Does Integration Change Gender Attitudes? The Effect of Randomly Assigning Women to Traditionally Male Teams*

Gordon B. Dahl† Andreas Kotsadam‡ Dan-Olof Rooth§

We examine whether exposure of men to women in a traditionally male-dominated environment can change attitudes about mixed-gender productivity, gender roles and gender identity. Our context is the military in Norway, where we randomly assigned female recruits to some squads but not others during boot camp. We find that living and working with women for 8 weeks causes men to have more egalitarian attitudes. There is a 14 percentage point higher fraction of men who think mixed-gender teams perform as well or better than same-gender teams, an 8 percentage point increase in men who think household work should be shared equally and a 14 percentage point increase in men who do not completely disavow feminine traits. Moreover, men exposed to mixed-gender teams are more likely to choose military occupations immediately after boot camp which have a higher fraction of females in them. But these effects do not persist once treatment stops. Treated men's attitudes converge to those of the controls in a 6-month follow up survey and there is no long-term effect on choosing fields of study, occupations or workplaces with a higher fraction of women in them after military service ends. Contrary to the predictions of many policymakers, we do not find that integrating women into squads hurt male recruits' performance or satisfaction with service, either during boot camp or their subsequent military assignment. These findings provide evidence that even in a highly gender-skewed environment, gender stereotypes are malleable and can be altered by integrating members of the opposite sex. But they also suggest that without continuing intensive exposure, effects are unlikely to persist.

JEL Codes: J16, J24

---

*We thank seminar participants at several universities and conferences for useful comments and suggestions, and especially Henning Finseraas. Special thanks to the Norwegian Defense Research Establishment (FFI), and in particular to Frank Steder and Torbjørn Hanson. This study could not have been conducted without their help and the support of FFI's Age Cohort Research project team. Thanks also to the soldiers and staff of the North Brigade who participated in the study. Ada Fuglset, Eirik Strømland, and Wiktoria Szczesna provided excellent research assistance. This research has IRB approval from The Norwegian Center of Research Data (Number 39028) and the ethics committee at the National Service Administration, and is funded by Norwegian Research Council Project number 287766. Data made available by Statistics Norway and the National Service Administration have been essential for this research.

†Department of Economics, UC San Diego, e-mail; gdahl@ucsd.edu

‡Ragnar Frisch Centre for Economic Research, Oslo; e-mail: andreas.kotsadam@frisch.uio.no

§Swedish Institute for Social Research, Stockholm University; e-mail: dan-olof.rooth@sofi.su.se

I Introduction

Despite women making up almost half of the labor force in most developed countries, occupational segregation remains high.[1] Integration may partly have stalled due to preferences of men and women for different types of work environments, including women's desire for family friendly jobs with more flexibility (Goldin, 2014b; Goldin and Katz, 2016). On top of this, however, it could also be the case that further integration is hindered by exclusionary gender stereotypes and norms, especially in traditionally masculine work environments. Historically, women were denied access to universities, fired from their jobs if they married or had children and barred from certain occupations within firms (Goldin, 2014a). Removing these explicit restrictions resulted in an inflow of women into a variety of fields in the decades after World War II. While few formal constraints remain today, implicit restrictions stemming from gendered attitudes could remain a barrier for female employment and advancement in certain jobs (Bertrand et al., 2016; Moss-Racusin et al., 2012; Stamarski and Son Hing, 2015), especially when stereotypes cause belief distortions (Bordalo et al. 2016).

In particular, men's attitudes about mixed-gender productivity, gender roles and gender identity could play a critical role in explaining the persistence of occupational segregation. Employers may not hire women in male-dominated fields because they believe doing so will lower worker morale, group cohesiveness or productivity (Harrell and Miller, 1997). Stereotypical attitudes about gender roles related to home production versus paid work could reduce women's willingness to invest in certain careers, and also influence which tasks women are assigned within the workplace (Alesina, Giuliano, and Nunn, 2013; Goldin and Katz, 2002; Goldin, 2004). Likewise, gender identity concerns such as those discussed in Akerlof and Kranton (2000), Bertrand, Kamenica, and Pan (2015) and Goldin (2014b) could contribute to occupational segregation. In this paper, we examine whether integration of women into a traditionally male-dominated environment can change these types of gender attitudes and alter men's willingness to pursue occupations with more women in them.

Understanding the link between gender attitudes and occupational segregation is

[1] In the second half of the 20th century, occupational sex segregation steadily declined in both Europe and the U.S., but stalled starting in the 1990s (Bettio and Verashchagina, 2009; Blau, Brummund, and Liu, 2013; Olivetti and Petrongolo, 2016). Around the same time, female labor force participation reached its peak and the convergence in wage earnings slowed (Blau and Kahn, 2003, 2017). Consistent with these trends, studies find that occupational sorting by gender accounts for a sizable portion (between 22 and 42 percent) of the gender wage gap (Goldin, 2014a). For additional evidence on and explanations for the slowing of women's progress, see Bailey and DiPrete (2016).

important for designing policies which aim to better integrate the workplace.[2] Yet estimating the causal link between the two is difficult due to reverse causality, self-selection and unobserved heterogeneity. Working in a male-dominated occupation could cause men to develop less egalitarian attitudes, but it is also possible that men with less egalitarian attitudes select into these types of occupations in the first place. Dynamically, it is difficult to know if some occupations become more integrated because attitudes are changing, or if increased integration is responsible for a shift in attitudes.

To overcome these empirical challenges, we set up a field experiment which randomly assigned women to teams which were traditionally all male. Our context is the military in Norway, where women make up less than 15 percent of the professional military force (a fraction which is similar to the U.S.). In cooperation with the Norwegian Defense Research Establishment, we randomized female recruits to some squads but not others during the eight weeks of boot camp. The experiment provided intensive exposure of treated males to females during these 8 weeks, as squads are typically comprised of 6 members who train as a team during the day and also live together in the same room. We conducted a baseline survey prior to the start of boot camp, conducted a survey near the completion of boot camp and have access to a survey 6 months into military service, each of which measured attitudes towards a variety of gender related questions.

Empirically, we find that living and working with women for eight weeks causes men to have more egalitarian attitudes by the end of boot camp relative to the controls. Our first result is that men who have women randomly assigned to their squads are 14 percentage points more likely to think mixed-gender teams perform as well or better than same-gender teams. This is a 26 percent increase relative to the control mean of 52 percent. Second, exposed men are 8 percentage points more likely to believe that men and women should share household work equally, relative to the control mean of 66 percent. Third, we find that men exposed to women are 14 percentage points more likely to not completely disavow their feminine side, relative to the control mean of 63 percent. These are sizable swings which move men closer to the attitudes of women: depending on the outcome, treatment reduces the gap in mean attitudes between men and women by between 31 and 46 percent.

Our results are robust to a variety of alternative specifications, including models based on first differences versus one-period lags, models using the full scale of responses

[2]For example, see Bayer and Rouse (2016), which argues that underrepresentation of women in the economics profession could be due to implicit attitudes and institutional practices, and proposes several programs to address these problems.

and varying squad sizes. We verify that observable male covariates are uncorrelated with the assignment of females to a team, and find no evidence that prior attitudes are affected by future exposure in a placebo test. These exercises provide empirical support for randomization.

The changed attitudes are accompanied by a change in occupational assignments during military service. Men exposed to mixed-gender teams during boot camp on average serve in jobs with more women in them: there is a 22% increase in the fraction of women in one's occupation for the treated group relative to the control group. Assignment to a military occupation is determined by a combination of soldier preferences, ability and staffing needs. In interviews with military officials we learned they try to fulfill the wishes of the soldiers as far as possible to maintain high motivation. Hence, we interpret this finding as largely reflecting the altered preferences of recruits for different occupations.

Given the changes in gender attitudes and occupation choices observed in the short run, a natural question is whether these effects persist after treatment ends. As background, it is important to note that squad assignments made during boot camp do not continue during subsequent service and that the new squads are comprised of individuals from a variety of occupations. Using a 6-month follow up survey, we find that treated men's gender attitudes converge to those of the controls. For the key outcome of mixed-gender team productivity, treatment prevents a deterioration of attitudes during boot camp, but by 6 months into service the gap between treatment and control disappears. Consistent with finding no long-run gap in gender attitudes, there is no evidence that treatment affects future education or occupation choices. Using linked administrative register data, we find that treated individuals are not more likely to enroll in a field of study, work in an occupation or be employed in an establishment with a higher fraction of women in the years following military service.

