MEMORANDUM

No 07/2012

Efficiency and Productivity in the Operational Units of the Armed Forces

Torbjørn Hanson

ISSN: 0809-8786

Department of Economics
University of Oslo

This series is published by the
University of Oslo
Department of Economics

P. O.Box 1095 Blindern
N-0317 OSLO Norway
Telephone: + 47 22855127
Fax: + 47 22855035
Internet: http://www.sv.uio.no/econ
e-mail: econdep@econ.uio.no

In co-operation with
The Frisch Centre for Economic Research

Gaustadalleén 21
N-0371 OSLO Norway
Telephone: +47 22 95 88 20
Fax: +47 22 95 88 25
Internet: http://www.frisch.uio.no
e-mail: frisch@frisch.uio.no

Last 10 Memoranda

No 06/12	Sheetal K. Chand *The Relevance of Haavelmo's Macroeconomic Theorizing for Contemporary Policy Making*
No 05/12	Tone Ognedal *In the Shadow of the Labour Market*
No 04/12	Michael Hoel *Second-best Climate Policy*
No 03/12	Tom Kornstad, Ragnar Nymoen and Terje Skjerpen *Macroeconomic Shocks and the Probability of Being Employed*
No 02/12	Erik Biørn *The Measurement Error Problem in Dynamic Panel Data Analysis: Modeling and GMM Estimation*
No 01/12	Erik Hernæs and Zhiyang Jia *Earning Distribution and Labour Supply after a Retirment Earnings Test Reform*
No 27/11	Jørgen Heibø Modalsli *Polarization, Risk and Welfare in General Equilibrium*
No 26/11	Jo Thori Lind and Dominic Rhoner *Knowledge is Power: A Theory of Information, Income, and Welfare Spending*
No 25/11	Kjell Arne Brekke and Tore Nilssen *The Market for Fake Observations*
No 24/11	Berhe Mekonnen Beyene *Determinants of Internal and International Migration in Ethiopia*

Previous issues of the memo-series are available in a PDF® format at:
http://www.sv.uio.no/econ/forskning/memorandum

Efficiency and productivity in the operational units of the armed forces[1][2]

by

Torbjørn Hanson

Norwegian Defence Research Establishment (FFI)

PhD student Department of Economics, University of Oslo

February 2012

Abstract: Most nations spend a considerable part of their gross domestic product (GDP) on defense. However, no previous study has addressed the productivity and efficiency of the core area of the armed forces, operational units, using Data Envelopment Analysis (DEA). Introducing a model for the production process of an operational unit, productivity and efficiency are estimated by DEA for units of one branch of the Norwegian armed forces. Small samples are a characteristic of DEA studies in the military, and the public sector in general, resulting in a lion's share of the units being estimated as fully efficient. We find that, by using the bootstrap technique to estimate confidence intervals, we can point at the uncertainty of the estimates and reduce the number of candidates for best practice.

Keywords: Military; Productivity; Efficiency; DEA; Bootstrap

JEL classification: D24, H40, C60

[1] I am grateful to Finn R. Førsund and Sverre A. C. Kittelsen for their comments on this paper.

[2] The paper is written within the research program "Kostnadseffektiv drift av Forsvaret (KOSTER III)" (Cost Efficiency in Defence) at the Norwegian Defence Research Establishment (FFI). Views expressed are those of the author and do not necessarily reflect the views of FFI or the University of Oslo.

1. Introduction

Most nations spend a considerable part of their gross domestic product (GDP) on defense. NATO has set a target for its member countries to allocate at least 2 % of GDP to defense objectives. The branches or services of the armed forces like army, navy and air force produce services which are classical examples of public goods not provided by markets. However, most resources are bought in the market place or have shadow prices set by markets. [3] The fact that services are not sold in markets leaves the armed forces without the information from a price mechanism in evaluating efficient use of resources or effective mix of services. Despite the absence of price information on services, the assessment of efficient resource allocations may still be carried out by other methods if physical information on the services is available.

In the efficiency literature Data Envelopment Analysis (DEA) is a well established non-parametric method for efficiency studies which can be employed without any information on market prices. [4] Previous studies of efficiency and productivity by DEA in the armed forces have solely been concentrated around various support functions like maintenance and recruitment. However, operational units, the core area of defense, have not been studied in the literature. The purpose of this paper is to show that studies of efficiency and productivity by DEA can be carried out also for the operational units of the armed forces.

One reason for the lack of studies could be difficulties in modeling the production process and output of the armed forces. What is the output of the armed forces, and where can the line between outputs and outcomes of defense be drawn? These questions are addressed in this paper by setting up a general model for the production process of an operational unit. The model is then specified for the units of one branch or service of the Norwegian armed forces, the Home Guard. [5]

[3] The armed forces may have distinct legal rights to draw upon the resources of society, e.g. conscripted personnel are not paid according to market prices.

[4] DEA is a non-parametric method for the estimation of production frontiers by a piecewise linear surface enveloping the observations from above. The initial DEA model presented in Charnes et al. (1978) built on the earlier work of Farrell (1957). Statistical interpretations and an overview of recent developments can be found in e.g. Fried et al. (2008).

[5] The principal task of the Home Guard is to protect important infrastructure, support national crisis management, strengthen the military presence throughout the country and provide support to the civil community (Norwegian defence Fact and Figures, 2010).

Studies of productivity and efficiency are of interest to the Home Guard for identifying potential benchmarks. However, the low number of units in the Home Guard (11) constitutes a relative small sample limiting the interpretation of the results, as a major part of the units appear fully efficient. From a review of the literature of DEA in the military we do have reasons to believe that small samples are a common phenomenon for studies of the sector. In order to reduce the number of units estimated as fully efficient and thereby also reducing the number of potential benchmarks, the estimation is supplemented by other methods.

Introducing the method of resampling enables a statistical interpretation of the results and the constructing of confidence intervals around the estimates. Additional information provided by the confidence intervals can reduce the number of potential benchmark candidates among the units significantly and contribute to the making of more informed decisions for picking benchmark units within the Home Guard. Further, confidence intervals for the Malmquist index let us consider also the significance of changes in productivity. The resampling for the efficiency scores and the Malmquist index is done by the bootstrap procedure developed in Simar and Wilson (1999).

The paper is structured as follows. In section 2 a literature review of DEA in the defense sector is given. Section 3 of the paper presents concepts and data. First, military activity is linked to the concepts of public service activities, drawing a line between output and outcome in the sector, before we set up a general model for the output of an operational unit. The model is specified for the Home Guard, a branch of the Norwegian Armed Forces. The estimates from the 11 original observations are presented in section 4 of the paper, before we introduce the bootstrap procedure resampling the data, and additional pseudo observations are generated. The developments in productivity for the Home Guard and its units are presented in the last part of the section. Finally, section 5 of the paper concludes and points at some topics for further research.

2. Literature review

Previous DEA studies of productivity and efficiency in the defense sector are solely concentrated around various service and support functions, like maintenance and the recruitment of soldiers. In the following we will discuss the studies in terms of the production model, sample size and the choice of output measures. The studies are characterized by

relatively small samples and a wide range of various input and output variables, resulting in low degrees of freedom. For most studies this has the effect of a lion's share of the units being estimated as fully efficient, resulting in a lack of information for identifying best practice units. An overview of DEA studies in the military, including the field of study and number of variables, is outlined in table 1.

Maintenance is the most frequently studied field in the literature of DEA in the defense sector. This field starts with Charnes et al. (1985) study of 14 aircraft maintenance units in the U.S. Air Force over a period of seven months. The four outputs in the model include hours of mission capable aircrafts, hours of non capable aircrafts due to maintenance problems, number of sorties flown and the number of completed jobs of a specific type. By

Table 1. Bibliography of DEA in the military.

