JEL: D24, H40, C60

1. Introduction

Public sector effectiveness is measured in order to investigate whether public spending has the intended effects on society. The effectiveness of public sector service provision is usually evaluated by the impact the services have on measurable outcome indicators. Attainable effects and outcomes are determined not only by the amount of spending, but also by the technology used in producing public services, the service mix and a set of environmental variables affecting the transformation processes related to outputs and outcomes. The present paper sets out to pinpoint the role of output mix in public sector effectiveness. Despite the fact that the role is frequently recognized in theory, few practical solutions have been offered for estimating the effects of different mixes of output and the effect seems to be misunderstood in practical applications ([13]).

Four main motivations for effectiveness studies are found in the literature: the identification of output-changes in the public sector to be used in the national accounts to follow the development of public sector productivity;[2] the evaluation of investment decisions; the evaluation of programs and systems;[3] and the measurement of effectiveness in order to make quality adjustments for outputs.[4] Without careful assessment of the problem in question, however, important reasons for inefficiencies may be left out of the analyses. Thus, a fourth motivation for empirical studies of effectiveness is made explicit in the present paper through the introduction of the measure of mix-effectiveness and its implications for the impact of managers and policy makers on effectiveness.

[2]The Atkinson Review [3] is an example of careful examination of outcome measures in the context of the UK National Accounts.

[3]See e.g. the WHO handbook on evaluating health systems [22].

[4][9] states that "The quality of the output lies in its results, i.e. in the outcome.The most appropriate way of adjusting for quality therefore is to investigate changes in outcome indicators".

The various motivations and problems behind studies of effectiveness tend to lead to the development of different efficiency concepts and methods for measuring effectiveness. The usual practice is to define effectiveness as the Farrell technical efficiency measure, but substituting outputs with outcome variables.[5] Of course, in principle the general efficiency concept can be applied over several transformation steps. However, if the transformation steps are not modeled explicitly, significant sources of inefficiency can be left unidentified. Hence, substituting outputs with outcomes in the standard Farrell measure includes the effect of output mix in the model, but leaves the effect of both output mix and input mix unidentified. Yet another strand uses outputs as standard inputs and outcomes as the outputs.[6] This approach clearly deals with output mix, but without modeling any output production technology explicitly. Hence, the main disadvantage of the approach is that evaluation of output mix is not based on an assumption about the efficient use of the technology involved. Thus, the conditions for true effectiveness based on efficient use of technology where all units are projected to the production frontier remain unidentified.

Another problem unaddressed by both approaches is the level of control decision makers have in transforming outputs into outcomes [13]. If the effects of variables outside the control of decision makers are substantial, the meaning of efficiency, measured relative to a frontier, becomes unclear. A clarification of the effectiveness concept is found in Førsund [13], who introduces a new measure of overall preference effectiveness. The measure is similar to the revenue efficiency measure, which allows for substitutions in outputs, decomposing effectiveness into technical efficiency and output mix effectiveness instead of output mix efficiency. This is carried out by introducing a preference function for outcomes where the preferences take the role of output prices in the standard problem, but for outcomes rather than outputs.

[5] See e.g. [26] and [18].

[6] See e.g. studies of the the education sector in [20], the non-profit organizations in [21], transportation economics in [6], and public libraries in [15].

If multiple outcomes exist, some sort of signal, i.e. prices, is needed in order to determine whether the desired effects have taken place so that effectiveness can be evaluated. As market prices for outcomes related to public sector service provision do not exist, the preferences of policy makers e.g. politicians or civil servants could replace the role of prices in such evaluations. Preferences are, however, not straight forward to implement as guidelines for the production of outputs. This leads to an asymmetry in information between the policy maker (principal) and the managers at the production units (agents) which is only partly dealt with by the often carefull instructions provided by policy makers. Instructions could be implicit in the form of budget grants for various activities or more explicit in the form of output and outcome targets. As we will show, this is important for how we understand and interpret effectiveness in general and any inefficiencies from output mix in particular. In fact, distinguishing between output mix effectiveness and technical efficiency enables us also to specifically address any inefficiencies to either the policy makers or the managers in control of a production unit.

The links between outputs, exogenous variables (environmentals) and outcomes are complicated, and the dynamic relationships involved have to be modeled. The transformation of outputs and environmental variables into outcomes is described through an outcome mapping function.[7] If data on environmentals are available, the outcome mapping function can in principle be estimated. However, in many applications the variables are stochastic, unmeasurable or have too little variation for use in econometrical analysis. Few practical solutions are offered in the literature for specifying a function. Any evaluation of mix effectiveness based on empirical data is made possible only by dealing with these problems and by constructing continuous outcome indicators.

The complexities related to the estimation of the outcome mapping function,

[7]The concept of *mapping* could be found to be more appropriate than *production* in describing the function, which is not to be confused with production functions in the traditional sense. This clarification is owed to Sverre A. C. Kittelsen.

or the relationships such a mapping function are to express, are frequently pointed out in the effectiveness literature e.g. in the fields of health, education and defense. In the literature on effectiveness in health systems, there are problems related to the lag between inputs and outcomes [22]. The outcome of the health system is not only a function of inputs this year, but also partly a function of inputs from previous years as well as variables outside the control of the health sector. Murray and Evans [22] point out that inefficiencies could be partly due to technical inefficiency and partly due to choosing the wrong mix of outputs. However, only the first source is covered in detail in their framework.

Perhaps in response to the various applications of effectiveness studies, the concepts of output and outcome are used differently in the literature. Studies of the education system have been largely concerned with school outputs [7]. Pritchett and Filmer [23] question the basic approach and assumptions behind the studies of education production functions, after many empirical studies that have found that resources are only tenuously related to measured achievement. However, they did so without referring to any explicit distinction between production of outputs and outcome mapping, which possibly could have explained the tenuous relationships. Environmental variables are obviously important also in education. Ruggiero [26] suggests that standard data envelopment analysis (DEA) models do not work in studies of efficiency in the educational system, due to the heterogeneity of decision making units arising from socioeconomic differences.[8] In estimating the efficiency of New York State school districts, the education of the adult population is used as an environmental variable in the model, representing all exogenous community characteristics that influence educational production [26]. This enables the evaluation of school districts with the same level of environmental conditions in a DEA model. However, because it is concerned with outcomes but links inputs directly to the outcomes, the distinction between the technical efficiency of the school managers and the output

[8]Ruggiero [25] points also to the role of environmental variables in the public sector in general.

mix effectiveness of the policy makers is not explored in the study. Time lags in the transformation of inputs to outcomes in education are small as long as the outcomes are defined as test scores in each grade. The lags between inputs and outcomes such as employment rates and starting wages could, however, be substantial.

In studies of effectiveness in the military, the same general problems found in health and education apply. There is a possibly substantial time lag between the production of military outputs and the realization of outcomes. [16] states that there is no clear connection between inputs and outcomes in the sense that a marginal change in defense budgets is unlikely to have an immediate impact on the status of outcomes such as peace or overall sovereignty. Furthermore, the outcome of a conflict or situation is also a function of variables outside the production process and the control of decision makers at a national level. Studies of the military involve some specific complexities. Outcomes in the military, as stated in policy documents, include among others sovereignty and a contribution to the collective defense of an alliance. The problem with a measure of sovereignty is its binary nature at an aggregated level. At the aggregated level, a country is either sovereign or it is not. But there can be minor violations of sovereignty in limited geographical areas, such as fighter planes crossing a neighboring country's airspace, and the concept of sovereignty could thus also be interpreted as a continuous measure. However, except for such minor incidents where variations occur on a daily basis, outcome measures as dependent variables have little variation overall. This leaves the general method in the literature [4, 31] of regressing efficiency scores on the exogenous variables as meaningless. Hence, because of the lag problems in the variables, other methods are needed for the estimation of outcome mapping functions in the military and many other fields of application.