Beyond the general lessons learned about how integration can alter gender attitudes and occupational choices, we also explore how treatment affects recruits' performance and satisfaction with service. We find no evidence that integrating women into squads hurt male recruits along these dimensions, either at the end of boot camp or during subsequent military service. Treated men feel as qualified for the military and are as satisfied with their service in both the short and longer run. Looking at real outcomes, treated men are equally likely to be promoted at the end of boot camp and have similar scores on performance evaluations assigned at the end of service. These findings are

particularly relevant for countries considering further integration of women into the military. Many policymakers have worried that allowing women to serve in the military would ruin its esprit de corps, causing lower performance and dissatisfaction within the ranks, but this appears not to be the case.[3]

We also asked questions about which sex makes the best leaders at various levels of command and find no effect of treatment, either in the short or longer run. To put this in context, note that we did not randomize the gender of these higher-level leaders, and that most of them are male. This indicates that exposure to rank and file women does not have a spillover effect on attitudes about higher up female leadership. This suggests that exposure to female leaders at the relevant level is required (Beaman et al., 2009; Bertrand et al., 2014). Indeed, using similar Norwegian military data, Finseraas et al. (2016) conducted a vignette experiment which asked about attitudes towards hypothetical females becoming squad leaders, and found that exposure at the squad level did make a difference.

Our paper is related to Fortin (2005), which finds that gender role attitudes across 25 OECD countries are associated with female employment rates and the gender wage gap, and Fortin (2015) which links changes in gender attitudes to the leveling off of female labor force participation in the U.S. Our study is also related to the literature on contact theory which explores how biases and beliefs of a dominant group are affected by exposure to members of a minority group. This theory predicts that mixing groups will break down stereotypes and encourage understanding, especially if the two groups interact at a personal level and are given equal status and common goals as is the case during boot camp (Allport, 1954; Pettigrew and Tropp, 2006).[4] Field experiments which randomly assign a college roommate of another race or ethnicity find a reduction in prejudice (Boisjoly et al. 2006; Burns, Corno, and La Ferrara 2016; Van Laar et al. 2005), and exposure to black students within a military setting results in less racial bias and a higher likelihood of rooming with a black student in the following year (Carrell, Hoekstra, and West, 2015). Similarly, Finseraas and Kotsadam (2017) and Finseraas et al. (2019) find that contact with immigrants in a military setting changes attitudes

[3]For examples, see "Marine Commander's Firing Stirs Debate on Integration of Women in Corps," New York Times July 12, 2015 and "Gender Integration of Marines Brings Out Unusually Public Discord," New York Times, September 18, 2015.

[4]Pollution theory, on the other hand, suggests that female employees will reduce the prestige and wages of previously male-dominated occupations (Goldin, 2014b). Stigma could increase if the rise in female employment is driven by a gender quota, although laboratory hiring experiments find that stigma is reduced when merit-based criteria are introduced (Evans, 2003; Resendez, 2002).

and increases trust in immigrants.[5]

Our paper contributes to the literature on contact theory by randomly assigning men to work and live together with women for an extended period and analyzing gender attitudes and real-world outcomes, similar to what has been done in race and ethnicity studies.[6] Unlike the race and ethnicity studies, however, our treated individuals have already likely had many repeat and personal interactions with members of the opposite group. This is because boys attend school with girls, grow up with sisters or have other female relatives, and live in the same neighborhoods. What is new about boot camp is the experience of working side-by-side with women in a traditionally male-dominated setting. Our paper is unique in its ability to study real-world outcomes and analyze effects long after the intervention ends, features not present in most contact theory studies (Paluck, Green, and Green, 2019).[7] We show that consequential measures of performance are not harmed by mixing men with women and that occupational choices are altered in the short, but not longer, run. We also find that effects do not persist in the longer run once intensive contact with females ends. This suggests that to maintain changed attitudes and behaviors, intensive exposure also likely needs to be maintained. We emphasize that it would be incorrect to infer that exposure to different groups cannot change attitudes in the longer run and in other settings. Our intervention, while intense during bootcamp, was relatively short compared to the military experience overall. And our setting is a highly male-dominated environment, where attitudes may be the most difficult to change permanently without continuing exposure.

The remainder of the paper proceeds as follows. Section 2 describes our setting and field experiment. Section 3 introduces our survey questions and describes the data. In Section 4, we present OLS estimates and discuss the validity of our experiment. Section 5 presents our main experimental results for gender attitudes at the end of boot camp.

[5]There is a related literature on the peer effects of randomly assigned roommates (e.g., Sacerdote, 2011; Stinebrickner and Stinebrickner, 2006; Zimmerman, 2003). In addition, studies have looked at what happens when whites serve with black jurors in criminal trials (Anwar, Bayer, and Hjalmarsson, 2012), athletes are mixed with different castes and religions on sports teams (Lowe, 2020; Mousa, 2020), students are exposed to classmates with different religions (Scacco and Warren, 2018) and wealthy students are exposed to poorer students (Rao, 2019).

[6]The U.S. Marine Corps has examined the performance of mixed-gender teams, but only in the context of rotating physical assignments which lasted a few hours, such as a two hour hike or rifle shooting (see "Marine Corps Study: All-Male Combat Units Performed Better than Mixed Units," NPR, September 10, 2015).

[7]To the best of our knowledge, the longest run outcome in the literature is found in Mousa (2020), which examines whether individuals exposed to mixed-religion soccer teams are more likely to be in mixed teams 6 months after treatment ends.

Section 6 discusses longer-run results and the final section concludes.

II Setting and Field Experiment

A Occupational Segregation and Attitudes in Norway versus other Countries

As background, we first document the amount of gender segregation and gender attitudes in Norway compared to other countries around the time of our experiment. Norway is often viewed as a progressive country when it comes to gender issues, but like the rest of Europe and the U.S., its workforce is highly sex segregated. In both the U.S. and Norway, 47 percent of the labor force was female as of 2014. Table I provides some examples from 2014 of both male-dominated and female-dominated occupations. In both countries, stereotypically female jobs such as kindergarten teachers, nurses and social workers are primarily held by women, while stereotypically male jobs such as firefighters, pilots and computer programmers are primarily held by males.[8] An especially relevant comparison is that women comprise 13% of the military in Norway, compared to 15% in the U.S. (not including civilian employees).

To provide a more holistic view of the amount of segregation, we calculated Duncan segregation indices using 4 digit occupational codes from Norwegian census data (Duncan and Duncan, 1955). This commonly used measure is calculated as the absolute difference in the fraction of men and the fraction of women in an occupation, summed over all occupations and multiplied by one half. In 2010 in Norway, we calculate the index to be 0.53 (based on 483 categories), which means that 53% of women would have to change occupations to make the occupational distributions of men and women identical. Norway's index is slightly higher compared to other EU countries when using common occupational classifications (Bettio and Verashchagina, 2009). Using U.S. census data, we calculate a Duncan index of 0.51 for the U.S. (based on 491 categories); while the U.S. occupational categories do not map directly to those of the EU countries and so cannot be directly compared, segregation is clearly high in the U.S. as well.

We next provide a comparison of gender role attitudes in Norway versus other countries, based on common questions asked in OECD countries. Using responses from the World Values Survey (WVS), Fortin (2005) documents that 80% of women and 64% of men in Norway believe that "working mothers can have a warm relationship

[8]The fraction female in these gendered-occupations are remarkably similar across the two countries, with a few exceptions. For example, pharmacists are heavily female in Norway but not so in the U.S., while architects are heavily male in the U.S. but not so in Norway.

with their children." The U.S. has somewhat more positive views about maternal labor force participation, with 83% of women and 73% of men agreeing with the statement. In both countries, the gender gap in attitudes is substantial. Another question in the WVS asks if "being a housewife can be fulfilling." Here, Norway diverges from the U.S., with 54% of women and 56% of men agreeing with the statement, compared to 76% of women and 76% of men in the U.S. For both of these attitude measures, Norway is in the upper middle of the distribution (i.e., somewhat more gender equal) compared to other OECD countries participating in the survey.

B Military Conscription and Service in Norway

Norway has a selective mandatory military draft. Since 2010, all 17 year old men and women are required to register and be screened for service by the military. The first step in the screening process is internet based and involves answering questions related to health, school, personality, motivation and desire to serve. Around 60,000 males and females complete this online form each year. In the second step of the screening process, approximately 20,000 18 and 19 year olds are selected for physical and cognitive testing, and an interview with a recruitment officer. Based on this screening, the military then selects a subset of individuals for military service. Around 8,000 to 10,000 individuals are chosen to serve.

Service is mandatory for men if chosen. But the screening process generally prioritizes candidates who express a motivation and willingness to be in the military, so de facto, most men are not coerced into service. Around 1 in 6 males end up being chosen for service. During the period of our study in 2014, service was voluntary for women, even though they had to participate in the screening process. Around 14 percent of the individuals sent to boot camp for training are women. Both men and women who end up serving in the military are therefore self-selected to be highly motivated. This type of selection is similar to what would be observed in a regular workplace, where individuals apply for and are hired for jobs for which they are a good fit. Of course, our setting differs from most workplaces in that individuals will be in a male-dominated military environment.

