Efficiency and productivity in the operational units of the armed forces: A Norwegian example[☆]

Torbjørn Hanson[a,1]

[a]*Norwegian Defence Research Establishment (FFI), Instituttveien 20, P O Box 25, NO-2027 Kjeller*

Abstract

Most nations spend a considerable part of their gross domestic product (GDP) on defense. However, no previous study has addressed productivity and efficiency in 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 nearly half of the units being estimated as fully efficient. We find that, by using the bootstrap technique to estimate confidence intervals, we can point to uncertainty in the estimates and reduce the number of candidates for best practice.

Keywords: Military, Productivity, Efficiency, DEA, Bootstrap
JEL: D24, H40, C60

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 alternative cost set by markets.[2] The fact

[☆]The paper is written within the research program Kostnadseffektiv drift av Forsvaret (Cost Efficiency in Defence) at the Norwegian Defence Research Establishment (FFI).

Email address: `torbjorn.hanson@ffi.no` (Torbjørn Hanson)

[1]I am grateful to Finn R. Førsund, Sverre A. C. Kittelsen, Tore Nyhamar, Espen Berg-Knutsen, Hanne Hanson, Sverre Kvalvik and three anonymous referees for their comments on this paper.

[2]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, but still have opportunity cost.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

that services are not sold in markets leaves the armed forces without information from a price mechanism in evaluating efficient use of resources or effective mix of services. Despite the absence of price information on services, 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.[3] Previous studies of efficiency and productivity by DEA in the armed forces have solely been concentrated on various support functions like maintenance and recruitment, reviewed in Section 2 of the present paper. However, operational units, the core area of defense, have not been studied in the literature. In fact, the field has not developed towards military operational applications since the pioneering work by Lewin and Morey (1981) and Charnes et al. (1984). The purpose and main contribution of this paper is to show that studies of efficiency and productivity by DEA can be carried out also for operational units of the armed forces.

There are at least three possible reasons for the lack of studies: Difficulties in modeling and measuring output in the military; heterogeneity leading to small populations of military units; and restricted data on performance of operational units. The most important, but perhaps also the easiest problem to overcome, is difficulties in modeling the production process and output of the armed forces. Hartley (2010, 2012) implicitly defines military output as aircraft squadrons, submarine or tank forces by acknowledging that defense markets have no market prices for their outputs, referring to the lack of prices on such forces. Furthermore, Hartley acknowledges that few published studies have estimated military productions functions, and those which have are using a cost-effectiveness approach. In this manner the present paper represents a methodological contribution by its pioneer work on output measures for operational units 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. In our model the emphasis is on production of troops and soldiers, and not outcomes of operations. In this way education and training activities are essential, resulting in the necessary proficiency levels (quality) for combat readiness being met for soldiers and units. A simple measure of number of soldiers, e.g. reported in The Military Balance, is of no value without adjusting for quality as measured by e.g. training level and available equipment.[4] The

[3]DEA is a non-parametric method for the estimation of production frontiers by a piecewise linear surface enveloping the observations from above in the standard case. 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).

[4]The Military Balance is an annual assessment of global military capabilities including number of personnel and equipment for 171 different countries, published by the International Institute for Strategic Studies. Publicly available combat readiness and quality of forces is of

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

model is specified for units of one branch or service of the Norwegian Armed Forces, the Home Guard.[5] This is where the largest number of homogeneous units are found in Norway, and where we are given access to a quality index. Specified for land force units the model is, however, straight forward to apply also for navy and air force units.

The second reason for the lack of studies on operational units is the low number of observations generated. This is due to limited demand for homogeneous units in most armed forces. Very few nations have the need nor the resources necessary for large scale duplication of military units. Studies of efficiency in operational units is therefore limited to small populations of homogeneous units. Studies of productivity and efficiency are of interest to the armed forces for identifying potential benchmarks. However, low number of units in most studies limits interpretation of the results, as on average nearly half of the units appear fully efficient. From a review of the literature on DEA in the military we do have reasons to believe that small samples are a common phenomenon for studies of the sector. This is the case also in our empirical example of all the eleven operational Norwegian Home Guard districts during the period 2008–2011. Panel data from four successive years gives us the opportunity to pool the data and increase the number of observations to 44. In order to reduce the number of units estimated as fully efficient and thereby reducing the number of potential benchmarks, the estimation could also be supplemented by other methods.

Introducing state of the art methods (bootstrapping) enables a statistical interpretation of results and construction of confidence intervals around the estimates. Additional information provided by 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 armed forces. Further, confidence intervals for the Malmquist index let us consider also the significance of changes in productivity. Resampling of efficiency scores and the Malmquist index is done by the bootstrap procedure developed in Simar and Wilson (1999). The convergence rate of the DEA estimator is sensitive to sample size and dimensionality (Simar and Wilson, 2000). The literature provides no rules of thumb for when the DEA-bootstrap is justified, but experiments in Simar and Wilson (2000) indicates that a sample size of $n = 10$ is perhaps too small to obtain meaningful results in applied studies. However, increasing the sample size even from $n = 10$ to $n = 25$ more than

course not found in such assessments.

[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 Ministry of Defence, 2010). The Home Guard consists of one Home Guard Staff, two school departments, and a number of operational districts located in all geographical regions in Norway. The personnel in a Home Guard district are mostly conscripted personnel with a full time job outside the military (300 to 1000 officers and 1500 to 4500 soldiers), except for the personnel in the District Staff who are full time employed in the Armed Forces (around 50 people).

DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

halves the range of confidence intervals in the experiment. These findings lead us to focus only on the pooled sample of size $n = 44$ when bootstrapping efficiency scores for the Home Guard. We argue that it is better to report standard errors for the DEA estimates rather than completely ignoring any uncertainty in the estimates.

Including several nations in the study could extend available data dramatically. However, data sources on performance of operational units are usually restricted and possibilities of collecting an extended data set consisting of data from several nations are few.[6] Additionally, heterogeneity in training standards and requirements between nations makes a wider study even more difficult. Restrictions on data and heterogeneity between nations represent a third possible reason for lack of studies. We have benefitted from close cooperation with the Armed Forces in access to data on military operational units. The scope of the paper is, however, limited to measuring the performance of operational military units and thereby offering a new tool for managers. Our results point to best practice candidates among the units, but explanations for the mechanisms behind any estimated differences in performance is outside the scope of the present paper.

