Methods Of Measuring The Economy, Efficiency And Effectiveness Of Public Expenditure

ANNEX 7:

August 2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

The PER normally seeks to facilitate and improve the implementation of the medium term effort to strengthen budget management, in terms of such predictability, efficiency and sustainability. It therefore evaluates budget performance against the approved allocation framework, costs, output, and outcome goals. The purpose is to determine whether funds are being spent according to plans and whether the spending units achieved the intended objectives.

This Annex documents a set of methods used to measure the economy, efficiency, and effectiveness of expenditure in the context of a PER. In considering efficiency, both technical and scale efficiency will be considered. Technical efficiency refers to the ratio of actual to potential output of the spending unit. Scale efficiency refers to the extent to which the spending unit takes advantage of opportunities to grow its output faster than it can grow its inputs, assuming the latter is growing at some fixed rate. The note motivates the measurement methods with the policy context of the PER, the baseline data needed, the necessity of the measures, the importance of proper coding of allocations, and background assumptions. Then, the various measurement methods are presented. Among the methods considered are DEA and stochastic frontier analysis. Additional assumptions are indicated in context.

2 PER Context

The PER normally includes background information to support interpretation of the measures. Table 1 provides an indication of the general social development information and perspectives of the evaluation.

Table 1: Baseline social perspectives of the PER

Indicators	*Baseline social perspectives of the PER*
Poverty and unemployment indicators, defining the social development challenge.	Such as: • Access to infrastructure / services (such as water / roads / energy / sanitation) adequacy measures • Health indicators and poverty measures • Education indicators and poverty measures • Aggregate poverty and unemployment assessments.
The statement of development imperatives for the next 3-5 years.	Such as: • Reducing poverty • Increasing equity
The specific numerical targets that define the development objectives.	Such as • 60% improvement in education measures over 3 years • 50% improvement in health indicators over 5 years • 80% reduction in disguised unemployment over 5 years
Annual targets and costs.	Such as • 20% improvement in education measures; Cost $50 million. • 10% improvement in health indicators; Cost $80 million. • 16% reduction in disguised unemployment; Cost $100 million.

Table 2 provides an indication of the baseline economic data and perspectives of the PER.

Table 2: Baseline economic data for the PER

Indicators	*Economic data for perspectives of the PER*	*Economic perspectives for the PER*

Economic activity, by industrial and economic classifications.	• Industrial sectors
	• Demand categories:
	o Consumption;
	o Investment;
	o exports;
	o imports;
	o government budget
		Such as
		• 2% annual growth of sector outputs
		• 3% growth of exports
		• 2% growth of effective consumption per dollar of imports

Economic infrastructure,	• Water, minimum thresholds to be met and demand profiles • Sanitation, minimum thresholds to be met and demand profiles • Roads, minimum thresholds to be met and demand profiles • Energy/electricity, minimum thresholds to be met and demand profiles	Such as • 60% increase of customers with adequate water supply; 3 years, $100 million • 50% improvement in sanitation indicators over 5 years; $80 million • 80% increase in customers with adequate electricity supply over 5 years; $66 million
Employment	Distribution by • Industry • Employment status • Gender. • Location	Such as: • 95% employment rate in 4 years; $150 million • Gender equity in pay and working conditions over 3 years; $100 million • 6% growth in rural employment over 5 years
Budget profile	In terms of • Revenue • Expenditure - recurrent expenditure and capital expenditure; wages, emoluments; materials and minor equipment; goods and services; transfers and gifts • Budget balance	Such as • Balanced budget • Increase in the share of minor equipment and materials to 5% or more of budget • Reduction of the share of transfers and gifts to 15% of budget while achieving poverty targets. 20% increase in number of persons graduating from springboard programs to full employment; Cost $100 million

3 Necessity of the measures

The measures of economy, efficiency and effectiveness are necessary to assess the productivity of resource use in government. Government spends its resources to deliver infrastructure and services to the population, and to encourage community development. As one of the largest spenders of national resources, it is on a continuous search for ways to prioritise budget allocations in a way that improves the relationship between expenditures inputs, actions, outputs, and their outcomes. In market operations, the prices of output can be used to value them, define the profit, and specify allocations guided by elegant calculations. However, no suitable market prices are available to value the outputs produced by government. One alternative is to use the methods identified in this Annex. The methods take account of the importance of public involvement in the planning, implementation and review of expenditures to deliver infrastructure and services.

4 Measurement and coding

Activity and input analysis of each allocation is the foundation of the PER. In a PER, all government's expenditures on infrastructure and services must be coded, listed and ranked on an overall index. This makes it possible to choose transparently which activities qualify for reprioritisation over the period.

5 Assumptions

The results of the above evaluations can be used to upgrade the efficacy of budget management. The results can be used to enhance the medium term expenditure strategy. They can lead to re-estimation of the activities in the budget, by updating the estimates of the cost of providing infrastructure that work and services that are in demand. The use of quantitative measures is intended to improve costing.

Allocations are translated into inputs, which include both resources and demand:

1. Labour - staff
2. Capital –space, equipment (such as desks and beds), and intermediate goods and services.
3. The populations to be served (demand)
4. The natural environment
5. Foreign exchange

Outputs include

1. Services
2. Infrastructure built
3. Surplus or profit

Thus, input and output prices are needed to compute efficiency. A basic assumption of the measurements presented is that projects are properly costed. In that context, the financial allocation is not an adequate guide. Economic costing is more appropriate, particularly with regard to the activities of the action units. Economic costs depend on:

1. The technology of the activity – technical efficiency
2. The scale of the activity – scale efficiency
3. Environmental effects related to use of natural resources
4. Psychological costs associated with failure to develop

Economic costs consider all resource costs, including the time used by full-time monthly paid staff and unpaid voluntary work. The amount of time worked must be recorded and valued, even if by imputation. Cost minimization does not mean shifting the expenditure of resources from paid work to voluntary work. Important in the total costs are the costs of using natural resources. This is an increasingly important aspect of environmental cost and should be estimated even when no market transaction are involved. Natural and environmental costs are important for infrastructure projects which often have significant environmental impacts.

Finally, some of the important costs of a project are linked to the failure of the project to deliver development outcomes, and to associated shortages of opportunity. Many of these issues can be incorporated into quantitative measures of efficiency and effectiveness.

6 Definitions and basic qualitative measures

Economy: The **economy** of use of inputs is a measure of how accurately the planned budget relates to actual spending and is used to procure transparently the best human resources and the best tangible and intangible assets.

This is fundamentally about whether proper procurement and accounting procedures are in place for disbursement, transfer, and virement of funds, with room for justifiable adjustments when circumstances change. The budget is spent with economy if 100% of the planned spending is actually achieved, while following official procurement rules. Inaccuracy normally draws attention to matters of planning as well as transfer, disbursement and virement procedures, but might also draw attention to capacity challenges, fraud or error. Variance should not exceed 10%.

Efficiency: This is an operational concept that mirrors the accountant's idea of value for money, whereby the best achievable relationship is maintained between actual infrastructure and services delivered and the potential that could be delivered.

Broadly, efficiency of the budget outputs is judged qualitatively by the extent to which specifications are achieved and delivered on time. If it takes 15 years to deliver what was planned for a 3 year delivery schedule, then the efficiency is 20%. If it takes twice as long, then it is 50%, and so on. If the work plan has 20 items and 16 are fully completed, then the efficiency rate is 80%.

A quantitative measure of technical efficiency of the budget output is the ratio of the output to the maximum possible output. If in a given time 100 units can be delivered per dollar of expenditure and only 80 units are delivered, then the efficiency is 80%. If the work plan delivers only 20 work items and 25 are possible, then the efficiency rate is 80%. Scale efficiency measures can also be computed. These indicate whether a resource growth program can be devised that leads to output growth faster than the rate of growth of the resources. If the real value of each input grows by 1% and causes the real value of all outputs to grow by 1.2%, then that is an indication that scale efficiency exists. It would be necessary to ensure that the growth of output is properly assigned to the inputs and not to external factors.

Effectiveness: This is a strategic or impact concept which requires the best possible relationship between an expenditure and the benefits it generates for the public over 3-5 years. The **effectiveness of impact** is measured by the extent to which the original problem has been solved, which in turn relates to whether funding goes to government responsibilities with the highest priority – matters of market failure. This can only be considered adequately if the total context of government spending is taken into account. Suppose money is spent on necessary infrastructure for a tourism project intended to grow the occupancy rate by 50%. Then, if the occupancy rate grew by 50%, the budget has a 100% effectiveness, and if it grew by only 10%, then the budget has a 20% impact; and so on. Similarly, if a sports facility is built and then used only 50% of what was planned – 50% occupied – then the effectiveness is 50%. Or, if a particular social group is targeted for subsidized employment under the condition that the beneficiaries must go to school while receiving the benefits, then if 20% of the number of members in the target group received the designed benefits, the measure of effectiveness is also 20%. If money is spent to improve education performance, according to national standards, then comparing actual performance against the checklist of standards will indicate if the expenditure is 100% effective or less.