As is true in many countries, Norway has been trying to increase female representation at all levels in the military. A goal around the time of our survey was to have one out of five positions in the military staffed by women by the year 2020 Strøm-Erichsen 2013. To help achieve this goal, in 2013 the Norwegian Parliament

passed a law extending the mandatory draft to women, making it the first European country to do so.[9] Mandatory conscription of women took effect in 2015 (a year after our study period). The Norwegian Chief of Defense, Admiral Haakon Bruun-Hanssen, heralded the change, saying "The new law means equal rights and duties for men and women... Now we have twice as many people to choose from. This will make it easier to direct motivated personnel and the right expertise to our different tasks and positions" (Norwegian Armed Forces, 2015).

Mandatory military service consists of two phases. The first is boot camp, which is also known as basic training, and lasts eight weeks. Its purpose is to prepare recruits physically, mentally and emotionally for service. It is an intense period of training, comprised of both field exercises and classroom time. Soldiers are grouped into squads which normally have 6 team members. Most tasks during boot camp are completed jointly, with squads learning how to function effectively as a team. In our setting, squad members both train as a unit and share living quarters. Moreover, they are not allowed to leave base during boot camp, so they spend most of their daytime and nighttime hours together.

The second phase occurs after the end of boot camp, when soldiers begin their mandatory service period which lasts approximately 10 months. The service period is less intense compared to boot camp. Individuals are assigned to different military occupations, based on soldier preferences, ability and staffing needs. Moreover, soldiers are assigned to completely new squads, which do not correspond to occupation, after boot camp. Being in the same squad no longer means sleeping in the same room. Moreover squad members interact less with each other, in part because some training is organized by occupation rather than squad and in part because individuals are allowed to spend evenings and weekends off-base.

After the end of mandatory military service, individuals can choose whether to apply for further employment in the military. The military uses an end of service evaluation to help determine a soldier's suitability for future service.

C Field Experiment and Survey Administration

Gender-mixed squads are increasingly being used throughout the Norwegian military to facilitate the integration of female soldiers into the military. In part to assess the effectiveness of such integration, the Norwegian Armed Forces and the Norwegian

[9]See "Norway becomes first NATO country to draft women into military," Reuters, June 14, 2013.

Defense Research Establishment have conducted a series of surveys starting as early as 2008 (see Hanson, Steder, and Kvalvik (2016) for a summary of findings). The surveys and accompanying reports focus on recruitment and motivation to serve, but have also probed attitudes towards gender integration among soldiers. The results of these surveys suggest the integration of women was largely successful, although the analysis is based on non-random assignment of women to squads.[10]

The Army's North Brigade has been at the forefront of integration: since 2010 all eight of its battalions have had mixed-gender sleeping quarters. To evaluate the effects of mixed-gender squads on men's attitudes, we convinced three battalions of the North Brigade, or about half of the contingent, to randomize soldiers into squads during boot camp in 2014. These three battalions are the Second Battalion of Northern Norway (Andre Bataljon Nord-Norge), the Artillery Battalion (Artilleribataljonen), and the Armored Battalion (Panserbataljonen). During boot camp, squads are comprised of 6 team members on average, with between 5 and 10 squads forming a troop and troops with a common function making up a battalion. In these battalions, the room also constitutes the squad during boot camp. Our field experiment did not introduce the use of mixed-gender squads, as male and female recruits had already been integrated into squads in the North Brigade for several years before our study. By the time we implemented our field experiment in 2014, the military had several years to iron out any logistical issues and assignment to a mixed-gender squad was not anything out of the ordinary.

Our baseline survey was conducted at a military base near Oslo, where recruits are given a battery of final qualification tests before service starts. Recruits are divided into groups of 20-30 members, and rotate among testing stations throughout the day. These testing stations assess a recruit's physical fitness, mental ability and psychological profile. We managed one of the stations, having recruits fill out an online survey questionnaire that we developed in consultation with the Norwegian Defense Research Establishment. In addition to our gender attitude questions, the survey asked questions about demographics, personality traits, leadership potential and military service. Recruits were told the survey was for research purposes only, and would not be a part of their official record or used by the military for screening or assignment

[10] Anthropological studies without random assignment by Hellum (2014) and Hellum (2017) find that gender mixed rooms increase feelings of sameness across gender and reduce gender essentialist notions, while Lilleaas and Ellingsen (2014) find that mixed rooms promote mutual understanding, de-sexualization and reduced sexual harassment. For a summary of qualitative studies, see Ellingsen, Lilleaas, and Kimmel (2016).

purposes. The survey included over 100 questions and took on average 18 minutes to complete.[11]

We took several steps to minimize concerns related to experimenter demand effects and social desirability bias. We administered the survey ourselves and told recruits we were academic researchers. Moreover, we emphasized that their individual answers would be anonymous and not shared with the military. Several features of our experimental setting also help to minimize social desirability concerns. First, treated recruits are not primed to be thinking about gender integration, as they were not made aware that women were randomly being assigned to some squads and not others as part of an experiment. Second, the survey was conducted in a large group setting, and not at the squad level (the level at which women are being assigned to teams). Third, integration of women into the military was not a new phenomenon, as it had already been going on for several years by the time of our field experiment. Finally, the survey questionnaire was not focused on gender issues, but instead asked a few gender questions interspersed among a longer list of unrelated questions. Despite all of this, we recognize that social desirability bias could play a role in explaining any results we find.

After taking the qualification tests and our baseline survey, the recruits were subsequently flown to northern Norway for the eight week boot camp. Before their arrival at boot camp, recruits were randomly assigned to squads, most of which have 6 individuals in them. Randomization occurred at the troop level, with officers using a template Excel spreadsheet which was programmed in advance to randomize name lists. While the military wanted at least 2 women per squad if possible, officers had some discretion to override this rule.

The randomizer worked as follows. First, officers entered the number of rooms, which normally hold 6 individuals each (rooms and squads are the same during boot camp). Then females were assigned randomly to rooms, with 2 females per room unless there was an odd number, in which case 3 women were assigned to a room. Finally, each room was randomly assigned male soldiers up to the specified room size. Since the number of individuals in a troop does not necessarily equal a multiple of 6, the troop officer could make manual adjustments to even out room sizes. At this stage, an officer could manually override the rule of at least 2 females per room; for example, he

[11]The survey was conducted in a classroom, with recruits being given an internet link to an online survey. Most soldiers used their own mobile phones to complete the survey, but tablets and paper versions of the survey were also available for use. While participation was voluntary, recruits had to stay in the room while the survey was being conducted and were not allowed to talk to each other.

could split three women among multiple rooms. Note that randomization is preserved with these manual adjustments, as room assignments are completed prior to the arrival of recruits at boot camp when the only information available is gender and name. During boot camp, it is possible that individuals are moved to different rooms, which is something we do not measure. When we asked the military, they said this happened rarely, if at all. A conservative interpretation of our estimates is that they capture intention to treat (ITT) effects, based on a recruit's initial assignment.

Near the completion of basic training, we conducted a second survey wave at each battalion's base in Northern Norway. Soldiers were gathered and completed a slightly modified version of the baseline survey, using phones or tablets as in the baseline survey. We administered this second survey ourselves and again emphasized that all answers would be anonymous. Six months into service, the military included some of our questions in a follow-up survey, which allows us to study persistence in effects. We were not directly involved in the administration of this third wave of questions. We were able to supplement these surveys with additional information from the military and Statistics Norway.

III Data and Survey Questions

A Baseline Descriptive Statistics

In the three battalions which randomly assigned women to squads during boot camp, there are a total of 20 troops. The smallest troop has 5 squads, while the largest has 10 squads. Three-fourths of squads in our data have a standard size of 6 members, although other squad sizes arise. During boot camp, the room also constitutes the squad, so we use the two terms interchangeably during boot camp. The most common reason for a nonstandard room/squad size is that the number of members within a troop is not equally divisible by 6. For our main sample, we focus on the 95% of rooms which have between 5 and 7 members in them.[12] In our main sample, we have a total of 153 rooms. Ninety-six of these have zero women, 11 have 1 women, 38 have 2 women, and 8 have 3 or 4 women in them.