The contribution of this paper is threefold. First, we suggest a new framework for measuring outcome mapping functions and measuring effectiveness in the public sector. The framework exploits the development of military scenarios as continuous outcome indicators in order to overcome the problems related to

stochastic environmental variables and to lags between the transformation processes. Objectives or goals are replaced by scenarios as outcome indicators, each representing a unique vector of fixed environmentals. This implies that we are using ex-ante knowledge about expected effects rather than ex-post observations in evaluating effectiveness. Because the indicators are constructed from ex-ante information, we are able to estimate an immediate effect from marginal changes in inputs and outputs on outcomes. Given that there are preferred outcomes, we can evaluate effectiveness by comparing actual performance in the scenarios with the ideal performance.

We illustrate the scenario framework by developing a model evaluating effectiveness in the military, a classic example of public goods provision. Empirical results are estimated based on input and output data from military units in the Norwegian Armed Forces. In our specification of the model, we let the manager at each military unit take control of a single output, and we let the policy maker, in the form of the Ministry of Defence, determine output mix implicitly thorough their budget grants to each unit. This setup lets us assign technical issues to the managers (officers) at the military units and output mix to the policy makers. Distinguishing between politicians and bureaucrats "doing the right things" and officers "doing things right", we point to reasons for inefficiencies in the Armed Forces.

The second contribution of the paper is to estimate the outcome mix effectiveness from the sample of units in the Norwegian Armed Forces. To our knowledge, there is no other empirical work estimating the impact of policy makers' choice of output mix on effectiveness.

In the defense sector, there is a long tradition of cost-effectiveness studies and studies of the effect of different force structures on a single objective. The third contribution of the paper is to extend the effectiveness studies in the military to the defense sector as a whole, rather than looking only at partial cost-effectiveness studies of single systems or weapon platforms.

The rest of the paper proceeds as follows. In Section 2, different measures for outputs and outcomes in the defense and public sector are discussed, before

the scenario framework for measuring effectiveness is presented in Section 3. In Section 4, a model for evaluating effectiveness in the public sector is presented. The model is further specified for the Armed Forces. Data and empirical results are presented in Section 5, and Section 6 concludes the paper.

2. Literature review: Output and outcome measures in the public sector

Most, if not all, concepts regarding efficiency and effectiveness are used interchangeably in the literature. The usual practice is to define a production set Ψ describing the physical attainable points of $x \in m$ input variables and $y \in p$ output variables

$$\Psi = \{(x, y) \in \Re^{m+p} \mid x \text{ can produce } y\} \tag{1}$$

In the following we use the definition for technical efficiency in production to illustrate the confusion in the literature. The Farrell input-oriented efficiency measure for unit j_0 is defined by

$$E_{j_0} = \min \theta_{j_0} \tag{2}$$

s.t.

$$\sum_{j=1}^{n} x_{i,j}\lambda_j \leq \theta_{j_0} x_{i,j_0}, \quad i = 1, ..., m \tag{3}$$

$$\sum_{j=1}^{n} y_{r,j}\lambda_j \geq y_{r,j_0}, \quad r = 1, ..., p \tag{4}$$

$$\sum_{j=1}^{n} \lambda_j = 1 \tag{5}$$

$$\lambda_j \geq 0 \tag{6}$$

However, the lion's share of the literature defines effectiveness as the Farrell technical efficiency measure, replacing outputs y in (4) with outcome variables $s \in S$, and another strand even uses outputs y as standard inputs in (3) and outcomes s as the outputs. Such interchangeable use of output and outcome

terms leads to inconsistency in the use of the concepts "efficiency" and "effectiveness". In principle, the general efficiency concept in (2) – (6) can be applied over several transformation steps. If outputs are considered in (4), we are evaluating the use of a technology. However, if outcomes are considered in (4) and the usual inputs such as capital and labor in (3), we are evaluating two different effects – the effect of technology and the effect of prioritizing the optimal output mix – but without the ability to isolate either of them. When outcomes are evaluated, both the technical output efficiency and the output mix effectiveness have to be modeled in order to get a full evaluation of the problem. Hence, if the two transformation steps are not modeled explicitly in effectiveness studies, a possible significant source of inefficiency is left unidentified – the decision makers' choice of output mix.

Førsund [13] shows that it is possible to express output mix effectiveness explicitly by introducing a new measure of overall preference effectiveness, decomposing effectiveness multiplicatively into technical efficiency and output mix effectiveness. This is carried out by introducing a preference function for outcomes s, $W(s) = w(s(y))$ where the preferences take the role of output prices in the standard problem, but now for outcomes rather than outputs. The Førsund [13] preference effectiveness measure and its decomposition is explained in more detail in Section 4.

The decision makers' degree of control over the transformation process could be used to distinguish between the concepts of efficiency and effectiveness. Transformation of inputs to outputs is controlled by decision makers. But transformation of outputs to outcomes is out of the reach of the decision makers as long as the transformation process is taking place outside the production unit, so that other environmental variables z enter the process. Transformation of outputs and environmental variables into outcomes could be described by an outcome mapping function $s = g(y; z)$.

Findings in the literature of e.g health, education and defense economics indicate that estimation of a function is difficult in practice. This is due to too little variation or long lags in the variables over the period of time studied or

simply a huge number of unobservable environmental factors. As we will return to below, a possible route to success in empirical studies lies in the construction of continuous outcome indicators. The first step towards an appropriate outcome indicator is, however, a consistent definition of outcome.

As noted, despite apparently similar methods and motivation, there is a huge variety in the use of concepts, definitions and models. The most detailed and documented definitions regarding effectiveness concepts are found in the national accounts literature. In this literature, outputs are broken down into two components: activities or processes, and quality [29]. Further, outcomes are divided into either direct or indirect outcomes. The direct outcome is closer to the production process, such as the stage of knowledge of students, while the indirect outcome is associated with, for example, higher earnings resulting from higher human capital [29].

In the present paper, inputs x, outputs y and outcomes s are referred to as follows. Inputs, including labor, capital and other intermediate inputs, are combined and transformed by way of a production technology. Outputs are countable actions in the case of services or countable physical units in the case of goods.[9] Outcome is defined as the state valued by consumers and thus aligned with the definition of indirect outcomes in [29].[10] That state could, for example, be freedom in the case of military production, life expectancy in the case of health and human capital in the case of education [9]. However, as the following short review of public sector measures demonstrates, constructions found in the national accounts literature and applied in this paper are used inconsistently by most other strands of literature. This inconsistency could result in studies

[9] Outcomes could be dependent on both output quantity and quality. A possible approach is to assume a multiplicative decomposition of the output into a quantity part and a quality part. In a seminal paper on hedonic cost functions[33] it is suggested to treat effective output as a function of a generic measure of physical output and its qualities. [16] models the output measure for operational military units as a function of quantity and quality. Quality aspects of the outputs are based on this approach in the empirical part of the present paper.

[10] The distinction of C- and D-outputs in Bradford et al. [5] is similar.

where a distinction between "doing things right" and "doing the right things" is ruled out by definition.