The paper is structured as follows. In Section 2 a discussion on previous DEA studies 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 operational units in the Home Guard, a branch of the Norwegian Armed Forces. Estimates from the 44 observations are presented in Section 4 of the paper, before we introduce the bootstrap procedure resampling the data, and additional pseudo observations are generated. 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. Data Envelopment Analysis in the military

Application of DEA in the defense sector started out in the early eighties with the studies by Lewin and Morey (1981) and Charnes et al. (1984) on recruitment and aircraft maintenance units respectively. While the military was introduced as a new and promising field for DEA studies at the time, it is somewhat striking that application to other areas and activities in the armed forces is still absent 30 years later. In fact, the approach in Charnes et al. (1984) constitutes the best example of measuring efficiency in operational units, the core activity of the armed forces, in the literature. An overview of DEA studies in the military, including the field of study and number of variables, is outlined

[6] However, Owen (1994) shows that international benchmarking of military manpower ratios is possible across 15 different countries.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Table 1: Bibliography of DEA in the military

Paper	Field and country	Inputs	Outputs	Observations
Lewin and Morey (1981)	Recruitment (USA)	10	2	43
Charnes et al. (1984)	Maintenance (USA)	8	4	42
Bowlin (1987)	Maintenance (USA)	3	4	21
Bowlin (1989)	Accounting and Finance (USA)	1	5	18
Ali et al. (1989)*	Recruitment (USA)	n/a	n/a	n/a
Roll et al. (1989)	Maintenance (Israel)	3	2	10-35
Clarke (1992)	Maintenance (USA)	4	2	17
Ozcan and Bannick (1994)	Hospitals (USA)	6	2	23-124
Bowlin (2004)	Civil reserve air fleet (USA)	4	7	37-111
Brockett et al. (2004)	Recruitment (USA)	1	10	n/a
Sun (2004)	Maintenance (Taiwan)	6	5	30
Farris et al. (2006)	Engineering design projects (Belgium)	4	1	15
Lu (2011)	Outlets (Taiwan)	4	2	31

*Paper not available online, nor through the author's library

in Table 1. Charnes et al. (1984) study 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 aircraft, hours of non capable aircraft due to maintenance problems, number of sorties flown and the number of completed jobs of a specific type. By introducing hours of mission capable aircraft and number of sorties flown as outputs possible measures of operational outputs are included in the model. This is in contrast to solely measuring the number of completed maintenance jobs of various types. However, for the measure to fully cover the operational unit, in form of a squadron of aircraft in this case, at least some measure of personnel (e.g. pilots) has to be introduced.

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. 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, related to a subjective scale based on judgment from the expert team.[7] The final model included, thus, two outputs: flying hours weighted by type of aircraft and the standard deviation of the daily number of sorties. Expert opinion is used also in the present paper in determining indicators for troop production. In the other three studies of military maintenance found in the literature operational outputs are limited to the average number of days during a month assigned vehicles are in serviceable condition (Clarke, 1992; Sun, 2004), while regular maintenance outputs in form of number of job orders are the only outputs in Bowlin (1987).

Another issue found in the defense sector is the small samples of military units to be studied. Small samples result in a low number of observations and low degrees of freedom in the models.

[7]The type of aircraft was introduced as a weighting factor of flying hours.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Operational units are likely to be more heterogenous between branches and services compared to units concerned with mere support functions. This is also supported by our findings in the literature review, e.g. hospitals are spread across the service components army, air force and navy (Ozcan and Bannick, 1994). In the studies of maintenance units in Charnes et al. (1984), Roll et al. (1989) and Bowlin (1987) window analysis is used to increase the number of observations.[8] In the Charnes et al. (1984) study w is set to three months for the $n = 14$ maintenance units. This procedure increases the number of observations to 42 (3x14) in each window. Correspondingly, Roll et al. (1989) run DEA for as few as five maintenance units. However, by windows of six time periods the number of observations is increased to 30 in each window. Despite efforts to increase the number of observations in past studies we found that on average 44 % of the units were estimated as fully efficient. Compared to previous studies our data on eleven units over four years seems promising. The total of 44 observations is more than any previous study covering some high end aspects of the armed forces.

Other applications of DEA related to the defense sector include recruitment (Lewin and Morey, 1981; Ali et al., 1989; Brockett et al., 2004), accounting and finance offices (Bowlin, 1989), defense hospitals (Ozcan and Bannick, 1994), engineering design projects (Farris et al., 2006), financial performance of airlines participating in the US Department of Defense's civil reserve air fleet (Bowlin, 2004) and provision of foods and products by military outlets (Lu, 2011). These studies have all in common the lack of operational-like outputs. The production is conducted within a military setting but outputs like project duration, inpatient/outpatient days, contracts signed and number of transactions are not at all unique to military activity.

3. Concepts and Data

At present there is no concise use of terms describing outputs and outcomes in the military, and we need to start with a proper definition of measures in order to refine our study to the concept of efficiency and not effectiveness. In general, when considering public service activities we can distinguish between two aspects, as described in Førsund (2013) 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

[8]The window analysis technique was first employed in Charnes et al. (1984). The technique is described in Charnes et al. (1994) as a moving average analogue. In each window of w periods a DMU is treated as if it where a different (artificial) DMU for all successive periods of the window. From this procedure the number of observation in a DEA is given by $n \cdot w$. It remains, however, unclear how artificial units can make any real improvements in DEA.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

"doing the right things". In the following we will discuss whether this model is straightforward 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 and sovereignty 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 or overall sovereignty. The main issue when applying the model to the defense sector is the distinction between and categorization of: (1) Outcomes; (2) Outputs; (3) Activities in the transformation process from inputs to outputs. In the following we discuss the concepts in light of applications in a military context and develop a model for operational output in the armed forces.[9]

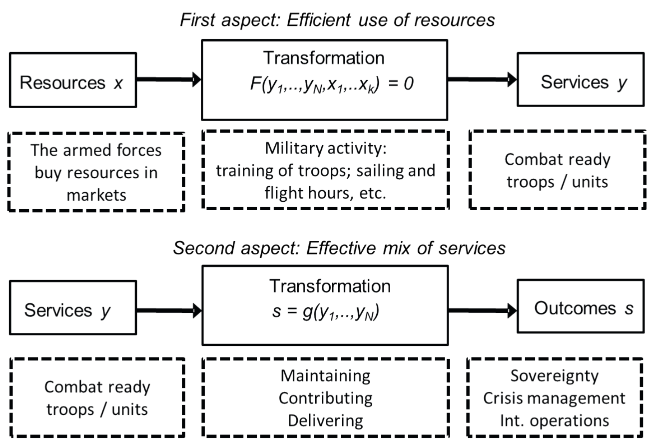

Figure 1: The two aspects of public service activities, and a specification for the armed forces

3.1. Output measures in the military

Traditionally, the national accounts approach of defining outputs equal to inputs in cost terms was used also for the defense sector.[10] However, there is an expanding literature on the concept of defense output and outcome. Hartley (2012) refers to defense outputs as a complex set of variables concerned with security, protection, risk management, including risks and conflicts avoided, safety, peace and stability. At the same time, defense outputs are also referred to as aircraft squadrons, submarine or tank forces. This shows that the use of

[9]However, transformation of outputs to outcomes is not covered in detail in the paper. Here, we confine by noting that trained and combat ready operational units serve to maintain sovereignty, contribute to crisis management and deliver troops to international operations.

[10]Without prices for the output, there are only two options for constant price measurement in national accounts: deflating inputs and direct volume measurement (Eurostat, 2001). Constant price measurement implies assuming that the change in the volume of inputs is representative for the change in the volume of output.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

the concepts output and outcome is somewhat diverse.

