, y^{t+1}, \bullet) - \ln D(x^t, y^t, \bullet) = \ln \frac{D(x^{t+1}, y^{t+1}, \bullet)}{D(x^t, y^t, \bullet)}. \tag{19}$$

Evaluating this expression at t and $t+1$, taking the average of the two logarithms and exponentiating, one obtains the standard expression of the Malmquist productivity index:

$$TFP = \left[\frac{D(x^{t+1}, y^{t+1}, t)}{D(x^t, y^t, t)}\frac{D(x^{t+1}, y^{t+1}, t+1)}{D(x^t, y^t, t+1)}\right]^{1/2}. \tag{20}$$

The explicit price information in the Solow residual (3) has been replaced by implicit shadow price information, derived from the shape of the frontier; see Coelli and Rao (2001). The Malmquist productivity index assumes constant returns to scale. The decomposition of the Malmquist index into technical change and efficiency change, see Eq. (17), is straightforward; see Färe et al. (1994).

The Malmquist productivity index is popular because of its simplicity. Moreover, it can be bridged with other important TFP growth indices. The Törnqvist productivity index is defined by the discrete-time approximation of (3) with value weights $w_k x_k / wx$ and $p_k y_k / py$ approximated by their arithmetic averages between periods t and $t+1$ and growth rates $\dot{x}_k / x_k$ and $\dot{y}_k / y_k$ approximated by the changes in the logs of x and y between periods t and

$t+1$. Caves et al. (1982) have shown that the Malmquist productivity index becomes a Törnqvist productivity index provided that the distance functions are of translog form with identical second-order coefficients and that the prices support cost minimization and profit maximization. The Fisher productivity index is also defined by a discrete-time approximation of (3), with the changes in the logs of x and y now evaluated at the prices in periods t and $t+1$ separately and then averaged arithmetically. Färe and Grosskopf (1996) have proved that the Malmquist productivity index approximates the Fisher productivity index under the assumption of profit maximizing behavior. Balk (2008) reviews comprehensively, including non-constant returns to scale.

A defect of the Malmquist, Törnqvist, and Fisher indices is that they are not transitive. The changes from periods t to $t+1$ and from periods $t+1$ to $t+2$ do not add to the change from periods t to $t+2$. A necessary and sufficient condition for transitivity is that the index between periods can be written as a ratio of values of a function evaluated in the two periods. This property is fulfilled for the efficiency change component of productivity growth, but not for the technical change component, unless technical change is Hicks neutral. However, Balk and Althin (1996) shows that a modification of the Malmquist index, averaging out between firm observations, is transitive.

8 Conclusion

The key concept in performance analysis is productivity, which is the output/input ratio. Both output and input are aggregates. The appropriate weights are shadow prices of the program that determines potential output. The latter is based on observed input-output pairs or a production function, corresponding with nonparametric and parametric performance analysis, respectively. Parametric performance analysis can be conceived as nonparametric performance analysis with an appropriate distribution of observations. Hence nonparametric analysis is more fundamental. Replacing output by potential output, productivity becomes optimal productivity. The ratio of actual productivity to optimal productivity is equal to efficiency. Performance may increase because of efficiency change, technical change, scale economies, or changes in the production environment. Technical change is a change in optimal productivity. All this can be grounded in economic theory, where optimality is defined in terms of consumer preferences. If consumers have Leontief preferences, with consumptions bundles

preferred to be in fixed proportions, which may vary between consumers, then performance analysis is freed from micro-consumer data requirements and shadow prices can be determined on the basis of production data and the proportions of final demand. Moreover, then the efficiency is measured by Debreu's coefficient of resource utilization and technical change by the Solow residual of net frontier output growth.

Acknowledgements I am grateful to a referee for detailed criticism that prompted numerous improvements.

References

Amores, A.F., and T. ten Raa. 2014. Firm efficiency, industry performance and the economy: Three-way decomposition with an application to Andalusia. *Journal of Productivity Analysis* 42 (1): 25–34.

Balk, B. 2008. *Price and quantity index numbers*. Cambridge: Cambridge University Press.

Balk, B.M., and R. Althin. 1996. A new, transitive productivity index. *Journal of Productivity Analysis* 7 (1): 19–27.

Blackorby, C., and R.R. Russell. 1999. Aggregation of efficiency indices. *Journal of Productivity Analysis* 12 (1): 5–20.