WHO has developed a framework for measuring the effectiveness of health systems, identifying five goals of the systems: health level, health distribution, responsiveness, responsiveness distribution, and fairness in financial contribution [22]. The effectiveness measure takes into account the health level in the absence of inputs, as the level of health will not be zero even without a health system. This is an important difference compared to most other effectiveness measures. Effectiveness is defined as the ratio of actual goal attainment to maximum goal attainment, where the minimum health level resulting from no inputs to the system (health level of the population without any interventions from a health system) is subtracted from actual goal attainment.[11]

Eurostat [9] defines output in education as the quantity of teaching (student-hours) received by the students, adjusted for quality, for example by test scores. Different approaches to and motives behind education outcomes are found: higher education as an information signal [34]; the cultivation of the virtues, knowledge, and skills necessary for political participation and reproduction of society [14]; graduates employment rates, starting salaries or acceptance rates into higher education [38]. Ruggiero [26] measures school district outcomes as the average test scores for a district in each of reading, mathematics and social studies. Drop-out rate is included as an outcome for grades with non applicable test scores. In order to control for exogenous community characteristics that influence educational production, the education of the adult population is used as an environmental variable in the model.

In defense economics, concepts of outputs and outcomes are used interchangeably. Hartley [17] points out that defense markets have no market prices for their outputs, referring to lack of prices on aircraft squadrons, submarine or tank forces. In this situation, forces or military units are considered outputs. Further, Hartley points out that few published studies have estimated military

[11] Note that the actual measure is for efficiency and not effectiveness.

production functions, and those which have are using a cost-effectiveness approach. The effectiveness in such studies are determined on a disaggregated level such as air defense alone. A study of air defense would typically compare the effectiveness of land-based air defense missiles relative to manned aircrafts and their representative costs. Crary et al. [8] is another example, sizing the US destroyer fleet by using the analytical hierarchy process to gather expert opinion and deriving a distribution for the effectiveness of a fleet with a particular mix of ships.

Traditionally, the national accounts approach of defining outputs equal to inputs was also used for the defense sector. However, there is an expanding literature on the concept of defense outputs and outcomes. Hartley [17] 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. In the terms used in this paper, such variables would be analogous to defense outcomes rather than outputs.

The UK Ministry of Defence has established a system for defining and measuring the readiness of its Armed Forces [17]. Readiness is a concept which could define, at least partly, the output of a military unit. Anagboso and Spence [2] 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 used to measure output. The steps include activities which measure specific things the armed forces do, and the 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 [2] 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.[12] Two possible measures of capability are identified: a manpower measure, and an equipment measure. Both measures would have

[12]Färe et al. [10] equate capability with health outcomes in a study of hospital efficiency. The production of medical services are referred to as intermediate outputs in the model.

both a quality and a quantity component. Suggested quality adjustments for manpower are rank, grade, manning balances, manning pinch points and other measures for critical levels of personnel. For equipment, an explicit quality adjustment taking into account quality changes over time is suggested as well as a readiness measure. Hanson [16] makes use of both a readiness concept, a quality-adjusted manpower measure and a quality-adjusted equipment measure when modeling the output and estimating the efficiency of one type of military operational unit. The empirical data from that study is used also in Section 5 of the present paper.

Our suggestion for ordering the steps from military inputs to outcomes, based on the various approaches found in the literature, is illustrated in Figure 1. The first two steps, the transformation of inputs x to outputs y, are carried out at the military unit level.[13] In the transformation process, personnel are trained and combined with equipment. The output is a military unit consisting of a given level of personnel and equipment, where the measures of personnel and equipment are quality-adjusted according to readiness and training level standards. Military outputs in the form of trained units differ from most other goods or services produced in the public sector in that they are not consumed directly by consumers. Hence, there is no demand for military unit output from consumers. In military scenario planning too an explicit demand for unit output is absent, as the emphasis is on the effects and not the production of the units. Thus, we have to opt for the effect of fully produced military units measured by capabilities in order to map outputs with outcomes.

Fully produced military units possess certain attributes which we refer to as generic capabilities d. A submarine force has the attribute of suppressing surface combatants (the attribute anti surface warfare in Figure 1), with a possible direct effect of sinking a ship and an indirect effect of deterring other ships from entering the waters. In this manner, the peace time effect from submarine force

[13] A military unit is defined in the UK Army Doctrine as "...the smallest grouping capable of independent operations with organic capability over long periods..." [37, p. 89].

production is deterrence, and in the case of combat there is an additional effect of eliminating enemy ships. These are the kind of effects demanded in a military scenario.

While the personnel and equipment compositions are unit specific, capabilities are not. Air defense is delivered, for example, both by land based and sea based units. The size or extent of a particular capability is given by a capacity measure [24]. The capability of air defense could be measured in numbers of e.g. SAMs[14] equivalents, where the required size of the capability (the capacity) in a given scenario might be four SAM equivalents. However, in order to operate a SAM it is necessary to provide a unit of trained personnel combined with a means of transportation such as vehicles or ships. We therefore let the size (capacity) of a capability be expressed as a function of unit output in the form of trained personnel and equipment, $d(y)$. This is in line with the suggestion of capability measures in Anagboso and Spence [2], only differing in the level at which the concept of military output is defined. Hence, capabilities are used in the present paper only as intermediate outcomes mapping military units to events in a scenario.[15] The use of capabilities is also found in the scenario-based planning framework used in many NATO countries [24]. However, in the long term scenario-based defense planning capabilities are used as inputs rather than outputs, realizing defense goals in the form of scenarios. In such studies, the cost minimizing defense structure fulfilling all capability requirements is derived. Hence, there are no budget restrictions or priorities between objectives specified in the studies.

Finally, at level (4) in Figure 1, scenarios operationalize objectives and serve as outcome indicators in our approach. In each scenario, various capabilities are required.[16] This lets us use the capability requirements to construct continuous

[14]Surface to Air Missile system, e.g. Norwegian Advanced Surface-to-Air Missile System (NASAM).

[15]Capability is closer to outcome than output [10].

[16]The links between capabilities and higher defense goals found in the long term defense planning approach [24], is taken advantage of in the model presented in this paper (level (3)

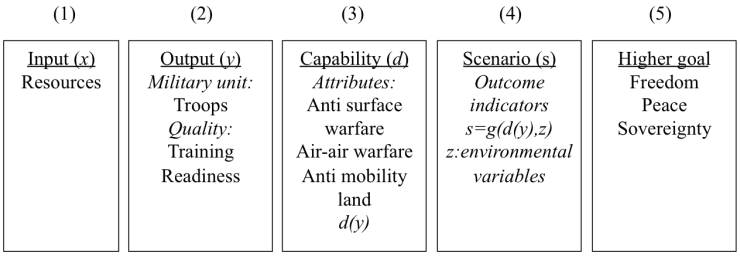

Figure 1: The levels from input to outcome in the military

outcome indicators for each scenario. As we modeled capability as a function of a unit's output y, we express an outcome indicator implicitly as a function of y

$$s = g(d(y), \bar{z}) \tag{7}$$

Going from level (1) – (5) in Figure 1 constitutes the basis of the model for effectiveness assessment presented in Section 4, including the scenario approach of mapping outputs and outcomes from level (2) – (4) presented in Section 3.

3. Effectiveness assessments and outcome mapping functions: A scenario framework

A frequent approach to studying effectiveness is to compare changes in outcome (often referred to as output in the studies) with changes in inputs (resources or products/outputs). When the reasons for inefficiencies are considered, at least two problems arise in this kind of study, both rooted in some sort of outcome mapping function estimation. The first problem is related to the impact of exogenous variables, named environmentals in the literature, on outcomes (but often referred to as outputs).[17] Second, for studies that also investigate the impact of output mix on effectiveness (but often referred to as efficiency), a separation of outputs and outcomes is shown and an outcome map-

– (5) in Figure 1).

[17]In principle, environmental variables can enter the transformation process both from inputs to outputs and from outputs to outcomes.

ping function has to be specified explicitly, including the role of environmental variables.