The baseline survey was asked of all recruits in Oslo, including the battalions which did not participate in the randomization of women to squads. We used an anonymous

[12]Our empirical results are robust to including all room sizes, as well as using only the standard room size of 6.

id number for each soldier to merge our survey data with administrative data from the military on muscle strength and cognitive skills. The cognitive skill measure is a general ability index (GAI) based on verbal comprehension and perceptual reasoning tests which have been shown to be correlated with a full-scale IQ test in other contexts. Summary statistics collected at baseline for both the entire sample and our main randomized sample are found in Appendix Table A1.

Our main sample for the three battalions which agreed to randomization includes 781 men and 119 women. The appendix table lists several background characteristics for these recruits. Two of these variables, muscle strength and general ability test scores, are of particular note, as the military uses this information to choose recruits. Men in our sample have above average values for both of these variables. Over half of men report having above average or far above average muscle strength relative to their male peers. Likewise, 51 percent of men score at the 6th stanine or above on the general ability test administered by the military (by construction, 40% of the population will score between the 6th and 9th stanine). Women in our sample have above average muscle strength, but lower general ability scores (only 25 percent score at the 6th stanine or above).

Basic demographic variables are also summarized in the tables, as are the fraction of missing values. For variables collected by the military, there are between 1 and 8 percent with missing values, depending on the particular subsample. Our survey questionnaire has somewhat higher missing values, with between 10 and 16 percent missing for most questions.[13]

B Gender Attitude Questions

The main purpose of our experiment is to assess how integrating women into squads affects male recruits' attitudes towards mixed-gender productivity. We are also interested in the broader issues of how female integration affects perceptions of gender roles and gender identity. We ask three gender related questions, both at baseline and after boot camp is over, to see how men's attitudes change. Two of these questions were also asked by the military in a follow-up survey 6 months into service. We refer to the baseline survey as wave 1, the end of boot camp survey as wave 2, and the later follow-up survey

[13]For the questions about sisters and brothers, 25 to 30 percent have missing values. This appears to be because some respondents without a sister (or without a brother) skipped the question if they had 0 sisters (or 0 brothers), even though "none" was explicitly an option. We include a dummy variable for missing values in our regressions so as to be able to include as many observations as possible.

as wave 3.

The distribution of attitudes at baseline for our gender attitude questions are found in Figure I. For our analysis, we will dichotomize the possible answers to create indicators to use as our main outcome variables. In each case, a value of 1 corresponds to a more gender equal viewpoint.

Our first question relates to whether mixed-gender teams underperform all-male teams in our military setting. The question asks respondents to give their view on the statement: "Teams perform better when made up of the same sex." Respondents were given a scale, beginning with 1 for "Completely disagree" up to 7 for "Completely agree." The distribution of answers for males at baseline for our main, randomized estimation sample is found in the top panel of Figure I. We create a dummy variable equaling 1 if the respondent disagrees with this statement, which we define as an answer of 1, 2 or 3 on the scale. Using this categorization, 63% of men disagree with the statement at baseline.

Our second question relates to gender roles. A commonly used method to assess attitudes regarding traditional gender roles is to ask about the division of household chores. We ask individuals their opinion on the statement: "It is important for men and women to share household work equally." Respondents could answer "Strongly agree," "Agree," "Neither agree nor disagree," "Disagree" or "Strongly disagree." The distribution of answers is plotted in the second panel of Figure I. We define a dummy variable for agreement which equals 1 if the individual either agrees or strongly agrees with the statement. Based on this dummy variable, two thirds of men agree with the statement.

Our final gender question concerns a somewhat different concept, namely an individual's self-perception of femininity. Instead of studying stereotypical attitudes about others, this question asks people about themselves. A series of questions regarding personality traits was asked in the survey, with respondents being asked how well certain statements described them. In this list was the statement "I am feminine." Respondents could answer "Does not fit at all," "Does not fit well," "Reasonable fit," "Fits well," or "Fits completely." We view this question as being related to gender identity, and the self-perception of being different from women, similar in spirit to Akerlof and Kranton (2000) and Bertrand, Kamenica, and Pan (2015). A similar idea, but related to racial identity, is discussed in Austen-Smith and Fryer (2005).

Figure I reveals that virtually no men think this statement is an apt description of

them. The relevant distinction is whether the respondent thinks the statement does not fit them at all versus whether it is merely a poor fit. We define a dummy variable which equals 0 for an answer of "Does not fit at all" and a 1 for all other answers. When the dummy variable is 0, we interpret this as a complete disavowal of femininity; roughly 58% of male respondents do not completely disavow their femininity by this measure in the baseline survey. It is important to keep in mind that a value of 1 does not mean the respondent feels they are feminine, but rather corresponds to an answer of "does not fit well" (versus "does not fit at all") for most individuals.

It is informative to compare men's answers to those of women. Appendix Figure A1 plots the distribution of female's attitudes to the same three questions at baseline. In contrast to men, only 10% of women think same gender teams outperform mixed-gender teams. Women are also more likely to believe household work should be shared equally among the sexes (88%). And finally, no women completely disavow their femininity, with most women acknowledging it but not subscribing to the description completely.

It is also informative to compare attitudes in our sample of military recruits to those in the general population. This is only possible for the question regarding household work, since that is the only issue for which we could find a comparable question. Surveys conducted by Kotsadam and Jakobsson (2011; 2014) in conjunction with Gallup asked a random sample of Norwegians "Is it important that women and men share responsibility for the household?" Respondents could answer on a scale, with 0 labeled as "No, not at all" and 10 as "Yes, of course." In Appendix Figure A2, we collapse responses from the Gallup surveys into 5 categories to enable an easier comparison to the household work question in our survey. While the questions are somewhat different, it appears men in our military sample have less gender egalitarian attitudes compared to the general population. In contrast, there is little difference in the distribution of attitudes for women in the military versus the general population. We conclude that males with less gender-equal attitudes select into military service, while female recruits are not a selected sample on this dimension.

C Additional Data

The surveys conducted by us and the military also asked about self-assessed military preparedness and satisfaction with service, as well as questions on whether men or women make better leaders at various levels in the military. We discuss these questions when we analyze them later in the paper.

In addition to the survey data, we have a set of outcomes collected by the military, including promotion outcomes (after the end of boot camp), occupations (assigned after boot camp) and service evaluations (conducted near the end of service). We likewise add administrative data from the Norwegian registers on education, occupation and workplace characteristics for the years after military service. For confidentiality reasons, the military worked directly with Statistics Norway to create a merged dataset for us.[14] The proposed analyses of these additional data, including coding choices, were described in a pre-analysis plan registered at the AEA RCT Registry (AEARCTR-0005987) before the data was received.[15] Summary statistics for the pre-existing variables we will use as controls for this merged dataset can be found in Appendix Table A2.

IV OLS and Experimental Validity

A OLS Estimates at Baseline

We begin by presenting OLS regressions of gender attitudes on background characteristics in Table II. These regressions are based on the entire sample of baseline survey respondents, regardless of whether they are in a battalion which participated in the field experiment. We have two measures of prior exposure to females: whether the male recruit has a large share of female friends, and whether he has a sister. We also include controls for muscle strength, general ability test scores and parental characteristics.

Column 1 regresses a dummy variable for whether same gender teams perform better, with a value of 1 for disagreement. Recruits who respond that more than 40% of their friends are females are 8 percentage points more likely to think mixed-gender teams perform better. In contrast, men with high muscle strength are 11 percentage points less likely to think mixed-gender teams have better performance. The other variables are not statistically significant. In column 2, the dependent variable is whether respondents agree it is important for men and women to share household work equally. In this regression, having a sister increases the probability of agreement by 6 percentage points. Men with above average general ability test scores are less likely to believe

[14] We received permission to merge whether an individual was in a treated squad during boot camp, as well as outcomes collected directly by the military, to the register data. However, in accordance with Norwegian privacy laws, we were not allowed to merge survey question responses to the register data.

[15] These data were added during the revision process; we did not create a pre-analysis plan for the other datasets we collected several years earlier. We did not know that it was possible to merge in military occupations when we wrote up our pre-analysis plan, but added this variable when we became aware of it.

household work should be shared equally, although it is not obvious why this pattern exists. Finally, in column 3, the outcome is whether the respondent does not completely disavow their femininity. Men with high muscle strength are more likely to reject their feminine side completely, while men with above average general ability test scores are less likely. Having parents who are divorced also significantly decreases the recruit's disavowal of a feminine side.

It is tempting to conclude that having female friends causes individuals to believe same gender teams perform better, but it is also possible that those who think mixed-gender environments are better are those who choose to have female friends. Having a sister, which predicts less stereotypical gender roles for household work, is arguably more exogenous, although endogeneous fertility could create a bias.[16]

B Internal Validity

As a test of random assignment, we explore whether the pre-determined covariates appearing in Table II can predict treatment. In Table III, we regress the treatment dummy, i.e., having a female in the squad, on these covariates. While those covariates have predictive power for gender attitudes, they are not statistically significant predictors of treatment status. The coefficient estimates are close to zero and not statistically significant. Moreover, the joint test for all the variables is insignificant, with a p-value of 0.73.[17] This test for covariate balance across treatment and controls corroborates random assignment.