The UK Ministry of defence has established a system for defining and measuring readiness for its Armed forces (Hartley, 2012). Readiness is a concept which could define, at least partly, the output of a military unit. Anagboso and Spence (2009) use the term *high level outcome* for peace and security, and point out that this level of outcome is difficult to measure. However, they suggest a number of intermediate steps between inputs and outcomes which could be utilized in output measurements. The steps include activities which measure specific things the armed forces do, and capabilities of the armed forces. In this setting, a capability is the ability of the forces to pursue a particular course of action. Anagboso and Spence (2009) find a capability approach more promising for measuring defense output, and they consider defense output to be the sum of the capabilities the armed forces provide. Two possible measures of capability are identified: a manpower measure, and an equipment measure. Both measures would have both a quality and a quantity component. Suggested quality adjustments for manpower are rank, grade, manning balances including training strength, and manning pinch points. For equipment, an explicit quality adjustment taking into account quality changes over time is suggested as well as a readiness measure. This line of reasoning corresponds with the more general method suggested for the public sector in Schreyer (2010); in the national accounts literature outputs are broken down into the two components activities/processes and quality. Further, outcomes are divided into either direct or indirect outcomes. The direct outcome is closer to the production process, such as the state of knowledge of students, while the indirect outcome is associated with for example higher earnings resulting from higher human capital.

We suggest a model for the output of operational units extending the methods in Anagboso and Spence (2009) and Schreyer (2010). UK Army Doctrine defines in UK Ministry of Defence (2010, p. 89) a military unit as "the smallest grouping capable of independent operations with organic capability over long periods".[11] Further, a unit is divided into sub-units i.e. troops. Such specialized sub-elements have to be trained both individually and together as a unit. In contrast to Anagboso and Spence (2009) we define output as sub-elements of the unit, like the number of a specific troop or force type, and not a given capability. In this manner, capabilities are referred to as possible intermediate outcomes realized by a certain combination of trained and combat ready troops rather than the direct output of a unit. The approach stresses the relevant peace time outputs of education and training, but not outcomes from operations.

An operational unit j, $j = 1, ..J$, is producing the output y^i, $i = 1, .., I$. For each effective output y^i there exist an output measure y_i operationalizing the output. The measure is modeled as a function of quantity l and quality P, e.g.

[11] Units typically comprise between 400 and 1000 people. In the marines, a unit is typically called a commando, in the navy units are warships, while the optimal grouping in the air force is a large squadron (UK Ministry of Defence, 2010). In the present paper output is specified for army or land force units. It's, however, straight forward to specify the model also for navy and air force units.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

a reinforcement troop where l is the number of soldiers in the troop and P their proficiency level

$$y^{i,j} = y_{i,j}(l_{i,j}, P_{i,j}) \tag{1}$$

Introducing a quantity and quality component is in line with suggestions for a measure in Anagboso and Spence (2009), however, here on a more disaggregated level. The idea of adjusting for quality in the output measure is not at all unique for the military. The importance of adjustments for quality variables in output measures is recognized in the Spady and Friedlaender (1978) seminal paper on hedonic cost functions. Instead of treating specific quality levels as separate goods, it is suggested to treat effective output as a function of a generic measure of physical output and its qualities.

Possible constructions of quality and quantity indices are numerous in the military. When established, measures are likely to vary among types of units, branches and nations. Typically, the quantity measure will consist of an index for essential personnel and equipment, while the quality measures consists of a measure for proficiency level. In military literature the concept of force multiplier effects is frequently used. The concept is a simple recognition that there are positive interaction effects among factors used in a military production function (Hurley, 2005). For example, considering a military production function consisting of the two variables soldiers and guns, the two factors are complementary. Any combination of soldiers and guns provides more force than only soldiers or only guns. We suggest that the same reasoning applies for a production function where quality is included; soldiers and equipment provides no force without a certain level of quality, i.e. a given training level. In Section 3.2 we will return to a an example of a fully specification of the output measure.

Making the link between outputs and capabilities simpler, one possible approach is to aggregate the outputs on a unit level, constructing an expression for capability production (intermediate outcome production function) as a function of a single unit output. When considering a capability function the concept of force multiplier effects is again important. Given an operational unit consisting of three different troop types: a command, a gun troop and an observation troop, the ability to eliminate enemy ground forces is dependent on the presence and performance of all troop types. Therefore, a multiplicative functional form is suggested if aggregation of troops is pursued, e.g. in order to construct a capability (intermediate outcome) measure.[12]

Besides some intended properties in the modeling of a single unit output, aggregation limits the dimensionality in a 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. Facing a small population of military units, the above discussion suggests consider-

[12]In an accompanying paper we suggest to use data on marginal costs in estimation of troop weights as coefficients in specification of a function. An alternative approach is to use military expert opinion in determining the weights, as pointed out in our review of previous studies in the military.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

ing aggregation as a possible solution. 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, hence contribution to capability as the intermediate outcome in the case of military. As the scope of the present paper is limited to studies of efficiency, and not effectiveness, we leave the use of such aggregated measures for further research.

3.2. Output specification for the Home Guard operational units

In the following we specify a model for the output of an operational unit in the Norwegian Home Guard based on equation (1). The main objectives for the Norwegian Home Guard are to protect the local population and 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 Home Guard 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. Both officers and soldiers are conscripted. However, the personnel in the District Staff are regulars employed full time in the Armed Forces. District Staff personnel, 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. 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. Objectives of the Home Guard have characteristics of outcomes rather than outputs, where the main capability is protection of local infrastructure. From tasks relevant to a single district, we have defined the dimensioning production to be certain types of troops at various proficiency levels. Three different outputs y^i, $i = 1, 2, 3$, are produced by each decision making unit (DMU): Rapid Reaction unit troops, Reinforcement unit troops, and District Staff. Troop size can vary among troops of the same type as well as the number of each troop type, and therefore also unit size. The measure of troop production is modeled from various indicators registered at district level and reported to the Home Guard Staff. Which indicators to use in defining troop production level is based on expert opinion from personnel at the districts and the Home Guard Staff. Using equation (1), the output measure y_i is a function of a personnel index l and the quality aspect represented by proficiency level index P, common for all $j = 1, ..., 11$ Home guard districts

$$y^i = y_i(l_i, P_i) = l_i P_i, \quad i = 1, 2, 3 \tag{2}$$

The proficiency index is constructed from the three levels of proficiency reported, where each level is met after passing given standards. Fully operational capacity

DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

is the highest level. The measure of personnel represents the size of the troop (number of personnel), constructed by a weighted sum of officers and soldiers. Weights are derived from data on marginal costs of training activities for officers and soldiers respectively. In absence of any empirically founded specification we let l and P enter multiplicatively. This is the standard specification in human capital and economic growth literature for modeling quality, knowledge or effectiveness of labor.[13]

A measure for equipment is left out due to its inconsistency over the period studied. Each district reports on equipment readiness. But the measure seems to be interpreted differently among the districts. However, the possible impact from the measure on unit output is minor in this case of less complex equipment and with levels proportional to the personnel measure.[14] This leaves us a straightforward interpretation of the troop output measure as the number of soldier equivalent quality adjusted personnel produced during a given period of time. As long as all Home Guard districts are given similar types of tasks, producing the three categories of outputs (troop categories) and using the same type of equipment, it is reasonable to assume the districts to be homogeneous in line with the assumptions for DEA.

3.3. Data and specification of the DEA model

The sample consists of yearly observations from eleven Home Guard districts over the years 2008–2011. No new equipment or operating conditions suggesting changes in technology are reported during the period studied, and it is reasonable to assume the technology to be stationary. This gives us 11 observations each year or 44 observations after pooling the data. As noted the data represents the largest pool of homogeneous units found in the Norwegian Armed Forces. Output data is collected from monthly and yearly district reports to the Home Guard Staff. The districts are similar regarding tasks and troop types, but different in size of personnel and geographical area. It was stated by the Home Guard staff that differences in geographical area may have some cost implications such as higher travel and transport expenses. This led to the hypothesis that units located in populated areas would outperform units located in rural areas due to their further access to available personnel, public logistics and shorter travels for personnel meeting for training. Furthermore, one unit (DMU-11) was restructured at the beginning of the period and expected to underperform during a transition period.