When environmental variables, such as GDP, population, territory, human capital, size of enemy forces etc., are introduced into a model, numerous studies apply a two-stage procedure to estimate the effect of such variables on outcomes (outputs) [32]. The first step is usually to estimate the efficiency of the system of interest, i.e. (2) – (6) where outputs are replaced with outcomes in (4), before regressing the efficiency scores on any environmental variables in a second step. Among empirical researchers, there are two well-established two-stage methods: those of Simar and Wilson [31] and Banker and Natarajan [4]. Simar and Wilson [31] suggest a two-stage method where bias-corrected efficiency scores from the first stage are regressed on environmental variables in a second stage. However, Simar and Wilson [31] did not advocate a two-stage approach but simply aimed at providing a coherent, well-defined statistical model where a second-stage approach would be appropriate [32]. Banker and Natarajan [4] find by simulation that a DEA-based two stage procedure with OLS in a second stage outperforms purely parametric methods, both one- and two-stages, in evaluating the impact of environmental variables on productivity. A huge limitation to empirical application of two-stage methods is the necessary assumption that environmental variables are orthogonal to the production frontier [32], i.e. environmentals can affect efficiency scores but not the frontier. Furthermore, necessary conditions for applying OLS in a second stage could be difficult to meet in many strands of the empirical literature, e.g. measurable and continuous outcome and environmental variables, variation in the variables over time, and that any effects are without substantial time lags. This argument is not a critique of the two-stage methods per se, but rather a recognition that alternative approaches are needed when the data (generating process) does not allow for regression-type approaches.

While a method for solving the problem regarding the impact of output mix is suggested in Førsund [13], simply assuming the existence of an outcome mapping function, neither that method nor the two-stage approaches offer a

practical solution in the case of variables that lack the variation necessary for regression analyses. In the standard application of the two-stage approach, the impact of the environmental variables on outcomes is estimated and some of the differences in efficiency scores among units are explained as well as important variables in an outcome mapping function being identified. However, in systems with little or no variation in outcome or environmental variables over the period of time studied, the estimation of efficiency scores and the following regression analyses are not meaningful.[18] When the outcome is more or less constant over time or between the systems studied, the observation with the lowest level of inputs will always end up as fully efficient. If the lag in the input to outcome relationship is longer than the length of the study, the effect of any changes in input will fall outside the study and the true relationship will remain unobserved.[19] For an outcome measure to be meaningful in this setting some variable representing fluctuations in the relevant data has to be introduced or other methods pursued.

Schreyer [29] suggests that a measure of the contribution to outcome should reflect the normal, or expected, effect of the output. Hence, normal, average or expected effects should be considered rather than ex-post effects. This line of reasoning is of particular relevance when outcomes are binary or have significant lags, where ex-post effects are difficult to identify. In the following, we suggest an alternative framework for estimating outcome mapping functions in situations where the two-stage approach fails due to lack of variation in the variables. The proposed framework is based on the type of scenarios found in long term defense planning under uncertainty, exploiting ex-ante information on effects. It is, however, applicable also in other fields for estimating outcome mapping functions. Each step is specified for the Armed Forces as an example, and some notation is introduced.

[18] e.g. defense variables in Hartley [17]

[19] e.g. health and education variables in Murray and Evans [22] and Worthington [38], respectively.

Scenario framework

Step 1: The first step in a scenario framework is to identify relevant policy objectives. In defense, such objectives are operationalized in white papers and other policy documents related to a nation's defense sector, and often stem from possible security threats. For example, in Norway such goals include upkeep of national sovereignty, national crises management, participation in international UN peace force operations, and similar more concrete activities (Norwegian defense facts and figures 2010).

Step 2: Second, all environmental variables z relevant to the policy objectives are identified. For example, from security threats such as threats to sovereignty, a set of environmental variables are identified. Further, a national sovereignty scenario is typically specific to a geographical area of a country and specific to the size of the enemy forces and type (from sea, land or air) of enemy assault, which are all environmental variables.

Step 3: The third step is to set up all relevant combinations of environmentals and objectives, where each combination represents a scenario s_j, $j = 1...J$. In each scenario the level of the otherwise stochastic environmental variables z are fixed. The fixed level for each variable represents the ambition of a scenario. In military planning, the relevant actors (possible enemies), their forces and capacities, and geographical areas of possible conflicts are combined with objectives, creating a set of scenarios. Each scenario now consists of at least one objective and a set of environmentals fixed at a given level. In Table 1, four different scenarios are set out for the military together with the required capabilities.

Step 4: The fourth step, uses the knowledge on how, first, outputs y affect capabilities $d(y)$, and, next, how the fixed environmentals z together with outputs, implicitly through capabilities, affect outcomes. This involves specifying the outcome mapping function in (7). Given the fixed levels of environmentals

$\bar{z}$, all stochastic elements are removed from the scenarios, resulting in a purely deterministic outcome mapping function. From the outcome mapping functions, continuous outcome indicators are constructed for each scenario. Of course, this is not relevant when a scenario plays out in reality. But policy makers are to be evaluated based on the information known prior to any actual events modeled in a scenario unfolding.

Outcome mapping function in the military

In the example of the armed forces, a military outcome mapping function is derived as an implicit function of outputs given the fixed environmentals in step four. Each scenario requires certain abilities from the defense structure, depending on the size of enemy forces, location at sea or land, etc. In a scenario at sea, an ability in demand is typically the ability to sink an enemy ship. As explained in Section 2, when such attributes are attached to military units they are referred to as capabilities and the size of the capability is referred to as capacity. In scenario development, requirements for capabilities and corresponding capacities are derived, which is what we model implicitly as a function of military unit output. An example is given in Table 1.

Here, two outputs y_1 and y_2 are mapped to four different capabilities in four different scenarios. Let y_1 be the output of a Submarine force with the capability of Anti Surface Warfare (ASuW) in the littoral zone[20] and ASuW in blue sea. Further, let output y_2 be produced by e.g. a Home Guard district mapped to the capability of High Asset Value Protection (HAVP) and Intelligence, Surveillance, Target Acquisition and Reconnaissance (ISTAR). This gives us four different mapping functions. Obviously, additional outputs to y_1 or y_2 will enter the functions, but here we treat any other outputs as constants for simplicity. This is the case also for environmental variables, which are treated as constants by definition in the mapping functions. Scenario s_1 and s_2 are

[20]The littoral zone is the part of a sea close to the shore, where some ships may not be able to operate.

Table 1: Relative scenario requirements by outputs and capabilities

		y_1		y_2	
		$d_1(y_1)$	$d_2(y_1)$	$d_3(y_2)$	$d_4(y_2)$
Scenario		ASuW1 [1]	ASuW2 [2]	HVAP [3]	ISTAR [4]
s_1	Strategic assault		40	80	30
s_2	Limited assault		30	70	
s_3	Coerced diplomacy	30			
s_4	Terror			70	

Note: Only illustrative numbers of relative use.

[1] Anti Surface Warfare

[2] Anti Surface littoral

[3] High Value Asset Protection

[4] Intelligence Surveillance, Target Acquisition and Reconnaissance

functions of both outputs, while scenario s_3 and s_4 are a function of only y_1 and y_2 respectively. Capabilities work here as the link between unit output and scenario requirements, enabling the specification of an outcome mapping function in our case.

Effectiveness assesments By inserting the various scenario outcome indicators into a preference value function, we get a single outcome measure representing the preferences of policymakers. From this setup, the evaluation of effectiveness for a given year is simplified to finding the mix of outputs that maximizes policymakers' preference value function for scenarios, given resource and technology constraints. The estimated effectiveness is now the ratio of the preference value of the optimal output mix and the preference value of the actual mix for that year. Planned or expected performance is evaluated rather than the often unobservable actual performance involving long time lags and stochastic variables. The full model for effectiveness assessment is presented in

Section 4.