In numerical evaluations of effectiveness, the important questions are still whether the actual expenditure and the infrastructure and services purchased:

1. Promoted equality among all groups in society?
2. Actually reached the target beneficiaries, especially the poor?
3. Delivered adequate citizen and community satisfaction?
4. Achieved the economic development goals?

Quantitative indicators for the first three of these questions can be obtained from Citizen Report Cards and Community Score Cards as well as by public expenditure tracking surveys. Answers can also be delivered by disaggregating the household data obtained from living standard measurement surveys. However, the answer to the fourth question must be evaluated with economic data from microenterprises and the industrial sectors. This outcome is related to the increase in effective consumption capacity. The basic measure is whether the expenditures on infrastructure and services increased the ratio of effective consumption per dollar of imports of the enterprises targeted by the expenditures. The goal might be to grow the ratio by 6% over the next 3 years. As explained in Annex 2, the main drivers of this indicator is the capital-labour ratio. The question can be reduced to whether: (i) the infrastructure and services promoted investment in skills and technologies, and in physical capital assets; and (ii) the acquisition of these assets caused the targeted firms to grow their resource productivity and build up their claims on foreign exchange while increasing their exports at a rate sufficient to achieve the 6% target.

These considerations lead to integrated measurements such as are reported in Table 3, relative to specific budget objectives. The measures should be followed by a concluding statement about the implications for the next year's budget. The measures assume a deliberate effort by the policy makers to target infrastructure and services to promote technical and scale efficiency as well as the growth of effective consumption capacity.

Consider allocations of the budget aimed at "increasing the supply of infrastructure from 38% to 70% of need in 3 years in specific districts where key exporting firms operate". The specific measures can be set out as follows:

Table 3: Integrated measure of Economy, Efficiency and Effectiveness

		Allocation Code or Project				
		0013	**0202**	**0003**	**0004**	**0505**
	Project Description	***Infrastructure construction, District 1***	***Classroom construction, District 2***	***Infrastructure construction, District 3***	***Classroom construction, District 4***	***Classroom construction, District 5***
1	**Budget**	10	20	30	40	50
2	Actual expenditure	8.5	17.8	28	32	48
3	Variance	1.5	2.2	2	8	2
4	% variance	15%	11%	7%	20%	4%
5	**Economy of inputs purchased:**					
6	Measures of the percentage of budget used	85%	89%	93%	80%	96%
7	**Efficiency of inputs purchased**					
8	Extent to which specification followed - number of planned work items completed vs number of items planned; or percentage of standards of delivery achieved	80%	70%	85%	90%	95%
9	Extent to which output delivered on time; measured as the % of planned time of delivery	50%	60%	70%	80%	90%
10	Actual output as percentage of potential output	80%	78%	95%	97%	75%
11	Efficiency Score (8+9+10)/2	70%	69%	83%	89%	87%
12	**Effectiveness of inputs purchased**					
13	Does the infrastructure solve the problem being addressed?	95%	96%	86%	97%	100%
14	Is the best education being provided?	100%	100%	100%	90%	90%
15	What fraction of the underemployed have used the asset as springboard to move into fulltime paid employment?	70%	75%	80%	90%	80%
16	Use rate of infrastructure built?	85%	90%	87%	95%	100%
17	Impact on achievement of targeted growth of effective consumption capacity over 3 years?	75%	65%	16%	12%	90%
18	Effectiveness Score (13+14+15+16+17)/5	85%	85%	74%	77%	92%
19	**Overall Score (6+10+16)/3**	**80%**	**81%**	**83%**	**82%**	**92%**

Similar tables should be constructed for all other key budget objectives, again often best expressed in terms of some percentage of need or demand. For example, the other economic development imperatives might translate to the following objectives:

1. Increase the supply of human capital from 58% to 88% of need within 5 years.
2. Improve business climate from an index of 58% to an index of 80% within 3 years.
3. Improve technical efficiency of exporters from 75% to 95% over 5 years.
4. Improve research and development capacity for growth of scale efficiency and export competitiveness among exporters 10% of need to 25% of need by 2020.
5. Increase access to external financing from 28% of business needs to 50% of business needs over the next 3 years.

The actual infrastructure and service needs specified in the coded budget allocations can then be monitored using tables such as Table 3 to monitor performance on the indicators.

Table 3 refers to measures of efficiency without indicating the complications that arise from the presence of multiple inputs and multiple outputs. Formally, efficiency compares the actual bundle of outputs from a given bundle of inputs with the maximum output that can be produced. Knowledge of the technology set available to the action unit is critical for efficiency measurement. In the multiple-input, multiple-output case, individual inputs and outputs need to be suitable aggregated. In the absence of market prices, the method of DEA provides a way to proceed with the measurement of efficiency.

6.1 Measuring Efficiency with DEA

Both DEA and stochastic frontier analysis identify a benchmark for use in comparing the performance of all other agencies. The benchmark is determined by the technology available. The comparison of the actual output produced with the maximum possible yields a measure of *technical efficiency and the associated scale efficiency*.

Public agencies procure multiple inputs and produce multiple outputs. Thus, input and output prices are needed to compute efficiency. DEA provides a way to proceed by computing ‘shadow prices’ when there are no actual market-based output or input prices, as is the case with public sector institutions. In education, health, and social protection, government organizations like hospitals, schools, or social relief agencies can be evaluated on this basis. In the case of agriculture, profit is the objective in the private sector and efficiency can be based on the ratio of profit to the maximum possible. This is economic efficiency and technical efficiency is a necessary condition for it. So, the impact of the government agencies delivering infrastructure and services can still be judged in terms of their technical efficiency. Assignment of cause matters. In agriculture, for example, the maximum output from a given set of inputs can vary randomly with the weather. Hence, if it is possible to control for such effects, then differences in efficiency can be ascribed to effort and ability.

6.1.1 DEA

DEA is a non-parametric method of choosing a benchmark and then measuring efficiency with multiple inputs, multiple outputs, and no market prices. The classic references are Charnes, Cooper, and Rhodes (1978, 1981)[1] but the roots of the method were laid down by Farrell (1957)[2], Debreu (1951) and Shepard (1953). DEA uses linear programming to compute 'weights' or 'shadow prices' as the alternative to market prices. Linear programing computes the shadow prices that maximize an objective function subject to the constraints identified. The data used for the DEA are the actual observed inputs purchased and the actual outputs produced by the spending units. The practical applications below are motivated by the following analytical framework.

6.1.2 Assumption of DEA

DEA assumptions are very strong. The method assumes that there is no random noise, measurement error, or outlier cases in the data. The data used to represent inputs and outputs are correctly known, and it does not matter how many variables are needed. There are no unique outputs or inputs. Correspondingly, if an output or input is zero, it has no significant effect on the measurement of the efficiency of a unit. Finally, it is assumed that if resources are unused, they can be disposed of without cost. The method is very useful when the number of decision-making units is small in a statistical sense, say less than 30 cases.

6.1.2.1 Objectives of Government and objectives of the DEA

The objective function in the DEA is specified in terms of the overall output/input ratio considering all the resources and outputs of all the decision-making unit consistent with the policy objectives of government. Thus, it is assumed that the decision-making units that come closest to the maximum efficiency possible will do the best in achieving a budget objective such as: "increase the supply of human capital from 58% to 88% of need within 5 years." Since benchmarking is the core of the method, if an international benchmark or 'best practice' ratio is available, then that simplifies the work involved. If only local data are available for benchmarking, then the procedure is as set out below.

6.1.2.2 Constraints

The constraints of the linear program are also output/input ratios pertaining to the decision-making units.

[1] Charnes, A., Cooper, W.W., and Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European Journal of Operations Research* 2: 429-444; (1981). Evaluating program and managerial efficiency: an application of Data Envelope Analysis to program follow through. *Management Science* 27: 668-697.

[2] Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Series A, 120: 253-90.