Available input data for the production process in the Home Guard is mainly cost data from the Home Guard district accounts. These include three categories of costs: Fixed personnel cost, such as ordinary wages; variable personnel cost, such as activity based payments, overtime pay and travel expenses; equipment

[13] See e.g. Romer (2006) referring to the product of knowledge and labor as *effective labor*.

[14] Main equipment for Home Guard personnel are assault rifles and basic infantry gear. Some troops have additional recoilles rifles (Carl Gustav) and 12.7 mm machine guns. Vehicles include lorries and light terrain vehicles. Equipment and vintages differ between troop types but not districts.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Table 2: Descriptive statistics for the initial three input and three output variables. Eleven Home Guard districts (DMUs) and 44 observations

Variable	Min	Max	Mean	Median	SD
x_1 Equipment cost	7.64	30.51	15.79	14.02	4.88
x_2 Fixed Personnel cost	10.86	39.24	24.52	23.34	6.86
x_3 Variable Personnel cost	8.83	24.71	16.63	15.83	4.66
y_1 District Staff	0.27	1.84	1	0.94	0.38
y_2 Rapid Reaction troops	0.37	2.17	1	0.92	0.44
y_3 Reinforcement troops	0.12	1.82	1	0.96	0.40

Note: Input data in mill. NOK. Output data normalized to mean values in order to maintain anonymity

cost, such as ammunition, spare parts and maintenance. Due to the lack of accrual accounts, activity-based troop specific expenditures are not perfectly correlated with 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. The measures included in the input variable are adjusted using the consumer price index. Data for the construction of output measures consist of data on the number of different personnel and proficiency levels. Proficiency levels P_i together with data on personnel are reported to the Home Guard Staff on a monthly basis. The indices are constructed from yearly averages.

This gives us in total six potential variables for the DEA model: The three inputs fixed personnel cost, variable personnel cost and equipment cost, while the outputs are District Staff, Rapid Reaction troops and Reinforcement troops expressed by equation (2). Descriptive statistics are found in Table 2. Output variables vary over years and Home Guard districts. Least variation is found in the district staff output y_1 as expected. Most of the variation in this variable results from an increasing trend over the four years. However, for the troop outputs there is a clear distinction between districts, reflecting the differences in the number of personnel in each district. The five highest y_2 values are found among the two districts with the highest number of personnel. Correspondingly are four out of five of the lowest values found among small districts measured by the number of personnel. The same pattern is found for the input variable fixed personnel cost x_2. But the equipment cost x_1 and variable personnel cost x_3 varies over both districts and years.

Model specification should be done with emphasis on degrees of freedom as the sample is limited to 44 observations when pooled. This is about the same number of observations found in Lewin and Morey (1981) and Charnes et al. (1984), 43 and 42 respectively. The rate of convergence for the DEA estimator depends on the sample size n as well as the number of variables in the model. Kneip et al. (2008) derived the rate of convergence for the DEA estimator to $n^{2/p+q+1}$ for the variable returns to scale case, and Park et al. (2010) the rate to

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

$n^{2/p+q}$ in the case of constant return to scale. Here, p and q refer to the number of input and output variables, respectively. This leads to the saying that the DEA estimator suffers from the "curse of dimensionality".

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).[15] 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. Further, we found that on average 44 % of the units were estimated as fully efficient in each of the studies. The same result is obtained for our data by cross sectional studies in a model of either three inputs (outputs) and one output (input).

Considering the "curse of dimensionality" and the high share of fully efficient units in previous studies we suggest emphasis on a more simple model capable of discriminating between units. Furthermore, we limit our studies to the pooled sample only, as in cross sectional studies even a model of four variables is found insufficient by the rule of thumb from Cooper et al. (2007). The number of observations in cross sectional studies is also too small for inference after bootstrapping efficiency scores, as we will return to in Section 4.

On the other hand 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 sometimes of fundamental importance. A positive correlation between two inputs might give less information to the analysis than if the inputs were uncorrelated. This is in particular the case if variables are complementary. But correlation could also appear from i.e. a certain input mix based on a manager's knowledge alone. One implication of this result is that the adequacy of a DEA model to some extent is an empirical question.[16]

Correlation between variables in our model is reported in Table 3. Low correlation (0.04) between District Staff output and variable personnel cost stands out. But this is simply due to the fact that personnel at the District Staff does not fully attend training activities. The highest correlation coefficient among outputs (0.58) is found between the two variables Reinforcement troops and Rapid Reaction troops. But the highest correlation is found among input variables. Increasing the degrees of freedom in the model we therefore aggregate the input side to a single variable x_4 measuring total cost. Aggregating the input side is, obviously, straight forward as each input variable is measured in monetary terms. The alternative of aggregating output variables involves the

[15]We have, however, not succeeded in finding neither any theoretical nor empirical evidence for the rules of thumb.

[16]Kittelsen (1999) shows that the extent of correlation is clearly important when testing the relevance of an additional input in the model.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Table 3: Correlation coefficients for all input and output variables. Input variable x_4 is the sum of x_1, x_2 and x_3. Eleven Home Guard districts (DMUs) and 44 observations

Variable	x_1	x_2	x_3	x_4	y_1	y_2	y_3
x_1 Equipment cost	1						
x_2 Fixed Personnel cost	0.67	1					
x_3 Variable Personnel cost	0.68	0.86	1				
x_4 Total cost	0.85	0.95	0.93	1			
y_1 District Staff	0.14	0.12	0.04	0.11	1		
y_2 Rapid Reaction troops	0.55	0.72	0.72	0.73	0.17	1	
y_3 Reinforcement troops	0.43	0.58	0.53	0.57	0.45	0.58	1

Table 4: Descriptive statistics for the four variables included in the DEA model, 44 observations

Variable	Min	Max	Mean	Median	SD
x_4 Total cost	27.33	89.34	56.94	54.94	14.95
y_1 District Staff	0.27	1.84	1	0.94	0.38
y_2 Rapid Reaction troops	0.37	2.17	1	0.92	0.44
y_3 Reinforcement troops	0.12	1.82	1	0.96	0.40

Note: Input data in mill. NOK. Output data normalized to mean values in order to maintain anonymity

more complex exercise of estimating weights for each troop type, either by cost shares or by expert opinion. Aggregation of input variables has a more straightforward cost function interpretation, and is in general preferred over output aggregation in studies of the public sector when prices on the latter is missing. Descriptive statistics for the model's four variables, including variable x_4, are reported in Table 4. For the total cost variable x_4 there is less variation over districts for high value observations, where two districts represent the six highest observations. Four different districts are represented among the six lowest observation.

4. Results

Using the DEA model including input variable x_4 and output variables y_1, y_2 and y_3 in Table 2 the Farrell technical efficiency scores are estimated for Home Guard units $j = 1, ..., 11$ for the years 2008–2011.[17] Productivity development over the four years we investigate by the Malmquist productivity index.[18] We

[17]The DEA specification is given in the Appendix.