Paper	**Field**	**Inputs**	**Outputs**	**Observations**
Lewin and Morey (1981)	Recruitment	10	2	43
Charnes et al. (1985)	Maintenance	8	4	42
Bowlin (1987)	Maintenance	3	4	21
Bowlin (1989)	Accounting and finance	1	5	18
Ali et al. (1989)*	Recruitment	n/a	n/a	n/a
Roll et al. (1989)	Maintenance	3	2	10-35
Clarke (1992)	Maintenance	4	2	17
Ozcan and Bannick (1994)	Hospitals	6	2	23
Bowlin (2004)	Civil reserve air fleet	4	7	37-111
Brockett et al. (2004)	Recruitment	1	10	n/a
Sun (2004)	Maintenance	6	5	30
Farris et al. (2006)	Engineering design projects	4	1	15
Lu (2011)	Outlets	4	2	31

*Paper not available

applying window analysis[6] the number of observations is increased to 42. A study of a similar production structure is done in Roll et al. (1989) for the efficiency of aircraft maintenance units in the Israeli Air Force. In this study DEA was run for five maintenance units in windows of six time periods, giving 30 observations in each window. The original production model consisted of three inputs and six outputs. However, the model was modified after studies of the relationship between the variables by a team of experts. This procedure led to reducing the number of outputs by specifying some of the outputs as a weighting factor for other outputs[7], related to a subjective scale based on judgment from the expert team. The final model included, thus, two outputs: flying hours weighted by type of aircraft and the standard deviation of the daily number of sorties.

Bowlin (1987) is another study of maintenance activities in the U.S. Air Force. The case here is real-property maintenance. Real property refers to land and land improvements such as buildings and appurtenances. The measured outputs are completed work orders, job orders and recurring work actions. Seven bases were studied using window analysis, increasing the sample from seven to 21 observations. Vehicle maintenance at 17 U.S. Air Force bases over a period of four years is studied in Clarke (1992). The production model consisted of two outputs, the average number of days during a month the assign vehicles are in serviceable condition and the number of trained mechanics. Another study of maintenance in the military is Sun (2004). Here five joint maintenance shops in the Taiwanese Army are studied over two periods of six months. With monthly data each period consists of 30 observations. The chosen output measures are similar to Clarke (1992), but the number of outputs is increased by including separate outputs for each vehicle type.

Accounting and finance offices at the U.S. Air Force base level are studied in Bowlin (1989). The production model consists of a single input, employee compensation, and five outputs measuring the number of various transaction types processed. A total number of 18 units are studied using cross-section analysis for a period of three years. The observations are also pooled for study of time variation.

[6] The window analysis technique was first employed in Charnes et al. (1982). The technique is described in Charnes et al. (1994) as a moving average analogue, where a DMU in each period is treated as if it where a different DMU. In the Charnes et al. (1985) study the size of the windows is set to three months, and each of the 14 maintenance units are represented as if they were a different DMU for each of the three successive months. This procedure increases the number of observations to 42 (3x14).

[7] The type of aircraft was introduced as a weighting factor of flying hours, using the subjective scale from the team of experts.

The efficiency of military recruitment is studied by DEA in Lewin and Morey (1981), Ali et al. (1989) and Brockett et al. (2004). The first study is comparing the performance of 43 U.S. Navy recruiting districts on quarterly data over a period of three years. The outputs are the number of contracts signed by school eligible and non-eligible male candidates. Brocket et al. (2004) combines regression analysis with DEA in the study of joint versus service specific advertising on military recruitment. The production model consists of a single output, total number of enlistments, three discretionary inputs and seven non-discretionary inputs.

Department of Defense hospital efficiency is studied in Ozcan and Bannick (1994). In this study 124 U.S. military hospitals are evaluated and compared with the performance of 3656 community hospitals over a period of three years. The output measures are total annual inpatient days and outpatient visits.

Other applications of DEA related to the defense sector are the evaluation of engineering design projects, the supply of supplementary food and products to soldiers and veterans, and the financial performance of reserve air fleet participants. Farris et al. (2006) study the performance of engineering design projects in the Belgian armed forces. In total 15 projects are evaluated, where project duration is chosen as a single output. Bowlin (2006) studies the financial performance of airlines participating in the US Department of Defense's civil reserve air fleet, and compares the performance to a group of non-member airlines. Lu (2011) studies the provision of food and products by 31 military outlets managed by the Taiwanese general welfare service ministry.

3. Concepts and data

Methodology

In general, when considering public service activities, we can distinguish between two aspects, as described in Førsund (2011) and outlined in figure 1. The first aspect is about the services produced by employing resources by the institution or state agency in question. The second aspect is about the effectiveness of these services, i.e. the impact the services make on the objectives that motivate producing the services in the first place. This distinction leads to the saying that efficiency is a question of "doing things right" and effectiveness is a question

of “doing the right things”. In the following we will discuss whether this model is straight forward to apply to the military.

Resources, in the upper left of figure 1, are easily definable and verifiable on an aggregate basis, as is the ultimate objective of providing the peace behind the lower right of the figure. However, there is no clear connection between the two endpoints in the sense that a marginal change in defense budgets is unlikely to have an immediate impact on the status of peace. The main issue when applying the model to the military is the distinction between and categorization of: (1) Outcomes; (2) Outputs; (3) Activities in the transformation process.

Schreyer (2010) discusses the different meanings of outcome in the literature on non-market services and follows the definition typical among national accountants, in line of Eurostat (2001), where outcome is used to describe a state that is valued by consumers. However, the desired outcomes of services from the armed forces are deduced from security policy on a national level and not from personal or consumer needs, which is the case in for example the provision of health and education services. In that sense, outcomes of military services differ from the Schreyer (2010) definition as they are valued by politicians rather than consumers.

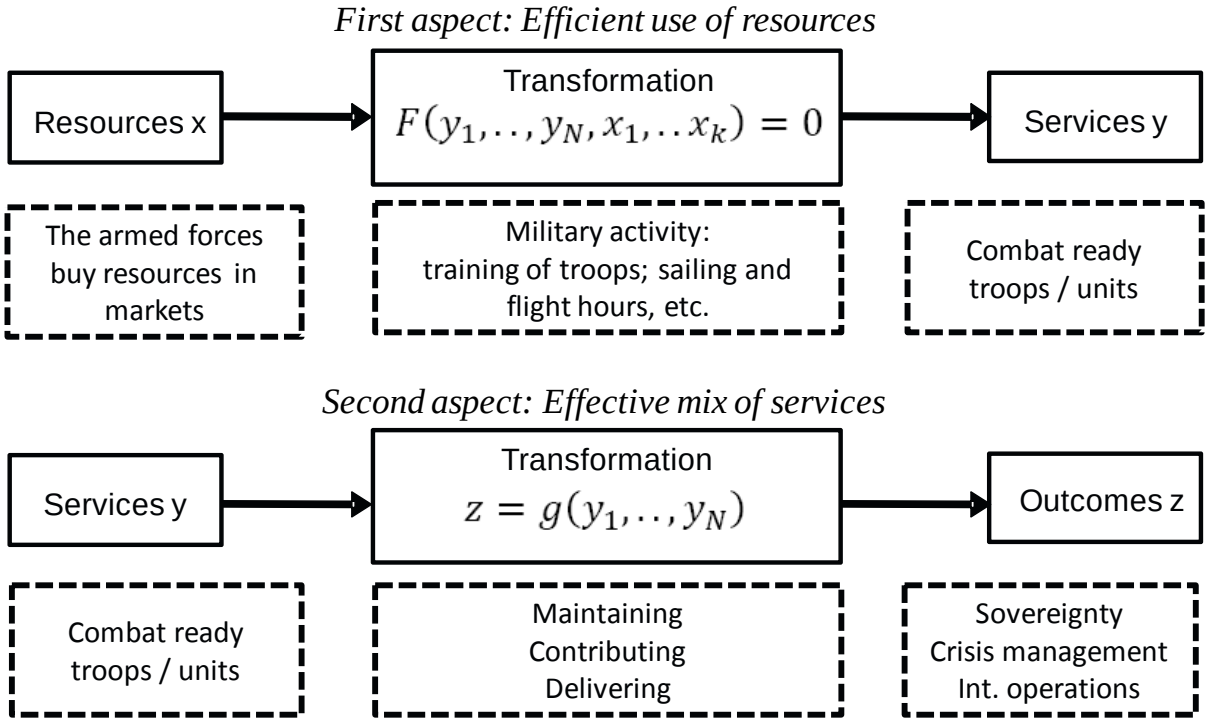


Figure 1. The two aspects of public service activities, and a specification for the armed forces

The definition of output we refine by breaking it down into two components: activities and the quality aspect associated with them. Activities are observable and countable actions by which services are delivered. The outputs are provided by military units. A military unit, the decision making unit (DMU) in our study, is defined in the UK Army Doctrine (2010: 89) as *"…the smallest grouping capable of independent operations with organic capability over long periods…"*.[8] Units are subdivided into sub-units. In our model a unit is subdivided into troops, typically of between 12 and 35 soldiers.

