UiO : **University of Oslo**

Torbjørn Hanson

Four essays on military economics: Efficiency, trust and risk preferences in the armed forces

Thesis submitted for the degree of Philosophiae Doctor

Department of Economics
Faculty of Social Sciences

Norwegian Defence Research Establishment (FFI)

2019

© **Torbjørn Hanson, 2019**

*Series of dissertations submitted to the
Faculty of Social Sciences, University of Oslo
No. 768*

ISSN 1564-3991

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

Acknowledgements

I am grateful to the Norwegian Ministry of Defence for financing my work under the project “Cost-efficiency in Defence” (KOSTER) at the Norwegian Defence Research Establishment (FFI). It is a pleasure to acknowledge the contribution of a number of people during the work of this thesis. First I would like to thank my supervisor Finn R. Førsund for his indispensable help, constructive feedback and comments. I am very grateful for the way he has guided me into the academic world, his network and introduced me to many great people in the field. Thanks also to my co-supervisor Sverre A.C. Kittelsen for useful comments and discussions. A special thanks to Petter Y. Lindgren for discussions and suggestions for the cover chapter, and to Erik A. Hanson for hosting me at the Mathematics department (UiB) and for his Matlab programming support. Thanks to my boss at FFI, Sverre Kvalvik, for great support and for excellent management of the project, and to Espen Berg-Knutsen for his support and motivation for starting up the phd. Thanks also to Tore Nilssen at the Department of Economics for guidance along the way. Øivind Schøyen has contributed with many useful comments, ideas and suggestion during his stay at FFI. Knut Are Aastveit has contributed with many important thoughts, suggestions and motivational words. I am grateful to both of them for their support. I would also like to thank my co-authors Henning Finseraas, Åshild A. Johnsen, and Gaute Torsvik for an inspirational partnership, and in particular Andreas Kotsadam for useful comments on my work. Thanks a lot to my family, friends, current and former colleagues at FFI, for their support throughout the writing of this thesis. Special thanks to my wife Hanne, my sons Aksel and Jonas, my mother Anne and father Tor for all their love and support. This achievement could not have been possible without you.

• **Torbjørn Hanson**

Oslo, June 2019

List of Papers

Paper I

Hanson, T. ‘Efficiency and productivity in the operational units of the armed forces: A Norwegian example’. In: *International Journal of Production Economics* **179** (2016), pp. 12–23.

Paper II

Hanson, T. ‘Estimating output mix effectiveness: An applied scenario approach for the Armed Forces’. In: *Omega* **83** (2019), pp. 39–49.

Paper III

Finseraas, H., Hanson, T., Johnsen, Å. A, Kotsadam, A., Torsvik, G. ‘Trust, ethnic diversity, and personal contact: A field experiment’. In: *Journal of Public Economics* **173** (2019), pp. 72–84.

Paper IV

Hanson, T. ‘Risk preferences and adaptive behavior at the workplace: Evidence from the military’. *Unpublished*.

Contents

Acknowledgements	i
List of Papers	iii
Contents	v
1 Introduction	1
1.1 Military economics	1
1.2 Production theory and the efficient use of resources in the armed forces	1
1.3 The military organization and its personnel	3
1.4 Translation from the military setting to society	5
1.5 Related literature	5
1.6 The main findings of the thesis	9
1.7 Challenges related to research ethics	10
References	11
2 Summary	15
2.1 Efficiency and productivity in the operational units of the armed forces: A Norwegian example	15
2.2 Estimating output mix effectiveness: An applied scenario approach for the Armed Forces	16
2.3 Trust, Ethnic Diversity, and Personal Contact: A field ex- periment	18
2.4 Risk preferences and adaptive behavior at the workplace: Evidence from the military	19
References	20
Papers	24
I Efficiency and productivity in the operational units of the armed forces: A Norwegian example	25
II Estimating output mix effectiveness: An applied scenario approach for the Armed Forces	39
III Trust, ethnic diversity, and personal contact: A field ex- periment	53

IV	Risk preferences and adaptive behavior at the workplace: Evidence from the military	69
-----------	--	-----------

Chapter 1

Introduction

1.1 Military economics

Military economics is a subfield in economics, which in its extended form concerns all aspects of the economics of military organizations, defense production, security, disarmament, conflict and peace. Military economics embraces both macro and micro orientations. The former concerns macro themes such as the relationship between macro variables (e.g. economic growth and unemployment), external threats, and military spending (see for instance Sandler and Hartley, 1995), while the latter concerns the micro (economic) foundation of military actors' behavior. The microeconomic version of military economics draws heavily on other microeconomic literatures, but adjusts the theories and methods to the specific technological, institutional, and cultural context in which military organizations, units and individuals operate in. Indeed, military organization and production share similarities with many other sectors of the economy, but the specific mix of characteristics arguably makes the microeconomics of military organizations unique.

This thesis places itself firmly within the microeconomic version of military economics. Specifically, the two first papers concerns the production of military outputs and outcomes, while the two latter papers deal with a distinct theme within labor and personnel economics, namely the effect on military conscripts attitudes and risk preferences by exposure to peers in a military setting. In order to contextualize the thesis contributions, I will therefore in the remainder of this background section discuss first the specific characteristics of the military production, and then the unique labor and personnel conditions and policies of military organizations. The rest of the Chapter first reviews the most important studies that the thesis papers build upon, develop and also challenge, before it turns to presenting the contributions of the thesis. Finally, in Chapter 2, I sum up each paper.

1.2 Production theory and the efficient use of resources in the armed forces

With slow growth rates, ageing populations, and burgeoning public debt figures, efficient use of public budgets has become a central concern for governments in mature economies world-wide. Even in petroleum-rich Norway, the government emphasizes the need to continuously strive to produce publicly provided services in a more efficient manner. The Nato target of two percent of gross domestic product (GDP), reinforced by President Trump on several occasions lately, places additional strain on the current and future priorities of governments in Western

Europe. In Norway, increases in the defense budget are empirically associated with explicit measures and targets for efficient use of resources within the defense sector. Furthermore, there is a mutual agreement between the parliament and the defense sector that increases in the budgets are legitimated through well documented efficiency gains.

Defense production is characterized by several economic aspects that make it particularly difficult for policymakers and other stakeholders to identify the armed forces' efficiency and effectiveness in transforming resources into preferred outcomes. Economics offer theory about these aspects, but also methods on how to overcome the difficulties in identifying and estimating the values involved. First of all, in contrast to many, if not most, goods and services in modern economies that are available in a market, the armed forces produce services which are classical examples of public goods not provided by markets. The fact that the services are not sold in markets leaves policymakers and citizens without information from a price mechanism in evaluating efficient use of resources. The lack of information resulting from the absence of prices is an aspect of production which the armed forces share with many other providers of public services. Consequently, we have seen the rise of a literature covering a wide range of methods for the assessment of efficient resource allocation if physical information on the services is available. Secondly, one key difference between most other publicly provided services, such as education, health care, and long-term care, is that defense production has no obvious (direct) recipients, e.g. patients or pupils. This means that the armed forces have few, if any, users that can provide feedback on the quantity and quality of services offered. Third, a key objective of the armed forces is to deter potential threats from hostile actions. In other words, one of the central benefits of an efficient and "powerful enough" defense production is to never be used. To evaluate whether the armed forces or other factors contribute to peace and stability is indeed a difficult task.

Fourth, Førsund, 2017 argues that "If outcomes are pure public goods it may be the case that the public does not demand the service outputs provided by the agency, but demand the outcomes themselves". This holds also for the production of military services. The public has preferences for the final outcomes the military organization contributes to, namely a nation's security, sovereignty, and peaceful relations in its core areas, but few citizens have demand for the specific troops, exercises, equipment, or the various activities at the service output level. Hence, a substantial asymmetry in information and competence exist between "experts" inside the defense sector and citizens. The public is to a large extent unable to evaluate what the armed forces do and how they perform. Because of the aforementioned, the evaluation of military production and whether the Forces deliver on their objectives, is left to a few number of internal "experts". When military output is "incomprehensible" and "unobservable" to the public, other "observable" factors could grasp the policy makers' attention. An example of such a factor is military presence in a limited geographical area. Military presence could for example result in more jobs within a local community and a boost to the local economy in general. Thus, the risk is that resource allocation within the armed forces is influenced extensively on matters included in consumers'

preferences, but not related directly to the armed forces objectives.

The final aspect related to military production and outcomes, also important for many other providers of public services, is the role of environmental variables. Environmental variables could make outcomes of military production outside the control of the production managers. This is also the case for outcomes in e.g. the health sector. The outcomes are not only reliant on the performance of domestic military organizations, but obviously as much on a powerful neighboring country or an alliance. Equivalently in the case of health care services, the outcome relies on the health status prior to intervention, as well as individual behavior, in addition to the health service outputs in question. I argue that economics provides important tools for evaluation of output, efficiency and effectiveness in the armed forces, and offers new insight into the above mentioned aspects. Important contributions from economics are among others Hartley, 2012 on defense output, the theoretical framework for measuring public sector effectiveness in Førsund, 2017, as well as two empirical contributions in the present thesis.

1.3 The military organization and its personnel

The armed forces are often among the largest employers in Nato countries, even compared to private sector counterparts, and plays a significant role in the labor markets. One possible implication of being such a substantial actor, is that the armed forces are likely to face the same fundamental challenges and opportunities as the society in large. Janowitz, 1960 argues that military personnel has to be integrated into society because it is crucial for the military organization's legitimacy to share and represent the values they are set to defend. Moreover, gender equal labor market outcomes and integration of i.e. ethnic minorities in the armed forces, are examples of topics which could be crucial for legitimacy. In Norway, the white papers state that "... the Armed Forces' legitimacy is conditional on a development of the forces in step with society at large".¹ In particularly, the white papers stress that the Armed Forces should represent the Norwegian population to a larger extent when it comes to the share of female soldiers and soldiers with a ethnic minority background. This diversity is also regarded, in the white papers, as important for handling the Armed Forces objectives.

Certainly the military setting offers some unique characteristics that make the armed forces potentially fertile ground for understanding human attitudes and behavior. Theories and methods from labor and personnel economics shed light on the military organization. In contrast to civilian sectors, the armed forces offer for instance employees long contracts, high degree of rotation of personnel to positions internally in the organization, vast geographical spread in base locations demanding employees to be flexible in terms of commuting and moving, a complex education system designed to educate employees within the military organization, and an ability to screen and select a large number of potential candidates that participate in the military organization for an extended

¹See St.meld. nr. 36 (2006–2007) (white paper).

1. Introduction

period through the system of compulsory conscription. The specific mix of economic characteristics of recruiting, retaining, motivating and retiring military personnel varies between nations, but the fact is that military organization is a peculiar institution in terms of labor and personnel economics.

Asch and Warner, 2001 seminal paper on compensation and personnel policy in the United States Armed Forces is one of the few studies that explains the economics behind peculiar features of the military organization: relatively equal pay structure and generous pension systems in addition to an emphasis on up-or-out system (in contrast to its (large) counterparts in the private sector). Fundamental to military organizations is the lack of lateral recruitment; most, if not all, military personnel are recruited at a young age, receive education in the organization and have their military career within the organization. The often criticized generous pension system in the United States Armed Forces makes economic sense in the context of attracting, retaining, and motivating high-ability personnel, because staying in the military organization rewards the winners in the competition, inducing efforts from all employees in order to achieve higher ranking.

While the United States Armed Forces no longer impose conscription, other modern nations continue to do so. Conscription offers the Armed Forces a unique opportunity to recruit personnel from all new cohorts entering the labor force after gathering information about the conscripts' abilities and motivation in a real world job environment over a substantial period of time. In other words, conscripted personnel represent an important "recruitment well" for e.g. the Norwegian Armed Forces. As being employed in the armed forces is an "experience good" (Asch and Warner, 2001) – individuals need to try employment before knowing if military service is in accordance with their preferences – conscription represents an important information-gathering process also on behalf of the conscripts.

Of specific interest for my two papers on risk and trust is the inner workings of the system of compulsory conscription. Here, thousands of young adults of both genders and with diverse backgrounds are recruited every year to spend most of their days together, including eating and sleeping, and exposed to a daily life in a military setting. The military hierarchy and decision-making culture places substantial boundaries on the conscripts' room for maneuver, while the systematic exposure to in-group peers every day make conscripts a unique group to study when interested in peer effects as well as attitudinal and behavioral effects of exposure to others.

Gender differences and integration of ethnic minorities are examples of labor market outcomes studied in economics. The Armed Forces represents a promising arena for studies on these subjects based on experimental methods including randomized trials. An example is random assignment of peers in huge cohorts of conscripted personnel or entering freshmen, doing service in a controlled environment with an explicit enforcing authority. In this context, soldiers of private rank have equal social status, share common goals and need to cooperate to solve their tasks. The soldiers do not know each other prior to service and could be randomly assigned to exposure to a given characteristic of their peers. Carrell,

Hoekstra, and West, 2015; Carrell, Sacerdote, and West, 2013 and Finseraas et al., 2016 are novel examples of how experimental economics contributes within a military setting. The first piece uses cohorts of entering freshmen at the U.S. Air Force Academy to study whether academic performance can be improved through the systematic sorting of students into peer groups. The same context is exploited in the second paper to study how white males' attitude toward African American changes based on the black peers to whom they are exposed. In Finseraas et al., 2016 a field experiment is used to show that male soldiers who are randomly assigned to share room and work in a squad with female soldiers do not discriminate in a vignette experiment. Thus, the third and fourth paper of the present thesis join the series of contributions in economics on how to understand gender differences and integration of ethnic minorities, based on experimental evidence. Other influential contributions include, among others, the series of papers using the Vietnam era draft lottery to construct instrumental variables (Angrist, 1990; Angrist, 1991; Angrist, Chen, and Song, 2011).

1.4 Translation from the military setting to society

It is perhaps of even greater interest if studies of subjects in the armed forces could result in new insight and policy implications beyond the military setting. Thus, this thesis set out to exploit the fact that the armed forces mirrors the general society on several topics, in order to shed light on some more general research questions: Can methods for evaluating efficiency and effectiveness in the Armed Forces help to overcome some of the problems in evaluating other public sector service providers, e.g. the police, health and education sector? Can empirical findings in the Armed Forces explain more general gender differences in labor market outcomes? Can studies of exposure to ethnic minorities during service contribute to the understanding of how diversity affects trust?

1.5 Related literature

Output measures and efficiency in the Armed Forces

In the first chapter I study how to measure efficiency in the operational units of the Armed Forces. In previous literature the use of sophisticated methods is almost absent on this topic. Studies of efficiency in the defense sector in general started out in the early eighties.² While the military was introduced as a new and promising field for Data Envelopment Analysis (DEA) studies at the time, application to the core area of defense, the operational units, is still absent more than 30 years later. In fact, the approach in Charnes et al., 1984,

²I emphasize on the literature using DEA. 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, Cooper, and Rhodes, 1978 built on the earlier work of Farrell, 1957. There are few examples of other relevant methods, such as Stochastic Frontier Analysis (SFA), applied in the current setting.

on the efficiency of aircraft maintenance units, constitutes the best example in the literature.

At present there is no concise use of terms describing outputs and outcomes in the military. This is probably an important reason for the lack of studies. 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. Anagboso and Spence, 2009 find a capability approach more promising for measuring defense output. In this setting, a capability is the ability of the forces to pursue a particular course of action. They consider defense output to be the sum of the capabilities the armed forces provide. The literature recognizes, however, the importance of including both a quantity and a quality component in a measure (Anagboso and Spence, 2009; Schreyer, 2010).

Very few nations have the need nor the resources required for large scale duplication of military units. Studies of efficiency in operational units is therefore limited to small populations of homogeneous units. The low number of units limits the interpretation of the results, as typically nearly half of the units appear fully efficient. I suggest that this is a common and growing problem in measuring efficiency in the public sector. For example police districts, hospitals and tax agencies tend to shrink in numbers but increase in size, which could result in relatively small samples of homogeneous units. However, additional information based on resampling (e.g. the Simar and Wilson bootstrapping methods) can reduce the number of units estimated as fully efficient. This way, more informed decisions are made possible.

Effectiveness in the public sector

Another related challenge to the study of efficiency in the armed forces, is the study of effectiveness in the public sector. The distinction between the concepts efficiency and effectiveness is often expressed in the literature as “doing things right” and “doing the right things”, respectively (see e.g. Drucker, 1977; Fitz-Gibbon and Tymms, 2002 and Førsund, 2017). Hence, the reasons behind any inefficiencies in providing public goods could be related to the mix of outputs as well as technical inefficiencies in the production of the outputs. It is, however, not straight forward to determine whether “the right things” are produced in the public sector for services without market prices. Thus, there is a huge literature on how to evaluate the effectiveness of public sector service provision (e.g. health, education, transport, libraries). The usual practice in empirical studies is to study either how resources (inputs) are transformed to outcomes, or how outputs (inputs) are transformed to outcomes. But a method for how to evaluate the two transformation steps simultaneously is developed only recently. Førsund, 2017 suggests a consistent method for distinguishing between the two sources for inefficiencies. It is, however, not straight forward how to apply the method when estimating the effectiveness of public services empirically. In the second paper I study how we can overcome the empirical challenges in evaluating effectiveness of the Armed Forces as an example, when the above mentioned method is applied.

In addition to the usual estimation of technical efficiency, the method suggested in Førsund, 2017 implies estimation of a function mapping outputs to outcomes, and a function representing the policy makers preferences for the various outcomes. The role of environmental variables, such as GDP, population, territory, human capital, size of enemy forces etc. is crucial in studying the link between outputs and outcomes. Numerous studies apply a two-stage procedure to estimate the effect of such variables on outcomes (outputs) (Banker and Natarajan, 2008; Simar and Wilson, 2007; Simar and Wilson, 2011). The first step is usually to estimate the efficiency of the system of interest, where outputs are replaced with outcomes, before regressing the efficiency scores on any environmental variables in a second step.

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. 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. Schreyer, 2010 suggests, however, 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. To my knowledge, this approach is yet to be implemented empirically.

Trust and diversity

The effects of diversity on trust are essential to understand: When people trust each other, transaction costs are reduced, organizations run better, the need for formal regulation is reduced, governments provide services more efficiently, policy promises become more credible, financial systems develop better, and military effectiveness increases (Algan and Cahuc, 2013; Guiso, Sapienza, and Zingales, 2008a; Guiso, Sapienza, and Zingales, 2008b; Guiso, Sapienza, and Zingales, 2011; Rosen, 1995). Some argue, however, that diversity can lead to less social trust and more tension and conflicts (Alesina and La Ferrara, 2000; Putnam, 2007). If migration and ethnic diversity have dismantling effects on the social fabric of societies, it becomes important to find out if and how public policy can mitigate such problems. For instance, can tensions be reduced and trust enhanced if governments create arenas where different ethnic groups regularly encounter each other? Can social contact build trust? In the third paper my co-authors and I shed light to these questions by studying how close personal contact with minorities affects trust in a field experiment in the Norwegian Armed Forces, where soldiers are randomly assigned to rooms with or without ethnic minorities.

There are at least three important limitations to the existing literature on the effect of diversity on trust: The first limitation concerns biases arising from endogeneity issues. The worry that the correlations between diversity and trust are driven by selection, reverse causality, or both looms large in the previous

literature. People self-select into neighborhoods and controlling for selection by including observables is likely to be insufficient. To date, there is no study using exogenous variation to identify the causal effect of diversity on trust. The second shortcoming of the previous literature is a conceptual conflation of exposure and contact. While living in an area with many immigrants increases exposure, it does not necessarily increase contact. A consensus has emerged in social psychology that shallow exposure need not produce the same beneficial effects, instead it is likely to cause opposite effects due to competition about jobs, resources, and cultural hegemony (see Pettigrew, 1998 for a review). Well identified studies have shown, however, that close personal contact reduces prejudice (e.g. Bauer, Fiala, and Lively, 2018; Boisjoly et al., 2006; Burns, Corno, and La Ferrara, 2016; Carrell, Hoekstra, and West, 2015; Dahl, Kotsadam, and Rooth, 2018; Finseraas and Kotsadam, 2017; Finseraas et al., 2016), illustrating the danger of conflating exposure and contact. The third limitation regards the measurement of trust. Most literature on the effects of diversity on trust relies upon survey questions on general trust. There is a debate about what these questions really measure, and some argue that they correlate with trustworthiness rather than trust (Glaeser et al., 2000; Sapienza, Toldra-Simats, and Zingales, 2013). The shortcomings mentioned above are all addressed in the third paper of the present thesis.

Gender differences in risk preferences

Gender differences are found in the general population for several attributes including risk preferences. But differences seem to disappear among professionals. The proposed explanation is that individuals self select into professions according to their risk preferences (Croson and Gneezy, 2009). For example, gender differences in risk taking are negligible for mutual fund investors when controlled for investor knowledge (Dwyer, Gilkeson, and List, 2002). Is adaptive behavior at the workplace another explanation for the lack of differences in some professions, and are peer effects a relevant mechanism driving such adjustments in preferences? In the fourth paper of my thesis, I investigate these questions by studying adjustments in risk preferences among conscripted soldiers in the Norwegian Armed Forces, where each individual is randomly exposed to other soldiers' risk preferences.

Dwyer, Gilkeson, and List, 2002 is an example of a study where risk preferences are estimated after the subjects are exposed to the job environment. This design does not control for environmental factors related to adaption as there is no baseline of risk preferences prior to exposure. Of course, the preferences of each individual could be estimated the first day at work, but this is likely to be difficult in practice as long as workers are employed at different points in time. Ahern, Duchin, and Shumway, 2014 is an example of a study where a baseline is established. However, studying peer effects for financial risk preferences among MBA students, the context is likely to differ somewhat from the context of any real workplace.

Ahern, Duchin, and Shumway, 2014 find a positive peer effect in risk preferences, i.e. when an individual's randomly assigned peers are relatively more

risk averse prior to starting MBA, the individual is relatively more risk averse after the first year. On the other hand, several theories suggest that individuals who experience greater risks in their environment may have less tolerance for risk in other areas (Gollier and Pratt, 1996; Guiso and Paiella, 2008; Lee, 2008; Li, 2011; Quiggin, 2003). Moreover, research in both economics and psychology points to a potential relationship between trauma and risk preferences. Callen et al., 2014 study the effect of exposure to violence in Afghanistan on preferences for certainty. They find that individuals exposed to violence exhibit an increased preference for certainty. There is also empirical evidence from lab experiments showing that subjects exposed to a higher risk environment exhibit higher levels of risk aversion than those who face a moderate or low risk environment (He and Hong, 2015).