6.1.2.3 General idea of the DEA Method

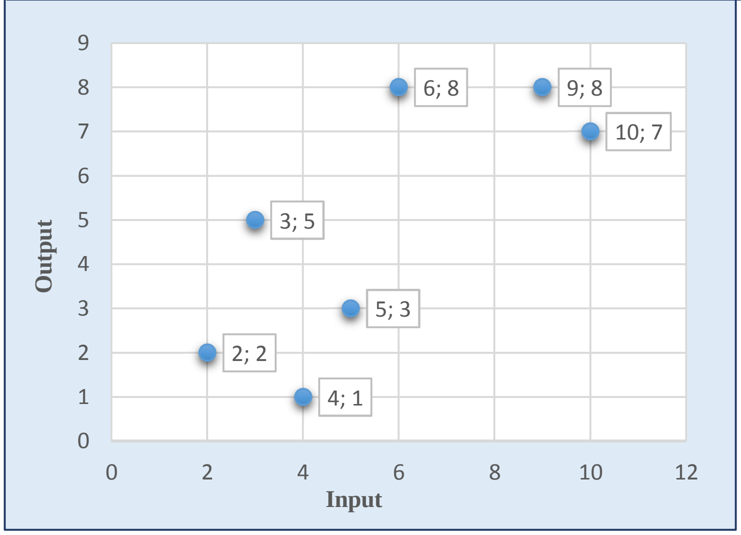
The ideas of the method are motivated and illustrated summarily with the data in Table 4. Assuming price weights are available, the gross inputs and outputs are reported along with the output/input ratios. The input-output combinations are graphed in Figure 1 but the ratios tell much of the story. The data indicate that the top performers are DMU1, DMU2, DMU3 and DMU4. The graph shows them 'enveloping' the others. They are the benchmark or efficient units. The other units are comparatively not efficient. They could either raise output or lower inputs (when compared to the benchmarks).

How they are interpreted is governed by the purpose of the PER, as set by the Ministry of Finance and the line ministries involved. If the purpose is to save, as in the case of the partner countries, then the appropriate interpretation is that the inefficient units could save on their allocations given to buy inputs. An input-oriented analysis is necessary in this case. For example, DMU5 could keep the value of output at 3 and lower the value of inputs to get to the output/input ratio of 1 of DMU1 or the 1.7 of DMU2. However, in a context where the budget is in balance, DMU5 could be guided to keep its value of inputs at 5 and raise output when compared to DMU2 and DMU3. This is an output-oriented analysis. Even if the budget is in deficit, the Ministry of Health might weigh in as the responsible ministry. It might insist that the quality and value of services are not yet to the level desired by government, so that DMU5 should keep its value of inputs at 5 and raise output when compared to DMU2 and DMU3. It might also benefit from the savings made from DMU6 being guided to save on its allocations. DMU6 is a special case. It has what is called a 'slack' in its rate of output compared to its 'peer' DMU1, so it cannot simply be instructed to save. Suppose it was instructed to save on its value of inputs bought with its allocation by reducing it to 2. Compared to DMU1, its immediate neighbour, it would still be missing 1 unit of output. So, perhaps it would also have to be guided by the Ministry of Health to make improvements on its output, perhaps by adopting the technology used by DMU1. The example illustrates why the PER is an inclusive process involving all stakeholders.

Table 4: Data for Illustration of DEA Rankings

	Input	*Output*	*Output/Input*
DMU1	2	2	**1.0**
DMU2	3	5	**1.7**
DMU3	6	8	**1.3**
DMU4	9	8	**0.9**
DMU5	5	3	**0.6**
DMU6	4	1	**0.3**
DMU7	10	7	**0.7**

Figure 1: Input Output Data Illustrated



6.1.2.4 Where will the prices come from?

Two different decision-making units may value inputs and outputs differently. DEA assigns the shadow prices for aggregation of inputs and outputs using a method that gives each decision-making unit *the best possible output/input ratio*. Starting with a specific unit:

1. The shadow prices assigned to the inputs and outputs of a decision-making unit give it the best possible output/input ratio. Therefore, it chooses them to reflect the importance the agency appears to have been placed on them. This is like solving the programming problem for the selected unit.

2. Simultaneously, though, DEA also assigns the same shadow prices to all the other decision-making units and compares the resulting output/input ratios with that for the one in focus.

3. If, at the assigned prices, the selected decision-making unit looks at least as good as any other, then it receives the maximum efficiency. If another decision-making unit looks even better, then since the prices were calculated to maximize the efficiency of the selected unit in first place, it will receive an efficiency score less than the maximum.

4. This method is repeated for all possible units.

5. The outcome is that each decision-making unit will be assigned an efficiency measure. The measure will be 1 if it is efficient and less than 1 if it is not. The entire set can then be ranked and all the true frontier cases will end up with ratings of 1. This is the substance of the label ‘technically or weakly efficient’.

6. If in addition, to being technically efficient, the decision-making unit also has no spare resources and no room to expand output any further– no slack –then it is called ‘strongly efficient’.

The importance of the multi-stakeholder perspectives can also be illustrated by considering in more detail how the actual weights are chosen when implementing steps 1 to 5 above. Suppose there are n firms producing s outputs $Y_i, i = 1 \dots s$ with m inputs, $X_i, i = 1 \dots m$. The shadow output prices are $\mu_r, r = 1 \dots s$. The shadow input prices are $v_i, i = 1 \dots m$. So, unit k uses the input bundle $X_k = (X_{k1}, X_{k2}, \dots X_{km})$ to produce the output bundle the output bundle $Y_k = (Y_{k1}, Y_{k2}, \dots Y_{ks})$. Algebraically, for each unit, the output-oriented linear 'fractional' programming problem is set up as follows:

1. $\max \text{AP}_k = \frac{\mu Y_k}{v X_k} = \frac{\sum_{r=1}^{s} \mu_{rk} Y_{rk}}{\sum_{i=1}^{m} v_{ik} X_{ik}}$

This value is maximized subject to two restrictions. The first is that the shadow prices must be non-negative, but free goods are permissible. The second is that the shadow prices are chosen such that, when all outputs and all inputs are aggregated using these prices, no firm's input-output bundle results in an overall average productivity greater than unity. Thus, no unit has an input/output bundle that causes the overall average productivity to be greater than 1. This implies that no unit has average productivity greater than 1. That is,

2. $\text{AP}_j = \frac{\sum_{r=1}^{s} \mu_{rk} Y_{kj}}{\sum_{i=1}^{m} v_{ik} X_{kj}} \leq 1, j = 1,2, \dots, k, \dots, n$
3. $\mu_{rk} \geq 0, r = 1 \dots s; v_{ik} \geq 0, i = 1 \dots m$

There will be many sets of shadow prices that satisfy these conditions but computer software can be used to find the one set that maximizes AP_k. The software available normally simplifies this problem to ensure a solution by multiplying each shadow price by an appropriate scaling factor, λ, that simplifies the search for the solution. This does not change either the objective function (Equation 1) or the constraints (Equations 2 and 3). It also implies the assumption of constant returns to scale, since the relationship between the input bundles and the output bundles also does not change. The clever choice of λ is:

4. $\lambda = \frac{1}{\sum_{i}^{m} v_{ik} X_{ik}}$

This is because it makes $\sum_{i}^{m} \lambda v_{ik} X_{ik} = 1$. Then, the above problem reduces to the following simple linear programming problem:

5. $\max \text{AP}_k = \sum_{r=1}^{s} \lambda \mu_{rk} Y_{rk}$

Subject to

6. $\text{AP}_j = \sum_{r}^{s} \lambda \mu_{rk} Y_{kj} - \sum_{i}^{m} \lambda v_{ik} X_{kj} \leq 0, j = 1,2, \dots, k, \dots, n$
7. $\sum_{i}^{m} \lambda v_{ik} X_{ik} = 1;\ \lambda \mu_{rk} \geq 0, r = 1 \dots s; \lambda v_{ik} \geq 0, i = 1 \dots m$

For example, if there are only two outputs and two inputs, the problem to be solved for the action unit is:

8.

$$\begin{array}{ccc}
\max \mathrm{AP}_k & = & p_{1k}Y_{1k}+p_{2k}Y_{2k} \\
\textit{subject to} & & \\
p_{1k}Y_{11}+p_{2k}Y_{21}-c_{1k}X_{11}-c_{2k}X_{21} & \leq & 0 \\
p_{1k}Y_{12}+p_{2k}Y_{22}-c_{1k}X_{12}-c_{2k}X_{22} & \leq & 0 \\
\cdots & \cdots & \cdots \\
p_{1k}Y_{1k}+p_{2k}Y_{2k}-c_{1k}X_{1k}-c_{2k}X_{2k} & \leq & 0 \\
\cdots & \cdots & \cdots \\
p_{1k}Y_{1n}+p_{2k}Y_{2n}-c_{1k}X_{1n}-c_{2k}X_{2n} & \leq & 0 \\
c_{1k}X_{1k}+c_{2k}X_{2k} & = & 1 \\
p_{1k}, p_{2k}, c_{1k}, c_{2k} & \geq & 0
\end{array}$$

The well-known simplex method can be used to solve the problem. The following are noted:

1. The shadow prices of inputs are chosen to cause the value of the observed input bundle of the unit under evaluation to equal unity ($\sum_i^m \lambda v_{ik}X_{ik} = 1$).