Military activity is likely to be observable, such as fighter planes spotted in the skies or navy vessels spotted at sea, either as a part of a training exercise or as a way of deterring and showing military presence. The activity is easy measurable in form of flight hours and number of hours in open seas. Output, as we define it at unit level, is among other factors the result of such activities. [9] If activities are intermediate outputs, these should neither be minimized nor maximized. Hence, the size of the activity is not an end in itself. The output at unit level is generally more complex and must be modeled based on several standards of activity.

Model for the output of a unit

In modeling output we look at variables for activity standards defining a combat ready unit or troop and a variable for the corresponding quality of those activities. Combat ready units are then the services realizing outcomes. The structure of the model is outlined in figure 2 and explained briefly in the following: The production process of an operational unit is formed to prepare the unit for its given tasks or operations. The tasks can be considered as the outcome of the unit production outlined in the first row of the figure. A typical task for an operational unit is to help maintaining sovereignty. In order to reach this outcome the operational units prepare a combination of equipment and personnel of a given standard. Further, first the troops and later the unit, are trained in order to reach a given proficiency level, the quality aspect of the service. For the unit to be combat ready both the standards for personnel and

[8] Units typically comprise between 400 and 1000 people. In the navy, a unit is typically called a commando, while the optimal grouping in the air force is a large squadron (UK Army Doctrine, 2010).

[9] For some units deterrence could be considered among the output variables. In general we are modeling deterrence rather as an outcome of the production of a military unit, leaving activities like sailing and flight hours as solely inputs in the production process. These training activities are often confused with output.

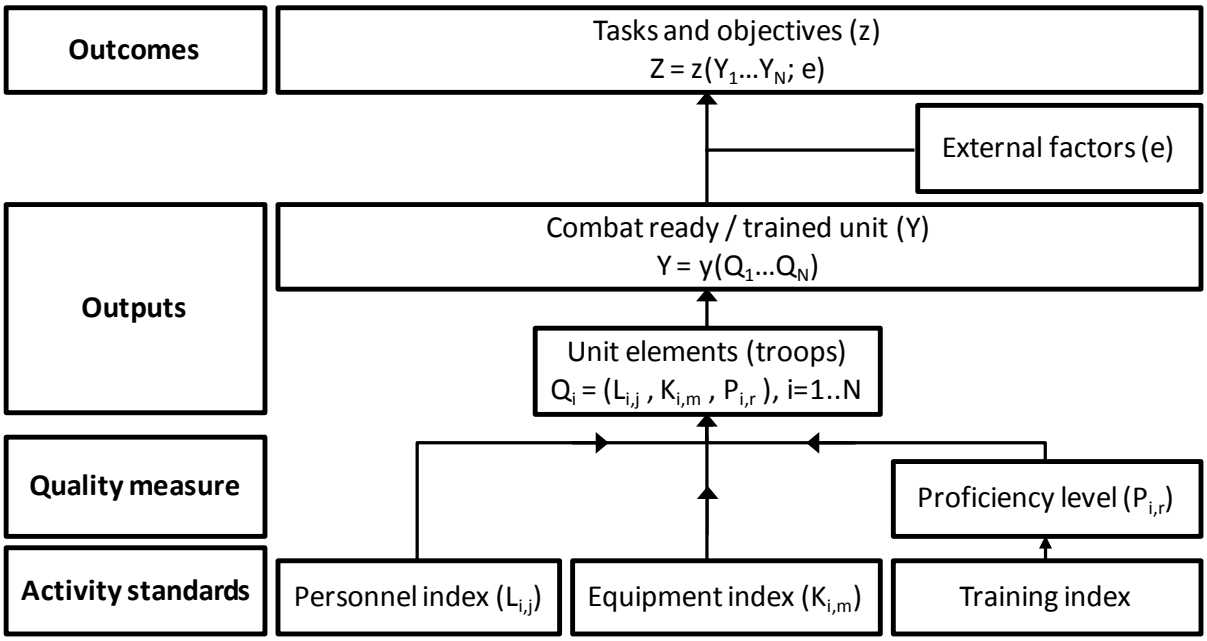


Figure 2. Model for the output of a unit

equipment, and a given proficiency level must be met. The activity standards are outlined in the bottom row in the figure.

Most units are divided into specialized sub elements, such as various troops, which have to be trained both individually and together as a unit. The break down of unit output into various troop types is outlined in the second row of the figure. The unit is internally dependent on the presence of all troop types in order to meet its tasks. In a cycle of for example one year the unit is to train the troops to a given proficiency level, maintain the proficiency level, and provide equipment of a specific type. The use of resources during a cycle is to be minimized within those main activities. We will return to a full specification of the model and motivate the aggregation of output below.

It is worth noting that the key factor in measuring the output of operational units is well defined and observable proficiency levels. The tasks or operations will vary among different types of units, but the concept of preparing and training will hold for all unit types in form of a fighter plane squadron as well as a submarine commando.

The very existence of the operational unit itself, and the fact that the possible threat to it is aware of this, makes the stream of services from the unit continuous and without a time span. However, the combination of equipment, personnel and their proficiency levels will vary over time.

By using the time dimension and proficiency levels we can measure the output as months on each level of proficiency. The maximum output for one troop, the disaggregation of a unit, is then 12 months at the highest proficiency level. For a unit there is no upper limit on output as long as there is no restriction on number of troops. However, a unit is typically designed to produce a given number of troops considered adequate with respect to its tasks. Let us say a unit consists of three troops that can be trained at three different proficiency levels. The output(s) of the unit is then given by the number of months at the various proficiency levels for each troop. Given that each of the three troops has to stay at least one month at a level before proceeding, the maximum output for the unit (three troops of 12 months) is three troop-months at level one, three troop-months at level two, and 30 troop-months at level three. The possibly drawback of this approach is the potential increasing number of outputs which follows several levels of both proficiency levels and different troop types. As we will show, this can be handled by aggregating the outputs at unit level.

The aggregation to a single output is mainly motivated by two considerations:

(1) Aggregation provides an intended property in the modeling of unit output. The unit is not only designed to produce a given number of troops, but also a simultaneous presence of certain troop types. This unit design, and the mutual dependence among troop types for handling the unit's tasks, makes the aggregation necessary for unit output as a concept. Only a balanced production of troops can lead to a meaningful statement of unit output. This property is also important when determining functional form in the unit output model. The same argument holds for the aggregation of processes in the troop output as a single variable cannot give a meaningful representation of the troop output.

(2) By aggregating outputs we limit the dimensionality in the DEA model. We believe that studies of the defense sector are challenged by a lack of observations, which is also underpinned by findings in our literature study of DEA in the military. In addition, the practice from national accounts can be taken into consideration. In national accounts data, output is aggregated into a volume index. Schreyer (2010) suggests a grouping of products according to their contribution to outcome. The criterion for grouping individual items is that they satisfy the same or similar consumer needs.

We are using a Cobb Douglas function in modeling the unit output. The choice of function is not based on any empirical estimates of the underlying technology, but rather for its mathematical characteristics. The Cobb Douglas function is simple and contains few

elements. The argument of simplicity is important for the specification of the model by military experts, as we will return to below. Given the requirement of a balanced production of troops in unit output, diminishing marginal returns to the factors is another important characteristic of the function.

The unit output *Y* is modeled as a function of its N number of lower level components *Q* such as troops in a land based unit, vessels in the navy and aircrafts in the air force.

$$Y = y(Q_1, .., Q_N) = {Q_1}^{a_1} {Q_2}^{a_2} \dots {Q_N}^{a_N}, a_1 + a_2 + \cdots + a_N = 1 \quad (1)$$

In equation (1) above the requirement for specific troops and compositions is modeled by a Cobb Douglas function, where the coefficient can be interpreted as weights for the various troop types. The weights are set by military experts and decision makers, or estimated from cost data.