Several risk-taking measures have been developed from various measurement traditions. The various measures are usually referred to as either self-reported propensity measures, assessing stated preferences, incentivized behavioral measures, eliciting revealed preferences, and frequency measures assessing actual risky activities. Frey et al., 2017 study the existence of a general factor of risk preferences by integrating the multiple risk-taking measures above. They find that risk preferences have a psychometric structure which involves both a general, stable component that can account for about half of the explained variance, and a series of facets capturing more specific aspects of risk preferences. One implication of the findings is to consider a broader range of risk measures, including both domain specific and more general risk preferences, if a general factor of risk preferences is to be considered. Dohmen et al., 2011 is another study of how well different methods elicited risk preferences. They found that a general risk question performed fairly well in a field experiment with real money stakes. When tested specifically for different domains, the best predictor of behavior was the corresponding domain-specific question, but the general risk question still out-performed more complex methods. A cost-benefit perspective could favor a general risk question over more complex methods. Charness, Gneezy, and Imas, 2013 conclude that a crucial disadvantage of complex methods is that, depending on the population, a significant number of subjects will fail to understand the procedure. Furthermore, Charness, Gneezy, and Imas, 2013 state that simpler methods have the advantage of being more straight forward and capable of eliciting sensible risk preferences from a broader set of individuals, but warn researchers from extrapolating from such measures to other domains. They conclude that researchers should choose the method that is best suited for the particular question asked as well as the sample studied.

1.6 The main findings of the thesis

In the first paper, I estimate the efficiency and productivity for eleven units of the Norwegian Home Guard. Based on the estimated efficiency scores I point to one single unit as candidate for best practice. Furthermore, I find that 64 % of the units have significantly improved productivity during the four years

studied. Comparing changes in cost to development in productivity, most of the units are characterized as improving in productivity while saving costs. Only one single unit is found to have a clearly unfavorable development, characterized as decreasing in productivity while costs are increasing. As the total use of resources has declined by 9.5 % in terms of modeled input, the results suggest a Home Guard capable of adjusting to the downward pressure on spending experienced during the period studied.

In the second paper, I find that the overall effectiveness in a sample of operational units from the Norwegian Armed Forces has decreased during the four years studied. The drop in effectiveness is not related to technical inefficiencies, but rather is the result of an inefficient mix of outputs by the policy makers. Allowing for differences in discounting, the model is estimated for two alternative specifications of the policymakers' preference function. The policy makers perform best if their preferences are nearsighted. The negative trend in output mix effectiveness is, however, common to all three specifications of the preference function.

We find, in the third paper, that close personal contact with minorities increases trust towards a generic immigrant. We replicate the result that individuals coming from more ethnically diverse areas trust minorities less, but random assignment to interact with minority soldiers removes this negative correlation.

Finally, in the fourth paper I find that soldiers on average change their domain specific risk preferences to dislike risk more during the service. There is, however, no change in the measure of general risk preferences. There is evidence for gender differences at first day of service for all of the four different risk measures (domain specific, profession related, physical and general risk measures). But the difference is not significant if drop outs during service are excluded from the sample. Moreover, I fail to reject the hypothesis that female and male risk preferences are identical during the period studied. Contrary to what I expected, I find a negative peer effect for domain specific risk preferences, i.e. on average soldiers diverge from the mean risk preference of their peers in which they share room with. One possible explanation of the result is that exposure to risky behavior from peers in the domain specific setting, makes the respondent more risk averse.

1.7 Challenges related to research ethics

Using data on individuals, as I do in this thesis, will always represent some challenges related to research ethics. In addition, data restricted due to confidentiality, as is often the case for micro data on military performance, could hamper any attempts on replicating some of the results in the present thesis.

The last two papers use various survey data collected from conscripted personnel in the Norwegian Armed Forces during their 12 months of service. When conducting such surveys researchers are obliged to obtain free and informed consent from the informants (soldiers). The informants shall be informed about

the research project, consequence of participation and that they can withdraw at any time without any negative consequences (NESH, 2010). We conducted three different surveys, the first day of service, after eight weeks, and finally after nine months. The first day of service starts at a military camp close to Oslo. At the camp, the soldiers go through a program of medical and psychological testing and they fill out the survey questionnaire. We stressed that the survey was independent research and unrelated to the screening process, and that members of the Norwegian Army did not have access to individuals' survey responses, participation nor withdrawals. The questionnaire was filled out by using a tablet. The soldiers could withdraw at any time without notice, and use the tablet for other purposes if they wanted.

Data on individuals shall not be stored any longer than the duration of the research project. Once the data has served its original purpose it shall be deleted (NESH, 2010). This principle could be in conflict with any future replication of my research results by other researcher. The same problem could occur also related to the restricted data on military performance. Data on inputs and outputs are available on an aggregated level, which makes a replication of estimated efficiency scores possible. However, output data on a micro level is only available to researchers holding the required security clearance. Thus, in practice, the data is not available to researchers outside Norway. This is one of the main challenges related to research ethics researchers face when studying data generated within the Armed Forces.

References

- Ahern, K. R., Duchin, R., and Shumway, T. (2014). "Peer effects in risk aversion and trust". In: *Review of Financial Studies*.
- Alesina, A. and La Ferrara, E. (2000). "Participation in heterogeneous communities". In: *The quarterly journal of economics* vol. 115, no. 3, pp. 847–904.
- Algan, Y. and Cahuc, P. (2013). "Trust and growth". In: *Annu. Rev. Econ.* Vol. 5, no. 1, pp. 521–549.
- Anagboso, M. and Spence, A. (2009). "Measuring defence". In: *Economic & Labour Market Review* vol. 3, no. 1.
- Angrist, J. D. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records". In: *The American economic review* vol. 80, no. 3, pp. 313–336.
- Angrist, J. D. (1991). "The draft lottery and voluntary enlistment in the Vietnam era". In: *Journal of the American statistical Association* vol. 86, no. 415, pp. 584–595.
- Angrist, J. D., Chen, S. H., and Song, J. (2011). "Long-term consequences of Vietnam-era conscription: New estimates using social security data". In: *American Economic Review* vol. 101, no. 3, pp. 334–38.

- Asch, B. J. and Warner, J. T. (2001). “A theory of compensation and personnel policy in hierarchical organizations with application to the United States military”. In: *Journal of Labor Economics* vol. 19, no. 3, pp. 523–562.
- Banker, R. D. and Natarajan, R. (2008). “Evaluating contextual variables affecting productivity using data envelopment analysis”. In: *Operations Research* vol. 56, no. 1, pp. 48–58.
- Bauer, M., Fiala, N., and Lively, I. (2018). “Trusting former rebels: An experimental approach to understanding reintegration after civil war”. In: *The Economic Journal* vol. 128, no. 613, pp. 1786–1819.
- Boisjoly, J. et al. (2006). “Empathy or antipathy? The impact of diversity”. In: *American Economic Review* vol. 96, no. 5, pp. 1890–1905.
- Burns, J., Corno, L., and La Ferrara, E. (2016). “Interaction, prejudice and performance. Evidence from South Africa”. In: *Working paper*.
- Callen, M. et al. (2014). “Violence and risk preference: Experimental evidence from Afghanistan”. In: *American Economic Review* vol. 104, no. 1, pp. 123–48.
- Carrell, S. E., Hoekstra, M., and West, J. E. (2015). *The impact of intergroup contact on racial attitudes and revealed preferences*. Tech. rep. National Bureau of Economic Research.
- Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013). “From natural variation to optimal policy? The importance of endogenous peer group formation”. In: *Econometrica* vol. 81, no. 3, pp. 855–882.
- Charnes, A., Cooper, W. W., and Rhodes, E. (1978). “Measuring the efficiency of decision making units”. In: *European journal of operational research* vol. 2, no. 6, pp. 429–444.
- Charnes, A. et al. (1984). “A developmental study of data envelopment analysis in measuring the efficiency of maintenance units in the US Air Forces”. In: *Annals of Operations Research* vol. 2, no. 1, pp. 95–112.
- Charness, G., Gneezy, U., and Imas, A. (2013). “Experimental methods: Eliciting risk preferences”. In: *Journal of Economic Behavior & Organization* vol. 87, pp. 43–51.
- Croson, R. and Gneezy, U. (2009). “Gender differences in preferences”. In: *Journal of Economic literature*, pp. 448–474.
- Dahl, G., Kotsadam, A., and Rooth, D.-O. (2018). *Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams*. Tech. rep. National Bureau of Economic Research.
- Dohmen, T. et al. (2011). “Individual risk attitudes: Measurement, determinants, and behavioral consequences”. In: *Journal of the European Economic Association* vol. 9, no. 3, pp. 522–550.
- Drucker, P. F. (1977). *An introductory view of management: instructor’s manual*. Harper and Row.
- Dwyer, P. D., Gilkeson, J. H., and List, J. A. (2002). “Gender differences in revealed risk taking: evidence from mutual fund investors”. In: *Economics Letters* vol. 76, no. 2, pp. 151–158.
- Farrell, M. J. (1957). “The measurement of productive efficiency”. In: *Journal of the Royal Statistical Society. Series A (General)* vol. 120, no. 3, pp. 253–290.

- Finseraas, H. and Kotsadam, A. (2017). “Does personal contact with ethnic minorities affect anti-immigrant sentiments? Evidence from a field experiment”. In: *European Journal of Political Research* vol. 56, no. 3, pp. 703–722.
- Finseraas, H. et al. (2016). “Exposure to female colleagues breaks the glass ceiling—Evidence from a combined vignette and field experiment”. In: *European Economic Review* vol. 90, pp. 363–374.
- Fitz-Gibbon, C. T. and Tymms, P. (2002). “Technical and ethical issues in indicator systems”. In: *education policy analysis archives* vol. 10, p. 6.
- Førsund, F. R. (2017). “Measuring effectiveness of production in the public sector”. In: *Omega* vol. 73, pp. 93–103.
- Frey, R. et al. (2017). “Risk preference shares the psychometric structure of major psychological traits”. In: *Science advances* vol. 3, no. 10, e1701381.
- Glaeser, E. L. et al. (2000). “Measuring trust”. In: *The quarterly journal of economics* vol. 115, no. 3, pp. 811–846.
- Gollier, C. and Pratt, J. W. (1996). “Risk vulnerability and the tempering effect of background risk”. In: *Econometrica: Journal of the Econometric Society*, pp. 1109–1123.
- Guiso, L. and Paiella, M. (2008). “Risk aversion, wealth, and background risk”. In: *Journal of the European Economic association* vol. 6, no. 6, pp. 1109–1150.
- Guiso, L., Sapienza, P., and Zingales, L. (2008a). “Social capital as good culture”. In: *Journal of the European Economic Association* vol. 6, no. 2-3, pp. 295–320.
- (2008b). “Trusting the stock market”. In: *the Journal of Finance* vol. 63, no. 6, pp. 2557–2600.
- (2011). “Civic capital as the missing link”. In: *Handbook of social economics*. Vol. 1. Elsevier, pp. 417–480.
- Hartley, K. (2012). “Conflict and Defence Output: An Economic Perspective”. In: *Revue d’économie politique* vol. Vol. 122, no. 2, pp. 171–195.
- He, T.-S. and Hong, F. (2015). “Exposure to risk and risk aversion: A laboratory experiment”. In: *Available at SSRN 2401438*.
- Janowitz, M. (1960). *The Professional Soldier: A Social and Political Portrait*.
- Lee, J. (2008). “The effect of the background risk in a simple chance improving decision model”. In: *Journal of Risk and Uncertainty* vol. 36, no. 1, pp. 19–41.
- Li, J. (2011). “The demand for a risky asset in the presence of a background risk”. In: *Journal of Economic Theory* vol. 146, no. 1, pp. 372–391.
- NESH (2010). *Guidelines for research ethics in the social sciences, law and the humanities*. National Committees for Research Ethics in Norway.
- Pettigrew, T. F. (1998). “Intergroup contact theory”. In: *Annual review of psychology* vol. 49, no. 1, pp. 65–85.
- Putnam, R. D. (2007). “E pluribus unum: Diversity and community in the twenty-first century the 2006 Johan Skytte Prize Lecture”. In: *Scandinavian political studies* vol. 30, no. 2, pp. 137–174.
- Quiggin, J. (2003). “Background risk in generalized expected utility theory”. In: *Economic Theory* vol. 22, no. 3, pp. 607–611.
- Rosen, S. P. (1995). “Military effectiveness: Why society matters”. In: *International Security* vol. 19, no. 4, pp. 5–31.

- Sandler, T., Hartley, K., et al. (1995). *The economics of defense*. Cambridge University Press.
- Sapienza, P., Toldra-Simats, A., and Zingales, L. (2013). “Understanding trust”. In: *The Economic Journal* vol. 123, no. 573, pp. 1313–1332.
- Schreyer, P. (2010). “Measuring the Production of Non-Market Services”. In: *Price Indexes in Time and Space, Contributions to Statistics*. Springer-Verlag Berlin Heidelberg.
- Simar, L. and Wilson, P. W. (2007). “Estimation and inference in two-stage, semi-parametric models of production processes”. In: *Journal of econometrics* vol. 136, no. 1, pp. 31–64.
- Simar, L. and Wilson, P. W. (2011). “Two-stage DEA: caveat emptor”. In: *Journal of Productivity Analysis* vol. 36, no. 2, pp. 205–218.

Chapter 2

Summary

2.1 Efficiency and productivity in the operational units of the armed forces: A Norwegian example

Despite the fact that most nations spend a considerable part of their gross domestic product (GDP) on defense, no previous study has estimated productivity and efficiency in the core area of the armed forces, operational units. As my review of the literature shows, previous studies using sophisticated methods are solely concentrated around support functions like maintenance and recruitment. The main purpose of the paper is to cover this gap in the literature.

I have identified three possible reasons for the lack of studies on efficiency of operational units: 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. At present there is no concise use of terms describing outputs and outcomes in the military. Hence, a proper definition of measures is needed in order to refine the study to the concept of efficiency and not effectiveness. In other words we have to come up with a concept for military unit output which deals with the distinction of outputs from outcomes in line with the saying that efficiency is a question of “doing things right” and effectiveness is a question of “doing the right things”. In my suggested framework the armed forces buy resources in markets and transform them to outputs in form of combat ready operational military units. This production process typically includes activities like training troops, sailing or flight hours. In this step an efficient use of resources is assured, while a second step deals with the effective mix of military units in order to realize outcomes such as sovereignty, crises management and other goals the armed forces are set to provide the society – the effectiveness of the armed forces.

Given this framework the output of an operational unit has to be specified in a manner which makes it possible to measure it empirically. Hence, for each effective output from operational units there exist an output measure operationalizing the output. I model the measure as a function of quantity, e.g. the number of troops, and quality, e.g. the proficiency level of each troop. 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 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.

2. Summary

After developing this generic measure of output for the operational military units, I specify the measure for units of the Norwegian Home Guard. 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 units (districts) are located in all geographical regions in Norway with tasks related to either naval, air force or land activities. A district consists of a District Staff and a number of various troop types. The straightforward interpretation of the troop output measure applied is: the number of soldier equivalent quality adjusted personnel produced during a given period of time.

I specify a Data Envelopment Analysis (DEA) model in order to estimate the Farrell technical efficiency score for 11 Home Guard units over the years 2008–2011. This results in 11 observations each year or 44 observations after pooling the data, assuming the technology to be stationary. Bootstrapping of the efficiency scores is used in order to obtain unbiased estimates. If emphasis is put on positive development in relative performance and to candidates which perform best in the last year of the period, I am able to identify a single candidate for best practice among the 11 units in the Home Guard. Considering the development in productivity I find that seven of the units (64 %) have significantly improved productivity during the four year period. 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. As the total use of resources has declined by 9.5 % in terms of modeled input, the results suggest a Home Guard capable of adjusting to the downward pressure on spending experienced during the period studied.

2.2 Estimating output mix effectiveness: An applied scenario approach for the Armed Forces

The ability of public sector policy makers to prioritize has a huge impact on the effectiveness of public service provision. Public services can take the form of final outputs demanded by consumers or of intermediate outputs contributing to a process of realizing the higher goals of society, in which I argue the latter is the case for the armed forces. In doing the right things, policy makers choose a mix of intermediate outputs maximizing their preference value for public service outcomes, while managers do things right when responsible for producing outputs efficiently. This distinction enables us to pinpoint important reasons for inefficiencies in the provision of public services. 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. The aim of the paper is to estimate output mix effectiveness where other methods fail due to long time lags and lack

of variation in the variation. The armed forces is a typical example of this case, as there is a potential long time lag from any changes in the defense budget might result in violation of the country's sovereignty. Furthermore, considering binary outcome variables like "peace or no peace" gives little meaning in evaluating the effectiveness of production in the defense sector. Moreover, externalities in the transformation process from outputs to outcomes, outside the control of managers and decision makers, is difficult to model. The size of enemy forces is a relevant example of such externalities for the armed forces.

The typical mistake in the literature is to either evaluate the transformation of resources (inputs) to outcomes without considering the mix of intermediate outputs, or to consider outputs as the inputs in the transformation from inputs (now outputs) to outcomes without evaluating the efficiency in the production of outputs. While the first model rule out the identification of possible inefficiencies from output mix by definition, the latter model could result in a biased measure of effectiveness if the production of an output is (technical) inefficient. To avoid the practice of considering the transformation from inputs to outcomes as a single process, I elaborate on the approach in Førsund, 2017. This implies distinguishing between two transformational steps, which requires the estimation of a function mapping outputs to outcomes in addition to the usual estimation of a production technology transforming inputs to outputs. The estimation of a mapping function is, however, not straight forward in the setting noted above of long time lags and lack of variation in variables. In response to this, I 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. Moreover, the approach involves replacing objectives with scenarios, each representing a unique vector of fixed environmental variables. When multiple objectives or outcomes are evaluated, a preference function has to be introduced representing politicians' or policy makers' preferences.

Using the Norwegian Armed Forces as an example I derive mapping functions for a sample of 12 operational units in the Submarine Force and the Home Guard. It was, however, difficult to present the scenarios to policy makers in a manner which could be used to generate data for estimation of preferences. I therefore relied on the assumption that all objectives are of equivalent importance to the policy makers as a baseline.

I 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, but rather is the result of an inefficient mix of outputs by the policy makers. I suggest that 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. 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 the baseline preference structure. The finding of a downward trend

2. Summary

is supported also by the farsighted preference structure. In fact the policy makers have the lowest relative performance in this preference structure. Sensitivity analysis suggests my results to be robust in trends, but sensitive in size.

2.3 Trust, Ethnic Diversity, and Personal Contact: A field experiment

Co-authored with Henning Finseraas, Åshild A. Johnsen, Andreas Kotsadam and Gaute Torsvik

The effects of diversity on trust are essential to understand: When people trust each other, transaction costs are reduced, organizations run better, the need for formal regulation reduces, governments provide services more efficiently, policy promises become more credible, and financial systems develop better (Algan and Cahuc, 2013; Guiso, Sapienza, and Zingales, 2008a; Guiso, Sapienza, and Zingales, 2008b; Guiso, Sapienza, and Zingales, 2011). Previous research has found a negative relationship between trust and the level of ethnic diversity in the respondents' area of residence, leading to the conflict theory (Alesina and La Ferrara, 2000; Dinesen and Sønderskov, 2015; Putnam, 2007). Furthermore, the constrict theory suggests that diversity may also reduce trust within the majority group. If migration and ethnic diversity have negative effects on the social fabric of societies, it becomes important to find out if and how public policy can mitigate such problems. As ethnic diversity increases in the Armed Forces along the trends in society, similar effects could also occur inside the defense sector, and in the worst case the policy could represent a cost in form of reduced military capabilities. However, conflict is not the only potential outcome of ethnic diversity. The contact theory (Allport, Clark, and Pettigrew, 1954) argues that personal contact with members of out-groups can reduce prejudice and misperceptions, and thereby increase trust. In fact the Armed Forces could represent a promising context for studies of the various effects of diversity on trust. In e.g. the army, soldiers of private rank have equal social status, they share the common goals of their unit, they need to cooperate to solve their tasks, and they have extensive contact with the soldiers they share room with which usually also is identical to their squad.

In the present paper my co-authors and I investigate the effect on ethnic majority individuals' in-group and out-group trust from personal contact with ethnic minority individuals. We study the effect by randomizing soldiers in the Norwegian Army to rooms during boot camp, implying that soldiers from the majority group (ethnic Norwegian soldiers) are randomized to share living quarters with at least one minority member, while others have only members of the majority group as roommates. At the end of the boot camp we ran a trust game with monetary stakes. This design handles a major limitation of previous studies of ethnic diversity and trust: the inability to control for selection and reverse causality. As to our knowledge this is the first study with a research design allowing for a causal identification of how close contact between majority

and minority individuals affects trust.

We find that individuals randomly assigned to close personal contact with minority soldiers send more money (indicating more trust) to the person with a name signaling minority origin, in line with the contact theory. Random contact with minority soldiers does not affect the amount sent to the person with a Norwegian name, in contrast to what constrict theory predicts. Next we find a negative association between immigrant share in the home municipality and trust in the out-group (in the minority member but not in the majority member), which is consistent with the conflict theory. We further find that the negative relationship between immigrant share in the home municipality and out-group trust is annulled for the soldiers that were randomly assigned to close personal contact with a minority soldier.

We conclude that social integration involving personal contact can reduce negative effects of ethnic diversity on trust. The policy implications of the results depend on subjective opinions on the external validity of the findings. In particular, three factors are important in this respect. Firstly, our sample consists of special representatives of the Norwegian population as only about one in six men serve military service. The military thereby select people based on ability and motivation. Secondly, the soldiers are exposed to a highly selected set of immigrants, likely to be better integrated. Thirdly, and perhaps most important, the setting under which contact occurred is very special. Although the context of our study is in part a necessity for deriving clear theoretical expectations, and while it assures a strong internal validity, it restricts external validity to contexts with some similarity to ours.

2.4 Risk preferences and adaptive behavior at the workplace: Evidence from the military

Gender differences in risk preferences are found in the general population, but tend to disappear among professionals. The absence of gender gaps among professionals could be due to self selection or adaptive behavior at the workplace. In this paper I introduce a research design which could identify the possible mechanism of adaptive behavior by including a baseline measure of risk preferences before individuals enter the "risky" profession. A better understanding of the mechanisms behind the gender gap in risk preferences is crucial for the Armed Forces in order to counter the disadvantage from a "natural limit" on the female population in recruiting otherwise qualified personnel. If risk preferences adapt during service, reducing the gender gap, this could have important implications for recruitment policies in general and for the Armed Forces in particular.

The Armed Forces represents a unique case allowing for the proposed research design: Thousands of new conscripts enter at the same time each year making a baseline measure for risk preferences possible. I developed a survey completed by conscripts during their first day of service while knowledge and experience of the military and the service is at its minimum. Results from this first wave are

2. Summary

considered my baseline. Eight weeks later, after completing the initial training period (boot camp), a follow-up survey is conducted in a second wave. Finally, after nine months, towards the end of the service, a third wave completes the survey period. Furthermore, the setting also allowed for testing "peer influence" as a relevant mechanism behind changes in risk preferences. As the soldiers were randomly assigned to rooms during the boot camp, studies of causal peer effects is possible.