A related test is to add pre-determined covariates into our regressions. If assignment is random, then the estimated treatment effects should not change appreciably. As we shall see in the next section, adding covariates does not appreciably change the estimates.

To further test for random assignment, we conduct a set of placebo tests. The idea of the tests are that prior attitudes should not be affected by future exposure to females if individuals are randomly assigned. Indeed, this is what we find in Appendix Table A3. Initial attitudes measured before boot camp in wave 1 are not affected by future

[16]For example, fertility stopping rules as a function of child gender could influence the presence and gender composition of siblings, and hence reflect underlying gender attitudes of parents (Dahl and Moretti, 2008).

[17]Although not reported, we also tried adding baseline attitude variables into this type of regression. The joint test for all of the variables, including both the variables listed in Table III and the baseline attitude variables, has a p-value of 0.85. Moreover, none of the estimated coefficients in this regression are individually significant and the group of baseline attitude variables is not jointly significant either.

treatment status for any of the three outcome measures.

Finally, we test for differential attrition after boot camp as well as after 6 month into military service. Not everyone finishes boot camp, with about 15 percent of recruits not completing the full 8 weeks according to the military. In addition, some recruits will not be available to take our survey for other reasons (e.g., sickness or another conflicting assignment) or could choose to skip the gender attitude questions when taking the survey. Together, these reasons contribute to slightly less than one third of our baseline sample with missing values for the gender attitude questions asked at the end of boot camp. There is additional attrition during the post-boot camp service period for similar reasons, so that roughly one-half of respondents have missing values for the gender attitude questions.

If attrition is random, this should not affect the causality of our estimates. But if women cause men to exit boot camp/military service or not complete the surveys, this could create a bias. In Appendix Table A4 we test whether treatment predicts attrition in either wave 2 (post boot camp) or wave 3 (6 months into military service). For each of our gender attitude questions and both survey waves, we regress a dummy variable for whether the outcome is missing. We find no evidence that being assigned to a squad with a female in it increases attrition. If anything, the estimates are negative. While we present estimates from separate regressions, it should be noted that having a missing value is highly correlated across the three attitude variables. It is therefore not surprising the estimates are similar. In a related check for nonrandom attrition, in Appendix Table A5 we show that attrition is not correlated with baseline answers in wave 1.

V Short-Run Experimental Results

A Short-Run Gender Attitudes

We are interested in the causal relationship between female integration in the squad and men's gender attitudes. Attitudes are measured both before and after treatment. We model attitudes for individual *i,* in squad *j,* in troop *k,* in wave 2 (i.e., at the end of boot camp), as:

$$y_{ij2} = \alpha_k + \beta y_{i1} + \theta F_{ij1} + \gamma x_{i1} + \epsilon_{ij2} \tag{1}$$

where α_k is a set of troop fixed effects, y_{i1} measures baseline attitudes in survey wave 1 (before treatment and assignment to a squad), F_{ij1} is a dummy variable for whether a

female is assigned to individual i's squad j in wave 1 and x_{i1} is a set of pre-determined control variables measured in wave 1.

In our regressions, we use binary indicators for attitudes in wave 2, with a value of 1 indicating a more gender equal attitude for our main outcomes. To control for initial, pre-treatment attitudes, we include a full set of dummy variables for the possible answers to the baseline questions.[18] Troop fixed effects are included since randomization occurs within troops. Since treatment is at the room/squad level, in all of our experimental regressions we report standard errors clustered by boot camp squad.

Table IV presents our main experimental estimates for the gender attitude questions in the short run. For each question, we report two sets of estimates. The first includes troop fixed effects and dummies for initial gender attitudes in wave 1.[19] The second adds in the covariates appearing in Appendix Table A1. As expected given randomization, adding in these additional covariates does not materially affect the estimates. Therefore, we focus our discussion on the columns in the table which include these pre-determined covariates.

Column 2 reports how having a female assigned to your team affects attitudes about mixed-gender productivity. Men exposed to female team members are 13 percentage points more likely to disagree with the claim that same gender teams perform better. Relative to the control mean of 52 percent, this is a 26 percent increase.

In column 4, the dependent variable is whether recruits agree that it is important for men and women to share household work equally. There is an 8 percentage point increase in egalitarian attitudes for this outcome after working and living with females for the 8 weeks of boot camp. This is a 13 percent increase relative to the control mean of 66 percent.

Column 6 shows how exposure changes attitudes related to gender identity. There is a 14 percentage point increase in the number of men who do not completely disavow their femininity, relative to an overall mean of 63 percent. It is important to remember that this change in attitudes does not represent a full embracement of being feminine; rather the relevant comparison is that most men are switching from a complete disavowal of femininity to a weaker disavowal of femininity.

When interpreting these treatment effects, it is useful to compare them to the

[18] We include a dummy variable for missings and combine adjacent response categories when creating the set of dummy variables if there are less than 10 responses in a category.

[19] Including initial attitudes substantially increases the explanatory power of the regressions; relative to only including troop fixed effects, the R-squareds rise from between .02 and .07 to between .19 and .36, depending on the outcome.

attitude changes of the control group, which consists of all-male squads. Even in the absence of being exposed to women, boot camp could change men's attitudes about gender issues because it is a new, masculine environment. Positive attitudes towards mixed-gender productivity fall by 11 percentage points for all-male squads between survey waves 1 and 2. Apparently, working intensively in an all-male team during boot camp results in more men thinking this type of arrangement is better than a mixed-gender team. The treatment effect of a 13 percentage point increase approximately cancels out this decline. In other words, being assigned to a mixed-gender squad prevents a deterioration in attitudes. So our findings could alternatively be interpreted as the negative effects of being segregated into an all male squad during boot camp relative to being mixed with women. In contrast, attitudes regarding household work and gender identity do not change much for the control group between waves 1 and 2. This implies that exposure to female squad members increases egalitarian attitudes along these dimensions.

These are large swings in gender attitudes. One way to see this is to compare the size of these estimates to the coefficients on various background characteristics in Table II. The magnitude of the effects are as large or larger compared to most of the coefficients, including having female friends or a sister. Another way to gauge just how large these effects are is to compare them to the gap which exists between men and women in these attitude variables. For the mixed-gender productivity question, there is a 31 percentage point gap in male versus female attitudes at baseline for our estimation sample, so the treatment effect is 42 percent as large. There is an 18 percentage point gap between men and women regarding the equitable sharing of household work. Exposure to women erases almost half of this gap. And finally, the gap in the gender identity variable is the largest, with a 44 percentage point difference between the sexes. Exposure to women reduces this gap by 31 percent.

Based on our experimental results, we conclude that men's gender attitudes can be changed in the short run via the intensive exposure to women provided by our field experiment. Experimenter demand effects or social desirability bias are unlikely to explain our results, for several reasons. First, treated recruits are not primed to be thinking about gender integration, as they are not made aware that women are randomly being assigned to some squads and not others as part of an experiment. Related, the survey was conducted in a large group setting, and not at the squad level. Moreover, integration of women into the military was not a new phenomenon, as it had

already been going on for several years by the time of our field experiment. Finally, the survey questionnaire was not focused on gender issues, but instead asked a few gender questions interspersed among a long list of unrelated questions.

B Short-Run Occupation Choices

Do the changed attitudes reported in Table IV correspond to observed changes in military outcomes? In particular, since treatment prevents a deterioration in attitudes related to mixed-gender teams, does treatment also result in an increased willingness to serve with women after boot camp relative to the controls? We answer this question by looking at whether treatment affects the fraction of women in one's military occupation. We run the following regression:

$$z_{i2} = \alpha_k + \theta F_{ij1} + \gamma x_{i1} + \epsilon_{ij2} \tag{2}$$

where z_{i2} is the fraction of women in a soldier's assigned occupation during wave 2, α_k are troop fixed effects, F_{ij1} is the treatment variable for having a female in one's squad during boot camp, and x_{i1} are predetermined control variables.

Soldiers are assigned to military occupations at the end of boot camp, and serve in these positions during their required service period of approximately 10 months. As a reminder, occupations are determined based on individual preferences in conjunction with the needs of the military. Example occupations with a large fraction of women in them include liaison soldier (17% female), administrative assistant (21%) and medical assistant (26%) while those with a smaller fraction include cannon soldiers (4%), rifle soldiers (5%) and storm troopers (6%).