The output of a lower level troop Q_i is in general modeled as a function of the various activity standards represented by a personnel index $L_{i,j}$, equipment index $K_{i,m}$ and the quality aspect represented by the proficiency level $P_{i,r}$, where *j*, *m*, and *r* represent different levels. Again the relationship can be modeled by a Cobb Douglas function in equation (2), with the coefficients set by military expert opinion.

$$Q_i = q_i\left(L_{i,j}, K_{i,m}, P_{i,r}\right) = {L_{i,j}}^{l} {K_{i,m}}^{k} {P_{i,r}}^{p}, l + k + p = 1 \quad (2)$$

Introducing a time span for the unit production of *T* periods gives from (1) the aggregated unit output *Y* in equation (3).

$$Y = \left(\sum_{t=1}^{T} Q_{1,t}\right)^{a_1} \left(\sum_{t=1}^{T} Q_{2,t}\right)^{a_2} .. \left(\sum_{t=1}^{T} Q_{N,t}\right)^{a_N}, a_1 + a_2 + \cdots + a_N = 1 \quad (3)$$

The Norwegian Home Guard

The main objectives for The Norwegian Home Guard are to protect the local population and the essential functions of society. To achieve these objectives the Home Guard has defined several tasks that include helping to maintain sovereignty, national crises management, the reception of allied reinforcement and contributing to the safety and security of society. The Home Guard consists of one Main Staff, two school departments, and a number of

operational districts located in all geographical regions in Norway with tasks related to either naval, air force or land activities. A district consists of the district staff and a number of various troop types. The personnel in a Home Guard district are mostly conscripted personnel with a full time job outside the military, except for the personnel in the district staff who are full time employed in the armed forces. The personnel in the district staff, around 50 people, are either officers or civilians. The number of conscripted personnel in a district can vary in the range of 300 to 1000 officers and 1500 to 4500 soldiers. The conscripted personnel are trained a given number of weeks a year.

We have modeled the production of the eleven districts performing land and air force activities. The objectives of the Home Guard have the characteristics of outcomes rather than outputs. From the tasks relevant for a single district, we have defined the dimensioning production to be (1) certain types of troops at various levels and (2) the training of officers. The officer training and troops at different levels are then aggregated to one single output.

Home guard model

The measure of troop production is modeled from various indicators registered at district level and reported to the Main Staff. Which indicators to use in defining troop production level is based on expert opinion from personnel at the districts and the Main Staff. The single output has the following decomposition: Three different types of troops are produced; High intensity troops (I-FO), reinforcement troops (RF-FO), and the district staff (DS). Troop size can vary among troops of the same type, and therefore also the size of the units. The variation in unit size is modeled by introducing an index, *s*, for the size of the unit. The unit output from three different troop types with a time span of 12 months follows from (3):

$$Y = y(s, Q_1, Q_2, Q_3) = s\left(\sum_{t=1}^{12} Q_{1,t}\right)^{a}\left(\sum_{t=1}^{12} Q_{2,t}\right)^{b}\left(\sum_{t=1}^{12} Q_{3,t}\right)^{c}, a + b + c = 1 \quad (4)$$

In general there is no time span for the production as the total stream of services from a unit is more or less continuous. We therefore use the status at the end of each month rather than the status of each year for a higher precision. A time span of 12 months corresponds with the yearly planning cycles in the unit and the input measure from the unit accounts. The number of troops of each type to be produced in a given district is set by the Main Staff. Each district

produces the given number of troops to the highest level possible given its budget restriction. In order to meet its tasks, a district has to produce all types of troops simultaneously. The use of inputs in each troop type varies as well as the test requirements for passing a given level. Troop types and levels are given a weight by the Main Staff to represent these differences. The single aggregated district output is then given by the number of I-FO equivalents, where each troop is scaled by a number representing its size, and officer training level.

Each troop type Q_i is produced at three different proficiency levels $P_{i,r}$, where each level r is completed by passing a standardized test. The equipment standard is given by an index $K_{i,m}$ and the personnel standard by an index $L_{i,j}$. The training level of officers O_i is measured by the number of training days. From equation (2) the troop production is modeled by a Cobb Douglas functional, with personnel and equipment standards as well as weights set by the Main staff:

$$Q_i = q_i\left(L_{i,j}, K_{i,m}, P_{i,r}, O_i\right) = {L_{i,j}}^{l} {K_{i,m}}^{k} {P_{i,r}}^{p} {O_i}^{o}, l + k + p + o = 1 \qquad (5)$$

Input

Available input data for the production process in the Home Guard is mainly cost data from the Home Guard district accounts. Three input variables are defined for the use of equipment and personnel: (1) Fixed personnel costs, such as regular wages, (2) variable personnel costs, such as activity based payments, overtime pay and travel expenses, (3) material costs, such as ammunition, spare parts and maintenance. Due to the lack of accrual accounts, the activity based troop specific expenditures are not perfectly conceding with the troop activity and output. Typically, expenses in year one are materializing in output for year one and the two subsequent years. In order to match the inputs with output, all troop specific expenditures are spread over three years.

Data / sample

The sample consists of yearly observations from eleven Home Guard districts over three years. This gives us 11 observations each year or 33 observations after pooling the data and assuming the technology to be stationary. The data is collected from monthly and yearly

Table 2. Descriptive statistics for the eleven units, 33 observations

Variable	Min	Max	Mean	Median	SD
Y: Output	6.1	46.1	23.9	22.6	8.8
X1: Variable personnel cost	8.4	23.6	16.3	15.3	4.7
X2: Fixed personnel cost	10.4	37.5	24.4	22.7	6.8
X3: Equipment cost	7.3	29.1	15.3	15.5	4.9

Table 3. Correlation coefficients among inputs and output

Variable	Y: Output	X1: Variable personnel cost	X2: Fixed personnel cost	X3: Equipment cost
Y: Output	1			
X1: Variable personnel cost	0.65	1		
X2: Fixed personnel cost	0.75	0.87	1	
X3: Equipment cost	0.47	0.68	0.68	1

district reports to the Main Staff. The districts are similar regarding tasks and troop types, but different in size of personnel and geographical area. Differences in geographical area may have some cost implications such as higher travel and transport expenses. The input variables are adjusted using the consumer price index.

4. Results

We are using an input oriented[10] DEA model to estimate the Farrell technical efficiency scores for both the yearly and pooled data. The productivity development over the three years we investigate by the Malmquist productivity index. [11]

We start out by assuming a variable returns to scale technology (VRS). The VRS technology is the most general case, and any challenges resulting from a limited number of observations will appear more clearly under this assumption compared to the assumption of constant returns to scale (CRS). The initial DEA run for each of the three years yields the sets of efficient units (efficiency score of 1) and the sets of inefficient units (efficiency score less

[10] In the time period of our study, reducing input while maintaining the output level was a stated focus in the management of the Norwegian armed forces.

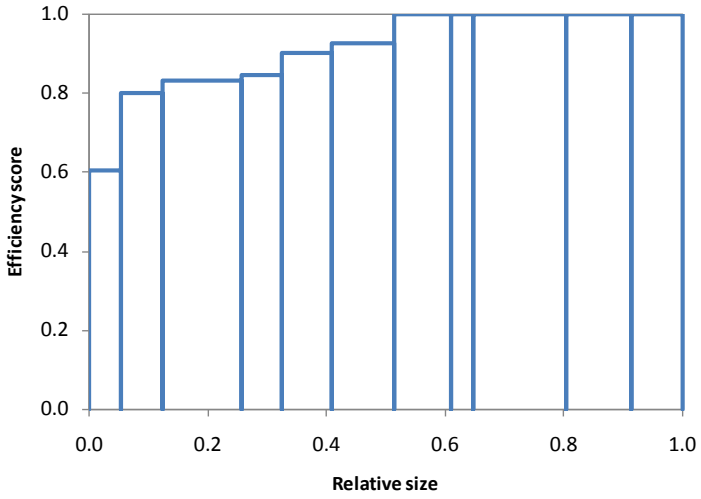
[11] All estimates and bootstraps were carried out by the FrischNP3.4 software, developed by The Ragnar Frisch Centre for Economic Research.

than one). About half of the units are estimated as efficient each year. The results are illustrated in figure 3 where the efficiency score for each unit is represented by a blue bar. The width of the bar represents the relative size of the unit measured by a weighted sum of the inputs. The share of efficient units is quite high, and it is difficult to identify best practice units on the basis of these results alone. However, compared to previous studies of the military the share of units estimated as efficient is not remarkably high. In our literature study we found that on average 44 % of the units were estimated as fully efficient in each of the studies.