4. Model

4.1. Background

The model sets out to evaluate the effectiveness of the Armed Forces and to track sources of any inefficiencies back to input mix and output mix, and the two types of agents: policy makers and managers. We assume that the policy makers are accountable for an output plan and the mix of outputs, while the managers are accountable for technical efficiency in the military units. This clear distinction between agents in the model is assured by letting each manager produce only one single output. In the multiple output case, managers are of course also possible sources of inefficiencies in output mix.

We draw upon a unique data set of 12 military units in the following. The sample consists of two types of military forces the Home Guard and the Submarine force. The Home Guard consists of 11 districts located in different geographical areas, producing the same type of output measured by the output indicator y_2. Home Guard units can produce military output by transforming inputs x_j, $j = 1 \dots J$, into output y_2 defined by the production set $G = \{f(x, y_2) \mid x \textit{ can produce } y_2\}$. There is one manager in each district. The Home Guard output is mapped into two different capabilities, HAVP and ISTAR, as shown in Table 1. The Submarine force consists of a single fleet which can produce military output by transforming inputs $x_j, j = 1 \dots J$, into output y_1 defined by the production set $H = \{f(x, y_1) \mid x \textit{ can produce } y_1\}$. The output is mapped into the capabilities ASuW1 and ASuW2 as shown in Table 1. There is a single manager in charge of the fleet. Data on the performance of the military units is studied in Section 5. Policy makers are given a budget for the defense sector B and resources are acquired at prices q. The output plan is executed implicitly by policy makers in the form of budget grants to each military unit.

The suggested model is best suited for evaluation of decisions in the short run, i.e. when the force structure is fixed and the flexible part of the budget is

the key decision variable. The reason is that the model is explicit on the level of output but the number of units is given. This assumption matches reality in the short run: Within a certain time interval, the size of a military unit measured by the number of personnel and equipment is usually fixed. As the marginal effect of quality on output is limited, major changes in output levels require changes in the military structure by either including or removing military units. Investing in new units to the structure is likely to also involve additional investments in equipment and infrastructure. Decisions on investments are, however, outside the scope of the suggested model. In the case of the Norwegian Armed Forces major changes in the force structure are usually the result of a revision carried out every fourth year. Minor changes in the force structure on a yearly basis, such as e.g. closing down a Home Guard district, occurs, on the other hand from time to time if approved by the parliament.

4.2. Policy makers' preferences

Preferences could replace prices when public sector outputs are transformed in to outcomes outside the market place. In our case of the military, defense goals are defined implicitly by the people through their elected representatives. The goals are further operationalized by policy makers i.e. politicians and officials at the Ministry of Defence, as explained in steps 1–3 in the scenario framework. Such policy makers are usually held accountable for effectiveness in their respective domains. Therefore, policy makers' preferences in relation to scenario outcomes are assumed in the model. First, goals are operationalized in scenarios s_j, $j = 1, ..., J$, before preferences for the scenarios are expressed assuming a Cobb-Douglas specification with constant elasticities

$$W = \prod_{j}^{J} s_j^{\omega^j}, \quad \sum_{j} \omega_j = 1 \tag{8}$$

The assumption implies that the policy makers prefer averages over extremes, in the sense that they will avoid leaving any single scenario completely uncovered. The weights ω_i have to be estimated.

Aggregating the various scenario outcome indicators using a preference value function, we get a single outcome measure representing the preferences of the policy makers. The weight of each scenario in such an aggregate measure represents the relative importance of each scenario for the policy makers responsible for the associated higher goals. Several methods have been used in the literature for the estimation of decision makers' preferences. Among others are surveys of health experts [19] and pairwise comparison of attributes for locating government agencies in Japan, conducted by expert opinion within the engineering community [35], both carried out by the analytic hierarchy process (AHP).[21] In studies of population health measures, various methods are used to estimate preferences for health outcomes [28], based on data from diverse general population samples. However, the valuation of defense outcomes in the form of scenarios are likely to require major branch-specific insight compared to the self valuation of the state of people's health. The use of general population surveys is therefore ruled out for relatively complex scenarios.[22]

4.3. Outcome mapping function and the military technology

The outcome mapping function describes the transformation from outputs y, implicitly through capability $d(y)$ and environmental variables z, to outcomes s (step 4 in the scenario approach). The general outcome mapping function is given by (7), but with environmentals z specified for each scenario in the scenario framework and taken out of the expression. Output y_l is produced to deliver capability d_l, $l = 1, ..., L$, in scenario j. The capability is fully delivered when the output level reaches $Y^*_{l,j}$, where $d_{l,j}(Y^*_{l,j}) \equiv 1$, $\delta d/\delta y > 0$ for $y < Y^*$

[21] AHP [27] is a technique for group decision making where a numerical value or weight is derived for an alternative or criteria at each level of a hierarchy. Building a hierarchy consisting of scenarios at the lowest level and goals at the top of the hierarchy, a pairwise comparison of each scenario's contribution to the overall goal provides that scenario's priority or weight.

[22] If data on decision makers behavior (preferences) is unavailable, an alternative approach is to incorporate preferences of the median voter for some higher objectives and decompose the general preferences to preferences for various scenarios.

and 0 for $y \geq Y^*$. Each capability has a scenario specific weight $v_{l,j}$, indicating the relative importance of that capability in a given scenario. The weight is determined by a risk analysis conducted by military expert opinion. Weights are the product of the probability that a capability is utilized in the scenario, and the consequence for the scenario outcome of not having the capability available when needed. For the capability Anti Surface Warfare (ASuW1 in table 1), the weight in scenario j is the product of the probability that an enemy ship has to be engaged in this scenario, times the impact on the scenario if the enemy ship is not engaged. The impact on the scenario without capability l when actually involved in the scenario is determined by the military experts rating the capability on a six point scale from no impact (0) to indispensable (5). This gives us the following outcome mapping function

$$s_j = \sum_{l=1}^{L} v_{l,j} d_{l,j}(y_{l,j}) \tag{9}$$

Assuming capability functions $d(y)$ linear in outputs, we express the outcome mapping function for scenario S_j as a function of outputs explicitly

$$\begin{aligned} s_j &= \sum_{l=1}^{L} v_{l,j} \left((\frac{y_l}{Y^*_{l,j}}) + (1 - \frac{y_l}{Y^*_{l,j}}) H(y_l - Y^*_{l,j}) \right), \qquad (10) \\ H(y_l - Y^*_{l,j}) &= \begin{cases} 1 & \text{if } y_l \geq Y^*_{l,j} \\ 0 & \text{if } y_l < Y^*_{l,j} \end{cases} \end{aligned}$$

The assumption $\delta d/\delta y = 0$ for $y \geq Y^*$ is assured by the Heavieside function[23] $H(\cdot)$ where no additional value is generated from outputs exceeding the capability requirements $Y^*_{l,j}$. When outputs exceed the requirements, $H(\cdot)$ equals 1, the y/Y^* terms are deleted and we are left with unity, a fully delivered capability.

Outputs y_i are the production of military units. We are using the output concept for operational military units developed in Hanson [16]. An output

[23]The Heavieside function is a unit step function, used for example to "turn off" and "turn on" variables or expressions over different intervals. See e.g. Adams and Essex [1].

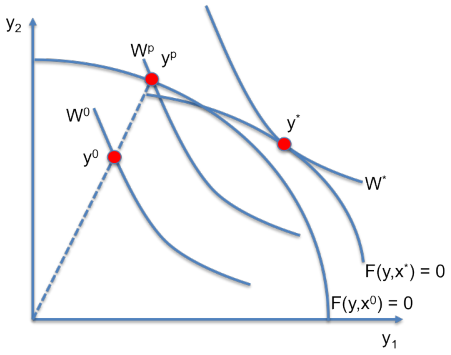

Figure 2: The realization of output mix effectiveness

index is created from the status of personnel, equipment and proficiency levels in each unit. A constant return to scale technology is assumed for the transformation of resources x to a military unit y_i given by $F(y, x)$.