The last two columns of Table IV report the results of running these regressions. Focusing on column 8, exposure to women during boot camp results in a 2.2 percentage point increase in the fraction of women in one's chosen military occupation. This is a non-trivial effect when compared to the control mean of 10% women in an occupation. In other words, treated individuals have a 22% higher fraction of women in their occupation compared to those in all-male boot camp squads. We interpret this as being driven largely by the preferences of treated men, although we recognize that it is also possible that military leaders believe that treated men interact better with women and therefore are more likely to place them in occupations with more women.

C *Short-Run Performance and Satisfaction*

Leaders within and outside the military have debated whether allowing women to serve in the military would ruin its esprit de corps, causing lower performance and dissatisfaction within the ranks. To test this hypothesis, the survey asked recruits' about their self-assessed preparedness and satisfaction with boot camp. Table V reports estimates from regressions which parallel those appearing in Table IV. Our first survey question is "I feel qualified for further military service." Our second question asked "Overall, how satisfied were you with military service?" As the table shows, we find no evidence that having a female in the squad affected a recruits' preparedness or satisfaction attitudes at the end of boot camp.

We have one short-run military performance outcome which the military collected for us. At the end of boot camp, or in some cases shortly thereafter, the military evaluates which recruits should be promoted to Vice Corporal, a junior rank just above Private. This rank designates a higher level of achievement and preparedness for future service, and often comes with a leadership appointment. Sixteen percent of recruits in our control group achieve this rank. If anything, treatment increases the probability of promotion, but the estimate is not statistically significant.

The survey also asked several other questions related to recruits' self-assessment of their boot camp experience. Appendix Table A6 presents regression results for the questions "Want to do military service," "Satisfaction with room," and "Plan to continue in the military." For each of these questions, the estimates at the end of wave 2 (when boot camp finishes) are close to zero and not statistically significant.

Based on the results in Table V and Appendix Table A6, we conclude that, contrary to the predictions of many policymakers, integrating females into squads during boot camp does not result in measurable drops in male recruits' preparedness, satisfaction or desire to serve.

D *Robustness and Heterogeneity*

Appendix Table A7 presents a series of robustness checks corresponding to Table IV. We show that our estimates are robust to using lasso regression to chose pre-determined control variables, the exclusion of baseline attitudes and other predetermined controls, probit regressions, first difference specifications and the use of the full range of responses instead of binary outcomes. Another set of robustness checks explores the effect of different numbers of females in the room and different room sizes, finding no statistical

evidence of heterogeneity. Details on how these robustness checks are implemented can be found in the Appendix.

We further explore how the distribution of answers responds to treatment in Appendix Figure A3. These graphs plot the change in attitudes, using the full linear scale and separately for treated and control individuals, after boot camp versus before the experiment began. As seen in each of the three panels of the figure, the distribution of changed attitudes is shifted to the right for treated soldiers compared to the controls for each of our gender attitude questions. Formal tests reject equal ordinal rankings for treatment and control observations in all three panels; these tests are reported below the graphs in Appendix Figure A3.[20]

In Appendix Table A8, we report heterogeneous effects based on treatment interactions with pre-determined characteristics of men. We find no statistical evidence of heterogeneity by men's muscle strength, or by men's general ability test scores. Likewise, we find no significant heterogeneity by whether the man has a high share of female friends, a sister or a highly educated mother. The same is true for heterogeneous interactions based on the man's initial gender attitude for the relevant outcome. We are quick to point out that heterogeneity could exist, but that our standard errors are too large to detect it. While it would also be interesting to explore heterogeneity by female's characteristics, we have even less precision to explore such effects.

E Attitudes towards Female Leadership

Before continuing, we explore attitudes on which sex makes the best leaders at various levels of command. Our first question on female leadership asks respondents: "Which sex do you believe is the best at leading a troop?" As a reminder, troops are groups of 5 to 10 squads. Our second question is: "Which sex do you believe are the best leaders at the highest level?" The third question is: "Which sex do you believe is the best at leading foreign operations?" Possible responses to each of these questions are "Men," "Equally good," and "Women." We create dummy variables which equal one for an answer of "Equally good" or "Women" for each of the questions. It is important

[20] We use the Goodman-Kruskal gamma to test the equality of ordinal rankings for the treated and control groups. Applied to our setting, it estimates the population parameter $(P_t - P_c)/(P_t + P_c)$, where P_t is the probability that a randomly selected treatment observation will have a higher value than a randomly selected control observation, and P_c is the reverse probability. The probability that the two observations will have the same value is also a possibility, and explains why $P_t + P_c \neq 1$. The statistic varies between -1 and 1, and is zero under the null hypothesis of equal ordinal rankings.

to recognize that a value of 1 largely represents "Equally good," as only a handful of either male or female respondents think that women are the best leaders.

Table VI estimates similar regressions as for our main gender attitude questions in Table IV. For all three female leadership questions, the estimates are close to zero and statistically insignificant in both the short run (wave 2, at the end of boot camp) and the longer-run (wave 3, 6 months later). Apparently, exposure to women at the squad level does not affect these attitudes. To put this in context, note that we did not randomize the gender of these higher-level leaders, and that most of the leaders at these levels are male. This suggests that exposure to rank and file women does not have a spillover effect on attitudes about higher up female leadership.

What may be required is exposure to female leaders at the relevant level. Indeed, using a similar Norwegian military setting, Finseraas et al. (2016) conducted a vignette experiment which asked about attitudes towards hypothetical females becoming squad leaders, and finds that exposure to females at the squad level does make a difference. Related work by Beaman et al. (2009) finds that prior randomized exposure to a female leader through gender quotas results in changes in voter attitudes regarding the effectiveness of female leaders and weakens gender stereotypes. In a similar vein, Bertrand et al. (2014) find that while an increase in female board members may have improved the representation of female employees at the very top, there is no evidence these gains trickled down.

VI Longer-Run Experimental Results

A Longer-Run Gender Attitudes

Given the changes in gender attitudes and occupation choices observed in the short run, a natural question is whether these effects persist after treatment ends. Likewise, a related question is whether there are any long-run effects on performance or soldiers' satisfaction with service.

To examine whether changed gender attitudes persist, we run similar regressions to equation 1, but using wave 3 survey attitudes as the outcome variables. To hone in on changes between the end of boot camp and after further military service, we also estimate regressions which use the difference between wave 3 and wave 2 attitudes as the outcomes. As a reminder, in terms of the timeline, wave 1 of the survey is conducted before boot camp, wave 2 is conducted near the end of boot camp, and wave

3 is conducted 6 months after boot camp. Figure II plots the means for treatment and control for each survey wave to help illustrate the attitudinal changes as individuals progress in their military service.

Table VII reports on the longer-run evolution of gender attitudes in a regression framework. Men treated during boot camp do not have statistically different attitudes regarding same gender team performance or gender identity 6 months into mandatory service, despite there having been large effects in the short run. The estimates for each of these attitudes are close to zero (see columns 1 and 3), although we note that we cannot rule out modestly-sized effects.[21] The question on household work was not asked in wave 3 of the survey, which was conducted by the military and not overseen by us. The null findings are confirmed when comparing the full distribution of possible changes in attitudes in Appendix Figure A4, with formal statistical tests failing to reject equality.

The differences specification in Table VII, columns 2 and 4, quantifies the magnitude of the deterioration in attitudes between the end of boot camp and 6 months into service. There is a 19 percentage point drop in the fraction of treated men who think that same gender teams perform better compared to the controls, and a similar 19 percentage point drop for the gender identity question. Comparing attitudes for the controls prior to boot camp (wave 1), immediately after boot camp (wave 2) and 6 months into military service (wave 3) yields further insight. The control group of men who are in all-male squads during boot camp see an 11% deterioration in beliefs that mixed-gender teams are effective by the end of boot camp. These same controls see a further 11% drop by 6 months into military service. In other words, treatment prevents a deterioration of attitudes during boot camp, but by 6 months into service the gap between treatment and control has completely disappeared. This suggests that military service is itself a strong treatment which affects attitudes. While the bootcamp exposure effects do not persist, if the exposure had been more persistent at a comparable scale for the entire military service, it might have had lasting effects.

B Longer-Run Education, Occupation and Workplace Choices

In Section B, we found that exposure to women during boot camp changed which occupations men served in during subsequent military service, with men being assigned

[21] If we re-estimate the short-run attitude regressions appearing in Table IV, but restrict the sample to individuals with wave 3 data, the estimates are similar and remain statistically significant. The estimate corresponding to column 2 is .1480 (s.e.=.0694) and to column 6 is .1778 (s.e.=.0600).

to occupations with more women in them. Using our linked data from Statistics Norway, we test whether this type of change persists in the longer-run, after the mandatory service period ends. We do this using variables for the fraction of women in an individual's chosen civilian field of study, occupation or workplace after mandatory military service has ended.