The surveys include four different risk measures to elicit domain specific, profession specific (police, firefighters, rescue, soldiers, etc.), physical (physical injury) and general risk preferences. I find that soldiers on average dislike domain specific and domain related risk more during the service. But there is no change in the measure of general risk preferences. Despite the findings of gender differences at baseline for all of the four different risk measures, I find no difference if we only consider the soldiers completing our survey after nine months of service. Given the high drop out rates during the first weeks of service, my results could suggest that self selection eliminates gender differences. On the other hand, the findings of a significant gender difference for domain specific risk preferences after nine months, shows that other mechanisms could be at work. I fail, however, to reject the hypothesis that the estimated coefficients for gender differences are equal in size during the period studied.

Furthermore, I find a negative peer effect for domain specific risk preferences, i.e. on average soldiers diverge from the mean risk preference of their peers in which they share room with. One possible explanation of the result is that exposure to risky behavior from peers in the domain specific setting, makes the respondent more risk averse. My results suggest that policy makers lower their odds for increasing the share of females if they rely on the assumption that risk preferences of female soldiers adapt in a favorable direction during service. Policies could rather raise awareness of possible gender differences in preferences during service and stress the impact of deliberated communication when it comes to undesirable behavior and events related to risk.

References

- Alesina, A. and La Ferrara, E. (2000). "Participation in heterogeneous communities". In: *The quarterly journal of economics* vol. 115, no. 3, pp. 847–904.
- Algan, Y. and Cahuc, P. (2013). "Trust and growth". In: *Annu. Rev. Econ.* Vol. 5, no. 1, pp. 521–549.
- Allport, G. W., Clark, K., and Pettigrew, T. (1954). "The nature of prejudice". In:
- Dinesen, P. T. and Sønderskov, K. M. (2015). "Ethnic diversity and social trust: Evidence from the micro-context". In: *American Sociological Review* vol. 80, no. 3, pp. 550–573.
- Førsund, F. R. (2017). "Measuring effectiveness of production in the public sector". In: *Omega* vol. 73, pp. 93–103.

- Guiso, L., Sapienza, P., and Zingales, L. (2008a). “Social capital as good culture”. In: *Journal of the European Economic Association* vol. 6, no. 2-3, pp. 295–320.
- (2008b). “Trusting the stock market”. In: *the Journal of Finance* vol. 63, no. 6, pp. 2557–2600.
- (2011). “Civic capital as the missing link”. In: *Handbook of social economics*. Vol. 1. Elsevier, pp. 417–480.
- Putnam, R. D. (2007). “E pluribus unum: Diversity and community in the twenty-first century the 2006 Johan Skytte Prize Lecture”. In: *Scandinavian political studies* vol. 30, no. 2, pp. 137–174.

Papers

Paper I

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

Torbjørn Hanson

Published in *Int. J. Production Economics*, 2016, volume 179, pp. 12–23.

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

Torbjørn Hanson

Norwegian Defence Research Establishment (FFI), Instituttveien 20, PO Box 25, NO-2027 Kjeller, Norway

ARTICLE INFO

Article history:

Received 11 May 2015

Received in revised form

18 May 2016

Accepted 20 May 2016

Available online 26 May 2016

Keywords:

Military

Productivity

Efficiency

DEA

Bootstrap

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.

© 2016 Elsevier B.V.. All rights reserved.

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.¹ The fact 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.²

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

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

E-mail address: torbjorn.hanson@ffi.no

¹ 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.

² 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).

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.³ The model is specified for units of one branch or service of the Norwegian Armed Forces, the Home Guard.⁴ 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 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.⁵ 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 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

³ 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 course not found in such assessments.

⁴ 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 (300–1000 officers and 1500–4500 soldiers), except for the personnel in the District Staff who are full time employed in the Armed Forces (around 50 people).

⁵ However, Owen (1994) shows that international benchmarking of military manpower ratios is possible across 15 different countries.

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) ^a	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

^a Paper not available online, nor through the author's library.

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.⁶ 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.

Operational units are likely to be more heterogeneous 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.⁷ 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 (3×14) 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.

⁶ The type of aircraft was introduced as a weighting factor of flying hours.

⁷ 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 were 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.

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 Fig. 1. The first aspect is about the services produced by employing resources by the institution or state agency in question. The second aspect is about the effectiveness of these services, i.e. the impact the services make on the objectives that motivate producing the services in the first place. This distinction leads to the saying that efficiency is a question of “doing things right” and effectiveness is a question of “doing the right things”. In the following we will discuss whether this model is straightforward to apply to the military.

Resources, in the upper left of Fig. 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.⁸

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.⁹ 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 the concepts output and outcome is somewhat diverse.

⁸ 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.

⁹ 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.

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

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”.¹⁰ 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 , $j = 1, \dots, J$, is producing the output y^i , $i = 1, \dots, I$. 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. a reinforcement troop where l is the number of soldiers in the troop and P their proficiency level

$$y^{ij} = y_{ij}(l_{ij}, P_{ij}) \quad (1)$$

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 (Spring 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.

¹⁰ 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 is, however, straight forward to specify the model also for navy and air force units.

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 provide no force without a certain level of quality, i.e. a given training level. In Section 3.2 we will return to an example of a full 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.¹¹

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 considering 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 Eq. (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 300–1000 officers and 1500–4500 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 Eq. (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 $j = 1, \dots, 11$ Home guard districts

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

The proficiency index is constructed from the three levels of proficiency reported, where each level is met after passing given standards. Fully operational capacity 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.¹²

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.¹³ 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

¹² See e.g. Romer (2006) referring to the product of knowledge and labor as effective labor.

¹³ Main equipment for Home Guard personnel are assault rifles and basic infantry gear. Some troops have additional recoils 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.

¹¹ 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.

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 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 Eq. (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 $n^{2/(p+q)}$ 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 (Bowlins, 1987).¹⁴ 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.¹⁵

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 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.

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

¹⁴ We have, however, not succeeded in finding neither any theoretical nor empirical evidence for the rules of thumb.

¹⁵ Kittelsen (1999) shows that the extent of correlation is clearly important when testing the relevance of an additional input in the model.

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.

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 $j = 1, \dots, 11$ for the years 2008–2011.¹⁶ Productivity development over the four years we investigate by the Malmquist productivity index.¹⁷ We 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.¹⁸ Four units are estimated fully efficient evaluated relatively to the intertemporal frontier, reported in Fig. 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

Fig. 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.

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.

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 Fig. 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 Fig. 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

¹⁶ The DEA specification is given in the Appendix.

¹⁷ All estimates and bootstraps were carried out by the FrischNP3.4 software, developed by The Ragnar Frisch Centre for Economic Research.

¹⁸ Using both an intertemporal frontier and a CRS assumption is equivalent to using a technical productivity measure with optimal scale as the reference frontier.

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)

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.

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

Fig. 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.

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 $q_{1-\alpha/2}$. The confidence interval is then given by

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

From Eq. (3) we have estimated the 95% confidence intervals for the pooled efficiency scores. The confidence intervals are outlined by the bold lines in Fig. 3.

The bars in Fig. 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 Fig. 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 Figs. 4 and 5. This way of reasoning adds more candidates for best practice. By considering bias corrected scores and their standard errors in Fig. 3 and Table 5 alone, DMU-6 is identified as the best performer. If we look at the performance of DMU-6 in Fig. 4, where the unit is outlined by a solid bar each year, its performance in 2010 is simply identified as the pooled sample. However, considering both 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 Fig. 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 Fig. 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 Fig. 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.

Fig. 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.

Fig. 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.

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 Fig. 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

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

$E_{i,1}$ 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) show that the Malmquist index provides an inaccurate measure of productivity change in the presence of non-constant returns to scale. Doubling all inputs and outputs from one period to the next, keeping input and output mixes constant, should not change productivity. Hence, the productivity measure should be homogeneous of degree 1.¹⁹ 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 Fig. 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.²⁰ 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 Fig. 6 where the Malmquist productivity score equals unity. Improved productivity for as many as seven units could also underline our indication from Fig. 4 of a catching up effect at the end of the four year period. The average unweighted

¹⁹ The efficiency measures in the index are homogeneous of degree +1 in outputs and –1 in inputs.

²⁰ In Førsund et al. (2005) this kind of diagram is named Edvardsen significance diagram.

Fig. 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.

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 Fig. 6. Any correlation between unit size and productivity change could be studied by looking at the width of boxes in Fig. 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 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 Fig. 7 units are evaluated according to their changes in spending and productivity. The figure

Fig. 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.

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 Fig. 6, and the productivity decreasing cost increase corresponds to the first unit in Fig. 6. Fig. 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.

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.²¹

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 represents 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

²¹ The pilot is documented in a research report (FFI-report 13/00064), available in Norwegian only, www.ffi.no.

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.

Acknowledgment

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.

Appendix A. DEA model

The Farrell input-oriented technical efficiency scores are estimated for the $j = 1, \dots, 11$ Home guard districts over the years $t = 1, \dots, 4$, giving the observations $k_{j,t} = 1, \dots, 44$. The efficiency measure for unit k_0 is defined by

$$E_{k_0} = \min \theta_{k_0} \quad (\text{A.1})$$

$$\begin{aligned} \text{s.t.} \\ \sum_{k=1}^{44} x_{4,k} \lambda_k \leq \theta_{k_0} x_{4,k_0} \end{aligned} \quad (\text{A.2})$$

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

$$\lambda_k \geq 0 \quad (\text{A.4})$$

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.

References

- 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. *Econ. Lab. Market Rev.* 3, 1.
- Bowlin, W.F., 1987. Evaluating the efficiency of us air force real-property maintenance activities. *J. Oper. Res. Soc.*, 127–135.
- Bowlin, W.F., 1989. An intertemporal assessment of the efficiency of air force accounting and finance offices. *Res. Gov. Nonprofit Account.* 5, 293–310.
- Bowlin, W.F., 2004. Financial analysis of civil reserve air fleet participants using data envelopment analysis. *Eur. J. Oper. Res.* 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. *J. Oper. Res. Soc.* 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.
- Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* 2 (6), 429–444.
- 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. *Ann. Oper. Res.* 2 (1), 95–112.
- Charnes, A., Cooper, W.W., Lewin, A.Y., Seiford, L., 1994. *Data Envelopment Analysis: Theory, Methodology and Applications*. Springer, Norwell.
- Clarke, R.L., 1992. Evaluating USAF vehicle maintenance productivity over time: an application of data envelopment analysis. *Decis. Sci.* 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, New York.
- Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap. In: *Monographs on Statistics & Applied Probability*. Chapman and Hall/CRC, Dordrecht.
- Eurostat, 2001. *Handbook on Price and Volume Measures in National Accounts*. European Communities, Luxembourg.
- Fried, H.O., Lovell, C.K., Schmidt, S.S., 2008. *The Measurement of Productive Efficiency and Productivity Growth*. Oxford University Press, New York.
- 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? *Natl. Inst. Econ. Rev.* 197 (1), 106–119.
- Førsund, F.R., 2010. Dynamic efficiency measurement. *Indian Econ. Rev.* 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).
- Farrell, M.J., 1957. The measurement of productive efficiency. *J. R. Stat. Soc. Ser. 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. *Eng. Manag.* 53 (3), 471–482.
- Griffell-Tatjé, E., Lovell, C.K., 1995. A note on the Malmquist productivity index. *Econ. Lett.* 47 (2), 169–175.
- Hartley, K., 2010. The case for defence. *Defence Peace Econ.* 21 (5–6), 409–426.
- Hartley, K., 2012. Conflict and defence output: an economic perspective. *Revue d'economie polit.* 122 (2), 171–195.
- Hurley, W., 2005. A clarification of the concepts of force multiplier and returns to force scale. *Defence Peace Econ.* 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. *Econom. 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 non-parametric frontier approach. *Afr. J. Bus. Manag.* 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 Peace Econ.* 5, 269–288.
- Ozcan, Y.A., Bannick, R.R., 1994. Trends in Department of Defense hospital efficiency. *J. Med. Syst.* 18 (2), 69–83.
- Park, B.U., Jeong, S.-O., Simar, L., 2010. Asymptotic distribution of conical-hull estimators of directional edges. *Ann. Stat.* 38 (3), 1320–1340.
- Pedraja-Chaparro, F., Salinas-Jimenez, J., Smith, P., et al., 1999. On the quality of the data envelopment analysis model. *J. Oper. Res. Soc.* 50 (6), 636–644.
- Roll, Y., Golany, B., Seroussy, D., 1989. Measuring the efficiency of maintenance units in the Israeli Air Force. *Eur. J. Oper. Res.* 43 (2), 136–142.
- Romer, D., 2006. *Advanced Macroeconomics*. McGraw-Hill Series Economics.

- McGraw-Hill, New York.
- Schreyer, P., 2010. Measuring the production of non-market services, in: *Price indexes in time and space, contributions to statistics*, Springer-Verlag Berlin, Heidelberg.
- Silverman, B.W., 1986. *Density Estimation for Statistics and Data Analysis* Vol. 26. CRC Press, London.
- Simar, L., Wilson, P.W., 1998. Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. *Manag. Sci.* 44 (1), 49–61.
- Simar, L., Wilson, P.W., 1999. Estimating and bootstrapping Malmquist indices. *Eur. J. Oper. Res.* 115 (3), 459–471.
- Simar, L., Wilson, P.W., 2000. Statistical inference in nonparametric frontier models: the state of the art. *J. Prod. Anal.* 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., 1978. Hedonic cost functions for the regulated trucking industry. *Bell J. Econ.* 9 (1), 159–179, Spring.
- Sun, S., 2004. Assessing joint maintenance shops in the Taiwanese army using data envelopment analysis. *J. Oper. Manag.* 22 (3), 233–245.
- UK Ministry of Defence, 2010. *Operations*. UK Army Doctrine Publication. The Development, Concepts and Doctrine Centre (DCDC).

Paper II

Estimating output mix effectiveness: An applied scenario approach for the Armed Forces

Torbjørn Hanson

Published in *Omega*, 2019, volume 83, pp. 39–49.

Estimating output mix effectiveness: An applied scenario approach for the Armed Forces^{☆☆☆}

Torbjørn Hanson¹

Norwegian Defence Research Establishment (FFI), Instituttveien 20, Kjeller, PO Box 25, NO-2027, Norway

ARTICLE INFO

Article history:

Received 24 August 2017

Accepted 2 February 2018

Available online 8 February 2018

JEL classification:

D24

H40

C60

Keywords:

Public sector

Overall preference effectiveness

Output mix efficiency

DEA

Scenario approach

Armed Forces

ABSTRACT

The ability of public sector policy makers to prioritize has a huge impact on the effectiveness of public service provision. Public services can take the form of final outputs demanded by consumers or of intermediate outputs contributing to a process of realizing the higher goals of society. In doing the right things, policy makers choose a mix of intermediate outputs maximizing their preference value for public service outcomes, while managers do things right when responsible for producing outputs efficiently. This distinction enables us to pinpoint important reasons for inefficiencies in the provision of public services. Taking advantage of the method of scenario based planning, a model for measuring effectiveness is developed for situations where traditional methods such as two-stage regressions fail due to long time lags and lack of variation in the variables. Scenarios take the role of outcomes in the modeling of outcome mapping functions, where each scenario represents a set of environmental variables. The model is specified for the provision of defense outcomes, where the lag between changes in input and impacts on outcomes are substantial. From a sample of 12 combat units in the Norwegian Armed Forces, producing different outputs, we find that inefficiencies in output mix can explain most of the changes in overall effectiveness over a four-year period of time.

© 2018 Elsevier Ltd. All rights reserved.

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.

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

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

^{☆☆} This manuscript was processed by Associate Editor Silva.

E-mail address: torbjorn.hanson@ffi.no

¹ Would like to thank Finn R. Førsund, Erik A. Hanson, Sverre A. C. Kittelsen, Petter Y. Lindgren and two anonymous referees for useful comments.

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 Eurostat [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”.

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 [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.

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 careful 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 em-

pirical 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, 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 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. Hanson [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 mea-

⁵ See e.g. [18,26].

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

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.

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

sure. 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 environmental. 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 through 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

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

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

$$E_{j_0} = \min \theta_{j_0} \quad (2)$$

s.t.

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

$$\sum_{j=1}^n y_{r,j} \lambda_j \geq y_{r,j_0}, \quad r = 1, \dots, p \quad (4)$$

$$\sum_{j=1}^n \lambda_j = 1 \quad (5)$$

$$\lambda_j \geq 0 \quad (6)$$

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 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 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. Cray 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 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 Fig. 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 de-

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. Hanson [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.

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

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.

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].

Fig. 1. The levels from input to outcome in the military.

mand 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 Fig. 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 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. SAMs¹⁴ 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 Fig. 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 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

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

Going from level (1) – (5) in Fig. 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 environmental variables 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 mapping 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. environmental variables can

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)–(5) in Fig. 1).

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

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 [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.

3.1. 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.

Table 1

Relative scenario requirements by outputs and capabilities.

Scenario		y_1 $d_1(y_1)$ ASuW1 ^a	y_2 $d_3(y_2)$ HVAP ^c	$d_4(y_2)$ ISTAR ^d
		$d_2(y_1)$ ASuW2 ^b		
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.

^a Anti Surface Warfare.

^b Anti Surface littoral.

^c High Value Asset Protection.

^d Intelligence Surveillance, Target Acquisition and Reconnaissance.

Step 3: The third step is to set up all relevant combinations of environmental variables and objectives, where each combination represents a scenario s_j , $j = 1 \dots J$. In each scenario the level of the other-wise 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 environmental variables 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 environmental variables z together with outputs, implicitly through capabilities, affect outcomes. This involves specifying the outcome mapping function in (7). Given the fixed levels of environmental variables $\bar{z}$, 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.

3.2. 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 environmental variables 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 zone²⁰ 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

¹⁸ E.g. defense variables in [17].

¹⁹ E.g. health and education variables in [22] and [38], respectively.

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

variables, which are treated as constants by definition in the mapping functions. Scenario s_1 and s_2 are 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.

3.3. Effectiveness assessments

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 $G = \{f(x, y_2) \mid x \text{ can produce } y_2\}$. 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 $H = \{f(x, y_1) \mid x \text{ can produce } y_1\}$. 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 , $j = 1, \dots, J$, before preferences for the scenarios are expressed assuming a Cobb–Douglas specification with constant elasticities

$$W = \prod_j s_j^{\omega_j}, \quad \sum_j \omega_j = 1 \quad (8)$$

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 environmental 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 , $l = 1, \dots, L$, in scenario j . The capability is fully

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.

delivered when the output level reaches $Y_{l,j}^*$, where $d_{l,j}(Y_{l,j}^*) \equiv 1$, $\delta d/\delta y > 0$ for $y < Y^*$ 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

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

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

$$s_j = \sum_{l=1}^L v_{l,j} \left(\left(\frac{y_l}{Y_{l,j}^*} \right) + \left(1 - \frac{y_l}{Y_{l,j}^*} \right) H(y_l - Y_{l,j}^*) \right), \quad (10)$$

$$H(y_l - Y_{l,j}^*) = \begin{cases} 1 & \text{if } y_l \geq Y_{l,j}^* \\ 0 & \text{if } y_l < Y_{l,j}^* \end{cases}$$

The assumption $\delta d/\delta y = 0$ for $y \geq Y^*$ is assured by the Heavieside function²³ $H(\cdot)$ 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 [16]. An output 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 $qx = B$.

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

The solution to the problem is illustrated in Fig. 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

Fig. 2. The realization of output mix effectiveness.

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].

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

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 Fig. 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 Farell [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 2008–2011. 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

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. [1].

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, $\omega_i = 0.25, i = 1, \dots, 4$, 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

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

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 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 [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. Simar and Wilson [30] showed how to estimate the sampling bias in DEA using the bootstrap method (resampling). Following the suggested procedure

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

Table 3

Overall preference effectiveness and its decomposition for different preference structures.

Measure	2008	2009	2010	2011
<i>Overall preference effectiveness</i>				
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
<i>Technical preference efficiency</i>				
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
<i>Output mix effectiveness</i>				
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

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 observations, 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.

$$y_2 = \alpha x_1^\alpha x_2^\beta x_3^{1-\alpha-\beta} \quad (13)$$

$$y_1 = d \sum_{i=1}^3 x_i \quad (14)$$

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

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

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 Fig. 2, 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 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 dynam-

ics 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.