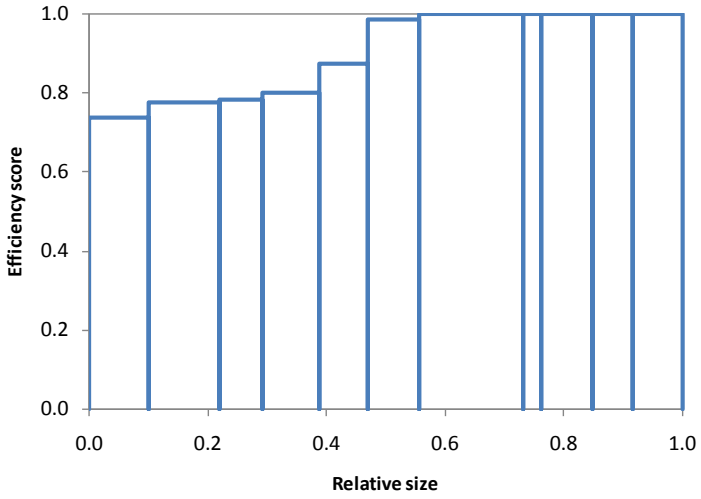
The high share of efficient units could be related to a low degree of freedom in the model. The literature suggests various rules of thumb to deal with the degrees of freedom. Cooper et al. (2007) give a rough rule of thumb, where the number of decision making units should be at least as great as the maximum of the product of inputs and output factors or three times the sum of the factors. Other rules in the literature suggest that there should be at least two observations for each input and output factor (Bowlin, 1987). In our DEA model of three inputs and one output the rule of thumb from Cooper et al. suggests a minimum of 12 observations. By this rule alone our yearly samples of eleven units are not sufficient. For comparison, applying the rule of thumb on previous studies of the military, two out of three studies fully exploit the dimensionality of the model.

Pedraja-Chaparro et al. (1999) stress that a mere count of number of factors in a DEA model is an inadequate measure of the dimensionality of the model. The correlation between inputs (or outputs) in DEA analysis is also of fundamental importance. A positive correlation between two inputs gives less information to the analysis than if the inputs were not correlated. One implication of this result is that the adequacy of a DEA model to some extent is an empirical question. Also, in deciding variable specification of a model correlation can be an important factor. Kittelsen (1999) shows that the extent of correlation is clearly important when testing the relevance of an additional input in the model.

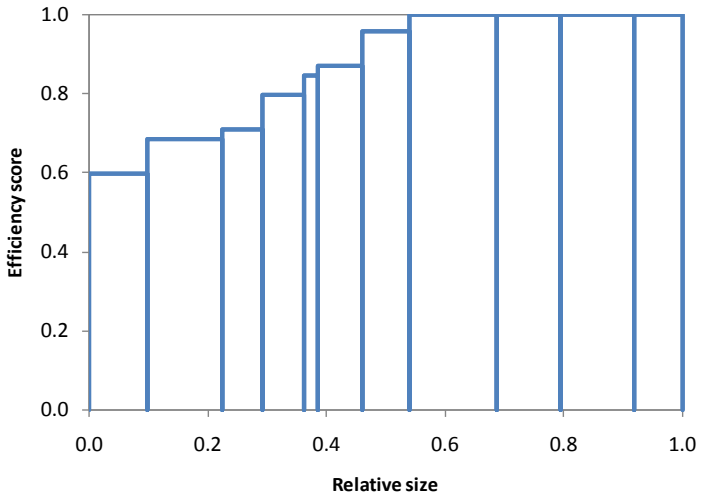
In order to further investigate the variation in efficiency among the units we can use the method of resampling. By resampling and generating additional data we can test whether it is possible to overcome some of the challenges related to the relative low number of units by pointing at the uncertainty related to the estimates.



Panel 1. Year 1



Panel 2. Year 2



Panel 3. Year 3

Figure 3. DEA efficiency scores (blue bars), VRS frontier. Relative unit size is represented by the weighted sum of inputs (width of the bars).

Resampling and bootstrapping

The estimates of efficiency which researchers are interested in involve uncertainty due to sampling variation. Efficiency is only measured relative to estimates of an underlying true production frontier, conditional on the observed data resulting from an unobserved data-generating process. The uncertainty of the estimated efficiency scores can be illustrated by a confidence interval. Simar and Wilson (1998) demonstrate that the key to statistically consistent confidence intervals lies in the replication of the unobserved data-generating process, and that this can be carried out by a bootstrap procedure.

Bootstrapping is a way of testing the reliability of the dataset, and works by generating artificial observations using resampling of the original dataset. The empirical distribution of the efficiency scores from the initial DEA run is used to estimate a smoothed distribution by a Silverman (1986) kernel density estimate (KDE) using reflection to avoid the accumulation of efficiency values of one. We then generate 2000 artificial observations by first projecting all inefficient observations to the DEA frontier and then drawing randomly an efficiency score for each unit from the KDE distribution.

Confidence intervals and the bias-corrected efficiency scores

In order to construct the confidence intervals for the estimates we follow the procedure in Simar and Wilson (2008). This involves sorting the values of the difference between the bootstrap estimates, $\hat{E}^*$, and the original estimated efficiency scores, $\hat{E}$, deleting ((α/2) x 100)-percent of the elements at either end of the sorted array, and then setting the endpoints equal to $c_{\alpha/2}$ and $c_{1-\alpha/2}$. The confidence interval is then given by

$$Prob\left(c_{\alpha/2} \leq \hat{E}^* - \hat{E} \leq c_{1-\alpha/2}\right) = 1 - \alpha \qquad (6)$$

The DEA method is based on enveloping the observations as tightly as possible from above. However, there might be potential realizations of the unknown technology not appearing as actual observations. This results in a frontier estimator that is pessimistically biased, and correspondingly efficiency scores which are optimistically biased. Simar and Wilson (1998) showed how to estimate the sampling bias in DEA using the bootstrap method.

From equation (6) we have estimated the 95 % confidence intervals for each of the three years. The confidence intervals are outlined by the red lines in figure 4 together with the bias corrected efficiency scores represented by the blue bars. The bars in the figures are sorted by the size of the original estimates for easy comparison with the corresponding panels of figure 3. The width of the confidence intervals varies significantly among the efficient units, giving us information on the uncertainty of the estimates. This information let us eliminate potential best practice units from around five to two units each year.

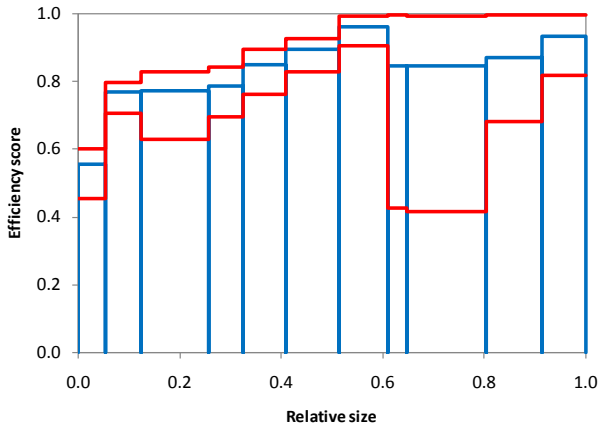
The bias-corrected efficiency scores are set out in panel 1 of figure 5 by the red lines. All units get a considerable downward shift in efficiency. The bias correction has the highest impact on one of the two units estimated as fully efficient in the original run, and leaves us with a single candidate for best practice unit. However, as noted by Efron and Tibshirani (1993), the bias-corrected estimator may have higher mean square error than the original estimator. In 20 out of 33 observations the mean square error of the corrected estimate is higher than the error of the uncorrected estimate. Simar and Wilson (1999) suggest that whether bias correction should be used is always an empirical question.

In our data a shift of the confidence intervals from the bias corrected estimate to the original estimate will imply an upper limit of the interval above one[12] for two of the observations, outlined by the red lines in panel 3 of figure 5. Obviously, an efficiency score above one does not make any sense, and we cannot do any reasonable inference, such as hypothesis testing, on the basis of this estimate. Despite such contradictive estimates of upper limits the results still provide useful information.

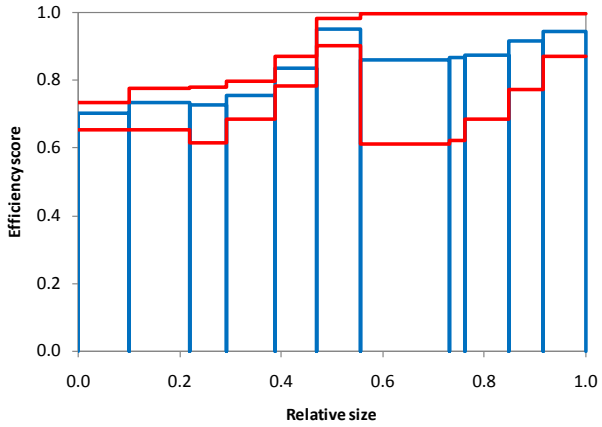
By looking at the standard deviation of the estimates we can say something about the uncertainty of the estimates. We therefore choose to present the results including bias correction, where the size of the confidence intervals represents the statistical uncertainty of the estimates. The original and corrected estimates together with standard errors are also presented in table 4.