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

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

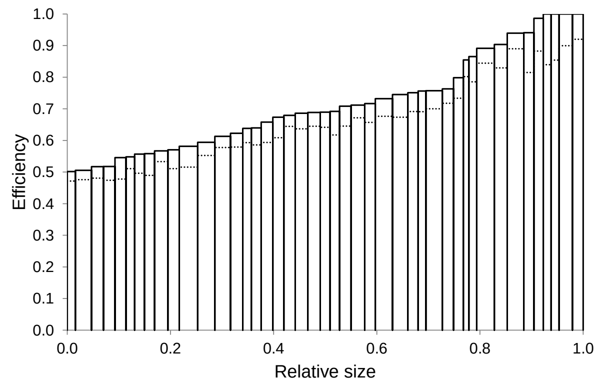

Figure 2: Estimated efficiency scores (bars) and bias corrected scores (dotted lines) for each unit and year. Data for eleven units over four years are pooled to 44 observations (44 bars). Width of bars proportional to unit size measured by the share of total cost

start out by pooling data for the four years increasing the number of observations to 44. Due to moderate number of observations the only meaningful scale specification is constant returns to scale (CRS), and no further test for scale specification is pursued.[19] Four units are estimated fully efficient evaluated relatively to the intertemporal frontier, reported in Figure 2. The width of the bars represents unit size measured by the share of total cost. This leaves us with multiple Home Guard districts as candidates for best practice.

But 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 DEA method is based on enveloping the observations as tightly as possible from above in the standard case. There might, however, 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. The uncertainty of 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. Pointing at the uncertainty of the estimates enables us to further investigate the group of best performing units and eliminate the most uncertain candidates for best practice among the Home Guard districts.

[19]Using both an intertemporal frontier and a CRS assumption is equivalent to using a technical productivity measure with optimal scale as the reference frontier.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

4.1. Resampling and bootstrapping the efficiency scores

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 kernel density estimate (KDE) using reflection (Silverman, 1986) to avoid the accumulation of efficiency values of one. We 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. The bias-corrected efficiency scores are set out by a dotted line inside each bar in Figure 2, where all units get a significant downward shift in efficiency. Bias correction has a considerable impact on two of the four units estimated as fully efficient in the original run (observation number three and four from the right in Figure 2), leaving us with two candidates (DMU-6 and DMU-8) for best practice unit from this illustration of the results.

Simar and Wilson (1999) suggest that whether bias correction should be used is always an empirical question, but without providing any explicit criteria for judging the results. However, mean square error (MSE) is considered when performance of their estimators is compared. Experiments show that inference from bootstrapping is sensitive to model specifications and sample size (Simar and Wilson, 2000). As noted in Section 3 the DEA efficiency estimator converges at the rate $n^{-2/p+q}$ in the case of CRS. Consequently its mean square error (MSE) is vulnerable to dimensionality measured by the number of input p and output q variables in addition to sample size. Simar and Wilson (2000) study the performance of their DEA bootstrap estimator by Monte Carlo simulation and show that small n and dimensionality worsen the performance in form of higher confidence interval range and MSE in bias estimates. In their experiment Simar and Wilson find that $n = 10$ is a very small sample size for DEA bootstrap, perhaps too small to obtain meaningful results in applied studies. Confidence intervals estimated from small n are, due to their excessive widths, not informative beyond drawing attention to the uncertainty in the estimated efficiency score. But Simar and Wilson (2000) find that increasing n to 25 reduces the average confidence intervals by almost half in their CRS specification. The performance of our results is discussed below.

As noted also by Efron and Tibshirani (1994), the bias-corrected estimator may have higher mean square error than the original estimator. We find that the bias-corrected estimator outperforms the original estimator in form of lower mean square error for the most efficient units, which support our finding of two units standing out as best practice candidates when biased-corrected scores are considered. However, in 7 out of 44 observations in Table 5, mean square error in the corrected estimate is higher than the error in the uncorrected estimate. But still the average MSE is slightly higher for uncorrected scores. Despite the lack of clear interpretation concerning performance of the corrected estimates, we still believe it is important to account for uncertainty in the estimator and choose to proceed with corrected estimates in the following.

Table 5: Results from the pooled sample, 44 observations

DMU-Year	Est.	Bias Corr. E.	SE	MSE	MSE Corr. E.	Diff
06-2010	1.000	0.920	0.039	0.008	0.006	-0.002
08-2011	1.000	0.900	0.042	0.012	0.007	-0.005
07-2011	1.000	0.854	0.079	0.027	0.025	-0.003
01-2010	1.000	0.840	0.069	0.031	0.019	-0.011
01-2011	0.986	0.883	0.057	0.014	0.013	-0.001
06-2011	0.941	0.815	0.071	0.021	0.020	-0.001
02-2008	0.939	0.890	0.033	0.004	0.004	0.001
05-2011	0.904	0.830	0.040	0.007	0.007	-0.001
02-2009	0.891	0.844	0.031	0.003	0.004	0.001
09-2011	0.865	0.785	0.043	0.008	0.007	-0.001
11-2008	0.855	0.802	0.031	0.004	0.004	0.000
04-2010	0.798	0.734	0.037	0.006	0.006	0.000
10-2009	0.763	0.717	0.029	0.003	0.003	0.000
02-2010	0.757	0.700	0.035	0.004	0.005	0.000
09-2010	0.756	0.691	0.032	0.005	0.004	-0.001
10-2011	0.751	0.691	0.030	0.004	0.003	-0.001
02-2011	0.745	0.673	0.041	0.007	0.007	0.000
08-2010	0.732	0.677	0.029	0.004	0.003	-0.001
04-2008	0.717	0.657	0.034	0.005	0.005	0.000
06-2009	0.712	0.671	0.025	0.002	0.003	0.000
03-2011	0.708	0.645	0.035	0.005	0.005	0.000
01-2009	0.692	0.617	0.046	0.008	0.009	0.001
07-2009	0.690	0.641	0.025	0.003	0.002	0.000
06-2008	0.689	0.645	0.028	0.003	0.003	0.001
10-2010	0.686	0.636	0.026	0.003	0.003	0.000
03-2008	0.679	0.644	0.023	0.002	0.002	0.000
01-2008	0.673	0.608	0.046	0.006	0.009	0.002
03-2010	0.658	0.594	0.031	0.005	0.004	-0.001
07-2008	0.640	0.586	0.029	0.004	0.003	0.000
04-2011	0.638	0.593	0.028	0.003	0.003	0.000
03-2009	0.622	0.579	0.023	0.002	0.002	0.000
05-2010	0.613	0.578	0.023	0.002	0.002	0.000
08-2008	0.594	0.552	0.025	0.002	0.003	0.000
08-2009	0.582	0.516	0.034	0.005	0.005	-0.001
04-2009	0.570	0.511	0.035	0.005	0.005	0.000
05-2008	0.567	0.533	0.023	0.002	0.002	0.000
11-2011	0.558	0.489	0.044	0.007	0.008	0.001
09-2008	0.557	0.496	0.037	0.005	0.005	0.000
09-2009	0.548	0.511	0.023	0.002	0.002	0.000
11-2010	0.546	0.478	0.044	0.007	0.008	0.001
07-2010	0.518	0.474	0.024	0.002	0.002	0.000
10-2008	0.517	0.481	0.024	0.002	0.002	0.000
05-2009	0.506	0.476	0.018	0.001	0.001	0.000
11-2009	0.501	0.472	0.019	0.001	0.001	0.000