References

- [1] Adams RA, Essex C. *Calculus: a complete course*. 8. Canada: Pearson Education; 2013.
- [2] Anagboso M, Spence A. Measuring defence. *Econ Lab Market Rev* 2009;3(1).
- [3] Atkinson T. Measurement of government output and productivity for the national accounts. Palgrave Macmillan; 2005.
- [4] Banker RD, Natarajan R. Evaluating contextual variables affecting productivity using data envelopment analysis. *Oper Res* 2008;56(1):48–58.
- [5] Bradford DF, Malt RA, Oates WE. The rising cost of local public services: some evidence and reflections. *Natl Tax J* 1969;22(2):185–202.
- [6] Chu X, Fielding GJ, Lamar BW. Measuring transit performance using data envelopment analysis. *Transp Res Part A* 1992;26(3):223–30.
- [7] Cooper ST, Cohn E. Estimation of a frontier production function for the South Carolina educational process. *Econ Educ Rev* 1997;16(3):313–27.
- [8] Crary M, Nozick LK, Whitaker L. Sizing the US destroyer fleet. *Eur J Oper Res* 2002;136(3):680–95.
- [9] Eurostat. *Handbook on price and volume measures in national accounts*. Luxembourg: European Communities; 2001.
- [10] Färe R, Grosskopf S, Lundström M, Roos P. Evaluating health care efficiency. In: *Evaluating hospital policy and performance: contributions from hospital policy and productivity research*. In: No. 18 in *Advances in Health Economics and Health Services Research*; 2008. p. 209–22.
- [11] Farrell MJ. The measurement of productive efficiency. *J R Stat Soc Ser A (General)* 1957;120(3):253–90.
- [12] Florens J-P, Simar L. Parametric approximations of nonparametric frontiers. *J Econ* 2005;124(1):91–116.
- [13] Førsund FR. Measuring effectiveness of production in the public sector. *Omega* 2017;73:93–103.
- [14] Gutmann A. *Democratic education*. Princeton University Press; 1999.
- [15] Hammond CJ. Efficiency in the provision of public services: a data envelopment analysis of uk public library systems. *Appl Econ* 2002;34(5):649–57.
- [16] Hanson T. Efficiency and productivity in the operational units of the armed forces: a norwegian example. *Int J Prod Econ* 2016;179:12–23.
- [17] Hartley K. Conflict and defence output: an economic perspective. *Rev Econ Polit* 2012;Vol. 122(2):171–95.
- [18] Karlaftis MG. A DEA approach for evaluating the efficiency and effectiveness of urban transit systems. *Eur J Oper Res* 2004;152(2):354–64.
- [19] Lauer JA, Lovell CK, Murray CJ, Evans DB. World health system performance revisited: the impact of varying the relative importance of health system goals. *BMC Health Serv Res* 2004;4(1):19.
- [20] Lovell CK, Walters LC, Wood LL. Stratified models of education production using modified DEA and regression analysis. In: Charnes A, Copper W, Lewin A, Seiford L, editors. *Data envelopment analysis: theory, methodology, and applications*. Kluwer Academic Publishers; 1994. p. 329–51.
- [21] Medina-Borja A, Triantis K. Modeling social services performance: a four-stage DEA approach to evaluate fundraising efficiency, capacity building, service quality, and effectiveness in the nonprofit sector. *Ann Oper Res* 2014;221(1):285–307.
- [22] Murray CJ, Evans DB. *Health systems performance assessment*. Office of Health Economics; 2006.
- [23] Pritchett L, Filmer D. What education production functions really show: a positive theory of education expenditures. *Econ Educ Rev* 1999;18(2):223–39.
- [24] RTO/NATO. *Handbook on long term defence planning*. 69; 2003. RTO-SAS-025
- [25] Ruggiero J. Nonparametric estimation of returns to scale in the public sector with an application to the provision of educational services. *J Oper Res Soc* 2000;51(8):906–12.
- [26] Ruggiero J. Performance evaluation in education. In: *Handbook on data envelopment analysis*. Springer; 2004. p. 323–46.
- [27] Saaty TL. *What is the analytic hierarchy process?* Springer; 1988.
- [28] Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. *Popul Health Metr* 2003;1(1):12.
- [29] Schreyer P. Measuring the production of non-market services. Price indexes in time and space, contributions to statistics. Berlin Heidelberg: Springer-Verlag; 2010.
- [30] Simar L, Wilson PW. Sensitivity analysis of efficiency scores: how to bootstrap in nonparametric frontier models. *Manage Sci* 1998;44(1):49–61.
- [31] Simar L, Wilson PW. Estimation and inference in two-stage, semi-parametric models of production processes. *J Econ* 2007;136(1):31–64.
- [32] Simar L, Wilson PW. Two-stage DEA: caveat emptor. *J Prod Anal* 2011;36(2):205–18.
- [33] Spady RH, Friedlaender AF. Hedonic cost functions for the regulated trucking industry. *Bell J Econ* 1978;9(1):159–79.
- [34] Stiglitz JE. The theory of screening, education, and the distribution of income. *Am Econ Rev* 1975;65(3):283–300.
- [35] Takamura Y, Tone K. A comparative site evaluation study for relocating Japanese government agencies out of Tokyo. *Socioecon Plann Sci* 2003;37(2):85–102.
- [36] Timmer CP. Using a probabilistic frontier production function to measure technical efficiency. *J Polit Econ* 1971:761–94.
- [37] UK Ministry of Defence. Operations. UK army doctrine publication. The Development, Concepts and Doctrine Centre (DCDC); 2010.
- [38] Worthington AC. An empirical survey of frontier efficiency measurement techniques in education. *Educ Econ* 2001;9(3):245–68.

Paper III

Trust, ethnic diversity, and personal contact: A field experiment

Henning Finseraas, Torbjørn Hanson, Åshild A. Johnsen, Andreas Kotsadam, Gaute Torsvik

Published in *Journal of Public Economics*, 2019, volume 173, pp. 72–84.

Trust, ethnic diversity, and personal contact: A field experiment

Henning Finseraas^a, Torbjørn Hanson^b, Åshild A. Johnsen^c,
Andreas Kotsadam^{d,*}, Gaute Torsvik^e

^aInstitute for Social Research, Norway

^bNorwegian Defense Research Establishment (FFI), Norway

^cNorwegian University of Life Sciences, Norway

^dRagnar Frisch Centre for Economic Research, Norway

^eUniversity of Oslo, Norway

ARTICLE INFO

Article history:

Received 28 September 2018

Received in revised form 8 February 2019

Accepted 12 February 2019

Available online xxxx

Keywords:

Integration

Field experiment

Ethnic diversity

Contact

Trust

ABSTRACT

We study how close personal contact with minorities affects in-group and out-group trust in a field experiment in the armed forces. Soldiers are randomly assigned to rooms with or without ethnic minorities. At the end of the recruit period, we measure trust by using a trust game. Results indicate that close personal contact with minorities increases trust towards a generic immigrant. We replicate the result that individuals coming from more ethnically diverse areas trust minorities less, but random assignment to interact with minority soldiers removes this negative correlation. We conclude that social integration involving personal contact can reduce negative effects of ethnic diversity on trust.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Western societies are becoming more diverse. Diversity can spur more innovation, creativity and economic growth (e.g., [Hunt and Gauthier-Loiselle, 2010](#); [Peri, 2012](#)), but some argue that it can also lead to less social trust and more tension and conflicts ([Alesina and La Ferrara, 2002](#); [Putnam, 2007](#)). The effects of diversity on trust are essential to understand: When people trust each other, transaction costs are reduced, organizations run better, the need for formal regulation reduces, governments provide services more efficiently, policy promises become more credible, and financial systems develop better ([Algan and Cahuc, 2013](#); [Guiso et al., 2008a,b, 2011](#)). If migration and ethnic diversity have dismantling effects on the social fabric of societies, it becomes important to find out if and how public policy can mitigate such problems. For instance, can tensions be reduced and trust enhanced if governments create arenas where different ethnic groups regularly encounter each other? Can social contact build trust? We speak to these questions by investigating the effects

on majority individuals' in-group and out-group trust from personal contact with minority individuals.

Several empirical studies find patterns that are consistent with what we denote conflict theory; diversity is associated with less trust ([Alesina and La Ferrara, 2002](#); [Dinesen and Sønderskov, 2015](#)).¹ [Putnam \(2007\)](#) extends the conflict perspective, arguing that ethnic diversity may not only lead to less trust between the majority and minority groups, it may also be detrimental to trust within the majority group. This hypothesis, which he labels constrict theory, is based on less diverse neighborhoods in the US having higher levels of in-group trust. These findings have spurred a debate on how diversity should be conceived and measured ([Abascal and Baldassarri, 2015](#)). Because more diverse US neighborhoods with lower trust levels are also poorer, more nonwhite, and less stable than the more

* Corresponding author.

E-mail addresses: henning.finseraas@samfunnsforskning.no (H. Finseraas), Torbjorn.Hanson@ffi.no (T. Hanson), asjo@nmbu.no (Å.A. Johnsen), andreas.kotsadam@frisch.uio.no (A. Kotsadam), gaute.torsvik@econ.uio.no (G. Torsvik).

¹ There is an extensive literature on the effects of ethnic diversity in other domains. For studies on diversity and economic outcomes, see for instance [Alesina et al. \(2003, 2016\)](#) and [Alesina and La Ferrara \(2005\)](#). For another strand of literature study diversity within organizations, see for instance [Lyons \(2017\)](#) on teams and national diversity, [Ottaviano and Peri \(2006\)](#) on productivity, [Hjort \(2014\)](#) on team productivity, and [Shore et al. \(2009\)](#) for a review. There is also a literature on ethnic diversity and provision of public goods, see [Beach and Jones \(2017\)](#) for a review.

homogenous neighborhoods, it is hard to disentangle the effect ethnic diversity has on trust from the effects of these other features of diverse societies.

A major limitation of previous studies of ethnic diversity and trust is the inability to control for selection and reverse causality. We randomize soldiers to rooms during boot camp, implying that soldiers from the majority group (ethnic Norwegian soldiers) are randomized to share living quarters with at least one minority member, while others have only members of the majority group as roommates. At the end of the boot camp we ran a trust game with monetary stakes. Soldiers play either against a person with an ethnic minority identity, or against a person from the majority group. This design allows us to test if close contact with individuals from a minority group causally affects the trust majority members show to a stranger with a minority identity. As far as we are aware, ours is the first study with a research design allowing for a causal identification of how close contact between majority and minority individuals affects trust.

Social conflict is not the only potential outcome of ethnic diversity. The more optimistic contact theory (Allport, 1954) argues that personal contact with members of out-groups can reduce prejudice and misperceptions, and thereby increase trust. There is evidence from well identified studies using random assignment, either of students (e.g. Boisjoly et al., 2006; Burns et al., 2016; Scacco and Warren, 2018) or within the military (Carrell et al., 2015; Dahl et al., 2018; Finseraas et al., 2016; Finseraas and Kotsadam, 2017), which shows that personal contact reduces prejudice and strengthens cooperation (Goette et al., 2006). Rao (2019) studies how random exposure to poor students affect wealthy students in terms of pro-social behaviour, discrimination, as well as academically. Alexander and Christia (2011) study how personal contact affect cooperation in a public goods game.^{2,3} Finseraas and Kotsadam (2017) is of particular relevance to our paper. They report the results of a similar field experiment as ours conducted on the previous cohort of soldiers (in 2014).⁴ Unlike us, they have no experimental outcome, but study the effects of minority contact on a set of survey outcomes. They find that contact makes treated soldiers more likely to state that immigrants have similar work ethics as Norwegians, but no effect on views on whether immigration is a net positive for society, or on support for equal rights to social assistance. We extend this work to study trust, and improve the research design by measuring the outcome in the lab rather than relying on self-reported outcomes. Moreover, we use our data to contrast and combine the conflict and contact perspective on ethnic diversity, by studying treatment heterogeneity according to previous exposure to diversity. In addition we examine heterogeneity on the ability of minority roommates (Carrell et al., 2015; Burns et al., 2016) and majority soldiers' initial attitudes towards immigration (Johnston et al., 2015).

We find that individuals randomly assigned to close personal contact with minority soldiers send more money to the person with a

name signaling minority origin (Ali). Random contact with minority soldiers does not affect the amount sent to the person with a Norwegian name (Morten). Next we find a negative association between immigrant share in the home municipality and trust in Ali (but not in Morten), which is consistent with the conflict perspective on diversity. We further find that the negative relationship between immigrant share in the home municipality and out-group trust is annulled for the soldiers that were randomly assigned to close personal contact with a minority soldier.

The rest of the paper is organized as follows: In Section 2 we present the field experiment. Section 3 illustrates the details of the trust game. Section 4 describes our data, while our empirical strategy is outlined in Section 5. The results are presented in Section 6, with discussion and concluding remarks following in Section 7. Instructions for the trust game and additional analyses are included in the Appendix.

2. Identifying effects of contact: the field experiment

The sample for the field experiment consists of incoming soldiers of the August 2015-contingent of the North Brigade of the Norwegian Armed Forces (NAF). The first day of service starts at a military camp close to Oslo. At the camp, the soldiers go through a program of medical and psychological testing and they fill out a survey questionnaire, which constitutes our baseline data.

After completing the program at the camp, soldiers board planes to Northern Norway to start their recruit period. When they arrive in Northern Norway, they are bussed to a number of different military camps. In the camps they are assigned to rooms for the eight weeks of the recruit period.

We provided the personnel officers in charge of room assignment with an excel sheet which they were instructed to use to randomize soldiers into rooms. The personnel officers enter the list of soldiers in the company and specify the size of the rooms in the camp, whereby the excel sheet randomizes soldiers into rooms. This room assignment for the 8-week long recruit period constitutes our main treatment variable (more on this below). Copies of the excel sheets were emailed to the Norwegian Defense Research Establishment (FFI) for verification. We only analyze data from companies for whom we could verify the randomization by them sending us the lists from the randomization program. The procedure allows for the construction of a treatment group consisting of soldiers with an ethnic Norwegian background who were randomized into a room with at least one soldier with an ethnic minority background (see definitions of majority and minority backgrounds in Section 4). The control group consists of soldiers who did not share the room with an ethnic minority soldier. A deviation from the randomization protocol was included so that when possible, women are allocated to rooms in pairs of two. We describe how we handle this in Section 5.

The recruit period is the basic training period, which is known for strict enforcement of military rules and regulations.⁵ During the eight weeks, the soldiers are to wear their uniform 24/7 and are not allowed to sleep outside of the base. The first extended leave is normally granted after completion of the basic training period. Because of the remote location of the bases, the soldiers basically spend all their time together. A normal day of boot camp starts with activities within the room, such as cleaning and preparing the room before inspection. Working hours are intense, usually 10–15 h a day. In addition, soldiers are expected to prepare their individual gear and equipment for the following day after duty ends. This leaves the soldiers with few opportunities for personal chores and socializing

² For a review of the overall literature covering ethnic diversity and social trust, see Dinesen and Sønderskov (2017), Kaufmann and Goodwin (2016), Meer and Tolsma (2014) and Schaeffer (2014).

³ One important distinction is between personal contact and shallow exposure. In a series of papers, summarized in Enos (2017), Enos investigates the causal effects of shallow contact. He finds increases in prejudice from random assignment to exposure to diversity on commuting trains (Enos, 2014) or waiting rooms (Enos and Celaya, 2018). Enos (2016) studies the electoral consequences of shallow exposure by using the demolition of a Chicago public housing project containing mostly black individuals to get exogenous variation in diversity in white voters' neighborhoods.

⁴ Finseraas and Kotsadam (2017) follow a pre-analysis plan which was submitted to the AEA RCT Registry. We intended to do the same, however, the Army's confirmation that we would get the necessary time slot in the soldiers' busy schedule came only a few days before the data collection. The late confirmation made it impossible to write a high-quality pre-analysis plan. However, coding of background covariates follow the pre-specified procedure in Finseraas and Kotsadam (2017). The pre-analysis plan of Finseraas and Kotsadam (2017) can be accessed at <https://www.socialscienceregistry.org/trials/507>.

⁵ After the recruit period the selection of soldiers for regular infantry and cavalry companies takes place and the soldiers move. Room composition is no longer random after this point.

outside their own room. The room is also important since it usually constitutes a squad within a platoon in the company. Thus, sharing room during the recruit period constitutes intense treatment in the form of personal contact.

According to contact theory, the positive effects of personal contact are expected to apply when certain criteria are met. The contact should take place in a context with equal status, shared common goals, be cooperative, and take place under some form of authority (Allport, 1954; Pettigrew, 1998). Finally, the setting should have friendship potential, which increases the probability of affective ties and willingness to learn about out-group members (Van Laar et al., 2005). The army context is one where the conditions of contact theory are likely to hold. Soldiers of private rank have equal social status within the army, they share the common goals of their unit, they need to cooperate to solve their tasks, and contact takes place in a context with an explicit, enforcing authority. Moreover, the army explicitly promote views of unity and equality among soldiers of the same rank. We conducted a trust game at the end of the eight weeks recruit period.

3. Measuring trust: the trust game

To measure the effect of close contact on trust, we use the trust game developed by Berg et al. (1995). In the standard trust game a person chooses how much to send to an anonymous other person, the responder. The amount sent is typically tripled and the responder decides how much to send back. The advantage of the trust game is that the game captures the essence of trust in economic exchange; there is money at stake for the trustor/sender and a substantial surplus is produced if the resources are handed over to the trustee/responder.⁶

In our case, the senders (each soldier) choose between sending 0, 25, 50, 75, or 100 NOK out of a total endowment of 100 NOK (12 US\$) to a responder.⁷ In turn, the experimenters (we), triple the chosen amount to the responder. The responder then decides how much of the received money to transfer back to the sender.

Sending behaviour in the standard trust game is assumed to reflect partly the sender's belief in the receivers trustworthiness, risk aversion, other-regarding preferences such as altruism (Sapienza et al., 2013), as well as inequality aversion and betrayal aversion (Fehr, 2009). However, sending money to an anonymous other does not allow for a distinction between in-group and out-group trust. We therefore use a modified version of the trust game where the senders either play the trust game with a responder with a typical Norwegian name (Morten), or with a responder with a name indicating a ethnic minority origin (Ali). By randomizing the names, we assure that all other factors that may influence how much a person sends does not vary systematically with the ethnic identity of the responder. Hence, with this design we can estimate to what extent beliefs about responder trustworthiness vary depending on the ethnicity of the responder, and most importantly, to what extent exogenous exposure to a minority member modifies the assessment of trustworthiness.⁸ Fershtman and Gneezy (2001) suggest complementing the trust game with a dictator game to further disentangle

the role of beliefs and social preferences. As we did not include a dictator game it should be noted that our results on sending behaviour may be affected by changed social preferences and not only changes in beliefs about trustworthiness.

We recruited responders at the University of Oslo before going to the military camps. We announced that we were recruiting participants to an experiment, and then recruited one student with a typical Norwegian name (Morten), and one student with a name indicating ethnic minority origin (Ali). Both responders/trustees were asked to make a back-transfer decision contingent on the amount that they would receive.⁹

In addition to the first name of the responder, the soldiers (senders) are told that the responder (trustee) is a real person living in the eastern part of Norway, that he has been recruited by us to take part in the study and that he received NOK 100 just for participating. The soldiers were also informed that the responder was aware of both the structure of the game and that the sender would be a soldier in the boot camp in Northern Norway. To be clear, the monetary incentives in the trust game are low, especially as there is only one person per session that is drawn. The main reason for choosing this incentive structure is that our collaborators only allowed us to offer limited incentives. Fortunately, we find that the amount sent is not different in different sessions with different numbers of participants (ranging from 46 to 100). Furthermore, it is unlikely to affect the difference in sending to Ali or Morten. A translated version of the instructions is included in Appendix Section A.

The experiment was conducted in September 2015 at two different military bases, Setermoen and Skjold, located 60 km apart in a rural area in the northern part of Norway. The Armed Forces decided when different groups of soldiers would attend our sessions based on the soldiers schedule for the given day. The experiment consists of two main treatments: the minority responder (Ali) and the majority responder (Morten). We randomized treatments within each session, so that about half of the soldiers in each session play the game with Morten, and half with Ali. One sender in each session is drawn randomly to be paid and the experiment was conducted using pen and paper.

We collect information at baseline and after 8 weeks, and at both times the soldiers are informed that any information provided will be treated confidentially, and that no person their platoon, military base, nor in the Armed Forces will be able to track their answers back to them. They are also informed that data collection has been approved by the ethics committee, and that they can contact out co-author Torbjørn Hanson (email address given) if they have any questions.

4. Data

4.1. Sample and treatment variable

We conducted the experiment on a subset of 656 subjects in 12 sessions in September 2015. The sessions ranged from 46 to 100 individuals. In our analysis sample we only include observations from companies which have confirmed that they followed our randomization protocol. The minority soldiers are not included in the analysis because we want to separate between the ones providing exposure

⁶ There is an ongoing discussion about what is captured by general trust questions, see Sapienza et al. (2013) for an overview. In particular, Glaeser et al. (2000) argue that the measures are correlated with trustworthiness rather than trust, while Thöni et al. (2012) suggest that the standard trust question is a proxy for cooperation preferences rather than beliefs about others' cooperation.

⁷ The service allowance for Norwegian recruits normally equals about NOK 170 a day, roughly USD 20 at the exchange rate at the time of the experiment.

⁸ The trust game has been used to investigate in-group/out-group trust across ethnicities/nationalities in several studies (see for instance Falk and Zehnder, 2013 and Fershtman and Gneezy, 2001). But exposure to the out-group is not random in these studies and they are therefore not able to identify the effect of ethnic diversity on trust.

⁹ We use the strategy method to obtain the back-transfer from the responder (see Stanley et al., 2011, for a similar set-up). There is a discussion in the literature if this method gives different results than the direct response method. In most cases it appears that the choice of method does not matter for the outcomes (Brandts and Charness, 2011). In our study the strategy method was the only viable option, and since we are only interested in the senders' decision, the way we extract the responders' return decision should be of second order importance. It is further unlikely that the strategy method induces a differential impact across our treatment and control groups.

Table 1
Background variables and balance across treatment.

	(1) Mino base Mean	SD	(2) Majo base Mean	SD	(3) Mino treat Mean	SD	(4) Majo treat Mean	SD
Dependent variable								
Amount sent in trust game (%)	69.80	(37.27)	71.35	(35.79)	79.41	(31.90)	70.98	(37.17)
Background characteristics								
Mother high education	0.60	(0.49)	0.63	(0.48)	0.63	(0.49)	0.68	(0.47)
Father high education	0.75	(0.43)	0.75	(0.43)	0.76	(0.43)	0.80	(0.40)
Mother works	0.91	(0.28)	0.94	(0.24)	0.94	(0.24)	0.93	(0.26)
Father works	0.99	(0.11)	0.99	(0.09)	0.98	(0.14)	1.00	(0.00)
Parents divorced	0.37	(0.48)	0.32	(0.47)	0.27	(0.45)	0.22	(0.42)
Plans education	0.69	(0.46)	0.67	(0.47)	0.69	(0.47)	0.64	(0.48)
Females	0.12	(0.33)	0.07	(0.26)	0.18	(0.39)	0.11	(0.31)
Females in room	0.24	(0.43)	0.28	(0.45)	0.26	(0.45)	0.26	(0.44)
Share non-west immig muni	0.07	(0.05)	0.07	(0.05)	0.06	(0.04)	0.08	(0.05)
Room size	6.10	(1.35)	5.97	(1.01)	5.98	(1.14)	6.18	(1.43)
Attitudes at base								
Immigration reduces trust	0.16	(0.37)	0.21	(0.41)	0.20	(0.40)	0.20	(0.40)
Immigrants' work ethic	3.84	(0.93)	3.77	(0.96)	3.47	(0.99)	3.61	(0.91)
Immigrants same rights	3.38	(1.10)	3.36	(1.05)	2.96	(1.23)	3.32	(1.16)
Lend money to roommate	0.91	(0.29)	0.91	(0.28)	0.86	(0.35)	0.87	(0.34)
General trust	6.91	(2.02)	6.86	(2.05)	6.45	(2.14)	6.65	(1.99)
Helpfulness	6.46	(1.92)	6.50	(1.79)	5.84	(1.80)	6.44	(1.64)
Fairness	6.95	(1.98)	6.81	(1.97)	6.73	(1.54)	6.76	(1.90)
N	245		240		51		56	

Note: *Mino base* denotes soldiers from majority rooms who played the trust game with the minority responder (Ali), *Majo base* denotes soldiers from majority rooms who played the trust game with the majority responder (Morten). *Mino treat* denotes soldiers who did share room with minorities who played the trust game with the minority responder (Ali), *Majo treat* denotes soldiers who did share room with minorities who played the trust game with the majority responder (Morten). Background characteristics: *Female*: Share of female soldiers. *Females in room*: Share of majority soldiers who also share room with female soldier (including female soldiers). *Municipality's imm. share*: Share of population in municipality with a non-western background. *Immigration reduces trust*: "Immigration leads to lower trust between the citizens of a country. Do you agree/disagree?" [0, 1]. *Immigrants' work ethics*: "In general, immigrants have poorer work ethic than Norwegians. Do you agree/disagree [1–5]?" *Immigrants same rights*: "During the first years of their stay in Norway, immigrants should receive lower social benefits than Norwegians. Do you agree/disagree [1–5]?" *Lend money to roommate*: If one of your roommates lost their wallet, would you lend them money? Unwilling/ willing [0, 1]. *General trust*: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people? Distrust/trust [1–10]." *Helpfulness*: "Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?, not helpful/helpful [1–10]. *Fairness*: "Do you think most people would try to take advantage of you if they got the chance, or would they try to be fair?, unfair/fair [1–10].

from the ones affected by it (see Angrist, 2014). Furthermore, there is no treatment variation at the extensive margin in the field experiment for the minority soldiers as they all live in rooms with majority soldiers. We also exclude soldiers with missing information on parents' birthplace. The analysis sample includes 592 soldiers, spread across 121 rooms.