4.4. Overall preference effectiveness

Solving the model, we maximize the preference function (8) subject to the outcome mapping function (10), the technology constraint given by the frontier function $F(y, x) = 0$ and the budget constraint $qx = B$.

$$\begin{aligned} &\max_{y} W(s) \\ &\text{s.t.} \\ &s = g(y) \\ &F(y, x) \leq 0 \\ &qx = B \end{aligned} \tag{11}$$

The solution to the problem is illustrated in Figure 2. In the figure we have two output dimensions, y_1 and y_2. The y_1 output is analogous to the output of a submarine fleet in the empirical data, and y_2 is analogous to Home Guard output.

The transformation between outputs at the frontier is shown by the curve $F(y, x_0)$ for the initial input vector x^0. We start out in y^0 and move towards the frontier. In y^p, all inefficiencies in the production of outputs are eliminated. However, this is not the solution to our optimization problem. The solution to the problem implies a different mix of outputs. We find the solution in y^*, where the contour curve W^* has a higher preference value than the contour curve W^p going through the point y^p. This is the realization of output mix effectiveness. Given a fixed budget, the production possibility sets will differ for different input vectors. In optimum the input vector x^* is different from the initial x^0, so is the production possibility set $F(y, x^0)$.

More formally, the two effects can be expressed by taking advantage of the overall preference effectiveness concept (OPE) introduced in [13].

$$OPE = \frac{W\left(S(y^0)\right)}{W\left(S(y^*)\right)} = \underbrace{\frac{W\left(S(y^0)|x^0\right)}{W\left(S(y^p)|x^0\right)}}_{\text{Technical efficiency}} \underbrace{\frac{W\left(S(y^p)|x^0\right)}{W\left(S(y^*)|x^*\right)}}_{\text{Mix effectiveness}} \tag{12}$$

Overall preference effectiveness is the ratio of the preference value for the initial output vector y^0 and the preference value of the optimal output vector y^*. Furthermore, we can decompose into the two effects of doing things right (technical efficiency) and doing the right things (output mix effectiveness) explained in Figure 2. In our model, where each manager is accountable for only one output and policy makers implicitly determine output mix by budget grants, the decomposition is equivalent to identify managers "doing things right" and policy makers "prioritizing the right things".

Here, the term tagged technical efficiency in (12) is in general not identical to the Farrell [11] definition of technical output efficiency. For the concepts to be analogous, sufficient assumptions are constant returns to scale for the preference function and the outcome production functions.

5. Results

5.1. Background and data

In estimating the model, we are using data from a sample of 12 different operational units in the Norwegian Armed Forces over the years 2008 to 2011. The units consist of eleven Home Guard districts and one Submarine force. The policy makers are to be evaluated on the mix effectiveness of outputs for the units, and the officers in command of the units (managers) are to be evaluated on the technical efficiency of their respective military unit. As there have been no explicit priorities or trade-offs between units selected for this study, the data might not be the best for testing hypotheses on the development of mix effectiveness. However, this does not mean that the sample was neglected by the policy makers during the period of time. Downscaling of the Home Guard prior to the July 22 terror, where the Home Guard had a significant role in securing the city centre of Oslo, and a decision on whether to extend the lifetime of the submarines or acquire new ones, were heavily debated at the time. A priority efficient defense sector should yield high output mix effectiveness for all samples, including this one.

5.2. Estimating the model

The sample of Home Guard and Submarine units spans over the four scenarios set out in Table 1. Each of the four scenarios is considered as an outcome indicator operationalizing some higher goals for the policy makers, and is a result of steps 1–4 in the scenario approach. The Submarine force delivers two capabilities (ASuW1 and ASuW2) in three different scenarios. The same is the case for the Home Guard districts (HVAP and ISTAR). This means that the units contribute to two common outcomes (s_1 and s_2). In the following we assume that the derived output requirement for capability l in scenario j, $Y^*_{l,j}$, is within the potential of the existing force structure. The Home Guard capabilities are assumed equally important independent of Home Guard district location.

As long as we have more than one goal or outcome indicator, a preference function must be applied. Estimating the parameters of equation (8) implies estimating the weight the decision makers assign to each scenario. Due to lack of data on policy makers preferences, we are limited in this study to supplementing a baseline (null hypothesis) of equal weights, $\omega_i = 0.25, i = 1, ...4$, with sensitivity analysis for alternative preference structures. We carry out two additional runs favoring different preference structures in addition to our null of equal weights for the scenarios. The four scenarios involved could be divided into two scenario preference groups: (a) encompassing but highly improbable (favoring s_1 and s_2) where weights are set to $\omega_1 = \omega_2 = 0.4$, $\omega_3 = \omega_4 = 0.1$ and (b) limited but more likely (favoring s_3 and s_4) and weights $\omega_3 = \omega_4 = 0.4$, $\omega_1 = \omega_2 = 0.1$. Group (b) corresponds to a preference structure where policy makers have preferences for avoiding current threats of relatively high probability, and we name this structure *nearsighted*. Correspondingly, group (a) is named *farsighted*, representing preferences for the long run and sovereignty on an aggregated level. Our null represents policy makers who state that all derived scenarios are of equal importance, or simply fail to prioritize between them. The basic interpretation of the suggested setup is to consider a constant preference structure over time. But in the long run preferences are likely to change. By considering various preference structures for at least two of the four years studied, we allow for interpretation of possible changes in preferences over time. The model is estimated separately for each of the preference structures as follows.

First, the four different outcome mapping functions, one for each scenario, are found by inserting in (10) the derived output requirements $Y^*_{l,j}$ and capability weights $v_{l,j}$ from step 4 in the scenario approach. A linear realization of capabilities in outputs is assumed for simplicity. After having specified the preference function (8) and the outcome mapping functions (10), at least two procedures could be followed in order to estimate a solution to (11). The first involves formulating (11) as an LP-problem and estimating it non-parametrically. But since $W(s)$ in (11) is nonlinear in our specification of the model, we suggest

rather using a mix of parametric and non-parametric approaches.

This involves a parametric approximation to the nonparametric part of the problem in the form of the technological restriction $F(y, x) = 0$. First, the production frontier is estimated by a nonparametric method before a parametric frontier function is estimated based on only efficient observations. Finally, the parametric expression for $F(y, x) = 0$ is inserted for the technology constraint when solving (11) numerically. Florens and Simar [12] argue that, since the production frontier is the locus of optimal production situations, we might get substantial improvements in bias, variance reduction, etc. if we use only efficient observations from a nonparametric method in the first step to estimate a parametric function in a second step. The expected order-m frontier is suggested to be used in the first step by Florens and Simar [12] due to its consistency and avoidance of dimensionality problems associated with other nonparametric estimators.[24]

Following the second procedure, we start out with a nonparametric estimate of the technical efficiency of the military units. DEA is used in the nonparametric first step to estimate the efficiency scores for the Home Guard and Submarine force respectively. The DEA model is input oriented, with one output, y_2, and three input variables for the Home Guard, x_j, $j = 1, 2, 3$. For the Submarine force, the lack of observations limits the model to a simple one input $x = \sum_j x_j$, $j = 1, 2, 3$, one output, y_1, model. Pooled data are used for the Home Guard (eleven units over four years, 44 observations in total) and the Submarine force (one unit over four years, four observations in total), respectively, due to the small samples.[25] The inputs are measured as operating costs for all units, and

[24] The origin of the procedure in which a parametric function is estimated in a second step, based on a non-parametric frontier estimation in a first step, is in [36]. This seems to be neglected in Florens and Simar [12]. The expected order-m frontier gives the expected maximum production among a fixed number of m firms using less than x inputs. The estimator is found to be superior to the FDH estimator and shifted OLS applied on experimental data.