2. As a result, the value of the output bundle itself ($\sum_{r=1}^s \mu_{rk}Y_{rk}$) becomes a measure of its average productivity. This idea underlies the use of relative outputs to measure efficiency in Table 1 (Row 10).

3. At the prices $\lambda\mu_{rk}$ and λv_{ik} no unit will have an observed input-output bundle that yields a positive surplus of revenue over cost. This fits the normal profile of a public sector unit creating infrastructure or delivering services in education, health or social protection.

4. If the imputed input prices of the resources used, including natural environmental resources, cause the imputed value of any input bundle to be less than the imputed value of the output bundle it produces, then the resources are being under-valued. The imputed input prices would have to be revised upwards.

5. If the imputed output prices create a value of the output bundle above the total imputed cost of the input bundle used, then the output bundle is over-valued and would have to be revised downward.

6. The method is sensitive to the assumptions about the returns to scale. The options are constant and increasing returns to scale. Constant returns to scale means that when all the resources increase in value at a given rate, the output value also grows at the same rate.

When constant returns to scale is assumed, the shadow prices will be chosen such that all units will generate zero surplus. This applies to all units, including the one being evaluated by the method set out above. Thus, the maximum value of the aggregate output, $Y_k = \sum_{r=1}^s \lambda\mu_{rk}Y_{rk}$, will be 1. This further implies that:

9. $\frac{Y_k}{Y_k^*} = Y_k = \sum_{r=1}^s \lambda\mu_{rk}Y_{rk}$

That is, the optimal solution of the linear programming problem is a measure of the output-oriented technical efficiency of the unit.

If the primary concern of the government is to save, then it is better to analyse the dual of the above problem. The dual sets up the search for a global minimum shadow price, θ, such that it is greater than or equal to the total cost of the inputs used by the n action units. Also, each input of all action units must be valued such that the total cost of the outputs of the n action units is greater than or equal to the real value of the output of the unit being evaluated. For the two-input two-output case above, that dual is therefore:

$$\begin{array}{llcc}
 & \min \theta & & \\
 & \mathit{subject\ to} & & \\
 & \gamma_1 Y_{11} + \gamma_2 Y_{12} + \cdots + \gamma_k Y_{1k} + \cdots + \gamma_n Y_{1n} & \geq & Y_{1k} \\
10. & \gamma_1 Y_{21} + \gamma_2 Y_{22} + \cdots + \gamma_k Y_{2k} + \cdots + \gamma_n Y_{2n} & \geq & Y_{2k} \\
 & \theta X_{1k} - \gamma_1 X_{11} - \gamma_2 X_{12} - \cdots - \gamma_k X_{1k} - \cdots - \gamma_n X_{1n} & \geq & 0 \\
 & \theta X_{2k} - \gamma_1 X_{21} - \gamma_2 X_{22} - \cdots - \gamma_k X_{2k} - \cdots - \gamma_n X_{2n} & \geq & 0 \\
 & \theta\ free, \gamma_j, j = 1,2 \ldots k, \ldots n & \geq & 0
\end{array}$$

Remember here that simplex is a search process. The intuition behind the search for the solution in this input-oriented case is linked to the intuition behind the output-oriented technical efficiency of firm k producing output bundle Y_k from the input bundle X_k. There, as observed in the motivating example and as indicated by Equation (9), the option adopted was to keep X_k fixed and rescale output to the maximum ${Y_k}^*$ producible from it. The scale factor can be defined as ϕ such l lies within the *technology set of all action units* and ϕ^{max} exists such that ${Y_k}^* = \phi^{max} Y_k$. The result is Equation (9), which defines the output-oriented technical efficiency of firm *k*. *Correspondingly*, in specifying the input-oriented technical efficiency of any action unit, it is necessary to determine whether the value of the inputs can be reduced without reducing the output (bundle). The immediate intuition is to look at $(X_k, \phi Y_k)$ and consider rescaling the inputs with $\frac{1}{\phi}$ to get the technology used as $(\frac{X_k}{\phi}, Y_k)$. One intuitive way to work with an input bundle, when there are no market prices to define the comparative importance of the inputs, is to use *equi-proportional* reduction in all the inputs. The inputs are scaled down but their proportions do not change. Then, when ϕ becomes ϕ^{max}, the input-oriented technical efficiency of firm k is necessarily $\frac{1}{\phi^{max}}$, *which is the minimum value of* $\frac{1}{\phi}$ such that $(\frac{X_k}{\phi}, Y_k)$ still lies in the relevant technology set of the action units.

Bearing in mind standard duality theory, the obvious clever first approximation of the solution via simplex comes from the form of the maximization problem. That is, the search must be for shadow prices such that ϕ and θ are related by:

$$11.\ P_{1k} Y_{1k} + P_{2k} Y_{2k} = \theta = \frac{1}{\phi}$$

Then, as ϕ grows θ declines and the minimization of θ is equivalent to the maximization of ϕ. That is, the search for a solution ends when the shadow prices are chosen such that (P_{1k}, P_{2k}) converges to (p_{1k}, p_{2k}) in the maximization problem (8).

Finally, observe that under the assumption of constant returns to scale, the input-oriented solution is identical to the output-oriented solution. To see this, take the rescaled output technology that gets to the maximum value of output $(X_k, \phi^{max} Y_k)$. Now, continue to scale up both the inputs and the output by some scale-factor λ keeping the technology the same. That is, consider $(\lambda X_k, \lambda \phi^{max} Y_k)$. Then, if we choose $\lambda = \frac{1}{\phi^{max}}$, the result must be $(\frac{X_k}{\phi^{max}}, Y_k)$, which (by Equation 11) makes the point the input-oriented solution, $\frac{1}{\phi^{max}}$, is identical to the output-oriented solution under constant returns to scale. The choice of approach can therefore be decided by computational ease and conceptual simplicity. Equation (10) has the advantage in this regard when transformed into a maximization problem.

6.2 Scale Efficiency Issues in DEA

To motivate the methods, consider the problem in Table 5 and its solution. There are six action units, with two inputs and two outputs. Shadow prices are needed to aggregate the inputs and output and compute efficiency. Consider evaluating the performance of DMU3.

Table 5: Illustrative Data for DEA with Potential for Scale Efficiency

	Ouput1 (Y1)	*Output2 (Y2)*	*Input1 (x1)*	*Input2 (x2)*
DMU1	4	2	2	3
DMU2	9	4	7	5
DMU3	6	3	6	7
DMU4	8	6	5	8
DMU5	7	5	8	4
DMU6	11	8	6	6

Bearing in mind problem (10) and solution (11), define $\phi = \frac{1}{\theta} = \frac{1}{P_{1k}Y_{1k} + P_{2k}Y_{2k}}$. Also, transform the shadow prices to $u_j = \frac{y_j}{\theta}$. With these variables, the problem being considered is now:

$$\begin{array}{rrcl} & \max \phi & & \\ & subject\ to & & \\ & u_1Y_{11} + u_2Y_{12} + \cdots + u_kY_{1k} + \cdots + u_nY_{1n} & \geq & \phi Y_{1k} \\ 12. & u_1Y_{21} + u_2Y_{22} + \cdots + u_kY_{2k} + \cdots + u_nY_{2n} & \geq & \phi Y_{2k} \\ & u_1X_{11} + u_2X_{12} + \cdots + u_kX_{1k} + \cdots + u_nX_{1n} & \leq & X_{1k} \\ & u_1X_{21} + u_2X_{22} + \cdots + u_kX_{2k} + \cdots + u_nX_{2n} & \leq & X_{2k} \\ & \theta\ free, u_j, j = 1{,}2 \ldots k, \ldots n & \geq & 0 \end{array}$$