We first calculate the fraction of women in every field of study, including both college majors and vocational training, in the entire Norwegian population using 4 digit Norwegian Standard Educational Codes. We then define the fraction of women in an individual's chosen education based on their first year of enrollment in higher education after 2014 (i.e., after mandatory military service is over). Seventy-eight percent of our sample pursues some type of further education.[22] We likewise calculate the fraction of women in every occupation in the entire Norwegian population using 4 digit International Standard Classification of Occupations Codes. We then define the fraction of women in an individual's chosen occupation in the first year they are employed after 2014. All but three of the individuals in our estimation sample held a job, even if just part time.[23] Finally, we calculate the fraction of women at a workplace based on the establishment an individual works at, using the job with the highest earnings in the first year they are employed after 2014.

As columns 5-7 of Table VII document, treatment during boot camp does not cause men to choose a field of study, occupation or workplace which has a higher fraction of women in it.[24] Treatment decreases the fraction of women in a chosen education by .3 percentage points, increases the fraction of women in a chosen occupation by 2.5 percentage points and increased the fraction of women in the workplace by 2.6 percentage points. These effect sizes are statistically insignificant, estimated with reasonable precision and small relative to the control group female shares of 52%, 45% and 39%, respectively.[25]

[22]We calculate shares using individuals enrolled in higher education born between 1991 and 1995. If there are fewer than 100 observations in a 4 digit education code, we collapse to the 3 digit level.

[23]The shares are calculated using all individuals that are working and between the ages of 18-65. If there are fewer than 100 observations in a 4 digit occupation code, we collapse to the 3 digit level. If an individual has more than one job, we assign their occupation based on the job with the highest earnings. If we focus on full-time employees by requiring earnings to be above one or two "basic amounts" as defined by the Norwegian government for social insurance programs, our results do not change.

[24]Table VII includes the control variables appearing in Appendix Table A1. Appendix Table A9 reports estimates without control variables, and Appendix Table A10 reports lasso estimates which choose which controls to include. The results are similar.

[25]In our pre-analysis plan, we also proposed a combined measure for education and occupation, since

We conclude that being treated by women during boot camp and subsequently serving in an occupation with more women during military service does not translate into longer-run education or occupation choices. This aligns with the pattern of deteriorating gender attitudes for treated men after intensive exposure during boot camp ends. For context in interpreting the results in Sections 6.1 and 6.2, recall that the military reassigns recruits to a new set of squads, which do not correspond to occupation, for the 10 month service period after boot camp. Treatment during boot camp was strong and intensive: treated men were exposed to 2 women in a team of 6 members, living and training together 24 hours a day, 7 days a week. In contrast, during the 10 month service period, individuals are exposed to a variety of influences. Individuals sometimes train with individuals in their occupation and sometimes perform activities together squad, but in a far less intensive manner. Individuals in the same squads do not necessarily live in the same rooms, and moreover, individuals are allowed to leave the base during evenings and on weekends. So even though treated individuals have a higher fraction of women in their military occupation on average, this occupational exposure is apparently not strong enough to make a difference during military service.

C Longer-Run Performance and Satisfaction

As a final set of longer-run outcomes, we study preparedness and satisfaction attitudes as well as military performance evaluations in Table VIII. Near the end of service, soldiers are asked the same two questions they were asked at the end of boot camp: "I feel qualified for further military service" and "Overall, how satisfied were you with military service?" Similar to what we found at the end of boot camp, there is no evidence for treatment affecting these measures in the longer run (columns 1 and 2).

To study military performance outcomes, we use the evaluations conducted by the military at the end of mandatory service. These evaluations are used by the military to decide which soldiers will be offered the option to continue with a military career. Recruits are graded in four different categories, receiving a rating of below expectations, at expectations, exceeds expectations and excellent. The four categories and their expectations are defined as:

- *Conduct*: The soldier performs the required tasks reliably, acts courteously, is punctual and shows respect for fellow human beings.

we thought we might not have enough observations for each separately. Using the combined variable yields an estimate of -.037 (s.e.=.097).

- *Cooperation and Communication*: The soldier solves tasks together with others, is open to the views of others, contributes to good communication and shows a willingness to cooperate.

- *Independence and Initiative*: The soldier takes initiative and is active as well as demonstrates an ability for independent judgment.

- *Overall Assessment*: Overall suitability for further military service; a minimum rating of at expectations is needed to qualify for international operations, the home guard or further training/employment by the military.

For each category we create a binary variable which equals 1 if the soldier is rated as exceeds expectations or excellent. We also create a summary variable which averages the mean of these 4 binary variables.

Results appear in columns 3-7 of Table VIII. We start with the average of the four evaluations in column 3, which has a control group mean of .47. Being treated during boot camp has a negligible effect on this outcome, with a coefficient estimate of .009 (s.e.=.043). The evaluations which individually measure conduct, cooperation, independence and overall performance are similarly not statistically affected by treatment. The estimates are precise enough to rule out large effects on performance.[26]

In Appendix Table A6, we report estimates for additional questions relating to the desire to serve in the military, satisfaction with one's room and plans to continue in the military. We find no evidence that any of these outcomes are affected by treatment in either the short or longer run. In Appendix Table A11, we explore commendations received during service for superior achievement in areas such as sharpshooting, loaded marches, and infantry runs.[27] There is some evidence for a negative treatment effect, although most of the coefficients are not statistically significant. We infer that negative award effects, if they do exist, are too small to have an impact on the end of service evaluations which are more consequential.

[26]If we instead measure the effect of receiving a rating of below expectations, the estimates are likewise insignificant.

[27]These can be earned anytime during military service, but not during boot camp. Sharpshooting medals are based on accuracy and speed; the loaded march medal is given for successfully marching 30 kilometers in wild terrain with full gear and an 11 kilogram backpack within a specified time; and the infantry medal is given for completing an 8-12 km orienteering obstacle course in wild terrain with shooting drills, but without a backpack, within a specified time. The skiing medal involves cross-country skiing and shooting and the fieldsport medal includes shooting, navigation, and running.

Taken together, the overall pattern of results points to little difference in stated or observed differences in performance in either the short or longer run. Likewise, treated individuals are just as likely to have been satisfied with their military service. We conclude that, contrary to the predictions of many policymakers, integrating females into squads during boot camp did not result in measurable losses in male recruits satisfaction with boot camp or service and did not alter their preparedness for or performance during military service.

VII Conclusion

This paper provides novel evidence on whether the integration of women into a traditionally male-dominated environment can change stereotypical gender attitudes of males. We overcome the difficulties of reverse causation, self-selection and unobserved heterogeneity by implementing a field experiment which randomly assigned some men to live and work with women during boot camp in the military. We find that this intensive interaction with women for 8 weeks causes men to become more egalitarian in their attitudes towards mixed-gender productivity, gender roles and gender identity, and moreover that it shifts their occupational choices towards jobs with more females in them.

These findings demonstrate that men's gendered attitudes are not fixed, but can change through interaction with women. Our field experiment was a strong intervention, changing both living arrangements and the working environment. In this setting, men and women were equal in rank and had to complete a similar set of tasks. Moreover, men and women were placed into teams which required cooperation to reach common goals, such as the completion of a training exercise. These features combine to create exactly the type of setting predicted by contact theory to result in changed attitudes.

However, we also find these short-run changes do not persist into the future, after intensive exposure to women stops. Gender attitudes during subsequent military service deteriorate for treated males, such that they are no longer different from the controls. Consistent with this finding, we find no effect on men's willingness to choose educations, occupations or workplaces which have a larger fraction of females in the longer run. We reiterate that this does not mean that exposure to women cannot change gender attitudes in the longer run and in other settings. Our intervention, while intense during bootcamp, was relatively short compared to the military experience overall.

Our findings have important implications for policies aimed at integrating the

workplace as well as for societal norms more broadly. Our results suggest that exposure is an important lever which can be used to overcome pre-existing priors regarding a woman's suitability for a job, as well as stereotypical attitudes which affect life outside the workplace. However, our results also suggest that to maintain such improvements, intensive exposure to females needs to continue or the gains are likely to disappear. Moreover, contrary to the concerns of some policymakers, our findings indicate that efforts to integrate women into the military can be achieved without destroying the camaraderie of service, team effectiveness or preparedness for future service.

Our study is the first to randomly assign men and women to live and work together for an extended period of time and study how attitudes and real-world outcomes change. This adds to prior work which has explored how extended exposure can change racial and ethnic attitudes. While the results of our field experiment provide new empirical evidence in support of contact theory in the gender domain, several questions remain. Do these results transfer to other settings, such as the integration of police forces, which have also been historically male-dominated? Can the attitudes of older individuals be changed via exposure, or are only young people's attitudes malleable? And finally, if intensive exposure were to continue, would the effects persist? These are interesting questions for future research.