Eliminating the units with the highest standard deviations can limit the number of potential best practice units substantially. In fact, this procedure reduces the span of best practice units to one single unit in the pooled sample outlined in figure 5.

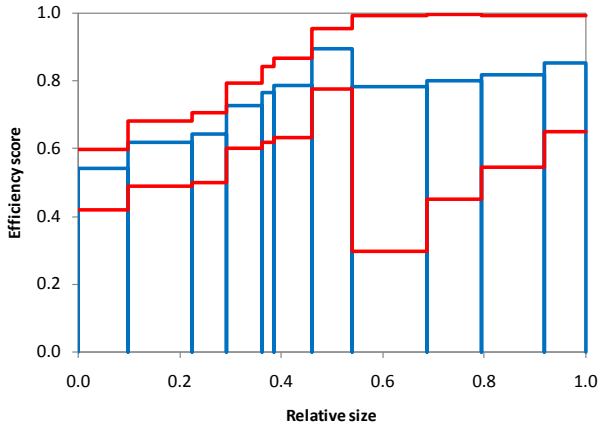
[12] Simar and Wilson (2008) discuss the problem of lower bounds for confidence intervals that are negative for the Shephard (1970) distance function. To our knowledge the problem of intervals in the range <0, 1>, or above one by the Farrell measure, is not discussed in the literature.



Panel 1. Year 1

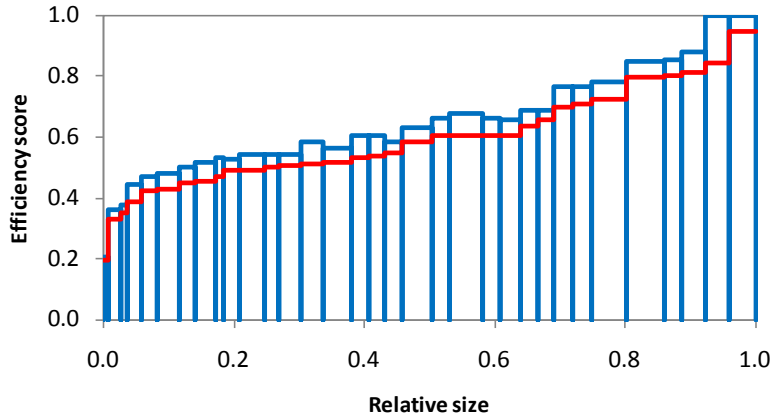


Panel 2. Year 2

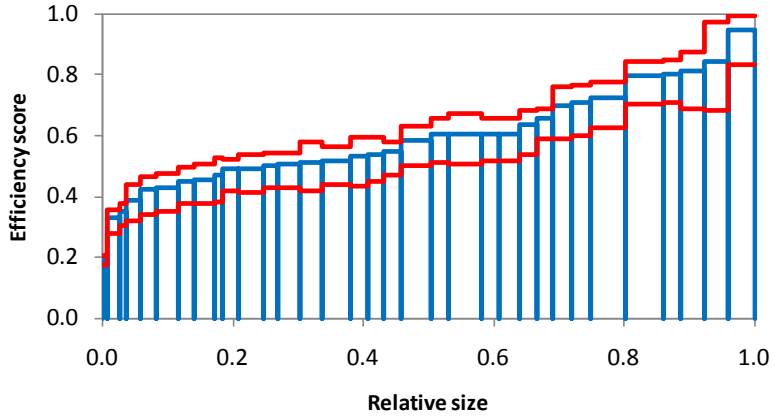


Panel 3. Year 3

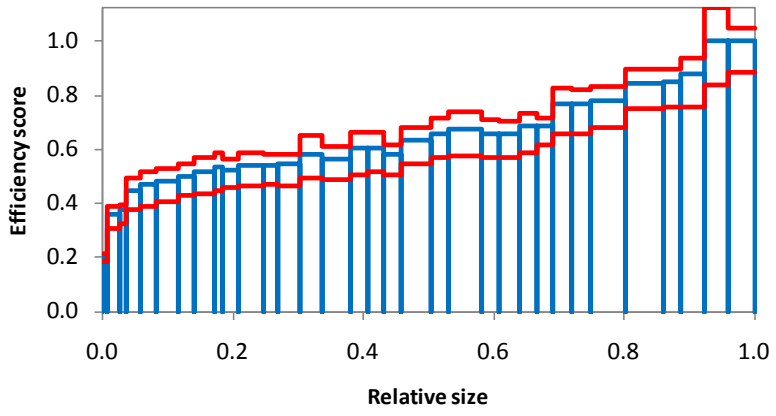
Figure 4. DEA bias-corrected efficiency scores (blue bars) and 95 % confidence intervals (red lines), VRS frontier. Relative unit size is represented by the weighted sum of inputs (width of the bars). Confidence intervals estimated by bootstrapping the efficiency scores.



Panel 1. Original (blue bars) and bias corrected (red lines) estimates



Panel 2. Bias-corrected estimates (blue bars) with 95 % confidence intervals (red lines)



Panel 3. Original estimates (blue bars) with 95 % confidence intervals (red lines)

Figure 5. DEA bias-corrected and original estimates. CRS frontier and pooled sample. Confidence intervals estimated by bootstrapping the efficiency scores.

Table 4. Results from the pooled sample, 33 observations

Estimate	Bias corr.	SE	MSE est.	MSE bias corr.
0.206	0.196	0.009	0.000	0.000
0.360	0.329	0.021	0.001	0.002
0.375	0.352	0.019	0.001	0.001
0.446	0.388	0.032	0.004	0.004
0.471	0.421	0.033	0.004	0.004
0.481	0.426	0.032	0.004	0.004
0.499	0.449	0.032	0.004	0.004
0.517	0.453	0.036	0.005	0.005
0.534	0.470	0.037	0.006	0.006
0.526	0.489	0.029	0.002	0.003
0.542	0.491	0.033	0.004	0.004
0.543	0.502	0.030	0.003	0.004
0.545	0.504	0.030	0.003	0.004
0.584	0.509	0.041	0.007	0.007
0.564	0.515	0.034	0.004	0.005
0.603	0.533	0.041	0.007	0.007
0.603	0.536	0.040	0.006	0.006
0.582	0.545	0.029	0.002	0.003
0.633	0.583	0.036	0.004	0.005
0.661	0.603	0.039	0.005	0.006
0.676	0.606	0.044	0.007	0.008
0.660	0.607	0.037	0.004	0.006
0.658	0.608	0.037	0.004	0.005
0.686	0.635	0.038	0.004	0.006
0.689	0.658	0.028	0.002	0.003
0.767	0.698	0.046	0.007	0.009
0.767	0.711	0.043	0.005	0.007
0.781	0.726	0.041	0.005	0.007
0.847	0.798	0.040	0.004	0.006
0.853	0.804	0.039	0.004	0.006
0.879	0.812	0.050	0.007	0.010
1.000	0.846	0.074	0.029	0.022
1.000	0.947	0.045	0.005	0.008