In terms of the data in Table 3, the problem is:

$$\begin{array}{lrcl} & \multicolumn{3}{c}{\max \phi} \\ & \multicolumn{3}{c}{\mathit{subject\ to}} \\ & 4\gamma_1 + 9\gamma_2 + 6\gamma_3 + 8\gamma_4 + 7\gamma_5 + 11\gamma_6 & \geq & 6\phi \\ 13. & 2\gamma_1 + 4\gamma_2 + 3\gamma_3 + 6\gamma_4 + 5\gamma_5 + 8\gamma_6 & \geq & 3\phi \\ & 2\gamma_1 + 7\gamma_2 + 6\gamma_3 + 5\gamma_4 + 8X\gamma_5 + 9\gamma_6 & \leq & 6 \\ & 3\gamma_1 + 5\gamma_2 + 7\gamma_3 + 8\gamma_4 + 4\gamma_5 + 6\gamma_6 & \leq & 7 \\ & \theta\ free, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6 & \geq & 0 \end{array}$$

Clearly, this can be transformed into a problem in which the outputs of DMU3 appear as coefficients of $-\phi$ on the left side of the output inequalities while the input quantities are constants on the right side of the input constraints:

$$\begin{array}{lrcl} & \multicolumn{3}{c}{\max \phi} \\ & \multicolumn{3}{c}{\mathit{subject\ to}} \\ & 4\gamma_1 + 9\gamma_2 + 6\gamma_3 + 8\gamma_4 + 7\gamma_5 + 11\gamma_6 - 6\phi & \geq & 0 \\ 14. & 2\gamma_1 + 4\gamma_2 + 3\gamma_3 + 6\gamma_4 + 5\gamma_5 + 8\gamma_6 - 3\phi & \geq & 0 \\ & 2\gamma_1 + 7\gamma_2 + 6\gamma_3 + 5\gamma_4 + 8X\gamma_5 + 9\gamma_6 & \leq & 6 \\ & 3\gamma_1 + 5\gamma_2 + 7\gamma_3 + 8\gamma_4 + 4\gamma_5 + 6\gamma_6 & \leq & 7 \\ & \theta\ free, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6 & \geq & 0 \end{array}$$

When the problem is solved with computer software, the solution is:

Table 6: Problem Solution

$\boldsymbol{\gamma_1}$	γ_2	γ_3	γ_4	γ_5	γ_6	ϕ
1	0	0	0	0	0.667	1.889

From the perspective of the challenges of the partner countries, the results can be interpreted as follows:

a. Referring to the technology $(X_k, \phi Y_k)$, all outputs of DMU3 can be increased by a factor of 1.889.

b. The associated technical efficiency measure is $\frac{1}{\phi^{max}} = \frac{1}{1.889} = 0.529$.

c. The valuations $\gamma_1 = 1$ and $\gamma_6 = 0.667$ indicate that the best option for DMU3 (call that DMU3$_{\text{optimal}}$) is to combine 0.667 of the input-output bundles of DMU6 with the input-output bundle of DMU1. Then, DMU3$_{\text{optimal}}$ would produce 11.34 units of y_1 (equal to .667*11+4) and 7.34 units of y_2 (equal to .667*8+2), using its 6 units of x_1 and 7 units of x_2.

d. If the optimal output of DMU3 is compared with this solution, it is clear that with no more inputs, its quantity of Y_1 could grow at most by a factor of 1.889.

> However, its quantity of Y_2 could grow by a factor as large as 2.445. This illustrates that the factor of 1.889 and the technical efficiency measure of 0.529 does not measure the full potential for increasing ALL outputs, but rather the minimum growth potential when the inputs are fully utilized. It is this maximum growth potential that is needed to grapple with the challenges of raising the effective consumption capacity of the economy.

Furthermore, the example in Table 4 indicates that it should also be possible to look for opportunities to save individual inputs while increasing the outputs. Overall, the examples motivate search for other possible opportunities to expand output or release resources and this refocuses attention on the assumption of constant returns to scale. The appropriateness of the assumption about returns to scale depends on the objectives of government and the facts on the ground about the possibilities for gaining benefits from scaling up the supply of infrastructure and services even if some of the proceeds must be exported. If the primary challenge of the economy is to solve the problem of development by growing exports, then the appropriate assumption is increasing returns. The push for increasing returns would be a matter of priority for the Planning Ministry. It would also provide the approach to searching for opportunities to expand output. This also requires that the output-oriented search problem be modified to allow the resources to exist on which production can rescale. These output growth capacities and available resources are usually referred to as ‘slack variables’.

Define the output slack variables at the solution as s_1 and s_2. Also define the input slack variables as ς_1 and ς_2. The slacks can vary to reflect returns to scale. In the system defined by Equation (11), the envelope (frontier) cases under variable returns to scale can be identified by adding the constraint that $\sum \gamma_j = 1$. However, the problem as redefined by Charnes, Cooper, and Rhodes (1979) and later Banker, Charnes, and Cooper (1984)[3] includes the very small number, ε, multiplied by the sum of the slacks in the objective function. This is to allow penalties (that make the output less valuable) in the objective function for strictly positive input and output slacks. Recall that in this case $\phi + \varepsilon(s_1 + s_2 + \varsigma_1 + \varsigma_2) = \frac{1}{\theta} = \frac{1}{P_{1k}Y_{1k}+P_{2k}Y_{2k}}$, so that the effect of $\varepsilon(s_1 + s_2 + \varsigma_1 + \varsigma_2)$ is to lower the shadow prices of output. The infinitesimally small ε is chosen (arbitrarily) by the analyst. The problems now has the form:

[3] Banker, R. D., A. Charnes, and W. W. Cooper. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30: 1078–1092.

$$
\begin{array}{rcc}
\max \widetilde{\phi} & = & \phi + \varepsilon(s_1 + s_2 + \varsigma_1 + \varsigma_2) \\
\textit{subject to} & & \\
u_1Y_{11} + u_2Y_{12} + \cdots + u_kY_{1k} + \cdots + u_nY_{1n} - \phi Y_{1k} - s_1 & \geq & 0 \\
15.\ u_1Y_{21} + u_2Y_{22} + \cdots + u_kY_{2k} + \cdots + u_nY_{2n} - \phi Y_{2k} - s_2 & \geq & 0 \\
u_1X_{11} + u_2X_{12} + \cdots + u_kX_{1k} + \cdots + u_nX_{1n} + \varsigma_1 & \leq & X_{1k} \\
u_1X_{21} + u_2X_{22} + \cdots + u_kX_{2k} + \cdots + u_nX_{2n} + \varsigma_2 & \leq & X_{2k} \\
\theta\ free, u_j, j = 1,2 \ldots k, \ldots n;\ s_1;\ s_2; \varsigma_1;\ \varsigma_2 & \geq & 0
\end{array}
$$

The objective function $\max \widetilde{\phi} = \phi + \varepsilon(s_1 + s_2 + \varsigma_1 + \varsigma_2)$ ensures that $\widetilde{\phi} > \phi^{max}$ when any slack variable is positive at the optimal solution.

With the penalty and the slacks taken into account, the original problem in equation (8) now has the form:

$$
\begin{array}{rcc}
\max \mathrm{AP}_k & = & p_{1k}Y_{1k} + p_{2k}Y_{2k} \\
\textit{subject to} & & \\
p_{1k}Y_{11} + p_{2k}Y_{21} - c_{1k}X_{11} - c_{2k}X_{21} & \leq & 0 \\
p_{1k}Y_{12} + p_{2k}Y_{22} - c_{1k}X_{12} - c_{2k}X_{22} & \leq & 0 \\
\ldots & \ldots & \ldots \\
16.\ p_{1k}Y_{1k} + p_{2k}Y_{2k} - c_{1k}X_{1k} - c_{2k}X_{2k} & \leq & 0 \\
\ldots & \ldots & \ldots \\
p_{1k}Y_{1n} + p_{2k}Y_{2n} - c_{1k}X_{1n} - c_{2k}X_{2n} & \leq & 0 \\
c_{1k}X_{1k} + c_{2k}X_{2k} & = & 1 \\
p_{1k}, p_{2k}, c_{1k}, c_{2k} & \geq & \varepsilon
\end{array}
$$

The difference between this problem and its earlier specification is that the lower bound of the shadow prices is now ε rather than 0. At the optimal solution, the output slack variables are defined by $s_1{}^* = Y_{1k}{}^* - \phi^{max}Y_{1k}$ and $s_2{}^* = Y_{2k}{}^* - \phi^{max}Y_{2k}$. Also the input slack variables at the solution are $\varsigma_1{}^* = X_{1k} - X_{1k}{}^*$ and $\varsigma_2{}^* = X_{2k} - X_{2k}{}^*$. The differences are a measure of the extent to which the DMU is scale-inefficient and can grow by increasing output. If the growth leads to an increase in output faster than inputs when the latter grow at the same fixed rate, then the decision making unit can exploit scale efficiencies by expansion. A decision-making unit will only get a rating of full efficiency at the optimal solution when $\phi^{max} = 1$ and $s_1 = s_2 = \varsigma_1 = \varsigma_2 = 0$. If any of these conditions fails at the optimal solution, efficiency will be less than 1 even if $\phi^{max} = 1$. Such a rating is not likely to be assigned under conditions of increasing returns to scale.