Note: Est. (uncorrected estimate of efficiency score); Corr. Est (bias corrected estimate); MSE (uncorrected mean square error); MSE Corr. E. (mean square error corrected estimate); Diff (difference between corrected and uncorrected MSE)

DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

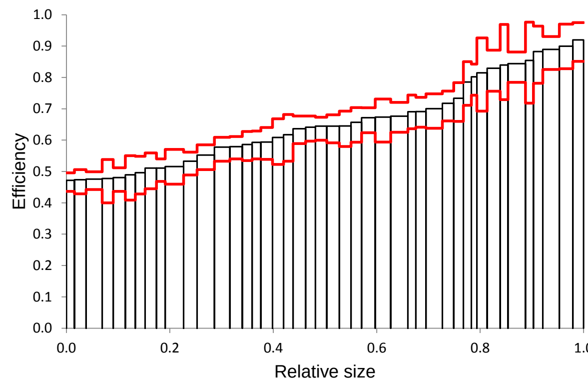

Figure 3: Bias corrected efficiency scores (bars) for each unit and year. Bold lines are 95 % confidence intervals. Data for eleven units over four years are pooled to 44 observations (44 bars). Width of bars proportional to unit size measured by the share of total cost

In order to construct confidence intervals for the estimates we follow the procedure in Simar and Wilson (1999), recently reviewed 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 $((\alpha/2)\times 100)$-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

$$Prob\left(c_{\alpha/2}\leq\hat{E}^*-\hat{E}\leq c_{1-\alpha/2}\right)=1-\alpha \tag{3}$$

From equation (3) we have estimated the 95 % confidence intervals for the pooled efficiency scores. The confidence intervals are outlined by the bold lines in Figure 3. The bars in Figure 3 are sorted by the size of bias corrected estimates. Bootstrapping seems to work fine for most efficiency scores when judged by range of confidence intervals in Figure 3, but four units stand out with a somewhat higher range (bar 4, 5, 7 and 9 from the right). This involves two of the best five performing units from the uncorrected scores, pointing at significant uncertainty related to those estimates. This information supports again our finding of best practice candidates for the pooled sample.

At the same time, it is reasonable to put more emphasis on best practice candidates which have a positive development in relative performance and to candidates which perform best in the last year of the period, as revealed in Figure 4 and 5. This way of reasoning adds more candidates for best practice. By considering bias corrected scores and their standard errors in Figure 3 and Table 5 alone, DMU-6 is identified as the best performer. If we look at the performance of DMU-6 in Figure 4, where the unit is outlined by a solid bar each year, its performance in 2010 is simply identified as the performance supporting its best practice candidature for the pooled sample. However, considering both

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

increasing and high performance at the end of the four year period suggests DMU-1 and DMU-8 as candidates.

For each of the three first years in Figure 4, where efficiency scores are sorted by year, there is a single unit pointed out as candidate for best practice, however a different DMU in each year. The spread in scores among the most efficient units is more narrow in the last year, without an obvious candidate for best practice. This could also indicate catching up in performance for some of the units towards the end of the period. From Figure 5, where the efficiency scores are sorted by unit for four successive years, we see that DMU 1, 5, 7, 8 and 9 are catching up towards the end of the period.

In comparison to the results of the original estimates presented in Figure 2, where the four units DMU 1, 6, 7 and 8 were estimated as fully efficient, bootstrapping made us reconsider DMU-1 and 7 due to bias and the size of their standard errors. Further, if we put more emphasis on positive development in relative performance and to candidates which perform best in the last year of the period, the candidature of DMU-6 is also weakened, and finally DMU-8 is suggested as the best practice candidate for the Home Guard.

As noted in Section 3.3 the Home Guard Staff stated the hypothesis that units located in populated areas would outperform units in rural areas. By looking at the units located in the less populated areas (DMU-9, DMU-10, DMU-11) we find DMU-11 as one of the best performers in 2008 and DMU-10 among the best in 2009. In addition, as we will return to in Section 4.2, two of the units have increased productivity during the period studied.

The unit restructured during the first years of the period (DMU-11) was expected to underperform by the Home Guard Staff. In Figure 5 we find support for a drop in performance for DMU-11 after 2008.

4.2. 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 during the period of four 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

$$M_i^s(1,2) = \frac{E_{i,2}}{E_{i,1}}, \quad i = 1, ..., N \tag{4}$$

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

When choosing a benchmark surface 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

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

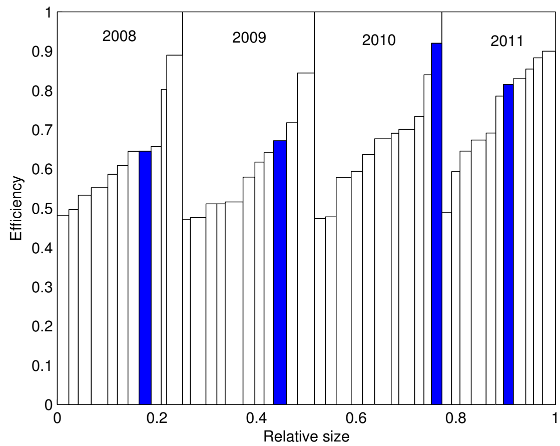

Figure 4: Yearly bias corrected efficiency scores (bars) per unit. 11 observations each year from the pool of 44 observations. Sorted by year and corrected efficiency score. Solid bars represent yearly performance of the unit with the highest efficiency score in a single year (2010) – DMU-6. Width of bars proportional to unit size measured by the share of total cost

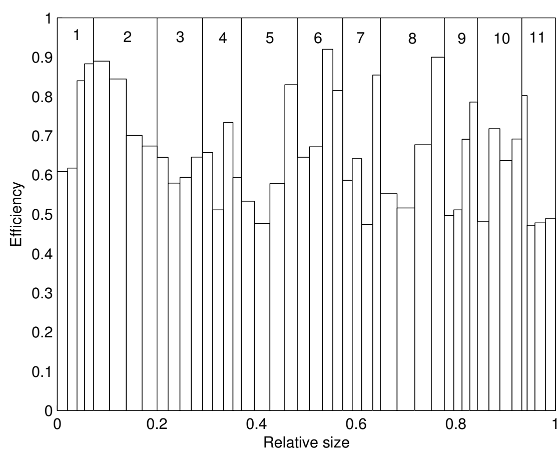

Figure 5: Bias corrected efficiency scores (bars) sorted by unit (DMU-1 – DMU-11). Four successive years for each unit. Width of bars proportional to unit size measured by the share of total cost

all inputs and outputs from one period to the next, keeping input and output mixes constant, should not change productivity. Hence, the productivity measure should be homogeneous of degree $|1|$.[20] This property makes the VRS specification unsuitable as a benchmark technology. Therefore, CRS is chosen as a benchmark in the Malmquist index. It is worth noting that using CRS just serves as a benchmark s for the productivity measure in (4), and no general assumptions of CRS technology are necessary. In addition, transitivity is ensured so the productivity of any pair of time periods can be calculated simply by multiplying the Malmquist productivity indices.