[25] Productivity and efficiency in the Home Guard during the period 2008–2011 is studied in [16]. No new equipment or operating conditions suggesting changes in technology were

divided into the variables equipment, personnel and activity-based personnel costs for the Home Guard units.

In the short run, capital is fixed and cannot be fully reallocated between units. This is also the case for some of the more specialized personnel resources in the short run. Hence, the substitution possibilities between Home Guard and Submarine forces could be incomplete for policy makers, influencing the mix effectiveness on a yearly basis. The aggregated output measure used, described in detail in Hanson [16], consists of quality-adjusted measures for manpower, equipment and training. Estimated DEA efficiency scores from the first operation, E_i for unit i, are presented in Table 2.

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. [30] showed how to estimate the sampling bias in DEA using the bootstrap method (resampling). Following the suggested procedure we could point at the uncertainty in the estimated efficiency scores for the Home Guard units. However, it is not meaningful to estimate bias corrected efficiency scores for the Submarine force due to the limited sample size. Hence, a further discussion of the uncertainty of the estimates is not possible from an empirical perspective. On the other hand, from a methodological perspective it is not straightforward how to estimate any biases in the effectiveness scores and how to address the uncertainty in the scores when uncertainty in preferences and technology are considered simultaneously. We leave this, however, to further research and the following results are reported without the estimation of confidence intervals.

In the second operation, the units are moved to the frontier by a proportional contraction of their inputs, xE_i, eliminating any inefficiency. From the 44 ob-

reported during the period studied, and it was considered reasonable to assume the technology to be stationary.

Table 2: Technical efficiency scores

DMU	2008	2009	2010	2011
HG-01	0.88	0.77	1.00	0.88
HG-02	0.78	0.85	0.68	0.57
HG-03	0.66	0.58	0.69	0.48
HG-04	0.58	0.50	0.69	0.60
HG-05	0.36	0.48	0.54	0.98
HG-06	0.77	0.52	1.00	0.79
HG-07	0.63	0.66	0.61	0.67
HG-08	0.54	0.54	0.56	0.87
HG-09	0.71	0.60	0.85	0.77
HG-10	0.47	0.60	0.45	0.67
HG-11	0.54	0.38	0.20	0.22
Sub. force	1.00	0.91	0.92	0.55

servations, a production function is estimated by OLS, assuming the functional form in (13). The estimated coefficients are $\alpha = 0.17$, $\beta = 0.28$ and $a = 0.0018$. All estimates are significant.

$$y_2 = a x_1^{\alpha} x_2^{\beta} x_3^{1-\alpha-\beta} \tag{13}$$

$$y_1 = d \sum_{i=1}^{3} x_i \tag{14}$$

The same procedure is followed for the Submarine force in (14), with a significant estimate of $d = 0.22$. The two production functions (13) and (14) are used together to find an expression for $F(y, x)$ in (11). Solving for y_2 in the maximization problem we get the expression

$$y_2 = 0.668(B - \frac{y_1}{0.22}) \tag{15}$$

The now parametric transformation function (15), given a budget B and outcome mapping functions from (10), is used as a constraint in maximizing the

policymakers' preference function (8) for the three different preference structures. This is the solution to (11), giving us the maximum preference score for each year. Finally, actual preference scores, derived from inserting actual output values in (10) and (8), are used together with the estimated maximum preference scores in order to derive the output mix effectiveness (OME) for each year. The results are outlined in Table 3.

5.3. Discussion

Decomposing the overall preference effectiveness scores into technical efficiency (TE) and output mix effectiveness (OME), we can now pinpoint reasons for inefficiencies in this sample for each of the preference structures. In our baseline preference structure, the main source of inefficiency at the start of the four-year period was related to technical aspects, the domain of managers at the operational units in our model. After a decrease in policy makers' mix effectiveness over the years, with only small corresponding changes in technical efficiency, the two sources of inefficiencies contribute about the same to overall effectiveness at the end of the period.

As improvement in technical efficiency is found difficult in practice, a clear policy implication of our results is further investigation into the policy makers' decisions on output mix and possible reallocations of resources. The results suggest a quick fix by simply reallocating production between the military production units. On the other hand, fixed inputs such as capital could somewhat limit reallocations and explain some of the inefficiencies from output mix in the short run. The solution to the maximization problem in (11), illustrated in Figure2, provides the optimal output levels and thus the necessary changes in the output mix. Corresponding changes in the use of resources are derived from the parametric production functions. A reallocation of funds by the policy makers from Home Guard (28 % budget decrease) to Submarine force (67 % budget increase) yields output mix effectiveness at the end of the period where the overall preference effectiveness is at its lowest. If the Home Guard capabilities are independent of geographical location the budget decrease could result

Table 3: Overall preference effectiveness and its decomposition for different preference structures

Measure	2008	2009	2010	2011
Overall preference effectiveness				
Baseline	0.678	0.650	0.660	0.597
Nearsighted	0.674	0.637	0.654	0.651
Farsighted	0.631	0.616	0.613	0.500
Technical preference efficiency				
Baseline	0.736	0.789	0.825	0.762
Nearsighted	0.674	0.757	0.799	0.782
Farsighted	0.805	0.823	0.852	0.744
Output mix effectiveness				
Baseline	0.920	0.824	0.799	0.782
Nearsighted	1.000	0.842	0.819	0.833
Farsighted	0.784	0.749	0.719	0.673
Total budget (B)	738	783	754	686

in closing down an inefficient unit, given that other factors are held constant. Budget decreases of such magnitude as 28 % are rare in the military, at least on an annual basis. However, the result could serve as guidance for future allocations if the total budget maintains its relatively moderate level. Also, further analysis could be necessary in order to reveal whether there are other reasons outside the model explaining the apparently inefficient allocation.

Testing the model on somewhat different preference structures, we find the same downward trend in output mix effectiveness over the four-year period of time. However, allowing for different discounting in policy makers' preferences, a higher level of mix effectiveness is found when limited but more likely scenarios are favored (s_3 and s_4). A possible interpretation of this finding is that policy makers are biased toward the likely and less hypothetical scenarios. Comparing estimates from the farsighted preference structure to baseline estimates does not support a preferences structure favoring less likely but more encompassing scenarios (s_1 and s_2) over the balanced baseline preference structure. From sensitivity analysis, we suggest that our results on evaluating the mix effectiveness of policy makers are robust in trend but sensitive in magnitude. For any further interpretations, preference functions have to be estimated based on real data, which is outside the scope of the present paper.

6. Conclusion

This study set out to examine empirically the impact of output mix on effectiveness in public service provision. A measure of overall preference effectiveness is used to isolate the effect of output mix from that of input mix on effectiveness. A clear distinction between output and input mix in our model enables a tracking of sources for inefficiencies back to managers "doing things right" in the production process and policy makers "doing the right things" in choosing the optimal mix of outputs. As far as we know, this is the first study to present any empirical evidence on the role of policy makers in effectiveness of public service provision.

Most empirical studies in the effectiveness literature model the transformation from inputs to outcomes as a single transformational process. Thus, identification of possible inefficiencies from output mix is ruled out by definition. We suggest, however, that distinguishing two transformational steps requires estimation of a function mapping outputs to outcomes in addition to the usual estimation of production technology transforming inputs to outputs. Realization of outcomes usually involves complicated dynamics from output and environmental variables. A two-stage method is used in the literature when the impact of environmental variables on outcomes or outputs are studied. In the second stage, environmental variables are regressed on a performance measure, e.g. technical efficiency scores. We recognized, however, that characteristics of outcome and environmental variables in many empirical examples make the use of regression-based approaches less meaningful. In response to this, we suggest a scenario approach for estimating outcome mapping functions. In the scenario approach, ex-ante knowledge about expected effects rather than ex-post observations are drawn upon in effectiveness assessments. This involves replacing objectives with scenarios, each representing an unique vector of fixed environmental variables.