We define minority background as being born in or having at least one parent being born in South-America, Asia, Oceania, or Africa, as in Finseraas and Kotsadam (2017). 4.2% of the experimental sample have minority background.¹⁰ On average, 18.2% of the experimental sample share a room with someone with a minority background.

The rooms vary in size, but the majority of the sample (72%) live in 6-person rooms. Of the 108 treated soldiers, 14 share room with two persons of a minority ethnic background, while the remaining 94 share room with one person of minority background. Since the rooms also vary in size, we have variation in the share of minority exposure in the room, ranging from zero to 40% (see Fig. D1 in the Appendix).

A high dismissal rate is normal during the recruit period. Importantly, we test and confirm that attrition in the panel is unrelated to treatment status (see Appendix Table B1 and the discussion there).

4.2. Control variables and balance

Table 1 presents means on background variables in the four groups which constitute our experiment. To examine balance across treatment status, we regress being treated in the field experiment

on pre-determined variables in Table 2. We include company fixed effects (9 in total) in all regressions, since room assignment is randomized within companies. We cluster the standard errors at the room level (121 rooms), since treatment occurs at the room level. Some coefficients in Table 2 are statistically significant, which is not surprising given the number of variables tested, but the F-test of joint significance produces a p-value of 0.49. Thus, we conclude that the background variables do not predict treatment status. In the regressions below we present results both with and without control variables.¹¹ In addition, we conduct a robustness check where we select control variables using a LASSO regularization approach (Belloni et al., 2014).

4.3. Representativeness of the sample

Norway has military conscription, but the military's demand for soldiers is lower than the size of the age cohorts. The soldiers are therefore positively selected on background characteristics such as grades in high school and physical capacity. Motivation for military service also weigh in when selecting soldiers and a majority of the soldiers are therefore doing military service voluntarily. Nonetheless, according to a previous survey, 34% of the soldiers are unsure of whether they would have served in the military if it was completely voluntary.

Finseraas and Kotsadam (2017) compare the soldiers of the 2014 contingent to a sample of men aged 18–30 years from the general population. They find that the soldiers have more liberal attitudes

¹⁰ Only Norwegian citizens are allowed to serve in the Armed Forces and hence our minorities are to a large extent second-generation immigrants. See Table D5 in the Appendix for details.

¹¹ In Appendix Tables C2 and C3, we report results from regressions of the treatment indicator interacted with responder in the trust game on the pre-determined variables (one-by-one). We again conclude that randomization has achieved balance.

Table 2
Living in an ethnically mixed room and pre-determined variables.

	(1) Treat	(2) Treat	(3) Treat	(4) Treat	(5) Treat	(6) Treat	(7) Treat	(8) Treat	(9) Treat	(10) Treat	(11) Treat	(12) Treat	(13) Treat	(14) Treat	(15) Treat	(16) Treat	(17) Treat	(18) Treat
Mother high education	0.04 (1.18)																	0.03 (0.59)
Father high education		0.04 (1.20)																0.03 (0.53)
Mother works			0.01 (0.22)															−0.01 (−0.10)
Father works				−0.01 (−0.06)														−0.03 (−0.20)
Parents divorced					−0.07* (−1.97)													−0.06* (−1.74)
Plans education						0.00 (0.09)												−0.02 (−0.55)
Sex of respondent							0.06 (0.86)											0.09* (1.97)
Females in room								0.00 (0.01)										−0.02 (−0.32)
Municipality's immigrant share									−0.00 (−0.33)									−0.01 (−0.44)
Room size										0.01 (0.19)								0.01 (0.24)
Lend money to roommate											−0.04 (−0.78)							−0.03 (−0.62)
General trust												−0.02 (−1.25)						−0.00 (−0.17)
Trust: helpfulness													−0.04** (−2.56)					−0.04** (−2.09)
Trust: fairness														−0.01 (−0.83)				0.01 (0.62)
Immigration reduces trust															−0.01 (−0.31)			−0.05 (−0.99)
Immigrants' work ethic																−0.03** (−2.02)		−0.03 (−1.52)
Immigrants same rights																	−0.03 (−1.49)	−0.01 (−0.57)
Observations	592	592	592	592	592	592	592	592	592	592	592	592	592	592	592	592	592	592
R-squared	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.08	0.07	0.11
Company FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Share treated	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181	0.181

Note: Each column presents the results from separate regressions with treated with minority (share room with minorities) as dependent variable and the pre-determined variable as control. Column 18 reports the coefficients when all pre-determined variables are included in one regression. The F-value from an F-test of the joint significance of all the pre-determined variables in the final column is 0.48 ($p = 0.4915$). Company and session fixed effects are included in all regressions. t-statistics (adjusted for room clustering) in parentheses. OLS regressions.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Table 3

Amount sent in the trust game and contact.

	(1) Mino	(2) Mino	(3) Majo	(4) Majo	(5) All	(6) All
Treated	9.92* (5.36)	11.48* (5.89)	2.65 (7.03)	6.33 (6.25)	1.72 (6.84)	4.19 (5.41)
Treated * Minority responder					7.52 (8.21)	8.78 (6.72)
Minority responder					−0.66 (3.01)	0.65 (3.40)
Observations	296	296	296	296	592	592
Company FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes	No	Yes
Mean trust	71.45	71.45	71.28	71.28	71.37	71.37
SD trust	36.54	36.54	35.99	35.99	36.23	36.23
Mean treated	0.17	0.17	0.19	0.19	0.18	0.18
SD treated	0.38	0.38	0.39	0.39	0.39	0.39

Note: The dependent variable is amount in NOK sent to the minority/majority responder in the trust game, which ranges from NOK 0 to 100 (about USD 12). *Treated* denotes soldiers from mixed rooms. Regressions (1)–(2) only include observations in treatment with the minority responder. Regressions (3)–(4) only include observations in treatment with the majority responder. Regressions (5)–(6) include observations from both treatments. All regressions include a dummy for whether the living quarters include female soldiers. Individual controls include whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, attitudes towards immigration, response to questions regarding trust, dummy variables for room size, and gender. Mean and standard deviation of dependent and independent variables are shown below. Robust standard errors in parentheses clustered on rooms (121 rooms). OLS regressions.

* $p < 0.1$.** $p < 0.05$.*** $p < 0.01$.

towards immigrants. With respect to the minority soldiers, we know that most of them are second-generation immigrants, and they are likely to be better integrated than a random sample of second-generation immigrants. For instance, the share having mothers that are working is higher for our soldiers than in the population of second-generation immigrants in general.

To check if our soldier sample differs from the general population with respect to trust we asked three general trust questions at baseline that also appear in the European Social Survey (ESS).¹² We compare the soldiers' answers to those of young Norwegian males aged between 18 and 30. We find that ethnic Norwegian soldiers are similar to the general population with one exception (the soldiers are more inclined to think that people are helpful). As compared to minorities in the ESS data, defined as in the army data, the minority soldiers report trust levels that are higher on both the generalized trust question and the question regarding whether most people try to be helpful. The results are displayed in Fig. D2 in the Appendix. Hence, we conclude that our sample of soldiers are slightly more trusting than the sample in the ESS surveys and the immigrant soldiers are more selected with respect to beliefs about trustworthiness. We return to the implications of this selection for the external validity of our results in the conclusion.

5. Empirical strategy

We estimate the following regression to identify the treatment effect:

$$Sent_{irt2} = \beta_1 Treated_r + \alpha_j + \gamma_s + \beta_1 X_{irt1} + \epsilon_{irt}, \quad (1)$$

where i indexes individuals, r rooms, t is time (either baseline 1 or follow up 2), j company, s session, $Sent_{irt2}$ is the amount sent, $Treated_r$ is a dummy equal to 1 if this person shares room with a minority soldier (or the share of minority soldiers in some specifications),

¹² The questions are the following: *Generally speaking, would you say that most people can be trusted or that you cannot be too careful in dealing with people?*; *Do you think most people would try to take advantage of you if they got the chance, or would they try to be fair?*; *Would you say that most of the time people try to be helpful, or that they are mostly just looking out for themselves?*

X_{irt1} is a set of individual level control variables measured at baseline (described in Section 4.2), and the error term, ϵ_{irt} , is clustered at the room level as treatment is at this level. The company fixed effects are included as the randomization was conducted within companies and session fixed effects are included to remove the influence of common experiences during the lab session. We add a vector of individual level controls, including demographics and attitudes at baseline. To make the models fully saturated, we partition the covariate space and add these control variables as indicator variables rather than using their multi-valued codings (Athey and Imbens, 2017). As the randomization procedure deviates by adding two women to a room when possible, and as female soldiers are more likely to be Norwegian, we control for having females in the room in all regressions. We show all results with and without the other individual level covariates.¹³ The regression is estimated separately for sending to the minority responder (Ali) and to the majority responder (Morten), but we also present results from an interaction model that tests the difference across the models.

6. Empirical results

6.1. Main results

The main results are presented in Tables 3 and 4. In Table 3 we present regression results of Eq. (1) using the amount sent in the trust game as the dependent variable and an indicator variable for treatment. The mean amount sent, reported at the bottom of the table, is similar across treatment states and is around 70 NOK, i.e. 70% of their endowment. This level of trust is comparable with what has been found in previous trust games in Scandinavia.¹⁴

¹³ We create an indicator for missing values in the controls and include the missing indicator in the regressions in order not to lose observations.

¹⁴ Johnson and Mislin (2011) conduct a meta study of the trust game literature. They find that there is large variation in how much subjects send on average, ranging from 22 to 96% of the total amount, but the average is equal to 50% of the endowment. The average fraction sent is 74% in the four studies from Sweden that are included in their review. Johnsen and Kvaløy (2016) find that Norwegian students on average send 71% of their endowment.

Table 4
Amount sent in the trust game and the extent of contact – share of minority soldiers in the room.

Sent amount in trust game	(1)	(2)	(3)	(4)	(5)	(6)
	Mino	Mino	Majo	Majo	All	All
Share of minority soldiers (std.)	4.23** (2.02)	4.42** (2.14)	0.46 (2.70)	1.83 (2.44)	0.24 (2.61)	1.13 (2.14)
Minority responder					0.70 (2.79)	2.26 (2.94)
Share of minority (std.) * Minority responder					3.64 (3.03)	4.04* (2.44)
Observations	296	296	296	296	592	592
Company FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes	No	Yes
Mean trust	71.45	71.45	71.28	71.28	71.37	71.37
SD trust	36.54	36.54	35.99	35.99	36.23	36.23
Mean Share of minority soldiers	−0.02	−0.02	0.02	0.02	0.00	0.00
SD Share of minority soldiers	0.98	0.98	1.02	1.02	1.00	1.00

Note: The dependent variable is amount in NOK sent to the minority/majority responder in the trust game, which ranges from NOK 0 to 100 (about USD 12). *Share of minority soldiers in room* is the standardized variable of share of minority soldiers within each room (mean 0, sd 1). Regressions (1)–(2) only include observations in treatment with the minority responder. Regressions (3)–(4) only include observations in treatment with the majority responder. Regressions (5)–(6) include observations from both treatments. All regressions include a dummy for whether the living quarters include female soldiers. Individual controls include whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, attitudes towards immigration, response to questions regarding trust, dummy variables for room size, and gender. Mean and standard deviation of dependent and independent variables are shown below. Robust standard errors in parentheses clustered on rooms (121 rooms). OLS regressions.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

In column 1 of Table 3 we see that individuals sharing room with a minority soldier send around 10 NOK more to the minority responder. In terms of magnitude, treatment increases transfers by 0.27 standard deviations (9.92/36.54). The estimate is similar when we add the controls (column 2). In both columns, the estimate is statistically significant at the 10% level. In order to investigate what effect magnitudes we can rule out, we use the equivalence testing approach of two one-sided t-tests (TOST).¹⁵ The results indicate that we can rule out any negative effects from treatment on trust in the minority candidate.

In columns 3–4, we present the same regressions but with sending to the majority responder as the dependent variable. The treatment effect is smaller and equal to 0.07 standard deviations (2.65/35.99), but it is not negative, as suggested by Putnam's (2007) constrict theory. However, using the equivalence testing approach we can not reject negative effects.¹⁶

The last two columns show the results of the interaction model. Here we find that the difference between those in the treatment group who played against the minority responder (Treated*Minority responder) and those in the control group who played against the majority responder (omitted group) is not statistically significant.

Our point estimate indicates that sharing room with a minority soldier increases trust (amount sent) towards an anonymous minority person named Ali with approximately 15%. This effect size is larger, but comparable to the effect of German citizenship on immigrants' trust towards ethnic Germans (Felfe et al., 2018). Kosfeld et al. (2005) find that treating individuals with oxytocin increases the

average amount sent by the trustor by 17%. A study that manipulates the image (the attractiveness) of the trustee finds that a person who is rated to be one standard deviation more attractive than the average sent 7.3% more than a person who is rated to be one standard deviation below the average (Wilson and Eckel, 2006). In a slightly different, dichotomous, trust game Burnham et al. (2000) show that labelling the receiver in the trust game as partner rather than an opponent, increases the fraction showing trust from 21 to 29%.

In Table 4, we investigate whether the share of minority soldiers among the roommates matters for how much the soldiers send to the minority/majority responder (Ali/Morten). The share of minority soldiers among the roommates vary from zero to 40% and has a standard deviation of 8%. We standardize the variable share of minority soldiers (to have a mean of zero and standard deviation of one) in order to simplify interpretation. We see in the first column that the share of minority soldiers in the room is positively associated with trust towards the minority responder. The estimate is similar when we add controls. In terms of magnitude, a one standard deviation increase in the share of minority soldiers increases transfers by 4.23 NOK (0.12 standard deviations, 4.23/36.54).

In columns 3 and 4 we find that intense contact with out-group members does not affect their trust in a stranger from the in-group, as the amount sent to the majority responder is not correlated with share of minority soldiers. In columns 5–6 we present the interaction models. We find that a one standard deviation difference in the share of minority soldiers in the room implies higher trust in an out-group stranger by about NOK 4 (effect size of 0.10, 3.64/36.23). The result is significant at the 10% level when the control variables are included.

6.2. Integrating the conflict and contact perspectives

Thus far we have found support for contact theory. Next we want to integrate our results with the conflict theory of diversity. Previous research has found a negative relationship between trust and the level of ethnic diversity in the respondents' area of residence (Alesina and La Ferrara, 2000; Dinesen and Sønderskov (2015); Putnam, 2007). We are particularly interested in whether contact can reduce this negative relationship. Carrell et al. (2015) find the largest effect of exposure for whites coming from American states with a low

¹⁵ These tests are equivalent to using a 90% confidence interval. We present the results from regressions (without controls) using the standardized trust measure with mean zero and standard deviation of one in Table F6 in the Appendix.

¹⁶ In Appendix Table E6 we show that we get similar estimates if we i) restrict the control vector to the variables for which we find significant imbalance in Table 2, or ii) estimate robust LASSO models on a fully saturated model—that is, we include all possible interactions between the covariates in Table 2—and use the post-double selection approach of Belloni et al. (2014) to select covariates. The estimated treatment effects in these models are 10.5 NOK (SE 5.2) and 8.9 NOK (SE 5.3) respectively. Thus, we do not believe that the treatment effect reflects imbalance between the treatment and control groups.

Fig. 1. Trust and previous exposure. *Note:* Amount sent in NOK on vertical axis, share of immigrants in the municipality on the horizontal axis. Trust in the majority responder represented by the black line/hollow dots, trust in the minority responder by the gray line/solid dots. Subjects in the main sample are grouped into equal-sized bins within each treatment based on the municipality's immigration share. 10 bins each treatment: Each dot represents the average trust and average immigration share for about 30 soldiers. 3 bins each treatment: Each dot represents the average trust and average immigration share for about 100 soldiers.

share of African Americans. Burns et al. (2016), however, find similar effects of exposure for South African students having been exposed to different degrees of racial heterogeneity during their high school education.

Fig. 1 shows how trust towards the majority/minority responder correlate with previous exposure to immigrants, measured by the share of non-western immigrants in the municipalities the soldiers' come from.¹⁷ There are 422 municipalities in Norway at the time of the experiment, and the soldiers in our main sample represent 197 different municipalities. Obviously, people were not randomly assigned to different municipalities or previous exposure, thus we do not interpret this relationship as a causal effect of immigrant share. In the left panel we group all soldiers within each treatment into equal-sized bins based on this immigrant share.

Trust in the majority responder (black line/hollow dots) is not correlated with municipality share of immigrants, but the figure shows that soldiers from municipalities with a high share of immigrants send less to the minority responder (gray line/dots). These results are consistent with the conflict hypothesis. In the right panel of the figure we group the sample into three equal-sized bins by treatment. Trust towards the minority responder is still considerably lower for the subjects from the most diverse municipalities.¹⁸

We investigate this relationship more formally in Table 5. The dependent variable is as before how much they send to the minority/majority responder (in NOK). *Municipality's immigrant share* gives the non-western immigration share in each soldier's home municipality. We have standardized the variable to simplify interpretation. In column 1 we see that an increase in the municipality's immigrant share by one standard deviation reduces trust towards the minority respondent by NOK 5.53, which corresponds to about 8% reduction

from the mean and a standardized effect size of 0.15 (5.53/36.54). The result is similar when we add control variables in column 2. We do not find any significant relationship between trust towards the majority responder and the municipality's immigrant share. In the interaction models we find that the relationship between amount sent to the minority and the majority responder is significantly different as a function of immigrant share. The standardized effect size is equal to 0.19 (6.77/36.23).

Our design with random assignment to rooms allows us to investigate if close contact with immigrants changes the relationship between the immigrant share in home municipalities and trust. The left panel of Fig. 2 depicts how trust in the minority responder varies with immigration share for those who lived in a treated (mixed room) and those who lived in a control room. In the right panel, we show sending to the majority candidate by treatment status. The patterns in the two panels are completely different: Close contact with a minority member breaks the negative association between municipality diversity and trust in out-group individuals. Minority shares in the municipality where one lives is not correlated with trust to an in-group member, and being treated does not reduce the level of in-group trust, rather the opposite.

In Table 6 we report the regression results for how the relationship between the share of immigrants in the home municipality and trust is affected by treatment. We find that within the control group, trust in the minority responder decreases in the share of immigrants. The coefficient is significant at the 10% level in column 1 and the standardized effect size 0.17 (6.05/36.54). In column 2 we add controls and the effect is statistically significant at the 5% level. This is not the case for the treatment group – the share variable and the interaction sums to about zero. The coefficient for treated is equal to 10.05, significant at the 10% level and 5% level when we add controls. The effect size is 0.28, and we can rule out any negative effects. We run the same regressions for those who played against the majority responder. We again see that share of immigration in the home municipality is uncorrelated with in-group trust and that treatment does not lower it. The effect size is very small and precise so that we

¹⁷ Immigrants include immigrants and Norwegian-born to immigrant parents in 2014. Non-western immigration encompass immigrants from countries outside of the EEA/EU, the US, Canada, Australia, and New Zealand.

¹⁸ In Section F in the Appendix we run all the analyses also at the postcode level. The results are very similar.

Table 5

The conflict hypothesis: Amount sent in the trust game and previous exposure.

	(1) Mino	(2) Mino	(3) Majo	(4) Majo	(5) All	(6) All
Municipality's immigrant share (std.)	−5.53* (2.81)	−6.48** (2.93)	1.19 (2.36)	1.62 (2.54)	1.05 (2.30)	1.23 (2.18)
Minority responder					0.39 (2.82)	1.97 (2.98)
Municipality's imm. sh. (std.) * Minority responder					−6.77* (3.60)	−8.00** (3.50)
Observations	296	296	296	296	592	592
Company FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes	No	Yes
Mean trust	71.45	71.45	71.28	71.28	71.37	71.37
SD trust	36.54	36.54	35.99	35.99	36.23	36.23
Mean ind. var. share	−0.03	−0.03	0.02	0.02	0.00	0.00
SD ind. var. share	1.00	1.00	1.00	1.00	1.00	1.00

Note: The dependent variable is amount in NOK sent to the minority/majority responder in the trust game, which ranges from NOK 0 to 100 (about USD 12). *Municipality's immigrant share*: Non-western immigration share in each soldier's home municipality (standardized, mean 0, sd 1). Regressions (1)–(2) only include observations in treatment with the minority responder. Regressions (3)–(4) only include observations in treatment with the majority responder. Regressions (5)–(6) include observations from both treatments. All regressions include a dummy for whether the living quarters include female soldiers. Individual controls include whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, attitudes towards immigration, response to questions regarding trust, dummy variables for room size, and gender. Mean and standard deviation of dependent and independent variables are shown below. Robust standard errors in parentheses clustered on rooms. OLS regressions.

* $p < 0.1$.** $p < 0.05$.*** $p < 0.01$.

can rule out positive and negative effects beyond very small ones. If anything, there is a positive interaction also for in-group trust and the coefficient is statistically significant in column 4 where controls are included.¹⁹

6.3. Heterogeneous effects

In this section, we investigate heterogeneous effects along two additional dimensions: the aptitude of the minority soldiers and the baseline attitudes of the majority soldiers. Do high aptitude minority soldiers affect their fellow roommates in a different manner compared to low-aptitude minority soldiers? Similar to Burns et al. (2016) and Carrell et al. (2015), we use a measure of the soldiers' academic achievement – self-reported average GPA during the last year of upper secondary school.²⁰

In Table 7 we test whether the GPA of the minority roommate affects the trust level of the majority soldiers in the trust game. We divide the treated soldiers into two equally sized groups: having a high or low GPA minority roommate (if there is more than one we use the average). We compare these two groups with the control group. We see in columns 1 and 2 that there is a positive and significant association between trust in the minority responder and living and working together with a minority soldier in the upper part of the academic distribution, while there is no treatment effect if the minority soldier has a low GPA. The effect size for high GPA minority roommate is equal to 0.44, and we can rule out negative effect sizes. The effect size for low GPA minority roommate is close to zero, and we cannot rule out negative effect sizes. With controls, the difference between high GPA and low GPA is statistically significant.²¹ We see

in columns 3 and 4 that the academic achievement of the minority roommate does not affect trust in the majority responder. Effect sizes are close to zero, and the effect of high aptitude minority soldiers on trust in the majority responder is quite imprecisely estimated.