Using a common reference surface in the estimation of productivity change, no assumptions about change in the underlying technology for the production of the Home Guard's output is needed. On the other hand, we also exclude the opportunity to decompose the productivity index in order to study reasons behind developments in productivity. But, as noted above, a limited number of observations makes an intertemporal frontier based on pooled data the only meaningful approach in the present study.

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 over the four year period is set out in Figure 6. Each unit is represented by a rectangle, where the horizontal line inside each rectangle is the bias corrected estimate, width of a rectangle represents the relative size of the unit (share of total cost) and the height a 95 % confidence interval estimated by the bootstrap technique.[21] Seven out of eleven units have significantly improved productivity during the four year period (DMU 1, 5, 6, 7, 8, 9 and 10), while three units have experienced a significant drop (DMU 2, 4 and 11) . Only one single unit is estimated to have no significant change in productivity (DMU-3), as its 95 % confidence interval includes the horizontal bar in Figure 6 where the Malmquist productivity score equals unity. Improved productivity for as many as seven units could also underline our indication from Figure 4 of a catching up effect at the end of the four year period. The average unweighted bias-corrected change in productivity is a growth of 24 %. As predicted by the Home Guard staff we find a significant drop in productivity for DMU-11 after being restructured, illustrated by the first rectangle in Figure 6. Any correlation between unit size and productivity change could be studied by looking at the width of boxes in Figure 6. Both significant reductions and improvements are found in each group of relatively small, average and large sized units.

Comparing the intervals for efficiency scores in 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

[20] The efficiency measures in the index are homogeneous of degree $+1$ in outputs and -1 in inputs.

[21] In Førsund et al. (2005) this kind of diagram is named Edvardsen significance diagram.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

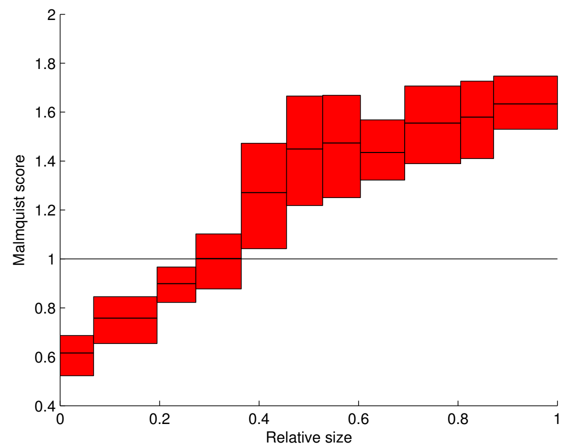

Figure 6: Bias corrected Malmquist productivity index estimates (horizontal lines inside rectangles) for each unit over the period 2008 – 2011. 95 % confidence intervals (rectangles) for each of the 11 units estimates. Estimates sorted by lower limit of confidence interval. Width of rectangles proportional to unit size measured by the share of total cost

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).

During the period studied, the Home Guard experienced a downward pressure on spending and reduced its use of inputs by 9.5 % in terms of the present model. In Figure 7 units are evaluated according to their changes in spending and productivity. The figure illustrates how the units adjusted to a downward pressure on costs. The nature of productivity changes is characterized by four quadrants in the figure: Productivity improving cost increase; productivity decreasing cost increase; productivity decreasing cost savings; and productivity improving cost savings (Førsund et al., 2006). Each unit is represented by a circle, where the circle's diameter is proportional to the units share of total cost. Nine out of eleven DMUs reduced their input compared to the beginning of the four year period. Almost all cost saving units are characterized as productivity improving cost savers. For the two only cost increasing units either no change or a decrease in productivity is estimated. The unit without significant change is easily identified as unit number four from the left in Figure 6, and the productivity decreasing cost increase corresponds to the first unit in Figure 6. Figure 7 indicates that the performance of the unit categorized as productivity decreasing cost increase (DMU-11) differs significantly from the performance of other units in the sample. This is suggested explained by the fact that DMU-11 was restructured during the period.

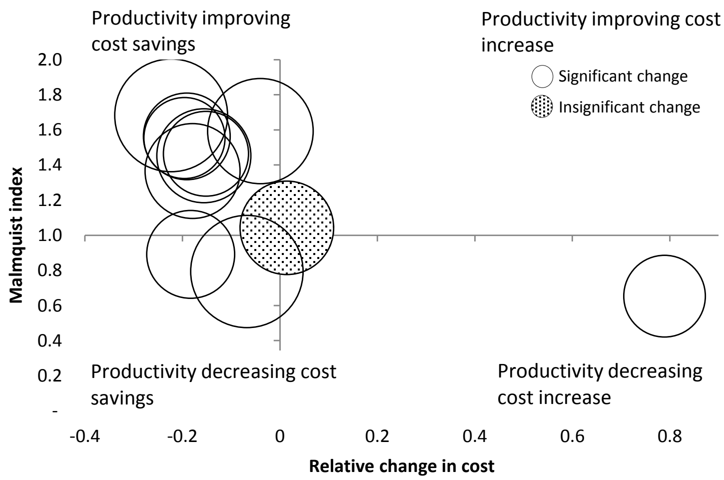

Figure 7: 4-quadrants of productivity and input changes. Circle diameters proportional to unit size measured by the share of total cost. Insignificant changes in productivity shaded

5. Conclusion

In addition to data issues a lack of output measures has limited the use of DEA in the military to various support services such as maintenance only. 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. Possible applications of DEA in the military is significantly expanded, and thereby the study offers an additional managerial tool to defense sector decision makers. Suggested applications of special interest to the military include, in addition to identifying "best-practice" units, also identifying possible improvements and efficiency gains by military unit or branch, and monitoring of productivity developments in light of current cost saving regimes and fluctuations in budget grants. The Norwegian Home Guard had no previous benchmarking system or other tools for comparing performance between units. After presenting the results to the Home Guard General Staff it was decided to run a one year pilot project where the model was implemented in the management of Home Guard units and a designated unit was evaluated intensively. The unit reported on the model variables to the General Staff. Results were discussed in a group of people representing the unit, the General Staff and researchers.[22]

Effective output is modeled as a generic function of physical output in form of troops (soldiers) and their quality within a military unit. The measure repre-

[22]The pilot is documented in a research report (FFI-report 13/00064, available in Norwegian only, www.ffi.no).

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

sents training and education outputs and not combat outcomes from operations. Outputs related to training activities are the relevant outputs in peace time and, due to its consistency over time, perhaps the only relevant measure to study productivity change. In defense sectors where quality measures and standards are established, application of our approach should be straightforward.

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 cross sectional studies about half of the units are efficient, leaving the Home Guard without clear candidates for best practice. After pooling the data 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.

Seven units improved their productivity significantly during the four year period studied. Comparing changes in cost to development in productivity, most of the units are characterized as productivity improving cost savers. Only one single unit is found to have a clearly unfavorable development, characterized as productivity decreasing cost increaser. This was, however, expected as the unit was restructured at the beginning of the period studied. As the total use of resources has declined by 9.5 % in terms of modeled input, our results suggest a Home Guard capable of adjusting to the downward pressure on spending experienced during the period studied. As there has been a downward pressure on military spending in NATO countries since 2009, with no clear sign of near reversal, extending the studies of productivity and efficiency to several other countries could be of interest as a benchmark and to compare possible peers between countries. Even though such studies could involve significantly larger armed forces than the present, lack of observations and established quality standards could still hamper extended studies.