Despite its contribution in isolating the role of output mix, our method comes with some possible obstacles when applied to empirical data. When multiple objectives or outcomes are evaluated, a preference function has to be introduced representing politicians' or policy makers' preferences. Testing the model on data from the Norwegian Armed Forces, we find difficulties in presenting the scenarios to policy makers in a manner which could be used to generate data for estimation of preferences based on the methods presented in the literature. Estimating baseline results, we therefore relied on the assumption that all objectives are of equivalent importance to the policy makers. This is in line with the response from policy makers in the defense sector when confronted with an inquiry about the ranking of objectives. Given the confusion in the literature concerning output and outcome concepts, we are not surprised if a similar confusion exists among policy makers. Thus, we believe it is difficult for

decision makers in real life situations to state their preferences for outcomes or outputs consistently. In principal, median voter preferences could replace the preferences of decision makers. But in the case of application to the military, outcomes in form of scenarios are relatively complex to evaluate and difficult to assess for the average voter without some branch specific knowledge.

We find that overall effectiveness in the sample of units from the Norwegian Armed Forces has decreased in the four-year period studied. The drop in effectiveness is not related to technical inefficiencies, which is the standard interpretation found in the literature, but rather is the result of an inefficient mix of outputs by the policy makers. Inefficient priorities could partly be explained by fixed inputs in the short run, as in the last year of the period a 9 percent decline in total budgets possibly increased the relative share of fixed costs at the end of the period. A limitation to inference based on the estimated results is, however, the difficulties in addressing the uncertainty in the effectiveness scores, both from an empirical (limited sample size) and from a methodological approach. Consistent methods for addressing the uncertainty in estimated overall preference effectiveness score is suggested as an important topic in further research.

Allowing for differences in discounting, the model is estimated for two additional specifications of the preference function. Favoring nearsighted scenarios, the downward trend in output mix effectiveness is maintained, but the level of mix effectiveness is higher compared to baseline preference structure. The finding of a downward trend is supported also by the farsighted preference structure. However, policy makers are performing worse in mix effectiveness in these scenarios, compared to both baseline and nearsighted preferences. Sensitivity analysis suggests our results to be robust in trends, but sensitive in size. We leave the specification and estimation of a preference function for defense objectives to further research. Any additional interpretation of our results we leave to policy makers themselves, as they determine which preference structure to be evaluated on. the estimated impacts of environmental variables an outcome production function ca

Bibliography

[1] Adams, R. A., Essex, C., 2013. Calculus: a complete course, 8th Edition. Pearson Education, Canada.

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

[3] Atkinson, T., 2005. Measurement of Government Output and Productivity for the National Accounts. Palgrave Macmillan.

[4] Banker, R. D., Natarajan, R., 2008. Evaluating contextual variables affecting productivity using data envelopment analysis. Operations Research 56 (1), 48–58.

[5] Bradford, D. F., Malt, R. A., Oates, W. E., 1969. The rising cost of local public services: some evidence and reflections. National Tax Journal 22 (2), 185–202.

[6] Chu, X., Fielding, G. J., Lamar, B. W., 1992. Measuring transit performance using data envelopment analysis. Transportation Research Part A: Policy and Practice 26 (3), 223–230.

[7] Cooper, S. T., Cohn, E., 1997. Estimation of a frontier production function for the South Carolina educational process. Economics of Education Review 16 (3), 313–327.

[8] Crary, M., Nozick, L. K., Whitaker, L., 2002. Sizing the US destroyer fleet. European Journal of Operational Research 136 (3), 680–695.

[9] Eurostat, 2001. Handbook on price and volume measures in national accounts. Luxembourg: European Communities.

[10] Färe, R., Grosskopf, S., Lundstrøm, M., Roos, P., 2008. Evaluating health care efficiency. In: Evaluating hospital policy and performance: contributions from hospital policy and productivity research. No. 18 in Advances in Health Economics and Health Services Research. pp. 209–22.

[11] Farrell, M. J., 1957. The measurement of productive efficiency. Journal of the Royal Statistical Society. Series A (General) 120 (3), 253–290.

[12] Florens, J.-P., Simar, L., 2005. Parametric approximations of nonparametric frontiers. Journal of econometrics 124 (1), 91–116.

[13] Førsund, F. R., 2017. Measuring effectiveness of production in the public sector. Omega 73, 93–103.

[14] Gutmann, A., 1999. Democratic education. Princeton University Press.

[15] Hammond, C. J., 2002. Efficiency in the provision of public services: a data envelopment analysis of uk public library systems. Applied Economics 34 (5), 649–657.

[16] Hanson, T., 2016. Efficiency and productivity in the operational units of the armed forces: A norwegian example. International Journal of Production Economics 179, 12–23.

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

[18] Karlaftis, M. G., 2004. A DEA approach for evaluating the efficiency and effectiveness of urban transit systems. European Journal of Operational Research 152 (2), 354–364.

[19] Lauer, J. A., Lovell, C. K., Murray, C. J., Evans, D. B., 2004. World health system performance revisited: the impact of varying the relative importance of health system goals. BMC health services research 4 (1), 19.

[20] Lovell, C. K., Walters, L. C., Wood, L. L., 1994. Stratified models of education production using modified DEA and regression analysis. In: Charnes, A., Copper, W., Lewin, A., Seiford, L. (Eds.), Data Envelopment Analysis: Theory, Methodology, and Applications. Kluwer Academic Publishers, pp. 329–351.

[21] Medina-Borja, A., Triantis, K., 2014. Modeling social services performance: a four-stage DEA approach to evaluate fundraising efficiency, capacity building, service quality, and effectiveness in the nonprofit sector. Annals of Operations Research 221 (1), 285–307.

[22] Murray, C. J., Evans, D. B., 2006. Health systems performance assessment. Office of Health Economics.

[23] Pritchett, L., Filmer, D., 1999. What education production functions really show: a positive theory of education expenditures. Economics of Education review 18 (2), 223–239.

[24] RTO/NATO, 2003. Handbook on long term defence planning. RTO-SAS-025 69.

[25] Ruggiero, J., 2000. Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. Journal of the Operational Research Society 51 (8), 906–912.

[26] Ruggiero, J., 2004. Performance evaluation in education. In: Handbook on data envelopment analysis. Springer, pp. 323–346.

[27] Saaty, T. L., 1988. What is the analytic hierarchy process? Springer.

[28] Salomon, J. A., 2003. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Population Health Metrics 1 (1), 12.

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

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

[31] Simar, L., Wilson, P. W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of econometrics 136 (1), 31–64.

[32] Simar, L., Wilson, P. W., 2011. Two-stage DEA: caveat emptor. Journal of Productivity Analysis 36 (2), 205–218.

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

[34] Stiglitz, J. E., 1975. The theory of screening, education, and the distribution of income. The American Economic Review 65 (3), 283–300.

[35] Takamura, Y., Tone, K., 2003. A comparative site evaluation study for relocating Japanese government agencies out of Tokyo. Socio-Economic Planning Sciences 37 (2), 85–102.

[36] Timmer, C. P., 1971. Using a probabilistic frontier production function to measure technical efficiency. The Journal of Political Economy, 761–794.

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

[38] Worthington, A. C., 2001. An empirical survey of frontier efficiency measurement techniques in education. Education Economics 9 (3), 245–268.