6.3 Using Software to Do the Estimates

The easiest software to use when doing DEA is Stata. Any recent version of the software will work, but the best available is Stata 14. Stata does DEA with a user-written command, **dea.ado**, which can be downloaded and installed into your version of the software (Ji and Lee, 2010)[4]. To do this, fire up your copy, type **net install st0193 and follow the instructions**. If you want help, type **help dea** to call up the help file. Once Stata is up and running, the execution code is:

[4] Ji, Y. and Lee, C. (2010). Data Envelopment Analysis. *Stata Journal* 10(2): 267-280.

```
>dea ivars = ovars [if] [in] [, options]
```

Options include

rts (crs/vrs/drs/nirs) specifies the returns to scale (default is **rts(crs)**)
ort (in/out) specifies the orientation **(**default **is ort(in)**)

where

crs - refers to constant returns to scale
vrs – refers to variable returns to scale, allowing for increasing returns
drs – refers to decreasing returns to scale
nirs – refers to non-increasing returns to scale

If these codes are typed into the command line, the program will run, but it is always good to write a small program <a do file, in Stata jargon> that includes the line of code. Here is an example of the data layout needed. The dataset is stored as **deadata.dta**:

Table 7: Data Layout Example

DMU	inp_x1	inp_x2	out_y1	out_y2	inp_c1	inp_c2	out_p1	out_p2
DMU1	2	3	4	2	1	2	3	3.4
DMU2	7	5	9	4	1	2	3	3.4
DMU3	6	7	6	3	1	2	3	3.4
DMU4	5	8	8	6	1	2	3	3.4
DMU5	8	4	7	5	1	2	3	3.4
DMU6	6	6	11	8	1	2	3	3.4

Here is a simple example of an output-oriented problem that actually runs repeatedly on the dataset **deadata.dta**:

```
>capture program drop deamod
>program deamod // data envelope analysis for ECLAC
>capture log close
>log using deamod, replace
>clear
>use C:\project\deadata.dta
>dea inp_x1 inp_x2 = out_y1 out_y2, rts(vrs) ort(o)
>end
>deamod
>log close
>exit
```

Note the following:

1. Inputs and outputs can be specified in any way necessary. Floor space can be treated as in input, just as the number of teachers or the number of nurses. Profits can be treated as an output, just as the amount of tomatoes sold in a given period.

2. As a general case, neither the amount and value of inputs nor the amount and value of output can be assumed fixed. Also, some inputs are imported and requires the use of foreign exchange. In partner countries, the amount of foreign exchange used or imports bought can be treated as a separate input. All such inputs can be distinguished in the methods represented above.
3. DEA does not explain the differences in the efficiencies observed among the DMUs. However, the results from the Stata/DEA program can directly feed to other Stata routines for further analysis, including regression analysis, to explain the differences in terms of the observed data on the DMUs. These other routines are commonly called ‘stage 2’ analyses, while the analysis presented above is referred to as ‘stage 1’. They are not developed in this manual.

6.3.1.1 Example 1 – output oriented DEA

Using the data stored a **deadata.dta,** we run a one-input, one-output example to build up the interpretation of the program output. Run the above program with the line of code:

```
>dea inp_x1 = out_y1, ort(o)
```

The result produced by Stata is formatted as follows:

Table 8: Stata results format

			ref:	ref:	ref:	ref:	ref:	ref:	islack:	oslack:
	Rank	theta	DMU1	DMU2	DMU3	DMU4	DMU5	DMU6	inp_x1	out_y1
dmu:DMU1	1	1	1	.	.	.	.	.	.	0
dmu:DMU2	4	0.64	2.25	.	.	.	.	.	.	0
dmu:DMU3	5	0.50	1.5	.	.	.	.	.	.	0
dmu:DMU4	3	0.80	2	.	.	.	.	.	.	0
dmu:DMU5	6	0.44	1.75	.	.	.	.	.	.	0
dmu:DMU6	2	0.92	2.75	.	.	.	.	.	.	0

1. Entries such as “.” mean that the value is virtually zero, too small to mention.
2. The first column to the left identifies the DMU being analyzed.
3. Column 2 reports the ranking of the DMUs in terms of their technical efficiency.
   a. DMU1 is ranked 1, DMU6 ranked 2, and so on.
4. Column 3 reports the technical efficiency computed (theta).
   a. DMU1 has an efficiency of 1, DMU6 has an efficiency of 0.92, and so on.
   b. Thus, DMU6 can increase its output by 8% without having to increase its use of its input, by adopting the approach of DMU1. Its total output could be increased to 3.24 units. Similarly, DMU2 can increase output by 36% with the same inputs if it adopts the methods of DMU1.
5. Column 4 reports the reference weights (lamdas) that are used to value the inputs, and with Column 11 (output slack) hold an important key to the interpretation. In this case

the reference weights all come from those of DMU1. Column 4 indicates that the output of DMU6 can beneficially increase by adding a maximum of 2.75 additional units to its output. However, because there is an output slack of 0, no more than an 8% increase is achievable.

6.3.1.2 *Example 2 – Input Oriented DEA*

As noted above, a PER concerned with savings should use an input-oriented DEA. Using the data in **deadata.dta,** run the above program with the line of code:

```
>dea inp_x1 inp_x2 = out_y1 out_y2, rts(crs) ort(i)
```

The results produced by Stata are:

Table 9: Stata results

1	2	3	4	5	6	7	8	9	10	11	12	13
			ref:	ref:	ref:	ref:	ref:	ref:	islack:	islack:	oslack:	oslack:
	Rank	theta	DMU1	DMU2	DMU3	DMU4	DMU5	DMU6	inp_x1	inp_x2	out_y1	out_y2
dmu:DMU1	1	1	1	.	.	.	.	0	.	0	.	0
dmu:DMU2	3	0.981818	.	.	.	.	.	0.818182	1.96364	.	1.59E-07	2.54545
dmu:DMU3	6	0.529412	0.529411	.	.	.	.	0.352942	.	9.54E-07	.	0.882354
dmu:DMU4	5	0.9	.	.	.	.	.	0.75	.	2.7	0.25	.
dmu:DMU5	4	0.954545	.	.	.	.	.	0.636364	3.81818	.	.	0.090909
dmu:DMU6	1	1	.	.	.	.	.	1	.	0	.	.

1. Column 2 indicates the rank of the DMU.
   a. DMU1 and DMU6 now have the same rank of 1.
   b. DMU2 now ranks 3, and so on.
2. Column 3 reports the technical efficiency measure (theta) on which the ranking is based.
   a. DMU1 and DMU6 have theta=1
   b. Both are strongly efficient because they have no slack inputs or output.
   c. Both are referents. A referents is a DMU that an inefficient DMU targets as a ‘fastest’ step to get to an optimum method.
   d. There are correspondingly two reference DMUs in the results.
   e. DMU 2 has a technical efficiency score of 0.981818, DMU3 has an efficiency of 0.529412, and so on.
3. Column 4 indicates that DMU1 is the reference for DMU3. DMU6 is the reference for the other inefficient DMUs.
   a. Thus, since DMU2 has an efficiency score of 0.981818, a 2% reduction in input would get it to the position implied by the weights of DMU1.

b. Using Column 6, the reference (output) weights (lambdas) for DMU3 are (0, 0, 0.529411, 0, 0, 0). Thus, a 47% reduction in inputs would improve the performance of DMU3 whatever other changes it makes. **These are the types of savings sought by the PERs**. The other possible changes are indicated by the slacks (Columns 10-13).

   i. Column 10 indicates that DMU3 has no slack on inp_x1 but by Column 11 has a positive but rather small slack of 0.00000095 on inp_x2.
   ii. By Column 12, DMU3 has no slack on out_y1 but has a slack of 0.882354 on out_y2.

c. Thus, the performance of DMU3 can be improved by subtracting a further 12% from out_y2, after having reduced all inputs by 47% without putting any other input or output in a worse position.

d. By comparison, using Column 7, the reference weights for DMU4 are (0, 0, 0, 0.75, 0, 0). DMU4 has an efficiency score of 0.9, **so a 10% saving all inputs would get it to the position implied by the weights of DMU6**. For the other changes,

   i. Column 10 indicates that DMU4 has no slack on inp_x1 but by Column 11 has a slack of 2.75 on inp_x2.
   ii. By Column 12, DMU4 has a slack of 0.25 on out_y1 but has no slack on out_y2.

e. Thus, the performance of DMU4 can be improved by reducing inp_x2 by 2.75 and subtracting a further 75% from out_y1, after having reduced all inputs by 47% without putting any other input or output in a worse position.