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 it is 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. Unfortunately, DEA offers no explanations for the inefficiencies in the Home Guard. However, estimated efficiency and productivity scores could serve as a basis for explaining Home Guard performance in a second step. We leave this important topic to further research.

Appendix A. DEA model

The Farrell input-oriented technical efficiency scores are estimated for the $j = 1, ..., 11$ Home guard districts over the years $t = 1, ..., 4$, giving the observa-

DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

tions $k_{j,t} = 1, ..., 44$. The efficiency measure for unit k_0 is defined by

$$E_{k_0} = \min \theta_{k_0} \tag{A.1}$$

s.t.

$$\sum_{k=1}^{44} x_{4,k}\lambda_k \leq \theta_{k_0} x_{4,k_0} \tag{A.2}$$

$$\sum_{k=1}^{44} y_{r,k}\lambda_k \geq y_{r,k_0}, \quad r = 1, 2, 3 \tag{A.3}$$

$$\lambda_k \geq 0 \tag{A.4}$$

The variables x_4 and y_1, y_2, y_3 are the observed input and output variables listed in Table 4. Composition of the frontier reference point for unit k_0 is given by the endogenous intensity weights λ_k. All values are constrained to be non-negative, and at least one input, one output and one intensity weight have all to be strictly positive. No further restrictions on $\sum_{k=1}^{44} \lambda_k$ implies constant returns to scale.

Bibliography

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

Anagboso, M., Spence, A., 2009. Measuring defence. Economic & Labour Market Review 3 (1).

Bowlin, W. F., 1987. Evaluating the efficiency of US Air Force real-property maintenance activities. Journal of the Operational Research Society, 127–135.

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 (3), 691–709.

Brockett, P., Cooper, W. W., Kumbhakar, S., Kwinn, M., McCarthy, 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., Diewert, W. E., 1982. The economic theory of index numbers and the measurement of input, output, and productivity. Econometrica, 1393–1414.

DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Charnes, A., Clark, C. T., Cooper, W. W., Golany, B., 1984. A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the US Air Forces. Annals of Operations Research 2 (1), 95–112.

Charnes, A., Cooper, W. W., Lewin, A. Y., Seiford, L., 1994. Data envelopment analysis: theory, methodology and applications. Springer.

Charnes, A., Cooper, W. W., Rhodes, E., 1978. Measuring the efficiency of decision making units. European journal of operational research 2 (6), 429–444.

Clarke, R. L., 1992. Evaluating USAF vehicle maintenance productivity over time: an application of data envelopment analysis. Decision Sciences 23 (2), 376–384.

Cooper, W. W., Seiford, L. M., Tone, K., 2007. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Springer.

Efron, B., Tibshirani, R. J., 1994. An introduction to the bootstrap. In: Monographs on Statistics & Applied Probability. Chapman and Hall/CRC.

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. Series A (General) 120 (3), 253–290.

Farris, J. A., Groesbeck, R. L., Van Aken, E. M., Letens, G., 2006. Evaluating the relative performance of engineering design projects: A case study using data envelopment analysis. Engineering Management 53 (3), 471–482.

Førsund, F. R., 2010. Dynamic efficiency measurement. Indian Economic Review 45 (2), 125–159.

Førsund, F. R., 2013. Measuring efficiency and effectiveness in the public sector. Memo Department of Economics, University of Oslo 2013 (16).

Førsund, F. R., Kittelsen, S. A. C., Lindseth, F., 2005. Efficiency and productivity of Norwegian tax offices. Memorandum, Department of Economics, University of Oslo.

Førsund, F. R., Kittelsen, S. A. C., Lindseth, F., Edvardsen, D. F., 2006. The tax man cometh-but is he efficient? National Institute Economic Review 197 (1), 106–119.

Fried, H. O., Lovell, C. K., Schmidt, S. S., 2008. The measurement of productive efficiency and productivity growth. Oxford University Press.

Grifell-Tatjé, E., Lovell, C. K., 1995. A note on the Malmquist productivity index. Economics letters 47 (2), 169–175.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Hartley, K., 2010. The case for defence. Defence and Peace Economics 21 (5-6), 409–426.

Hartley, K., 2012. Conflict and defence output: An economic perspective. Revue d'economie politique 122 (2), 171–195.

Hurley, W., 2005. A clarification of the concepts of force multiplier and returns to force scale. Defence and Peace Economics 16 (6), 463–465.

Kittelsen, S. A. C., 1999. Monte Carlo simulations of DEA efficiency measures and hypothesis tests.

Kneip, A., Simar, L., Wilson, P. W., 2008. Asymptotics and consistent bootstraps for dea estimators in nonparametric frontier models. Econometric Theory 24 (06), 1663–1697.

Lewin, A. Y., Morey, R. C., 1981. Measuring the relative efficiency and output potential of public sector organizations: an application of data envelopment analysis. Int. J. Policy Anal. Inf. Syst. 5 (4).

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

Norwegian Ministry of Defence, 2010. Norwegian defence Fact and Figures. Oslo.

Owen, N., 1994. How many men do armed forces need? an international comparison. Defence and Peace Economics 5, 269–288.

Ozcan, Y. A., Bannick, R. R., 1994. Trends in Department of Defense hospital efficiency. Journal of Medical Systems 18 (2), 69–83.

Park, B. U., Jeong, S.-O., Simar, L., 2010. Asymptotic distribution of conical-hull estimators of directional edges. The Annals of Statistics 38 (3), 1320–1340.

Pedraja-Chaparro, F., Salinas-Jimenez, J., Smith, P., et al., 1999. On the quality of the data envelopment analysis model. Journal of the Operational Research Society 50 (6), 636–644.

Roll, Y., Golany, B., Seroussy, D., 1989. Measuring the efficiency of maintenance units in the Israeli Air Force. European Journal of Operational Research 43 (2), 136–142.

Romer, D., 2006. Advanced Macroeconomics. McGraw-Hill Series Economics. McGraw-Hill.

Schreyer, P., 2010. Measuring the production of non-market services. In: Price Indexes in Time and Space, Contributions to Statistics. Springer-Verlag Berlin Heidelberg.

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016

Silverman, B. W., 1986. Density estimation for statistics and data analysis. Vol. 26. CRC press.

Simar, L., Wilson, P. W., 1998. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management science 44 (1), 49–61.

Simar, L., Wilson, P. W., 1999. Estimating and bootstrapping Malmquist indices. European Journal of Operational Research 115 (3), 459–471.

Simar, L., Wilson, P. W., 2000. Statistical inference in nonparametric frontier models: the state of the art. Journal of Productivity Analysis 13, 49–78.

Simar, L., Wilson, P. W., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives. Oxford University Press New York, NY.

Spady, R. H., Friedlaender, A. F., Spring 1978. Hedonic cost functions for the regulated trucking industry. Bell Journal of Economics 9 (1), 159–179.

Sun, S., 2004. Assessing joint maintenance shops in the Taiwanese Army using data envelopment analysis. Journal of Operations Management 22 (3), 233–245.

UK Ministry of Defence, 2010. Operations. UK Army doctrine publication. The Development, Concepts and Doctrine Centre (DCDC).

Dette er en postprint-versjon / This is a postprint version.
DOI til publisert versjon / DOI to published version: 10.1016/j.ijpe.2016.05.016