Productivity development

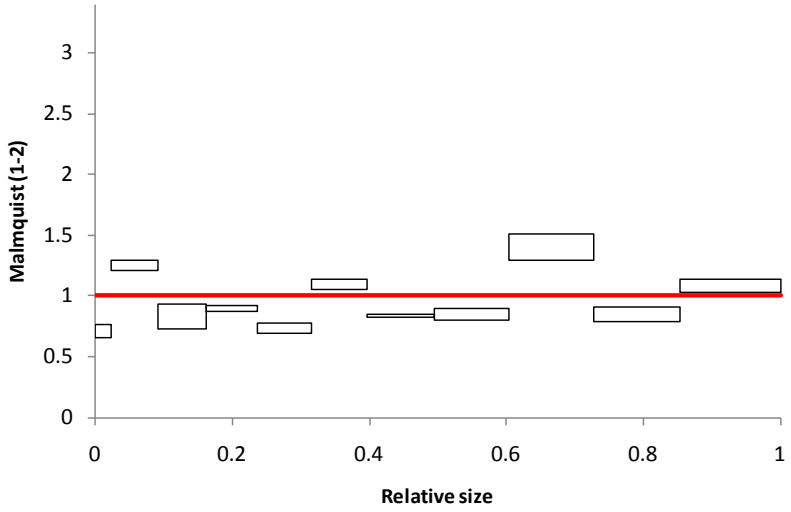
In order to study the development of productivity for the Home Guard districts, the Malmquist index (Caves et al., 1982) is estimated for changes in productivity between each of the three years. The Malmquist productivity index is based on the ratio of Farrell (1957) efficiency measures for two different time periods, 1 and 2, where the efficiency is measured against the same benchmark frontier technology *s*. Since the benchmark frontier is the same this relative measure has the interpretation of productivity change. The standard Malmquist index for a unit *i* is defined as

$$M_i^s(1,2) = \frac{E_{i,2}}{E_{i,1}}, i = 1, \dots, N \qquad (7)$$

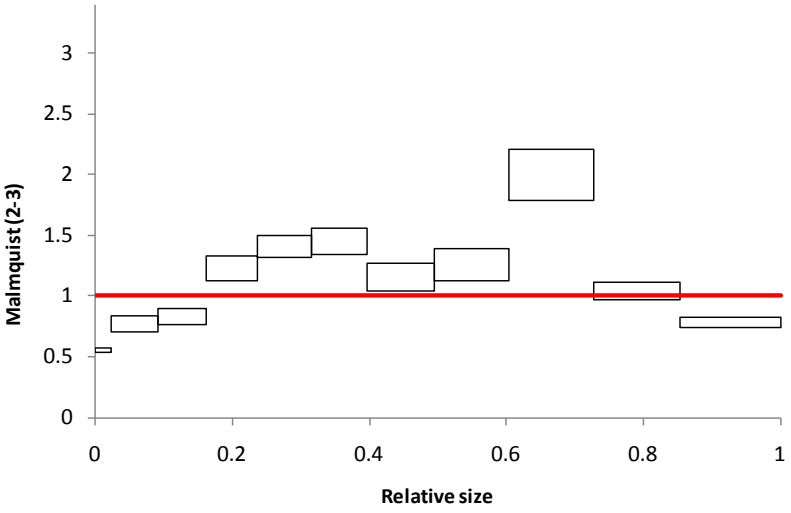
$E_{i,1}$ and $E_{i,2}$ are the Farrell efficiency scores for period 1 and 2 respectively.

When choosing benchmark technology for productivity measurement (at least) the two following considerations have to be taken into account (Førsund, 2010): the desired homogeneity property of the productivity index, and comparability of productivity changes between different periods. Grifell-Tatjé and Lovell (1995) shows that the Malmquist index provides an inaccurate measure of productivity change in the presence of non-constant returns to scale. Doubling all inputs and outputs from one period to the next, keeping input and output mixes constant, should not change productivity. Hence, the productivity measure homogeneity should be of degree 1. This property makes the VRS specification unsuitable as a benchmark technology. Therefore, CRS is chosen as benchmark technology in the Malmquist index. It is worth noting that using CRS just serves as a benchmark technology for the productivity measure, and no general assumptions of CRS are necessary. In order to compare productivity between different periods the index has to be circular, which is achieved by keeping the benchmark technology *s* in (7) fixed for all periods (Berg et al., 1992).

Due to the short time span of three years are we assuming that no change in the underlying technology for the production of the Home Guard's output has taken place. Simar and Wilson (1999) introduced the bootstrapping of Malmquist indices to allow researchers to speak in terms of whether changes in productivity are significant in a statistical sense. The productivity development for the units is set out in figure 6 and 7. Each unit is represented by a box, where the width of the box represents the relative size of the unit and the height of the box represents a 95 % confidence interval estimated by the bootstrap technique. The boxes are sorted by ascending relative size. With only three years of observations it is difficult to interpret trends. However we cannot reject a hypothesis of productivity change between two subsequent periods for all eleven units except for one unit in one case, the box on the red line in panel 2 of figure 6.



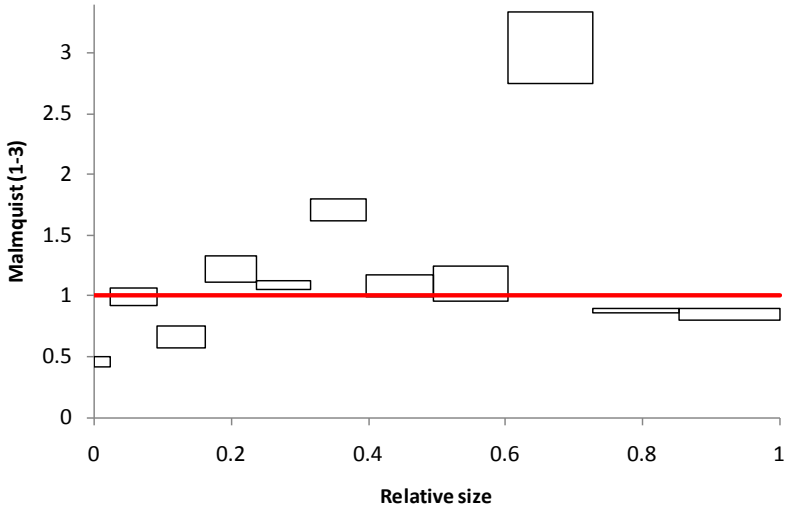
Panel 1. Year 1 – year 2



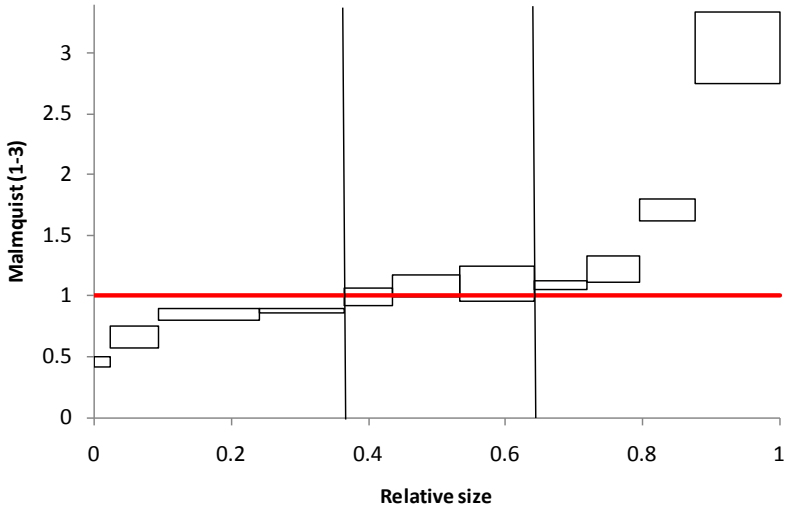
Panel 2. Year 2 – year 3

Figure 6. Significance testing of productivity change. Height of boxes: 95 % confidence intervals for the Malmquist productivity index. Width of boxes: relative unit size. Boxes sorted by relative size.

Only four units had significant changes in the same direction in both periods; two units with significant increase and two units with significant decrease. The changes in productivity from the first to the last period are outlined in panel 1 of figure 4.5. When considering changes from the first to the last period, both the relative small and big units seem to have a significant decrease in productivity. The midsized units have either no significant change or significantly increased productivity. In panel 2 of figure 4.5 the units are divided into three sectors, units with significant decrease in productivity, units with insignificant productivity



Panel 1. Year 1 – 3. Boxes sorted by relative size.



Panel 2. Year 1 – 3. Boxes sorted in three sectors: significant decrease, insignificant change and significant increase.

Figure 7. Significance testing of productivity change. Height of boxes: 95 % confidence intervals for the Malmquist productivity index. Width of boxes: relative unit size.

change and units with significant increase in productivity[13]. The units in each group are sorted, respectively, by ascending values of the upper limit of the confidence interval, ascending values of the mid value of the confidence interval and ascending values of the lower limit of the confidence interval. Four units have a decrease, three units experience no change and four units have a significant increase in productivity. The average unweighted change is a growth of 13.5 %.

[13] In Førsund et al. (2009) this kind of diagram is named *Edvardsen significance diagram*.

Compared to the confidence intervals for efficiency scores in both the CRS and VRS model, the width of the intervals for the Malmquist scores are relatively narrow. Going from the VRS model to the pooled CRS model and the Malmquist index, threefolding the number of observations, increases the accuracy of the estimates. The CRS assumption, also behind the estimates of the Malmquist index, limits the number of observations on the frontier, increasing the accuracy further. Comparing the intervals of the CRS model above to the Malmquist scores, the relatively narrow interval of the Malmquist scores could be explained by the difference in overall bias between the measures. Since the Malmquist indices are defined as ratios of distance functions, the overall bias of the Malmquist indices may be somewhat less than for individual distance function estimates, as the terms in both the numerator and denominator are biased in the same direction (Simar and Wilson, 1999).