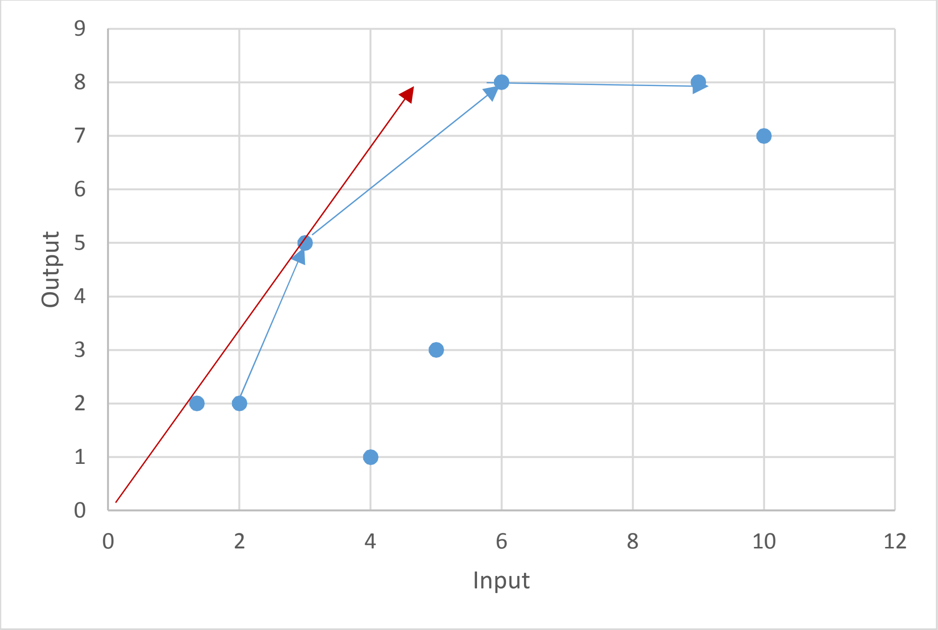
6.3.2 Scale Efficiency

To measure potential for scale efficiency, variable returns to scale must be understood. An understanding of variable returns to scale, and in particular increasing returns to scale, is motivated by observing conditions when the DMU is not exhausting all opportunities to attain its constant returns position. From an input-oriented standpoint, potential for scale efficiency is measured by identifying and computing the following:

1. The constant-returns to scale frontier. In the one-input one-output case of Figure 2 with reference to the data in Table 1, this is the line from the origin through the outermost point on the frontier.
2. The variable returns to scale frontier. In Figure 2, this is the frontier graph, similar to that in Figure 1.
3. The amount by which all inputs can be reduced to produce the same output as the DMU moves from the variable returns to scale frontier to the constant returns to scale frontier.

From the coordinates in Figure 2 below, this potential is about 0.7 or 35% of the input (2 units) that can be reduced without affecting the scale of output.

Figure 2: Illustration of Potential for Scale Efficiency



The principle is readily illustrated with the **coelli_table6.4.dta** data made available by Ji and Lee (2010)**.** To access the data run the program provided above with the lines of code appropriately substituted:

```
> use coelli_table6.4.dta
…
>dea i_x = o_q, rts(vrs) ort(i)
```

The code replicates the results of Ji and Lee (2010).

1	2	3	4	5	6	7	8	9	10
			ref:	ref:	ref:	ref:	ref:	islack:	oslack:
	rank	theta	A	B	C	D	E	i_x	o_q
dmu:A	1	1	1	.	.	.	.	.	.
dmu:B	5	0.625	.5	.	0.5	.	.	.	.
dmu:C	1	1	0	.	1	.	.	.	.
dmu:D	4	0.9	.	.	0.5	.	0.5	.	.
dmu:E	1	1	.	.	0	.	1	.	.

Technical efficiency is reported by the program from an input-perspective, as in the example immediately above:

1. Column 1 identifies the decision-making unit.
2. Column 2 reports the ranking of the unit in efficiency terms, using the efficiency score (theta) reported in column 3.
    a. There are three efficient decision-making units, DMU A, DMU C and DMU E, each with theta of 1.
3. Columns 4 to 8 indicate the referents of the other inefficient decision-making units.
    a. Column 8 indicates that DMU E is the referent for DMU D and Column 4 indicates that DMU A is a referent for DMU B. The coefficients in these columns indicate the amount by which output or input can be reduced if slacks exist, above what is indicated by the technical efficiency score.
    b. DMU D has an efficiency ranking of 3 based on a score of 0.9. As above, this means that DMU D can reduce its use of input i_x by about 10% and obtain the same output, if it adopts the management and technical methods of its referent, DMU E (Column 7).
    c. DMU B has an efficiency score of 0.625, and so without adjusting its output it can save 37.5% of the amount of input i_x currently used by adopting the management and technical methods of its referent, DMU A.
4. Columns 9 and 10 indicate that no DMU operates with a slack on input or a slock on output. So the only gains in technical efficiency come from savings on the inputs.

Since variable returns to scale is assumed in the program (**rts(vrs)**), the Stata program reports additional information on the potential scale efficiency of the units as follows:

Code: VRS Frontier(-1:drs, 0:crs, 1:irs)					
1	2	3	4	5	6
	CRS_TE	VRS_TE	NIRS_TE	SCALE	RTS
dmu:A	0.5	1	0.5	0.5	1
dmu:B	0.5	0.625	0.5	0.8	1
dmu:C	1	1	1	1	0
dmu:D	0.8	0.9	0.9	0.888889	-1
dmu:E	0.833333	1	1	0.833333	-1

1. Column 2 reports the technical efficiency score if constant returns to scale is assumed.
2. Column 3 reports that technical efficiency score if variable returns to scale is assumed.
3. Column 4 reports the technical efficiency score if non-increasing returns to scale is assumed.

   a. Non-increasing returns would be either constant returns to scale or decreasing returns to scale.
   b. Decreasing returns to scale exist when an increase in all inputs at a fixed rate causes output to fall faster.

4. Column 5 reports the scale efficiency score of the decision-making units. As indicated with reference to Figure 2 above, this is a measure of the amount by which all inputs can be reduced to produce the same output as the DMU moves from the variable returns to scale frontier to the constant returns to scale frontier.
   a. Note too that under constant returns to scale, technical efficiency is measured by the ratio of output to inputs. This is especially simple when there are only one input and one output.
   b. Scale efficiency is also simply measured as the ratio of the constant returns technical efficiency to the variable returns to scale technical efficiency.
5. Column reports whether the DUM operates on the increasing returns segment of the frontier, on the constant returns segment, or on the decreasing returns segment.

From the Stata report below, the program output indicates the following with respect to the details of the observed scale efficiencies:

6. DMU A operates on the increasing return to scale segment of the frontier and has a scale efficiency measure of 0.5. It can reduce its input use by about 50% if it moves to the constant returns to scale frontier and adopts the methods of DMU C. This also means that it could increase its output rate faster than its input rate by scaling up both its input and its output along the frontier to the same point where DNU C operates.
7. DMU B operates on the increasing return to scale segment of the frontier and has a scale efficiency measure of 0.8. It can reduce its input use by about 20% if it moves to the constant returns to scale frontier and adopts the methods of DMU C.
8. DMU D operates on the decreasing return to scale segment of the frontier and has a scale efficiency measure of 0.888889. It can reduce its input use by about 21% if it moves to the constant returns to scale frontier and adopted the methods of DMU C. Since it is on the decreasing return segment of the frontier, it can also improve its efficiency by reducing both its input and its output until it operates at the point occupied by DMU C.
9. DMU E operates on the decreasing return to scale segment of the frontier and has a scale efficiency measure of 0.833333. It can reduce its input use by about 17% if it moves to the constant returns to scale frontier and adopted the methods of DMU C.