These results are similar to the results of Carrell et al. (2015) who find that the effect of sharing a room with a high aptitude black student has a larger positive effect than sharing a room with a low-aptitude black student on future cross-racial roommate matches in The United States Air Force Academy. Burns et al. (2016) study roommate matching of white and black students in South Africa. They use their baseline data to measure pre-existing beliefs about the academic ability of black students and hence, have a more refined measure of belief updating potential based on the actual ability of the roommate. They measure beliefs and prejudice by two different implicit association tests, one regarding academic abilities and one regarding positive and negative attributes in general. They find that only white students who are positively surprised change their implicit bias about the relative academic ability of blacks. For the more general bias, there was a reduction for all exposed white students, irrespective of the academic ability of the roommate. We do not have any measure on initial beliefs about trustworthiness for our soldiers, but we can investigate whether the treatment is affecting different people differently depending on their baseline attitudes towards immigration and immigrants.

In the baseline survey we collected two questions related to attitudes towards immigration: *In general, immigrants have poorer work ethic than Norwegians*; and *During the first years of their stay in Norway, immigrants should receive lower social benefits than Norwegians*. We create an index based on these questions and define those who overall disagree with these statements as liberals (257 individuals). The ones who overall agree with these statements are classified as conservatives (335 individuals). We realize that this labelling is somewhat inaccurate since the index captures a mix of concerns, beliefs and policy positions on immigration, but we use it in lack of better alternatives.²²

¹⁹ The coefficient for immigrant share times treatment changes a lot from column 3 to column 4. This hints to a correlation between immigrant share in the municipality and observables, which is not surprising as immigrant share is not random. In Table E7 we show correlations between our baseline variables and immigrant share and we note that the latter correlates with parental education and employment as well as with attitudes regarding immigration and trust.

²⁰ This ranges from 1 (fail) to 6 (best). Appendix Fig. D3 shows the distribution for the average GPA for minority and majority soldiers in our sample. There is less spread in the variation of GPA for minority soldiers but otherwise the distributions are very similar.

²¹ We show in Appendix Table E8 that results are similar if we use a continuous measure of minority soldier GPA.

²² The results are very similar if we also include a third question: *Immigration leads to lower trust between the citizens of a country*. As this question is even further away from any policy or perception of immigrants we choose to keep it out, however.

Fig. 2. Amount sent in the trust game and share of immigrants. *Note:* Amount sent in NOK on vertical axis, share of immigrants in the municipality on the horizontal axis. Gray solid line captures those who are treated (live in mixed rooms) and are in the minority responder treatment. Gray dashed line captures those who are in the control group (homogenous rooms) and in the minority responder treatment. Black solid line captures those who are treated (mixed room) and in the majority responder treatment. Black dashed line captures those who are in the control group (homogenous rooms) and in the majority responder treatment. Subjects in the main sample are grouped into equal-sized bins within each treatment based on the postal code immigration share. 10 bins each treatment: Each dot represents the average trust and average immigration share for around 25 subjects in the control group, and around 5 subjects in treatment group. 3 bins each treatment: Each dot represents the average trust and average immigration share for about 80 subjects in control, and 17–19 subjects in treated.

In Table 8 we interact treatment status with the baseline attitudes. In column 1, we see that non-treated conservatives from homogenous rooms do not differ from liberals from homogenous rooms, i.e. those expressing sceptical views in the survey do not send significantly less to the minority responder in the trust game

(but we can not reject negative effects). Living in a mixed room increases trust in the minority responder significantly for liberals, but not for conservatives, as the interacted effect from being treated and conservative is negative, and of similar magnitude as the coefficient for treated. Hence, treatment affected the liberals, but did not

Table 6

Integrating the conflict and contact hypotheses: Amount sent, treatment status, and share of immigrants in the home municipality.

	(1) Mino	(2) Mino	(3) Majo	(4) Majo
Treated	10.05* (5.16)	11.91** (5.67)	3.21 (6.96)	7.40 (6.00)
Municipality's immigrant share (std.)	-6.05* (3.07)	-7.72** (3.11)	0.78 (2.38)	-0.05 (2.62)
Municipality's immigrant share(std.) * Treated	5.58 (5.61)	12.64* (6.82)	1.88 (6.66)	10.49* (6.05)
Observations	296	296	296	296
Company FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes
Mean trust	71.45	71.45	71.28	71.28
SD trust	36.54	36.54	35.99	35.99
Mean ind.var. share	-0.03	-0.03	0.02	0.02
SD ind.var. share	1.00	1.00	1.00	1.00

Note: The dependent variable is amount in NOK sent to the minority/majority responder in the trust game, which ranges from NOK 0 to 100 (about USD 12). *Municipality's immigrant share:* Non-western immigration share in each soldier's home municipality (standardized, mean 0, sd 1). Regressions (1)–(2) only include observations in treatment with the minority responder. Regressions (3)–(4) only include observations in treatment with the majority responder. All regressions include a dummy for whether the living quarters include female soldiers. Individual controls include whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, attitudes towards immigration, response to questions regarding trust, dummy variables for room size, and gender. Mean and standard deviation of dependent and independent variables are shown below. Robust standard errors in parentheses clustered on rooms. OLS regressions.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Table 7
Minority roommates' GPA and trust.

	(1) Mino	(2) Mino	(3) Majo	(4) Majo
High GPA	15.86** (7.35)	19.66*** (7.02)	3.34 (15.64)	11.64 (13.07)
Low GPA	2.12 (7.17)	−0.89 (8.25)	0.58 (6.26)	1.79 (6.46)
Observations	296	296	296	296
Company FE	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes
Mean trust	71.45	71.45	71.28	71.28
SD trust	36.54	36.54	35.99	35.99
Mean mino. GPA in room	4.12	4.12	4.13	4.13
SD mino. GPA in room	0.58	0.58	0.51	0.51

Note: The dependent variable is amount in NOK sent to the minority/majority responder in the trust game, which ranges from NOK 0 to 100 (about USD 12). *High GPA* is an indicator variable equal to 1 for those whose minority roommate has a high GPA score (equal to or above sample average), *Low GPA* is an indicator for those whose minority roommate has a low GPA score (below sample average). Regressions (1)–(2) only include observations in treatment with the minority responder. Regressions (3)–(4) only include observations in treatment with the majority responder. All regressions include a dummy for whether the living quarters include female soldiers. Individual controls include whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, attitudes towards immigration, response to questions regarding trust, dummy variables for room size, and gender. Mean and standard deviation of dependent and independent variables are shown below. Robust standard errors in parentheses clustered on rooms. OLS regressions.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

affect those expressing less liberal views at baseline. The effect size for treated is 0.63 (22.92/36.54), and we can rule out that it is negative. The effect size for the interaction between treated and holding conservative views is 0.52 (18.87/36.54), and we can rule out that it is positive.

In column 4, we see that non-treated conservatives do not trust the majority responder more or less than liberals. Living in a mixed room does not affect the level of trust of liberals – captured by the *Treated* dummy – and the effect is not statistically significantly different for conservatives – as captured by the interaction term. As individuals from municipalities with many immigrants are on average more liberal, see Table E9 in the Appendix, we also control for the municipality's immigration share in columns 2 and 5. We further add individual level controls in columns 3 and 6. The interpretation of the results remains the same.

The heterogeneity results we uncover are not completely consistent with any single model of prejudice reduction. The original formulation of the contact theory proposes that contact will reduce prejudice because negative stereotypes will be corrected when majority members have contact with representative minority members (Pettigrew, 1998). Thus, a possible prediction is then that those holding negative views will react more strongly to treatment (see Carrell et al., 2015, 11, for evidence consistent with this mechanism). However, later research in cognitive psychology on how people process new information questions the importance of learning as a mechanism for why contact might reduce prejudice (Pettigrew, 1998, 70). According to this line of research, people holding negative predispositions will not react to contact unless the information shock is very large. Pettigrew (1998) lists three other mechanisms, which can explain why contact causes less prejudice.

Table 8
Attitudes towards immigration/immigrants and trust.

	(1) Mino	(2) Mino	(3) Mino	(4) Majo	(5) Majo	(6) Majo
Conservative	4.94 (5.06)	3.58 (5.00)	3.97 (5.22)	−0.15 (4.58)	−0.15 (4.65)	−0.16 (5.91)
Treated * Conservative	−18.87* (10.20)	−16.99 (10.41)	−20.17* (11.18)	7.17 (10.81)	6.79 (10.73)	7.69 (12.32)
Treated	22.92*** (8.34)	21.02** (8.42)	24.14*** (8.79)	−1.92 (11.73)	−1.04 (11.53)	2.39 (11.18)
Municipality's immigrant share (std.)		−5.10* (2.77)	−6.47** (3.01)		1.12 (2.41)	1.86 (2.58)
Observations	296	296	296	296	296	296
Company FE	Yes	Yes	Yes	Yes	Yes	Yes
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	No	No	Yes	No	No	Yes
Mean trust	71.45	71.45	71.45	71.28	71.28	71.28
SD trust	36.54	36.54	36.54	35.99	35.99	35.99
Share conservative	0.57	0.57	0.57	0.56	0.56	0.56

Note: The dependent variable is amount in NOK sent to the minority/majority responder in the trust game, which ranges from NOK 0 to 100 (about USD 12). *Treated* denotes soldiers from mixed rooms. *Conservative* is equal to one for the half of the sample who holds more conservative views towards immigration. *Municipality's immigrant share*: Non-western immigration share in each soldier's home municipality (standardized, mean 0, sd 1). Regressions (1)–(3) only include observations in the minority responder treatment (Ali), (3)–(6) only include observations in the majority responder (Morten) treatment. All regressions include a dummy for whether the living quarters include female soldiers. Individual controls include whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, response to questions regarding trust, dummy variables for room size, and gender. Mean and standard deviation of dependent and independent variables are shown below. Robust standard errors in parentheses clustered on rooms

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

One is a reshaping of the view of the in-group. This mechanism is apparently not at work here, as we find no effect of treatment for those playing with Morten. A second mechanism is the generation of affective ties; Contact creates friendships which spill-over to positive views on the out-group. This mechanism is plausibly at play in our case but does not explain why only liberals are affected, unless they are the only ones becoming friends with the minority soldier. The third mechanism is changing behaviour in response to contact (Scacco and Warren, 2018). The claim is that behavioural change happens prior to attitudinal change, and will then cause attitudinal change if there is a dissonance between behaviour and attitude. The trust game measures behaviour, thus this mechanism is clearly in play for the liberal part of the sample. However, contact will not decrease prejudice, since it is those with liberal views that respond to treatment, thus few of those responding to treatment will experience dissonance between behaviour and attitude.

7. Concluding remarks

Based on previous literature on the relationship between trust and ethnic diversity, we would expect that increasing diversity leads to lower trust (Alesina and La Ferrara, 2000; Dinesen and Sønderskov, 2015; Putnam, 2007). There are, however, three important limitations to the existing literature that we address using a combined lab and field experiment.

The first limitation concerns biases arising from endogeneity issues. The worry that the correlations between diversity and trust are driven by selection, reverse causality, or both looms large in the previous literature. People self-select into neighborhoods and controlling for selection by including observables is likely to be insufficient. To date, there is no study using exogenous variation to identify the causal effect of diversity on trust. Our research design, involving a randomized field experiment, allows us to make causal inferences.

The second shortcoming of the previous literature is a conceptual conflation of exposure and contact. While living in an area with many immigrants increases exposure, it does not necessarily increase contact. A consensus has emerged in social psychology that shallow exposure need not produce the same beneficial effects, instead it is likely to cause opposite effects due to competition about jobs, resources, and cultural hegemony (see Pettigrew 1998 for a review). We investigate the correlations between ethnic diversity in soldiers' municipality of upbringing and trust, and replicate the frequently found result that there is less out-group trust among people from more diverse areas. Well identified studies have shown, however, that close personal contact reduces prejudice (e.g. Bauer et al., 2018; Boisjoly et al., 2006; Burns et al., 2016; Carrell et al., 2015; Dahl et al., 2018; Finseraas and Kotsadam, 2017; Finseraas et al., 2016), illustrating the danger of conflating exposure and contact. Our field experiment takes place in a setting which should produce the beneficial effects of contact, and we show that contact overturns the negative correlation between exposure and trust.

The third limitation regards the measurement of trust. Most previous literature on the effects of diversity on trust relies upon survey questions on general trust. There is a debate about what these questions really measure, and some argue that they correlate with trustworthiness rather than trust (Glaeser et al., 2000; Sapienza et al., 2013). We instead measure trust by conducting a trust game, which provides us with a behavioural measure of trust. More importantly, the generalized trust questions do not separate between in-group and out-group trust. We focus on majority individuals and let them send money to either Ali (signaling out-group) or Morten (in-group). This allows us to test if the correlations as well as the effects are different for in- and out-group trust.

We find that close contact increases trust in Ali, in particular for those from municipalities with a high share of immigrants. The policy implications of the results depend on subjective opinions on the external validity of the findings. In particular, three factors are important in this respect. Firstly, our sample consists of special representatives of the Norwegian population. While military service was mandatory for men in Norway until 2015 (from 2016, it is mandatory for both men and women), conscription is based on need, and only about one in six men serve. The military thereby select people based on ability and motivation. When we compare our sample to other young Norwegians they seem relatively similar, but they are somewhat more progressive with respect to their attitudes towards immigrants and they seem somewhat more trusting. Secondly, the soldiers are exposed to a highly selected set of immigrants. They are mostly second-generation immigrants, and even as compared to second-generation immigrants in general, they are likely to be better integrated. For instance, their mothers are more likely to work than the mothers of second-generation immigrants in general. Thirdly, and perhaps most important, the setting under which contact occurred is very special. Although the context of our study is in part a necessity for deriving clear theoretical expectations, and while it assures a strong internal validity, it restricts external validity to contexts with some similarity to ours. The structure of contact at workplaces, in classrooms, and in team sports are weaker and less streamlined which might imply that treatment effects from direct contact might be weaker than what we find. We strongly urge future studies to vary these different components in order to create a more general knowledge.

Acknowledgments

We would like to thank the Norwegian Defense Research Establishment (FFI), and in particular Frank Steder. This study could not have been conducted without the help of FFI and their project "Research on Cohorts". We thank Eliana La Ferrara, Eva Mørk and seminar participants at Carlos III in Madrid, IAS in Norrköping, Linköping University, the University of Oslo, the Norwegian University of Life Sciences, Statistics Norway, the ESOP Workshop: Learning in Social Situations, the ESA European meeting 2016 in Bergen, the AEA Annual Meeting in Chicago, INAS 2017 in Oslo, RC28 2017 at Columbia University in New York, and the Nordic Conference on Behavioral and Experimental Economics in Oslo for helpful comments. Grant numbers 236801 and 270772 (Research Council of Norway) is acknowledged. The project is part of the research activities at the Centre for the Study of Equality, Social Organization, and Performance (ESOP) at the Department of Economics, University of Oslo. ESOP is supported by the Research Council of Norway. We thank research assistants Kristine D. Ruud and Øyvind Bruer-Skarsbø.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jpubeco.2019.02.002>.

Reference

- Abascal, Maria, Baldassarri, Delia, 2015. Love thy neighbor? Ethnoracial diversity and trust reexamined. *Am. J. Sociol.* 121 (3), 722–782.
- Alesina, Alberto, Devleeschauwer, Arnaud, Easterly, William, Kurlat, Sergio, Wacziarg, Romain, 2003. Fractionalization. *J. Econ. Growth* 8 (2), 155–194.
- Alesina, Alberto, Harnoss, Johann, Rapoport, Hillel, 2016. Birthplace diversity and economic prosperity. *J. Econ. Growth* 21 (2), 101–138.
- Alesina, Alberto, La Ferrara, Eliana, 2000. Participation in heterogeneous communities. *J. Econ. Lit.* 115 (3), 847–904.
- Alesina, Alberto, La Ferrara, Eliana, 2002. Who trusts others? *J. Public Econ.* 85 (2), 207–234.

- Alesina, Alberto, La Ferrara, Eliana, 2005. Ethnic diversity and economic performance. *J. Econ. Lit.* 43 (3), 762–800.
- Alexander, Marcus, Christia, Fotini, 2011. Context modularity of human altruism. *Science* 334 (6061), 1392–1394.
- Algan, Yann, Cahuc, Pierre, 2013. Trust and growth. *Annu. Rev. Econ.* 5 (1), 521–549.
- Allport, Gordon W., 1954. *The Nature of Prejudice*. Addison-Wesley, Reading.
- Angrist, J.D., 2014. The perils of peer effects. *Labour Econ.* 30, 98–108.
- Athey, Susan, Imbens, Guido W., 2017. The econometrics of randomized experiments. *Handbook of Economic Field Experiments*. 1. Elsevier, pp. 73–140.
- Bauer, Michal, Fiala, Nathan, Levely, Ian, 2018. Trusting former rebels: an experimental approach to understanding reintegration after civil war. *Econ. J.* 128 (613), 1786–1819.
- Beach, Brian, Jones, Daniel B., 2017. Gridlock: ethnic diversity in government and the provision of public goods. *Am. Econ. J. Econ. Pol.* 9 (1), 112–136.
- Belloni, Alexandre, Chernozhukov, Victor, Hansen, Christian, 2014. Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* 81 (2), 608–650.
- Berg, Joyce, Dickhaut, John, McCabe, Kevin, 1995. Trust, reciprocity, and social history. *Games Econ. Behav.* 10 (1), 122–142.
- Boisjoly, Johanne, Duncan, Greg J., Kremer, Michael, Levy, Dan M., Eccles, Jacque, 2006. Empathy or antipathy? The impact of diversity. *Am. Econ. Rev.* 96 (5), 1890–1905.
- Brandts, Jordi, Charness, Gary, 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. *Exp. Econ.* 14 (3), 375–398.
- Burnham, Terence, McCabe, Kevin, Smith, Vernon L., 2000. Friend-or-foe intentionality priming in an extensive form trust game. *J. Econ. Behav. Organ.* 43 (1), 57–73.
- Burns, Justine, Corno, Lucia, La Ferrara, Eliana, 2016. Interaction, prejudice and performance. Evidence from South Africa. Working Paper.
- Carrell, Scott E., Hoekstra, Mark, West, James E., 2015. The impact of intergroup contact on racial attitudes and revealed preferences. NBER Working Paper No. 20940. <http://www.nber.org/papers/w20940>.
- Dahl, Gordon, Kotsadam, Andreas, Rooth, Dan Olof, 2018. Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams. National Bureau of Economic Research.
- Dinesen, Peter Thisted, Sønderskov, Kim Mannemar, 2015. Ethnic diversity and social trust: evidence from the micro-context. *Am. Sociol. Rev.* 80 (3), 550–573.
- Dinesen, Peter Thisted, Sønderskov, Kim Mannemar, 2017. Ethnic diversity and social trust: a critical review of the literature and suggestions for a research agenda. *Oxford Handbook on Social and Political Trust*. Oxford University Press, pp. 175–204.
- Enos, Ryan D., 2014. Causal effect of intergroup contact on exclusionary attitudes. *Proc. Natl. Acad. Sci.* 111 (10), 3699–3704.
- Enos, Ryan D., 2016. What the demolition of public housing teaches us about the impact of racial threat on political behavior. *Am. J. Polit. Sci.* 60 (1), 123–142.
- Enos, Ryan D., 2017. *The Space Between Us: Social Geography and Politics*. Cambridge University Press.
- Enos, Ryan D., Celaya, Christopher, 2018. The effect of segregation on intergroup relations. *Journal of Experimental Political Science* 1–13.
- Falk, Armin, Zehnder, Christian, 2013. A city-wide experiment on trust discrimination. *J. Public Econ.* 100, 15–27.
- Fehr, Ernst, 2009. On the economics and biology of trust. *J. Eur. Econ. Assoc.* 7 (2–3), 235–266.
- Felfe, Christina, Kocher, Martin G., Rainer, Helmut, Saurer, Judith, Siedler, Thomas, 2018. More opportunity, more cooperation? The behavioral effects of birthright citizenship on immigrant youth. Working Paper.
- Fershtman, Chaim, Gneezy, Uri, 2001. Discrimination in a segmented society: an experimental approach. *Q. J. Econ.* 351–377.
- Finseraas, Henning, Johnsen, Åshild A., Kotsadam, Andreas, Torsvik, Gaute, 2016. Exposure to female colleagues breaks the glass ceiling: evidence from a combined vignette and field experiment. *Eur. Econ. Rev.* 90, 363–374.
- Finseraas, Henning, Kotsadam, Andreas, 2017. Does personal contact with ethnic minorities affect anti-immigrant sentiments? Evidence from a field experiment. *Eur. J. Polit. Res.* 56 (3), 703–722.
- Glaeser, Edward L., Laibson, David I., Scheinkman, Jose A., Soutter, Christine L., 2000. Measuring trust. *Q. J. Econ.* 115 (3), 811–846.
- Goette, Lorenz, Huffman, David, Meier, Stephan, 2006. The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. *Am. Econ. Rev.* 96 (2), 212–216.
- Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2008a. Alfred marshall lecture social capital as good culture. *J. Eur. Econ. Assoc.* 6 (2–3), 295–320.
- Guiso, Luigi, Sapienza, Paola, Zingales, Luigi, 2008b. Trusting the stock market. *J. Financ.* 63 (6), 2557–2600.
- Guiso, L., Sapienza, P., Zingales, L., 2011. Civic Capital as the Missing Link. In: Benhabib, J., Jackson, M.O. (Eds.), *Handbook of Social Economics*. A. Bisin. vol. 1A.
- Hjort, Jonas, 2014. Ethnic divisions and production in firms. *Q. J. Econ.* 129 (4), 1899–1946.
- Hunt, Jennifer, Gauthier-Loiselle, Marjolaine, 2010. How much does immigration boost innovation? *Am. Econ. J. Macroecon.* 2 (2), 31–56.
- Johnsen, Åshild A., Kvaløy, Ola, 2016. Does strategic kindness crowd out prosocial behavior? *J. Econ. Behav. Organ.* 132 (Part A), 1–11.
- Johnson, Noel D., Mislin, Alexandra A., 2011. Trust games: a meta-analysis. *J. Econ. Psychol.* 32 (5), 865–889.
- Johnston, Christopher D., Newman, Benjamin J., Velez, Yamil, 2015. Ethnic change, personality, and polarization over immigration in the American public. *Public Opin. Q.* 79 (3), 662–686.
- Kaufmann, E., Goodwin, M., 2016. *The Diversity Wave: A Meta-analysis of Ethnic Diversity, Perceived Threat and Native White Backlash*.
- Kosfeld, Michael, Heinrichs, Markus, Zak, Paul J., Fischbacher, Urs, Fehr, Ernst, 2005. Oxytocin increases trust in humans. *Nature* 435 (7042), 673.
- Lyons, Elizabeth, 2017. Team production in international labor markets: experimental evidence from the field. *Am. Econ. J. Appl. Econ.* 9 (3), 70–104.
- Meer, Tom van der, Tolsma, Jochem, 2014. Ethnic diversity and its effects on social cohesion. *Annu. Rev. Sociol.* 40, 459–478.
- Ottaviano, Gianmarco IP, Peri, Giovanni, 2006. The economic value of cultural diversity: evidence from US cities. *J. Econ. Geogr.* 6 (1), 9–44.
- Peri, Giovanni, 2012. The effect of immigration on productivity: evidence from US states. *Rev. Econ. Stat.* 94 (1), 348–358.
- Pettigrew, Thomas F., 1998. Intergroup contact theory. *Annu. Rev. Psychol.* 49 (1), 65–85.
- Putnam, Robert D., 2007. E pluribus unum: diversity and community in the twenty-first century. *Scand. Polit. Stud.* 30 (2), 137–174.
- Rao, Gautam, 2019. Familiarity does not breed contempt: generosity, discrimination and diversity in Delhi schools. *Am. Econ. Rev.* 109 (3), 774–809.
- Sapienza, Paola, Toldra-Simats, Anna, Zingales, Luigi, 2013. Understanding trust. *Econ. J.* 123 (573), 1313–1332.
- Scacco, Alexandra, Warren, Shana S., 2018. Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria. *American Political Science Review* 1–24.
- Schaeffer, Merlin, 2014. *Ethnic Diversity and Social Cohesion: Immigration, Ethnic Fractionalization and Potentials for Civic Action*. Ashgate Publishing, Ltd.
- Shore, Lynn M., Chung-Herrera, Beth G., Dean, Michelle A., Ehrhart, Karen Holcombe, Jung, Don I., Randel, Amy E., Singh, Gangaram, 2009. Diversity in organizations: where are we now and where are we going? *Hum. Resour. Manag. Rev.* 19 (2), 117–133.
- Stanley, Damian A., Sokol-Hessner, Peter, Banaji, Mahzarin R., Phelps, Elizabeth A., 2011. Implicit race attitudes predict trustworthiness judgments and economic trust decisions. *Proc. Natl. Acad. Sci.* 108 (19), 7710–7715.
- Thöni, Christian, Tyran, Jean Robert, Wengström, Erik, 2012. Microfoundations of social capital. *J. Public Econ.* 96 (7–8), 635–643.
- Van Laar, C., Levin, Shana, Sinclair, Stacey, Sidanius, Jim, 2005. The effect of university roommate contact on ethnic attitudes and behavior. *J. Exp. Soc. Psychol.* 41 (4), 329–345.
- Wilson, Rick, Eckel, Catherine, 2006. Judging a book by its cover: beauty and expectations in the trust game. *Polit. Res. Q.* 59 (2), 189–202.