5. Conclusions

We have developed a model which makes it possible to analyze the productivity and efficiency by DEA also for the operational units of the armed forces. By aggregating activity standards and quality measures the model enables a meaningful and measurable expression for the output of an operational unit.

The sample consisting of observations from only eleven units of the Norwegian Home Guard puts some limits on the model and the interpretation of results. From the original estimates around 40 % of the units are efficient, leaving the Home Guard without clear candidates for best practice. We have overcome this problem by pointing at the uncertainty concerning the estimates, eliminating some candidates for best practice. The uncertainty of the results is found from resampling the original estimates using the bootstrap technique, giving us confidence intervals and bias-corrected efficiency scores.

The mean square error increases for a majority of the corrected estimates. However, a confidence interval centered at the original estimates leaves us with upper limits of the intervals above one for a few units. This puts some limits on the interpretations, such as the possibility of testing hypothesis. We therefore suggest the results to be presented with emphasis on the standard error of the estimates. By evaluating the standard error of the estimates we can reduce the number of best practice candidates each year from around five,

to one or two units. In the pooled sample we were able to reduce the candidates to a single unit.

It is our impression that small samples not only occur by chance in some sectors, but rather is a characteristic of some parts of the public sector. In order to study a wide variation of public sector activities is it relevant to look further into the problems of small samples. We believe that a continued emphasis on methods which enables a statistical assessment of the uncertainty of efficiency estimates is important. Therefore, further contributions on the estimation of confidence intervals are needed, to avoid limited interpretation due to the restricted bounds of the intervals.

References

Ali, A. I., Charnes, A., Cooper, W. W., Divine, D., Klopp, G. A. and Stutz, J. (1989): "*An Application of Data Envelopment Analysis to Management of US Army Recruitment Districts*", Applications of Management Science, A Research Annual, in R. L. Schultz (Ed.)

Berg, S. A., Førsund, F. R. and Jansen, E. (1992): "*Malmquist Indices of Productivity Growth During the Deregulation of Norwegian Banking, 1980 – 1999*", Scandinavian Journal of Economics, 94, 211-228.

Bowlin, W. F. (1987): "*Evaluating the Efficiency of US Air Force Real-Property Maintenance Activities*", Journal Of The Operational Research Society, (GB) 38(2), 127-35.

Bowlin, W. F. (1989): "*An Intertemporal Assessment of the Efficiency of Air Force Accounting and Finance Offices*", Research in Government and Nonprofit Accounting, 5, 293-310.

Bowlin, W. F. (2004): "*Financial analysis of civil reserve air fleet participants using data envelopment analysis*", European Journal of Operational Research 154, 691-709.

Brockett, P.L., Cooper, W. W., Kumbhakar, S. C., Kwinn, M. J. and McCarty, D. (2004): "*Alternative Statistical Regression Studies of the Effects of Joint and Service Specific Advertising on Military Recruitment*", Journal of The Operational Research Society 55 (10), 1039-1048.

Caves, D. W., Christensen, L. R., and Diewert, W. E. (1982): "*The Economic Theory of Index Numbers and the Measurement of Input, Output and Productivity*", Econometrica, 50, 1393-1414.

Charnes, A., Clark, C.T., Cooper, W. W., and Golany, B. (1985): "*A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the U.S. Air Forces*", Annals of Operation Research 2 (1), 95-112

Charnes, A., Cooper, W. W., Lewin, A. Y., and Seiford, L. M. (1994): "*Data Envelopment Analysis: Theory, Methodology and Applications*", Kluwer Academics publishers.

Charnes, A., Cooper, W. W., Divine, D., Klopp, G. A. and Stutz, J (1982): "*An Application of Data Envelopment Analysis Recruitment Districts*", CCS 436, Center for Cybernetic Studies, The University of Texas at Austin

Clarke, R. L. (1992): "*Evaluating USAF Vehicle Maintenance Productivity Over Time: An Application of Data Envelopment Analysis*", Decision Sciences 23 (2), 376.

Cooper, W. W., Seiford, L. M. and Tone, K. (2007): "*Data Envelopment Analysis: A Comprehensive Text with Models, Application, References and DEA-Solver Software*", Springer Science+Business Media, LCC, New York.

Efron, B. and Tibshirani, R.J. (1993): "An introduction to the Bootstrap", Chapman and Hall, Inc., New York.

Eurostat (2001): "*Handbook on price and volume measures in national accounts*", Luxembourg: European Communities.

Farrell, M. J. (1957): "The measurement of productive efficiency", Journal of the Royal Statistical Society A 120, 253-281.

Farris, J. A., Groesbeck, R. L., Van Aken, E. M. and Letens, G (2006): "*Evaluating the Relative Performance of Engineering Design Projects: A Case Study Using Data Envelopment Analysis*", IEEE Transactions On Engineering Management 53 (3).

Fried, H.O., Lovell, C. A. K., and Schmidt, S. S. (2008): "*The Measurement of Productive Efficiency and Productivity Growth*", Oxford University Press.

Førsund, F. R. (2010): "*Dynamic Efficiency Measurement*", Indian Economic Review 45 (2), 125-159.

Førsund, F. R. (2011): "*Measuring efficiency in the public sector*", Unpublished working paper, FFI

Førsund, F. R., Edvardsen, D. F., Kittelsen, S. A. C. and Lindseth, F. (2009): "*Productivity of Tax Offices in Norway*", Memorandum No 14/2009, Department of Economics, University of Oslo.

Grilfell-Tatjé, E. and Lovell, C. A. K. (1995): "A note on the Malmquist productivity index", Economics Letters 47, 169-175.

Kittelsen, S. A. C. (1999): "*Monte Carlo Simulations of DEA Efficiency Measures and Hypothesis Tests*", Memorandum No. 09/99, Department of Economics, University of Oslo.

Lewin, A. Y. and Morey, R. C. (1981): "*Measuring the relative efficiency and output potential of public sector organizations: An application of data envelopment analysis*", International Journal of Policy Analysis and Information Systems 5 (4).

Lu, W. M. (2011): "*Benchmarking management in military organizations: A non-parametric frontier approach*", African Journal of Business Management 5(3), 915-923.

Norwegian defence Fact and Figures (2010). Oslo: Norwegian Ministry of Defence.

Ozcan, Y. A. and Bannick, R. R. (1994): "*Trends in Department-of-Defense Hospital Efficiency*", Journal of Medical Systems 18(2), 69-83.

Pedraja-Chaparro, F., Salinas-Jimenez, J. and Smith, P. (1999): "*On the Quality of the Data Envelopment Analysis*" Journal of the Operational Research Society, 50, 636–644.

Roll, Y., Golany, B. and Seroussy, D. (1989): "*Measuring the Efficiency of Maintenance Units in the Israeli Air Force*", European Journal of Operational Research Volume 43, Issue 2, 27 November, 136-142.

Schreyer, P. (2010): "*Measuring the Production of Non-Market Services*", in *Price Indexes in Time and Spice. Contribution to Statistics,* L. Biggeri, G. Ferrari (eds.), Springer-Verlag Berlin Heidelberg 2010.

Shephard, R.W. (1970): "*Theory of Cost and Production Function*", Princeton: Princeton University Press.

Silverman, B.W. (1986): "*Density Estimation for Statistics and Data Analysis*", Monographs on Statistics and Applied Probability, London: Chapman and Hall.

Simar, L. and Wilson, P. W. (1998): "*Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models*", Mangement Science 44 (1).

Simar, L. and Wilson, P. W. (1999): "*Theory and Methodology: Estimating and bootstrapping Malmquist indices*", European Journal of Operational Research 115, 459-471

Simar, L. and Wilson, P. W. (2008): "*Statistical Inference in Nonparametric Frontier Models: Recent Developments and Perspectives*", in *The Measurement of Productive Efficiency and Productivity Growth,* H. O. Fried, C. A. Knox Lovell, and S. S. Schmidt (eds.), Oxford University Press.

Sun, S. (2004): "*Assessing joint maintenance shops in the Taiwanese Army using data envelopment analysis*", Journal of Operations Management 22, 233-245.

UK Army Doctrine Publication (2010): "*Operations*", www.mod.uk/dcdc