	VRS Frontier						
	1	2	3	4	5	6	7
	dmu	o_q	i_x	CRS_TE	VRS_TE	SCALE	RTS
1	A	1	2	0.5	1	0.5	irs
2	B	2	4	0.5	0.625	0.8	irs

3	C	3	3	1	1	1	-
4	D	4	5	0.8	0.9	0.888889	drs
5	E	5	6	0.833333	1	0.833333	drs

6.3.3 Limitations of DEA

As with numerical methods generally, the main limitations of DEA arise from failure of the assumptions. Random noise, measurement error, or outliers are normal in data. It is not simply appropriate to assume that an outlier is also a best practice. The data used to represent inputs and outputs are often not well-understood and the measure of efficiency is very sensitive to the number of variables needed. The main problem here stems from the fact that the method does not have to specify a relationship between the inputs and the outputs of the decision-making units. It only requires that input and output combinations are known for each unit. Anything beyond the simple logic of aggregated "labour' and 'capital' begins to add inconsistencies of measure and interpretation. Unique outputs or inputs are typical in education and healthcare, and even in agriculture. Notwithstanding these limitations, DEA is the main approach used to measure efficiency in public decision-making units.

6.4 Stochastic Frontier Analysis

The PER Team should be aware that there exists a method called stochastic frontier analysis that provides a way to address two of the limitations of DEA – the presence of random noise, measurement error, and outliers on the one hand and the relationship between the inputs and outputs, on the other. However, it does so without solving the fundamental problem of the need for price weights to aggregate the multiple inputs and multiple outputs of public institutions.

Stochastic frontier analysis is a statistical method that is based on regression rather than mathematical programing. It uses regression and the input and output bundles described under the DEA methods above to estimate a production function. Then, it uses the random errors generated by the estimation process to measure efficiency. In particular, the method uses the estimated production function to specify a technical, cost, or profit frontier against which the units of the analysis are compared and their degree of efficiency measured. The representation here is output-oriented, because it is assumed that the partner countries are concerned with exploiting opportunities to increase the rate of output faster than the scale of inputs.
A production function is a technical correspondence that indicates the amount of output that is generated by a given amount of the inputs available. The correspondence is normally labelled F. Using the notation for the input and output bundles set out above, a production function for DMU$_i$ would normally be written as:

6. $Y_i = F(X_i)$

Here, $X_i = (K_i, H_i)$, where K stands for capital and H for number of workers. If data on hours worked are available, then those are preferred. However, both of these variables can be disaggregated, and indeed can be represented as they are in the DEA analysis. For example, capital can refer to the space used by the DMU and the amount of computers available to staff in the DMU. Labour can refer to the number of doctors and the number of nurses.

In stochastic frontier analysis, the variables are all expressed in their natural logarithms. Thus, one econometric model for the production function is usually written as:

7. $y_i = \beta_0 + \beta_1 k_i + \beta_l h_i + \varepsilon_i$

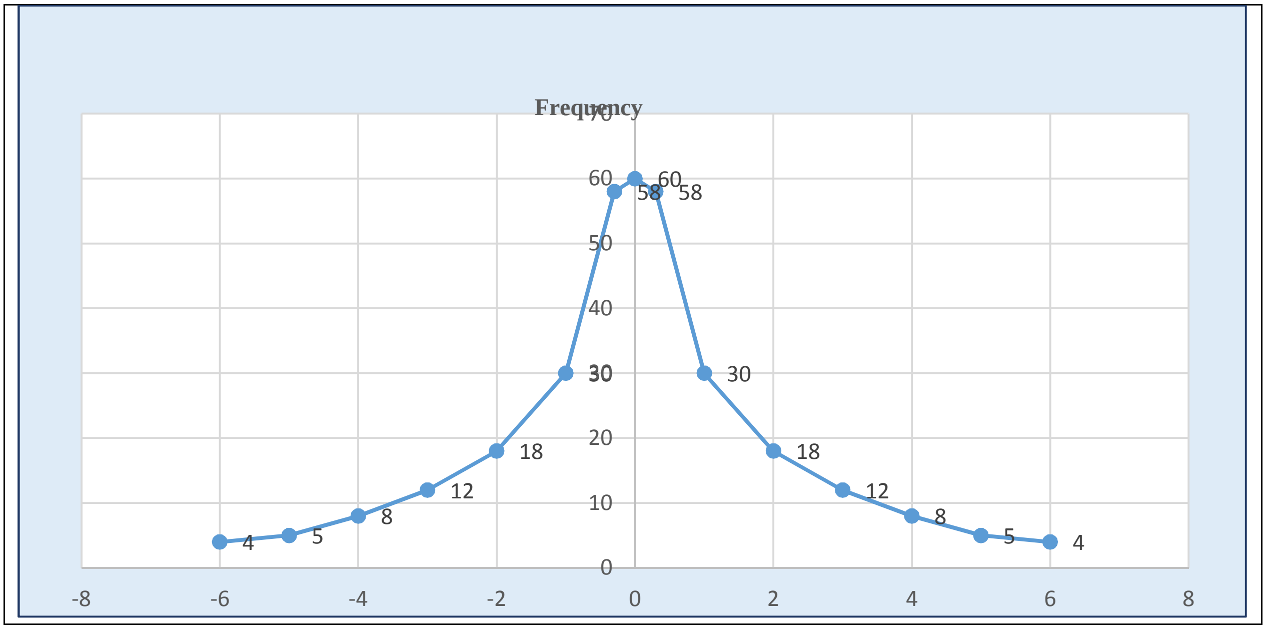
The β_j are the parameters to be estimated with regression analysis and ε_i is the error term that is the focus of the method. The error term measures the deviation between the actual observed output and the output predicted by the estimated equation component $\beta_0 + \beta_1 k_i + \beta_l h_i$ on the right hand side.

In stochastic frontier analysis, the error term is decomposed into two parts and written:

8. $\varepsilon_i = v_i - u_i$

The term v_i is a random variable representing random measurement error that is characterized as independently, identically, and normally distributed, and in particular as distributed independently of u_i. The term u_i is a non-negative random variable that is assumed to account for the degree of inefficiency of the unit observed, or its distance from the frontier observed. The term u_i is not normally distributed, and particular is sometimes assumed to be characterized by the upper half of a normal distribution (half-normal distribution). So, it is “positively skewed”. Figure 3 represents a normal distribution, so the upper half of the graph in Figure 2 illustrates a half-normal distribution.

Figure 3: Example of Normal and Half-Normal Distribution of Error



The regression to be estimated becomes:

$$9.\quad y_i = \beta_0 + \beta_k k_i + \beta_h h_i + v_i - u_i$$

In this regression model, the u_i now measures how far the DMU operates below its production frontier. If it is assume that inputs are properly allocated in relation to their costs (allocative efficiency), then the u_i is a measure of technical inefficiency such as might result from factors such as managerial inefficiency, outmoded equipment, or inadequate staffing of the DMU. If allocative efficiency cannot be assumed, then the u_i is possibly a measure of both allocative and technical inefficiency.

The efficiency of the DMU (Eff_i) is measured as:

$$10.\ Eff_i = \frac{\exp(-u_i)}{E(-u_i|v_i-u_i)}$$

Here,

1. exp is the exponential function
2. $E(-u_i|v_i - u_i)$ is the expected (mean) value of $-u_i$ given the observed values of $v_i - u_i$.

The sum of the estimated coefficients is used to measure scale efficiency. If $\beta_k + \beta_h > 1$, then the unit can increase output faster than it can grow its inputs and is deemed to be scale inefficient. Correspondingly, a good measure of the degree of scale efficiency is $\frac{1}{\beta_k+\beta_h}$. When the ratio takes the value of 1, the DMU is fully scale efficient and all potential for increasing returns have been exhausted.

All of this is easily done in Stata 14. The important lines of code are:

```
>frontier depvar [indepvars] [if] [in] [weight] [, options]
>test indepvar1+indepvar2=1
>predict double u_h, u
```

1. The line of code **frontier depvar [indepvars] [if] [in] [weight] [, options]** estimates the coefficients of the production function used to generate the frontier, generates the frontier, and calculates the values of v_i and u_i.
2. The line of code "**test indepvar1+indepvar2=1"** will test the sum of the coefficients for the degree of returns to scale. If the null hypothesis is upheld, then constant returns exists. If not, then increasing returns exist.
3. The line **"predict double u_h, u"** will produce the predicted level of inefficiency ($\widehat{u_i}$).

For example, if the logarithm of the output and inputs of the DMU is labeled **loutpu lkstock** and **lemploy,** then the lines of code would be:

```
>frontier loutpu lkstock lemploy, distribution (hnormal)
>test lkstock+lemploy=1
>predict double u_h,u
```

Then, equation (20) can be applied to get the measure of efficiency.

7. What to Do if No Financial Data

Most of these indicators require access to financial records on the actual outputs and the actual inputs used during each time period for each decision-making unit in the set of units being evaluated. In the absence of such records, a questionnaire should be designed, executed, and analysed, guided by **Annex 8**.