Paper IV

Risk preferences and adaptive behavior at the workplace: Evidence from the military

Torbjørn Hanson

Published in *Unpublished*,

IV

Risk preferences and adaptive behavior at the workplace: Evidence from the military

Torbjørn Hanson*

June, 2019

Abstract

Gender differences in risk preferences are found in the general population. But differences seem to disappear among professionals due to self selection or adaptive behavior. We study gender differences in risk preferences at the workplace by establishing a baseline measure the first day soldiers in the Norwegian Armed Forces enter service. After nine months of service we find that soldiers dislike risk more. Randomized intense collaborative exposure to peer soldiers is found as a mechanism for change in preferences. Despite a significant gender difference in risk preferences at baseline we find only weak evidence for changes in the gap during service.

*Phd-candidate at University of Oslo, and researcher at Norwegian Defence Research Establishment (FFI). E-mail: tbh@ffi.no. Acknowledgements: I am grateful to Andreas Kotsadam, Kari R. Strand, Petter Y. Lindgren and Sverre Kvalvik, for comments on this paper, and to Uri Gneezy for his comments on an early draft.

1 Introduction

Gender differences are found in the general population for several attributes including risk preferences. But differences seem to disappear among professionals. The proposed explanation is that individuals self select into professions according to their risk preferences (Croson and Gneezy, 2009). For example are gender differences in risk taking negligible for mutual fund investors when controlled for investor knowledge (Dwyer et al., 2002). We argue on the other hand that absence of gender gaps among professionals is due not only to self selection but also to adaptive behavior at the workplace.¹ By simply including a baseline measure of risk preferences before individuals enter the "risky" profession we can distinguish between the effects of self selection and adaptive behavior.

Croson and Gneezy (2009) studied the literature on gender differences and suggest that gender gaps in the labor market could partly be explained by the fact that women prefer jobs that involves less risk, are more socially virtuous and less competitive. They found from the many different tasks covered by the literature that certain tasks such as intellectual and physical risks produced larger gender differences than e.g. health risks like smoking. The phenomenon is explained by the claim that males are more likely to see a situation involving risk as a challenge that calls for participation while females tend to respond to these situations as threats that encourage avoidance (Arch, 1993). However, evidence is more mixed when it comes to financial risk (Croson and Gneezy, 2009).

The Armed Forces could represent a promising setting for natural experiments. Thousands of new conscripts enter at the same time each year making a baseline measure for risk preferences possible. In the Norwegian Army more than 2000 conscripts enter at the same day and location. We designed a survey completed by conscripts during their first day of service while knowledge and experience of the military and the service is at its minimum. Results from this first wave are considered our baseline. Eight weeks later, after completing the initial training period (boot camp), a follow-up survey is conducted in a second wave. Finally, after nine months, towards the end of the service, a third wave completes the survey period. The surveys include four different risk measures to elicit domain specific, profession specific (police, fire

¹Nguyen (2011) explores the possibility that working in certain occupations may influence people's risk preferences, and finds that the longer the worker has been working in the same environment, the more his references for risk are adopted to it.

fighters, rescue, soldiers, etc.), physical (physical injury) and general risk preferences. We find that domain specific risk preferences changed significantly from the first to the third wave, indicating soldiers adapting to military service. However, the baseline gender differences persist during the period and we fail to reject the hypothesis that women (minority group) fully adapt to male (majority group) risk preferences.

Adaption at the workplace could occur through several mechanisms. E.g. Ahern et al. (2014) found peer effects in risk preferences for first year MBA students divided into section groups, while e.g. Stinebrickner and Stinebrickner (2006) studied peer effects in academic skills among college roommates. Carrell et al. (2013) also studied peer effects in academic performance but in a military setting. We believe social interactions between conscripts which are roommates, sharing a room in the barracks (roommate is often also equivalent to team mate), could explain possible changes in risk preferences. If present, peer effects could be exploited in order to influence risk preferences towards a desired level. While gender differences in competitive behavior are proposed counter attacked by affirmative action (Niederle et al. (2013); Balafoutas and Sutter (2012)), we argue that policies related to peer effects should be investigated in the case of risk aversion preventing females from enlistment.

Estimation of causal peer effects relies, among other factors, on subjects being randomly assigned to rooms. We take advantage of the fact that officers were instructed to use a randomizer designed by researchers for room assignments to study peer effects on risk preferences.² Our findings indicate significant peer influence on the soldiers' domain specific risk preferences.

At the time of our study only men were conscripted in Norway, but enrollment and suitability assessments were mandatory also for women.³ This means that the female population in our study is self selected to service and unrepresentative for the female population in general. This is not the case to the same extent for our male population. Only one out of six in each cohort is needed for duty. This means that men are selected in a larger extent according to their ability and motivation. If not fully representing the general male population, we still expect our male population

²Finseraas et al. (2019) find significant peer effects in trust among roommates in data from the same subsample. In previous contingents in the Norwegian Army, Finseraas et al. (2016) and Dahl et al. (2018) find effects from exposure to female soldiers while Finseraas and Kotsadam (2017) from exposure to ethnic minority soldiers.

³Enrollment and suitability assessments are carried out in order to screen people not suitable for service and to measure motivation for service. Recently, also women are conscripted and the first female conscripts did service from August 2016. Of course, this reform is of great interest to research concerns.

to be closer to the general than our purely self-selected female population. Given that military service is associated with more risk, relative to each respective subject's relevant options, risk aversion could hamper the Armed Forces in recruiting sufficient number of females or exclude otherwise top candidates for recruitment to permanent positions. However, application for enlistment is handed in close to the end of the 12 months service. This means that relevant risk preferences for enlistment (recruitment) appear not until after around ten months of service. Timing of risk preferences has considerable policy implications for the Armed Forces. If risk preferences adapt during service, a gap at the beginning of service has little impact on whether the subject is recruited for permanent service or not. A potential gap in risk preferences before the service starts should then be of particular interest to the military, and by altering its risk reputation (the level of risk subjects expect from service) the Armed Forces can exploit expectations about the 12 months service to recruit more of top candidates to permanent positions.

The main contribution of the present paper is twofold: By introducing a baseline measure for risk preferences we are able to study any adaption of risk preferences for subjects within a professional setting. Second, we design a study testing for peer effects as a source to adapting risk preferences at the workplace, where previous studies concentrate on student environments only.

The paper proceeds as follows: In Section 2 we present our research design and field experiment. Section 3 illustrates the survey for eliciting risk preferences and the empirical strategy for estimating peer effects. We describe the data in Section 4. The empirical results are presented in Section 5. Finally, Section 6 concludes and points to further research. Additional descriptive statistics and regression tables are included in the appendices.

2 Identifying adaption in risk preferences: Research design

As noted by Croson and Gneezy (2009) the findings of no gender gap in risk preferences for risky professions could also be a result from adaptive behavior to the requirements of the job. By introducing a baseline measure for risk preferences prior to any influence from workplace environments we can track changes in risk preferences related to job environments. The military is perhaps one of very few workplaces

where estimation of such a baseline is possible in practice. We estimate a baseline (wave 1) for more than 1700 conscripts at their first day of service and follow up at two later events (wave 2 and 3) during service to study any changes in preferences after exposure to job environment. This is in contrast to previous studies, i.g. Dwyer et al. (2002), where risk preferences are estimated after the subject is exposed to the job environment and where it is very difficult, if not impossible in practice, to control for environmental factors related to adaption.

The sample for measuring baseline risk preferences consists of incoming soldiers of the August 2015-contingent of the North Brigade of the Norwegian Army. The first day of service starts at a military camp close to Oslo. At the camp, the soldiers go through a screening program of medical testing and interviews before they board planes to Northern Norway and start the basic training period at their designated Army unit. The program also includes a questionnaire, which constitutes our baseline measure of risk preferences (wave 1). To attend the program soldiers are divided into groups of 20 – 30 respondents. For the questionnaire each group is gathered in a class room and the soldiers are asked to use their smart phone or tablet to go online and complete a survey.⁴ The instructor stresses that the survey is not a part of any official tests, and that results are used for research purposes only and are anonymous to representatives of the Armed Forces

When arriving their designated Army unit the soldiers are assigned to rooms for the eight weeks of the recruit period. An Excel-sheet is programmed by researchers in advance for the officers to randomize from name lists. We provided the personnel officers in charge of room assignment with an excel sheet which they were instructed to use to randomize soldiers into rooms. First, officers declare room size for all rooms. Second, all females are assigned randomly to declared rooms, but with at least two females in each mixed room. Finally, each room is filled with randomly assigned male soldiers. This design, with randomized room assignment, lets us study any causal peer effects from roommates on the soldiers risk preferences.

The recruit period is the basic training period before soldiers specialize in given tasks and are selected to regular companies. Unfortunately, we were not allowed to randomize room assignments during the regular service. Thus, studies of peer effects are limited to exposure from the first eight weeks of service. The recruit period is subject to strict enforcement of military rules and regulations. The soldiers are to wear their uniform 24/7, and the first extended leave is not granted until completion

⁴A paper form was available as backup in case of technical problems.

of the recruit period. Sharing room constitutes intense treatment in the form of personal contact with other members of the room since it also usually constitutes a squad within a platoon in the company. A normal day of boot camp starts with activities within the room, such as cleaning the room before inspection. Working hours are intense, usually 10–15 hours a day in addition to preparation of individual gear after duty ends. This leaves the soldiers with few opportunities for personal chores and socializing outside their own room. At the end of the eight week recruit period we conducted the first following up survey (wave 2) including the questionnaire for measuring risk preferences. Soldiers are gathered according to troop or company affiliation, and complete a slightly modified version of the survey on their electronic device.

After nine months of service we conducted the third follow-up survey (wave 3) including the same questionnaire. Again, soldiers are gathered according to troop or company affiliation, and complete the survey on their electronic device at hand. At this time of service soldiers are fully educated in individual tasks and their unit is usually trained to interact with other units. Thus, the soldiers are used to the military environment and have a fair impression of what a career in the Armed Forces would look like. However, they have not yet applied for a regular position after the end of duty. In this sense the soldiers are at their peak of service, fully educated and those interested in an extended military career still competing for future options within the Armed Forces.

3 Eliciting risk preferences: The survey

Risky behavior is required for soldiers, and in extreme scenarios resulting in death for oneself or others. But such outcomes are of course rare in peace time practice for the Norwegian Armed Forces.⁵ However, during training activities risk is involved which could result in physical damages and permanently reduced physical and mental conditions.⁶ We argue that risk preferences for financial outcomes are relevant in this

⁵A survey among conscripted personnel in the Norwegian Army shows that soldiers on average take on more risk after joining the Army than prior to service (FFI-report 16/01012: <https://www.ffi.no/no/Rapporter/16-01012.pdf>. Available in Norwegian only.)

⁶Statistics from the Norwegian Armed Forces Joint Medical Services shows that 838 injuries were registered in the Norwegian Armed Forces in 2017. About half of the reported injuries were defined as a minor injury. In addition to various types of accidents are injuries related to cold climate a frequent

setting only implicitly through the unknown or at least ambiguous effect from service on future income.⁷ However, other risk correlated preferences, such as preferences for competition, could affect the decision for joining the military voluntarily. The fact that girls prefer competing other girls over boys (Croson and Gneezy, 2009), is to some extent taken into consideration by the Army in form of offering girls-only "boot camps" as preparation for service and female-only troops during the selection process in the Special Forces. Choosing methods for eliciting risk preferences, risk related to physical and mental (perhaps also social aspects) outcomes are given the highest weight as we believe such outcomes are the most important for subjects deciding on joining the Army.

Dohmen et al. (2011) studied how well different methods elicited risk preferences. They found that a general risk question performed fairly well in a field experiment with real money stakes. When tested specifically for different domains, the best predictor of behavior was the corresponding domain-specific question, but the general risk question still out-performed more complex methods. Complex and time consuming methods are difficult to implement in our setting where a huge number of soldiers are ran through medical tests and interviews during their first day of service (wave 1). A cost-benefit perspective clearly favors a general risk question over more complex methods in our setting. In addition Charness et al. (2013) conclude from a literature survey that the main disadvantage of complex methods is that, depending on the population, a significant number of subjects will fail to understand the procedure. This is of particular relevance to our setting of soldiers stressed from several tests trying to cope with a completely new environment.

Furthermore, Charness et al. (2013) state that simpler methods have the advantage of being more straight forward and capable of eliciting sensible risk preferences from a broader set of individuals, but warn researchers from extrapolating from such measures to other domains. They conclude that researchers should choose the method that is best suited for the particular question asked as well as the sample population being used. The finding that estimated risk preferences and parameters may not

source.

⁷Evidence on labor-market effects from conscription and drafts are mixed. Bauer et al. (2012) found that, after controlling for healthy men being selected into service, conscription had no significant effect on earnings. In contrast Angrist (1990) found that earnings of American white Vietnam veterans were approximately 15 percent less than the earnings of comparable nonveterans. However, risk and possible damage related to duty in Vietnam is of course very far from perspectives relevant concerning Norwegian conscripts.

have predictive power for behavior across domains is stressed also in Dohmen et al. (2011). They found that a standard lottery measure had little predictive power for real world, risky behavior in such domains as employment, car driving and health. Manski (2000) suggests that surveys provide a more direct measure of attitudes than outcomes.

Frey et al. (2017) study to what extent there exist a general factor of risk preferences by integrating multiple risk-taking measures from various measurement traditions. Their results suggest that risk preferences has a psychometric structure which involves both a general, stable component that can account for about half of the explained variance, and a series of facets capturing more specific aspects of risk preferences. Thus, the inclusion of measures of both domain specific and more general risk preferences in the present study is in line with the suggested structure of risk preferences.

We study the outcome of a questionnaire measuring general, domain specific, profession specific and physical (domain related) risk preferences based on five different questions. The questionnaire is included in a survey of in total 90 questions on demographics and various other topics, used in each of the three waves. We measure general risk preferences by evaluating the question "How strongly do you dislike risk" on a seven points scale from "not at all" to "very strongly". Domain specific risk preferences are evaluated based on a question directed towards subjects' behavior in a real world situation likely experienced during service: The subject's platoon has decided to take a shortcut and cross an icy lake on their snowcat. Additional information is given in form of a weather report indicating that ice quality has suffered significantly from a recent rise in temperature. The subject indicates on a seven points scale how strongly he/she dislikes the risk involved by crossing the icy lake. Because soldiers could update their expectations about outcomes associated with risky behavior through learning during service, we include a question where the respondents are asked to state the consequence of an accident where a snowcat goes through the ice. Outcomes are stated on a four points scale from "Very small. You are picked up immediately by other soldiers" to "Catastrophic. You drown or freeze to death very quickly". In order to evaluate risk preferences related to a profession which involves risky operations, but not limited to a career in the Armed Forces, we ask respondents if they "would consider a profession involving risky missions, such as special forces in the police, fire fighter, rescue or the armed forces". As noted earlier, military service involves a significant risk of physical injury and we argue

that preferences related to avoiding physical injury is domain relevant. We evaluate the risk preference for avoiding physical injury by the question "How strongly do you dislike the risk of injury related to physical activity" on a seven points scale from "not at all" to "very strongly".

3.1 Mechanism for adjustments in preferences: Peer effects

One possible environmental determinant for risk preferences is peer influence. As noted in Section 2, sharing a room constitutes intense treatment in the form of personal contact with other room members, which usually also represents the members of their squad. This means that roommates are likely to spend most of the time during service together and could influence attitudes of their peers. Ahern et al. (2014) study peer effects for financial risk preferences among MBA students and is, to our knowledge, the only study on this subject in the literature which also randomize the individual's exposure to peers. Furthermore, peer effects related to risk are also found in the health domain, e.g. Card and Giuliano (2013) study risky behavior (sex, smoking, marijuana use, truancy) of best-friend pairs, and in a school environment (Eckel et al., 2012). Ahern et al. (2014) find that when an individual's randomly assigned peers are relatively more risk averse prior to starting MBA, the individual is relatively more risk averse after the first year. Adjusting to a relatively more risky environment makes peer influence more relevant for conscripts compared to the usual MBA or college setting studies. In addition, we argue that military service represents a relevant workplace environment where the respondents face real risk and make on the job decisions based on their risk preferences. This is in contrast to studies of risk preferences among students, where decisions are made outside any real workplace environment.

Causal peer effects operates through endogenous and contextual effects. Manski (1993) shows that only under restrictive conditions can one identify whether an individual's choice is driven by peer choices, endogenous effects, or to exogenous characteristics of peers in form of contextual effects. To isolate one effect requires an exogenous shock to attitudes of a subject's peers, while leaving the subject unaffected. To identify true causal peer effects we randomly assign soldiers to rooms. This removes possible correlated group effects caused by self-selection.

Room assignments must be orthogonal to the underlying determinants of risk pref-

erences in order for the randomization to eliminate unobserved group effects. Therefore, we control for the soldiers' socioeconomic background characteristics. Shocks specific to soldiers within a single unit, team or room could result in biased estimates. This could be handled by using results from our first wave, which reflects predetermined risk preferences the first day of service before any room-specific shocks could occur. Using predetermined measures for risk preferences also allows for causal interpretation of results by handling the Manski (1993) reflection problem. Using a standard linear-in-means model we can estimate peer effects by

$$RPM_{j,i,s,2} = \alpha_p + \beta_1 R\bar{P}M_{j,-i,s,1} + \beta_2 X_{i,s,1} + \epsilon_{i,s} \quad (1)$$

$RPM_{j,i,s,t}$ is risk preferences of subject i in peer group s measured by survey question j in the second wave ($t=2$) at the end of the initial training period or after nine months of service ($t=3$). The risk preferences of subject i is a function of predetermined average peer group risk preferences excluding subject i 's preferences measured in wave 1, $R\bar{P}M_{j,-i,s,1}$, and the socioeconomic individual-level control variables, $X_{i,s,1}$, measured at baseline ($t=1$). We interpret β_1 as the peer influence on risk preferences. Platoon fixed effects, α_p , are included since randomization occurred at the platoon level, while standard errors will be clustered on rooms as treatment is at the room level.

4 Data

4.1 Sample and background characteristics

We conducted the survey on 1741 conscripted soldiers during the period from August 2015 to April 2016, which constitutes our main sample. However, a high dismissal rate is normal during the recruit period. This reduces the number of respondents in later waves. Further, due to restrictions on the number of questions in the survey, some questions were given only to a random subsample of respondents. Based on the composition of respondents to each of the four questions in each wave we define various subsamples. We define our main subsample as the panel of respondents where we have observations for the domain specific measure and the profession specific measure in all three waves, $n=788$. The subsample for the randomly drawn general measures is of size $n=844$. Based on the survey we identified the following background characteristics for the soldiers: age, whether mother/father work, mother's/father's level

of education, whether parents are divorced and the soldier’s educational plans.⁸ In addition we control for test scores from the screening process.⁹ Descriptive statistics for background variables are given in Table A.1 in the Appendix. The share of females in the sample is around 15 percent. We note that 66 percent of male and 86 percent of female respondents plan for higher education. This significant gender difference is consistent with the enrollment statistics for higher education in the general population in Norway. There is also a significant gender difference when it comes to minority background, where the share is 3.4 and 2.6 percent for male and female respondents, respectively.

Due to the fact that not all soldiers were assigned randomly to rooms during the recruit period, our sample for studies of peer effects is limited to the companies which followed the protocol for randomization. This leaves us with a sample of 596 respondents in wave 2. In wave 3 the sample is reduced to 329 and 244 for the domain specific and profession specific measure, respectively.