Caves, D.W., L.R. Christensen, and W.E. Diewert. 1982. The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica* 50 (6): 1393–1414.

Cella, G., and G. Pica. 2001. Inefficiency spillovers in five OECD countries: An interindustry analysis. *Economic Systems Research* 13 (4): 405–416.

Chakravarty, Satya R. 1992. Efficiency and concentration. *Journal of Productivity Analysis* 3: 249–255.

Charnes, A., W.W. Cooper, and E. Rhodes. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 419–444.

Coelli, T., and D.S.P. Rao. 2001. Implicit value shares in Malmquist TFP index numbers. CEPA Working Paper No. 4/2001, University of New England, Armidale, Australia.

Debreu, G. 1951. The coefficient of resource utilization. *Econometrica* 19 (3): 273–292.

Diewert, W. Erwin. 1983. The measurement of waste within the production sector of an open economy. *Scandinavian Journal of Economics* 85 (2): 159–179.

Färe, R., and S. Grosskopf. 1996. *Intertemporal production frontiers: With dynamic DEA*. Boston: Kluwer Academic Publishers.

Färe, R., S. Grosskopf, B. Lindgren, and P. Roos. 1994. Productivity developments in Swedish hospitals: A Malmquist output index approach. In *Data Envelopment Analysis: Theory, Methodology and Applications*, ed. A. Chames, W.W. Cooper, A. Lewin, and L. Seiford. Boston: Kluwer Academic Publishers.

Färe, R., S. Grosskopf, and V. Zelenyuk. 2002. Finding common ground: Efficiency indices. UPEG Working Paper 0305, Presented at the North American Productivity Workshop at Union College, Schenectady, NY.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of Royal Statistical Society* 120 (3): 253–290.

Førsund, F.R., and N. Sarafoglou. 2002. On the origins of data envelopment analysis. *Journal of Productivity Analysis* 17: 23–40.

Grifell-Tatjé, E., C.A.K. Lovell, and J.T. Pastor. 1998. A quasi-Malmquist productivity index. *Journal of Productivity Analysis* 10 (1): 7–20.

Houthakker, H.S. 1955. The Pareto distribution and the Cobb-Douglas production function in activity analysis. *Review of Economic Studies* 23 (1): 27–31.

Johansen, L. 1972. *Production functions: An integration of micro and macro, short run and long run aspects*. Amsterdam, The Netherlands: North-Holland.

Levhari, D. 1968. A note on Houthakker's aggregate production function in a multifirm industry. *Econometrica* 36 (1): 151–154.

Milgrom, P.R. 1981. Rational expectations, information acquisition, and competitive bidding. *Econometrica* 49 (4): 921–943.

Muysken, J. 1983. Transformed beta-capacity distributions of production units. *Economics Letters* 11 (3): 217–221.

Samuelson, P.A. 1951. Abstract of a theorem concerning substitutability in open Leontief models. In *Activity Analysis of Production and Allocation*, ed. T.C. Koopmans, 142–146. New York: Wiley.

Sato, K. 1975. *Production functions and aggregation*. Amsterdam, The Netherlands: North-Holland.

Sikdar, C., D. Chakraborty, and T. ten Raa. 2005. A new way to locate comparative advantages of India and Bangladesh on the basis of fundamentals only. In *Essays on international trade, theory and policy for the developing countries*, ed. R. Acharyya, 169–197. Kolkata: Allied Publishers.

Solow, R.M. 1957. Technical change and the aggregate production function. *The Review of Economics and Statistics* 39 (3): 312–320.

ten Raa, T. 1995. The substitution theorem. *Journal of Economic Theory* 66 (2): 632–636.

ten Raa, T. 2005. *The economics of input-output analysis*. Cambridge: Cambridge University Press.

ten Raa, T. 2008. Debreu's coefficient of resource utilization, the Solow residual, and TFP: The connection by Leontief preferences. *Journal of Productivity Analysis* 30 (3): 191–199.

ten Raa, T., and P. Mohnen. 2001. The location of comparative advantages on the basis of fundamentals only. *Economic Systems Research* 13 (1): 93–108.

ten Raa, T., and P. Mohnen. 2002. Neoclassical growth accounting and frontier analysis: A synthesis. *Journal of Productivity Analysis* 18 (2): 111–128.

Zieschang, K.D. 1984. An extended Farrell technical efficiency measure. *Journal of Economic Theory* 33 (2): 387–396.