5 Empirical results

5.1 Baseline preferences

At first day of service prior to any adapting behavior we expected to find a small, if any, gender gap in risk preferences compared to the population in general due to self selection among female soldiers. We find that male soldiers on average dislike risk less compared to female soldiers, see the descriptive statistics presented in Table A.2 in the Appendix. In fact, the regression results in Table 1 indicate a significant gender difference at baseline in all of the four different risk measures. But the gender difference in general risk preferences is not significant when including the soldiers’ background characteristics in the regression. The results are, however, somewhat different for domain and profession specific preferences if we only consider the panel of both wave 1 and 3 respondents, and control for background characteristics (see Table 2). In this case the gender difference is not significant at baseline, but only after nine months for the domestic measure. For the profession risk measure we find

⁸Several factors are found to be correlated with risk preferences in the economics literature, such as parental education (Dohmen et al., 2009), ethnicity and immigration status (Gruber, 2000; Barsky et al., 1997), cognitive skills (Dohmen et al., 2010) and age (Dohmen et al., 2009).

⁹The soldiers’ cognitive skills and vocabulary are tested.

no significant gender differences when the controls are included.¹⁰ We elaborate on the findings from the panel of respondents in the following Section.

5.2 Changes in risk preferences

In Table 2 and 3 we investigate the claim that adjustments in risk preferences at the workplace eliminate gender differences.¹¹ If risk preferences were to adapt during service we expect to observe effects from changes in measured preferences on the estimated gender difference in subsequent waves. We find only weak evidence for the claim in one of four different risk measures.

At baseline we found no significant gender differences in domain specific risk preferences for the panel of soldiers, cf. Table 2. Table 3 shows the changes in preferences from baseline to wave three for female and male soldiers. We see that both male and female change their preferences significantly in the same direction, and dislike the domain specific risk more strongly after nine months of service.¹² In contrast to the baseline results we now find evidence of a significant gender difference in preferences, see regression (2) in Table 2. Despite the significant gender difference in wave 3, there is no evidence supporting the claim that gender differences change from wave 1 to 3, as the Wald test fail to reject the hypothesis that gender coefficients are equal in size between the two waves ($p=0.49$).

Considering the profession specific measure we find that a significant lower share of male soldiers are positive to a position in the police, military etc. involving risky missions. We find, however, no significant change in the preferences of female soldiers. Despite a gender difference in means equal to .12 (about 1/4 sd. see Table A.3 in the Appendix), testing for changes in the estimated gender coefficients from wave 1 to wave 3, we fail to reject the hypothesis that the two coefficients are equal ($p=0.21$).

In contrast to wave 1, we find no significant gender difference in the preferences for avoiding physical damage in wave 3. The change in the estimated gender difference is significant between the two waves (Wald test, $p=0.02$), but only at the six percent level when including background variables in the regressions (Wald test, $p=0.055$). The change is due to converging preferences for male and female soldiers. Male soldiers dislike physical risk more, while female soldiers adjust their preferences in the opposite direction, see Table 3. Thus, we find weak evidence for the claim

¹⁰See Table B.5 in the Appendix for wave 3 results with and without control variables.

¹¹See also Table A.3 and B.5 in the Appendix.

¹²Descriptive statistics for risk preferences at wave 3 are found in Table A.3 in the Appendix.

Table 1: Regression: Baseline (wave 1) gender differences

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Domain	Domain	Profession	Profession	Physical	Physical	General	General
Female	0.270** (.098)	0.314** (.003)	0.104*** (.032)	0.107** (.033)	0.488*** (.141)	0.362* (.151)	0.312* (.128)	0.238 (.139)
Risky_cons_w1	0.844*** (.047)	0.868*** (.047)						
Observations	1689	1664	1689	1665	842	831	844	833
Battalion FE	no	yes	no	yes	no	yes	no	yes
Individual controls	no	yes	no	yes	no	yes	no	yes
Mean preferences	5.34	5.34	0.67	0.67	2.77	2.76	3.33	3.33
SD preferences	1.59	1.59	0.47	0.47	1.49	1.49	1.35	1.36

Note: The dependent variable is risk preference for each of the risk measures prior to service. *Female* denotes female soldiers and represents any gender differences in the dependent variable. *Risky_cons_w1*: the respondent's judgement of the outcome of the risky event. Regressions (2), (4), (6) and (8) include individual control variables: age, whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans and test scores from screening. Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS regressions.

Table 2: Regression: Wave 1 and 3 gender differences for panel and random sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dom._w1	Dom._w3	Prof._w1	Prof._w3	Phys._w1	Phys._w3	Gen._w1	Gen._w3
Female	0.263 (.262)	0.389* (.164)	0.0213 (.047)	0.0862 (.049)	0.362* (.151)	0.0291 (.165)	0.238 (.138)	0.0660 (.136)
Risky_cons_w1	0.893*** (.068)							
Risky_cons_w3		0.771*** (.069)						
Observations	758	758	767	767	831	768	833	770
Battalion FE	yes	yes	yes	yes	yes	yes	yes	yes
Individual controls	yes	yes	yes	yes	yes	yes	yes	yes
Mean preferences	5.33	5.53	0.65	0.61	2.77	2.99	3.33	3.36
SD preferences	1.56	1.69	0.48	0.49	1.48	1.60	1.35	1.34

Note: The dependent variable is risk preference for each of the risk measures at baseline (wave 1) and towards the end of service (wave 3), respectively. *Female* denotes female soldiers and represents any gender differences in the dependent variable. *Risky_cons_w1/_w3*: the respondent's judgement of the outcome of the risky event in wave 1 / wave 3. All regressions include individual control variables: age, whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans and test scores from screening. Standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS regressions.

that adjustments in risk preferences related to physical injury eliminate the gender difference.

For the general risk measure we found no significant gender difference in wave 1 after including the controls, cf. Table 1. The results in Table 3 suggest that very little happens to the soldiers general preferences during nine months of service. Including the control variables does not alter the results. We find no evidence of gender differences in wave 3, and we fail to reject the hypothesis that estimated gender differences are equal in the two periods (Wald test, $p=0.17$).

Table 3: Test statistics: Changes in risk preferences by gender

w3 - w1	Domain	Profession	Physical	General
Total	0.20**	-0.05**	0.22**	0.03
Female	0.36*	-0.01	-0.13	0.20
Male	0.16*	-0.05**	0.29***	0.07
Obs.	788	797	1714	1718
Female	124	127	265	266
Male	645	651	1397	1400

*Note: Difference in means ($w3 - w1$) for male and female soldiers. General and Domain, Profession and Physical specific measures. t -test mean $w3 = \text{mean } w1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Unpaired test for the Physical and General measure (random samples).*

5.3 Peer effects

The results for peer effects are presented in Table 4. The table shows regression results of Equation (1) using the domain specific risk measure (column 1–4) and the profession specific measure (column 5–8) as the dependent variables. Results are reported for wave 2 and 3, with and without control variables. We find evidence of peer effects on domain specific risk preferences after nine months of service, but no effect after the first eight weeks during the recruit period. There are, however, no significant peer effects on the profession specific risk preferences at any of the two points in time. The findings indicate that influence from peer preferences during the eight weeks of boot camp has a long term effect and forms the preferences towards the

end of service. A possible interpretation of the results is that the relevant peer groups are formed during the recruit period, and that the personal contact within rooms are less intense after the recruit period. Due to the fact that we do not have information on room assignments after the recruit period, we cannot rule out the possibility that room assignments persisted for an extended period and that the initial peer influence is underestimated. Furthermore, domain specific risk preferences do not seem to change already in wave 2 (See the Appendix Table A.4). Thus we did not expect to find any peer effects at this stage. The sign of the estimated peer effect is negative. This means that there is a negative relationship between the respondents risk preferences after nine months and mean roommates' risk preferences at baseline. On average, the more the roommates like risk when they enter the Armed Forces, the more the respondents dislike risk after nine months of service. This finding is in contrast to what we expected and to the findings in Ahern et al. (2014). One possible mechanism is that soldiers have some negative experiences with roommates taking relatively high risk during the recruit period, and they respond by being more cautious later on. In other words, exposure to risky behavior from peers, makes the respondent more risk averse. Several theories suggest that individuals who experience greater risks in their environment may have less tolerance for risk in other areas (Gollier and Pratt, 1996; Li, 2011; Quiggin, 2003; Guiso and Paiella, 2008; Lee, 2008). Moreover, research in both economics and psychology points to a potential relationship between trauma and risk preferences. Callen et al. (2014) study the effect of exposure to violence in Afghanistan on preferences for certainty. They find that individuals exposed to violence exhibit an increased preference for certainty. There is also empirical evidence from lab experiments which show that subjects exposed to a higher risk environment exhibit higher levels of risk aversion than those who face a moderate or low risk environment (He and Hong, 2015).

Table 4: Regression: Peer effects

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dom. w2	Dom. w2	Dom. w3	Dom. w3	Prof. w2	Prof. w2	Prof. w3	Prof. w3
Peer domain w1	-0.135 (.087)	-0.128 (.090)	-0.234 (.128)	-0.297* (.127)				
Peer prof. w1					-0.0395 (.111)	0.0285 (.137)	-0.0315 (.152)	-0.0575 (.149)
Female	0.116 (.204)	0.0857 (.253)	0.846*** (.234)	0.737** (.244)	0.111* (.052)	0.0362 (.060)	0.115 (.091)	0.163 (.110)
Risky cons. w2	0.826*** (.072)	0.814*** (.074)						
Risky cons. w3			0.702*** (.113)	0.772*** (.112)				
Peer Risky cons. w1	0.260 (.164)	0.254 (.163)	0.557* (.246)	0.628* (.243)				
Observations	597	592	329	326	596	358	246	244
Company FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Individual controls	No	Yes	No	Yes	No	Yes	No	Yes

Note: The dependent variables are domain and profession specific risk preference for each of the two risk measures at the end of the recruit period (w2) and after nine months of service (w3). *Peer domain w1* denotes the baseline mean domain specific risk preference for the peers (room mates) excluding respondent *i*. *Peer prof. w1* denotes the baseline mean profession specific risk preference for the peers (room mates) excluding respondent *i*. *Female* denotes female soldiers and represents any gender differences in the dependent variable. *Risky cons. w1*: the respondent's judgement of the outcome of the risky event. Regressions (2), (4), (6) and (8) include individual control variables: age, whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans, test scores from screening and dummy variables for gender mixed rooms. Standard errors in parentheses clustered on rooms, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS regressions.

6 Concluding remarks

The literature suggests that the lack of gender differences in risk preferences found in some professions could stem from self selection, adaption at the workplace or a combination of the two mechanisms. There is, however, a limitation to the existing literature when it comes to study design for identifying the mechanism at work. Further, studies addressing changes in risk preferences for workers within a given profession, tend to rely on observations of students rather than workers facing the real work environment when making decisions based on their risk preferences.

We speak to both of the shortcomings by establishing a baseline for risk preferences before the workers enter their real workplace environment, and study any changes in preferences during a time period of nine months. We find that soldiers on average dislike domain and domain related risk more during the service. But there is no change in the measure of general risk preferences. Despite the findings of gender differences at baseline for all of the four different risk measures, we find no difference if we only consider the soldiers completing our survey after nine months of service. Given the high drop out rates during the first weeks of service, our results could suggest that self selection eliminates gender differences. On the other hand, the findings of a significant gender difference for domain specific risk preferences after nine months shows that other mechanisms could be at work. We fail, however, to reject the hypothesis that the estimated coefficients for gender differences are equal in size during the period studied.

We suggest peer influence as a source for adjustments in risk preferences during service. By randomizing room assignments we are able to investigate any causal effects from peers risk preferences. In contrast to what we expected we find a negative peer effect for domain specific risk preferences, i.e. on average soldiers diverge from the mean risk preference of their peers in which they share room with. One possible explanation of the result is that exposure to risky behavior from peers in the domain specific setting, makes the respondent more risk averse.

The existence of a gender gap in preferences at baseline and towards the end of service could represent two different concerns for policy makers. First policymakers have to make sure that a significant share of the female population in the relevant age cohorts are recruited to duty and complete service. This concern is to a large extent handled by gender neutral conscription. Second, in order to actually increase the share of females in the Armed Forces the pool of female soldiers willing to accept

extended service and a permanent position has to be maximized towards the end of service while job applications are handed in. In this respect our results suggest that policy makers lower their odds if they rely on the assumption that risk preferences of female soldiers adapt in a favorable direction during service. Policies could rather raise awareness of possible gender differences in preferences during service and stress the impact of deliberated communication when it comes to undesirable behavior and events related to risk.

Carrell et al. (2013) show that policies based on the evidence of peer effects can diverge from the intended effects even in a military setting. Based on our findings we suggest extra caution in designing policies involving mechanisms based on peer influence in risky environments, and we urge further studies to investigate the magnitude of a possible negative peer effect in contexts also outside the military.

We show that gender differences in risk preferences occur even when women self select into a risky professions. We argue that a gender gap in risk preferences could hamper the recruitment and retention of female workers in general and female soldiers in particular. The gap could even set a natural limit on the share of female workers in a given profession. We have, however, little evidence on the impact of the gender gap on the actual number of females recruited and the pool of candidates in the female population. Further studies of risk preferences including a control group outside the military could provide important insight on this topic.

References

- Ahern, K. R., R. Duchin, and T. Shumway (2014). Peer effects in risk aversion and trust. *Review of Financial Studies*.
- Angrist, J. D. (1990). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American economic review* 80(3), 313–336.
- Arch, E. C. (1993). Risk-taking: a motivational basis for sex differences. *Psychological Reports* 73(1), 3–11.
- Balafoutas, L. and M. Sutter (2012). Affirmative action policies promote women and do not harm efficiency in the laboratory. *Science* 335(6068), 579–582.

- Barsky, R. B., F. T. Juster, M. S. Kimball, and M. D. Shapiro (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *The Quarterly Journal of Economics* 112(2), 537–579.
- Bauer, T. K., S. Bender, A. R. Paloyo, and C. M. Schmidt (2012). Evaluating the labor-market effects of compulsory military service. *European Economic Review* 56(4), 814–829.
- Callen, M., M. Isaqzadeh, J. D. Long, and C. Sprenger (2014). Violence and risk preference: Experimental evidence from afghanistan. *American Economic Review* 104(1), 123–48.
- Card, D. and L. Giuliano (2013). Peer effects and multiple equilibria in the risky behavior of friends. *Review of Economics and Statistics* 95(4), 1130–1149.
- Carrell, S. E., B. I. Sacerdote, and J. E. West (2013). From natural variation to optimal policy? the importance of endogenous peer group formation. *Econometrica* 81(3), 855–882.
- Charness, G., U. Gneezy, and A. Imas (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization* 87, 43–51.
- Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic literature*, 448–474.
- Dahl, G., A. Kotsadam, and D.-O. Rooth (2018). Does integration change gender attitudes? the effect of randomly assigning women to traditionally male teams. Technical report, National Bureau of Economic Research.
- Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2010). Are risk aversion and impatience related to cognitive ability? *The American economic review* 100(3), 1238–1260.
- Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association* 9(3), 522–550.
- Dohmen, T. J., A. Falk, D. Huffman, U. Sunde, and J. Schupp (2009). Individual risk attitudes: Measurement, determinants and behavioral consequences.

- Dwyer, P. D., J. H. Gilkeson, and J. A. List (2002). Gender differences in revealed risk taking: evidence from mutual fund investors. *Economics Letters* 76(2), 151–158.
- Eckel, C. C., P. J. Grossman, C. A. Johnson, A. C. de Oliveira, C. Rojas, and R. K. Wilson (2012). School environment and risk preferences: Experimental evidence. *Journal of Risk and Uncertainty* 45(3), 265–292.
- Finseraas, H., T. Hanson, Å. A. Johnsen, A. Kotsadam, and G. Torsvik (2019). Trust, ethnic diversity, and personal contact: A field experiment. *Journal of Public Economics* 173, 72–84.
- Finseraas, H., Å. A. Johnsen, A. Kotsadam, and G. Torsvik (2016). Exposure to female colleagues breaks the glass ceilingevidence from a combined vignette and field experiment. *European Economic Review* 90, 363–374.
- Finseraas, H. and A. Kotsadam (2017). Does personal contact with ethnic minorities affect anti-immigrant sentiments? evidence from a field experiment. *European Journal of Political Research* 56(3), 703–722.
- Frey, R., A. Pedroni, R. Mata, J. Rieskamp, and R. Hertwig (2017). Risk preference shares the psychometric structure of major psychological traits. *Science advances* 3(10), e1701381.
- Gollier, C. and J. W. Pratt (1996). Risk vulnerability and the tempering effect of background risk. *Econometrica: Journal of the Econometric Society*, 1109–1123.
- Gruber, J. (2000). Risky behavior among youths: An economic analysis. Technical report, National Bureau of Economic Research.
- Guiso, L. and M. Paiella (2008). Risk aversion, wealth, and background risk. *Journal of the European Economic association* 6(6), 1109–1150.
- He, T.-S. and F. Hong (2015). Exposure to risk and risk aversion: A laboratory experiment. *Available at SSRN 2401438*.
- Lee, J. (2008). The effect of the background risk in a simple chance improving decision model. *Journal of Risk and Uncertainty* 36(1), 19–41.
- Li, J. (2011). The demand for a risky asset in the presence of a background risk. *Journal of Economic Theory* 146(1), 372–391.

- Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies* 60(3), 531–542.
- Manski, C. F. (2000). Economic analysis of social interactions. Technical report, National bureau of economic research.
- Nguyen, Q. (2011). Does nurture matter: Theory and experimental investigation on the effect of working environment on risk and time preferences. *Journal of Risk and Uncertainty* 43(3), 245–270.
- Niederle, M., C. Segal, and L. Vesterlund (2013). How costly is diversity? affirmative action in light of gender differences in competitiveness. *Management Science* 59(1), 1–16.
- Quiggin, J. (2003). Background risk in generalized expected utility theory. *Economic Theory* 22(3), 607–611.
- Stinebrickner, R. and T. R. Stinebrickner (2006). What can be learned about peer effects using college roommates? evidence from new survey data and students from disadvantaged backgrounds. *Journal of public Economics* 90(8), 1435–1454.

A Descriptive statistics

Table A.1: Descriptive statistics: Demographics

	(1)	(2)	(3)	(4)
	All obs.	Panel	Rnd. phys.	Rnd. gen.
Female (%)	15.6	16.1	15.4	15.5
Age	19.57	19.5	19.8	19.6
	(.89)	(0.84)	(.88)	(.87)
Male	19.63	19.6	19.6	19.6
Female	19.25	19.3	19.2	19.2
Higher edu. plans (%)	68.9	74.6	69.1	69.3
Male	65.8	71.8	66.1	66.3
Female	86.2	88.7	85.9	86.0
Mother higher edu. (%)	53.6	54.8	52.9	53.0
Male	54.0	55.3	53.3	53.4
Female	51.3	52.4	50.9	50.8
Father higher edu. (%)	62.0	62.7	62.8	62.9
Male	61.3	61.4	61.4	61.5
Female	65.6	69.4	70.9	71.1
Parents divorced (%)	30.6	29.3	30.4	30.5
Male	30.7	29.3	30.5	30.7
Female	30.5	29.3	29.9	29.7
Minority background (%)	3.3	3.1	4.3	4.2
Male	3.4	3.6	4.2	4.1
Female	2.6	0.8	4.6	4.6
Obs.	1786	769	842	844

Note: Column (1) includes all observations in wave 1. Column (2) includes the panel of respondents in both wave 1 and 3. Column (3) and (4) include the baseline random sample from the physical and general risk measure, respectively. Standard deviation for the age variable in parenthesis.

Table A.2: Descriptive and test statistics: Baseline risk preferences (wave 1)

	Dom. (all)	Dom. (pan)	Prof. (all)	Prof. (pan)	Physical	General
Total	5.34 (1.59)	5.35 (1.56)	0.67 (0.47)	0.66 (0.48)	2.77 (1.49)	3.33 (1.35)
Female	5.43 (1.52)	5.40 (1.55)	0.75 (0.43)	0.72 (0.45)	3.18 (1.43)	3.60 (1.46)
Male	5.32 (1.59)	5.32 (1.57)	0.65 (0.48)	0.64 (0.48)	2.70 (1.49)	3.28 (1.33)
Gender diff.	0.11	0.08	0.20***	0.08	0.48***	0.32*
Obs.	1691	769	1689	778	842	844
Female	262	124	263	127	130	131
Male	1429	645	1426	651	712	713

Note: Mean values for all observation (all) in wave 1, and panel (pan) including respondents both in wave 1 and 3. *Standard deviation* in parentheses. t-test mean female = mean male, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.3: Descriptive and test statistics: Wave 3 risk preferences

	Dom. (all)	Dom. (pan)	Prof. (all)	Prof. (pan)	Physical	General
Total	5.56 (1.68)	5.53 (1.69)	0.61 (0.49)	0.61 (0.49)	3.00 (1.60)	3.36 (1.34)
Female	5.79 (1.49)	5.77 (1.51)	0.71 (0.45)	0.71 (0.46)	3.06 (1.50)	3.40 (1.25)
Male	5.52 (1.72)	5.49 (1.72)	0.59 (0.49)	0.59 (0.49)	2.98 (1.62)	3.35 (1.36)
Gender diff.	0.27*	0.28*	0.12***	0.12***	0.08	0.05
Obs.	812	769	821	778	820	822
Female	133	124	135	127	135	135
Male	679	645	686	651	685	687

Note: Mean values wave 3 for all observation (all) and panel (pan) including respondents both in wave 1 and 3. *Standard deviation* in parentheses. t-test mean female = mean male, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A.4: Test statistics: Changes in risk preferences w1-w2 by gender

w2 - w1	Domain	Profession	Physical	General
Total	-0.01	-0.05***	0.15*	0.10
Female	-0.07	-0.05*	-0.36*	0.01
Male	-0.01	-0.05***	0.25***	0.12*
Obs.	1417	1414	1631	1632
Female	216	217	248	249
Male	1168	1164	1330	1330

Note: Difference in means (w2 - w1) for male and female soldiers.

General and Domain, Profession and Physical specific measures.

t-test mean w2 = mean w1, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Unpaired test for the Physical and General measure (random samples).

B Other regression tables discussed in the text

Table B.5: Regression: Wave 3 gender differences, panel and random sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Domain	Domain	Profession	Profession	Physical	Physical	General	General
Female	0.395** (.142)	0.389* (.155)	0.119** (.044)	0.0862 (.048)	0.0753 (.142)	0.0291 (.159)	0.0477 (.047)	0.0660 (.133)
Risky_cons_w3	0.727*** (.727)	0.771*** (.770)						
Observations	769	758	778	767	820	768	822	770
Battalion FE	no	yes	no	yes	no	yes	no	yes
Individual controls	no	yes	no	yes	no	yes	no	yes
Mean preferences								
SD preferences								

Note: The dependent variable is risk preference for each of the risk measures after nine months of service. *Female* denotes female soldiers and represents any gender differences in the dependent variable. *Risky_cons_w3*: the respondent's judgement of the outcome of the risky event. Regressions (2), (4), (6) and (8) include individual control variables: age, whether mother/father work, mother's/father's education, whether parents are divorced, the soldier's educational plans and test scores from screening. Robust standard errors in parentheses, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS regressions.