References

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *The Quarterly Journal of Economics 115*(3), 715–753.

Alesina, A., P. Giuliano, and N. Nunn (2013). On the origins of gender roles: Women and the plough. *The Quarterly Journal of Economics 128*(2), 469–530.

Allport, G. W. (1954). *The Nature of Prejudice*. Reading: Addison-Wesley.

Anwar, S., P. Bayer, and R. Hjalmarsson (2012). The impact of jury race in criminal trials. *The Quarterly Journal of Economics 127*(2), 1017–1055.

Austen-Smith, D. and R. Fryer (2005). An economic analysis of "acting white". *Quarterly Journal of Economics 120*(2), 551–583.

Bailey, M. J. and T. A. DiPrete (2016). Five decades of remarkable but slowing change in U.S. women's economic and social status and political participation. *The Russell Sage Foundation Journal of the Social Sciences 2*(4), 1–32.

Bayer, A. and C. E. Rouse (2016). Diversity in the economics profession: A new attack on an old problem. *Journal of Economic Perspectives 30*(4), 221–242.

Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009). Powerful

women: Does exposure reduce bias? *The Quarterly Journal of Economics 124*(4), 1497–1540.

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (2014). Breaking the glass ceiling? The effect of board quotas on female labor market outcomes in Norway. *National Bureau of Economic Research Working Paper 20256*.

Bertrand, M., P. Cortes, C. Olivetti, and J. Pan (2016). Social norms, labor market opportunities, and the marriage gaps for skilled women. *National Bureau of Economic Research Working Paper 22015*.

Bertrand, M., E. Kamenica, and J. Pan (2015). Gender identity and relative income within households. *The Quarterly Journal of Economics 130*(2), 571–614.

Bettio, F. and A. Verashchagina (2009). Gender segregation in the labour market: Root causes, implications and policy responses. European Commission's Expert Group on Gender and Employment.

Blau, F. D., P. Brummund, and A. Y.-H. Liu (2013). Trends in occupational segregation by gender 1970-2009: Adjusting for the impact of changes in the occupational coding system. *Demography 50*(2), 471.

Blau, F. D. and L. M. Kahn (2003). Understanding international differences in the gender pay gap. *Journal of Labor Economics 21*(1), 106–144.

Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature 55*(3), 789–865.

Boisjoly, J., G. J. Duncan, M. Kremer, D. M. Levy, and J. Eccles (2006). Empathy or antipathy? The impact of diversity. *American Economic Review 96*(5), 1890–1905.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Burns, J., L. Corno, and E. La Ferrara (2016). Interaction, prejudice and performance: Evidence from South Africa.

Carrell, S. E., M. Hoekstra, and J. E. West (2015). The impact of intergroup contact on racial attitudes and revealed preferences. *National Bureau of Economic Research Working Paper No. 20940*.

Dahl, G. B. and E. Moretti (2008). The demand for sons. *The Review of Economic Studies 75*(4), 1085–1120.

Duncan, O. D. and B. Duncan (1955). A methodological analysis of segregation indexes. *American Sociological Review 20*(2), 210–217.

Ellingsen, D., U.-B. Lilleaas, and M. Kimmel (2016). Something is working–but why? Mixed rooms in the Norwegian army. *NORA-Nordic Journal of Feminist and Gender Research 24*(3), 151–164.

Evans, D. C. (2003). A comparison of the other-directed stigmatization produced by legal and illegal forms of affirmative action. *Journal of Applied Psychology 88*(1), 121.

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 ceiling: Evidence 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*.

Fortin, N. M. (2005). Gender role attitudes and the labour-market outcomes of women across OECD countries. *Oxford Review of Economic Policy 21*(3), 416–438.

Fortin, N. M. (2015). Gender role attitudes and women's labor market participation: Opting-out, aids, and the persistent appeal of housewifery. *Annals of Economics and Statistics/Annales d'Économie et de Statistique* (117/118), 379–401.

Goldin, C. (2004). The long road to the fast track: Career and family. *The Annals of the American Academy of Political and Social Science 596*(1), 20–35.

Goldin, C. (2014a). A grand gender convergence: Its last chapter. *The American Economic Review 104*(4), 1091–1119.

Goldin, C. (2014b). A pollution theory of discrimination: Male and female differences in occupations and earnings. In *Human capital in history: The American record*, pp. 313–348. University of Chicago Press.

Goldin, C. and L. F. Katz (2002). The power of the pill: Oral contraceptives and women's career and marriage decisions. *Journal of Political Economy 110*(4), 730–770.

Goldin, C. and L. F. Katz (2016). A most egalitarian profession: Pharmacy and the evolution of a family-friendly occupation. *Journal of Labor Economics 34*(3), 705–746.

Hanson, T., F. Steder, and S. Kvalvik (2016). Hva motiverer til tjeneste i forsvaret? En innledende kvantitativ analyse av holdninger og adferd i brigade nord. Technical report, FFI.

Harrell, M. C. and L. L. Miller (1997). New opportunities for military women effects upon readiness, cohesion, and morale. Technical report, RAND Corporation.

Hellum, N. (2014). Sminkedritt over hele vasken – en kvalitativ feltstudie av kjønnsblandede rom og maskulinitetskultur i Forsvaret. Technical Report 2156, FFI.

Hellum, N. (2017). Not focusing on whether it's a spout or a handle: An anthropological study on even gender balance among conscripts in a Norwegian Air Force battalion. Technical Report 01196, FFI.

Kotsadam, A. and N. Jakobsson (2011). Do laws affect attitudes? An assessment of the Norwegian prostitution law using longitudinal data. *International Review of Law and Economics 31*(2), 103–115.

Kotsadam, A. and N. Jakobsson (2014). Shame on you, John! Laws, stigmatization, and the demand for sex. *European Journal of Law and Economics 37*(3), 393–404.

Lilleaas, U.-B. and D. Ellingsen (2014). *Likestilling i Forsvaret Fortropp, baktropp og kamparena, (Gender equality in the Armed Forces: Battleground, rear guard and vanguard)*. Cappelen Damm Akademisk.

Lowe, M. (2020). Types of contact: A field experiment on collaborative and adversarial caste integration. *CESifo Working Paper 8089*.

Moss-Racusin, C., J. Dovidio, V. Brescoll, M. Graham, and J. Handelsman (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences 109*(41), 16474–16479.

Mousa, S. (2020). Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq. *Science 369*(6505), 866–870.

Olivetti, C. and B. Petrongolo (2016). The evolution of gender gaps in industrialized countries. *Annual Review of Economics 8*, 405–434.

Paluck, E. L., S. A. Green, and D. P. Green (2019). The contact hypothesis re-evaluated. *Behavioural Public Policy 3*(2), 129–158.

Pettigrew, T. F. and L. R. Tropp (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology 90*(5), 751–783.

Rao, G. (2019). Familiarity does not breed contempt: Generosity, discrimination, and diversity in Delhi schools. *American Economic Review 109*(3), 774–809.

Resendez, M. G. (2002). The stigmatizing effects of affirmative action: An examination of moderating variables. *Journal of Applied Social Psychology 32*(1), 185–206.

Sacerdote, B. (2011). Peer effects in education: How might they work, how big are they and how much do we know thus far? In *Handbook of the Economics of Education*. Elsevier.

Scacco, A. and S. S. Warren (2018). Can social contact reduce prejudice and discrimination? Evidence from a field experiment in Nigeria. *American Political Science Review 112*(3), 654–677.

Stamarski, C. and L. Son Hing (2015). Gender inequalities in the workplace: The effects of organizational structures, processes, practices, and decision makers' sexism. *Frontiers in Psychology 6*, 1400.

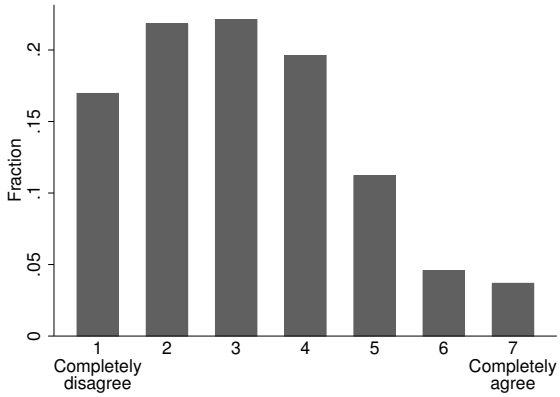
Stinebrickner, R. and T. R. Stinebrickner (2006). What can be learned about peer effects using college roommates? Evidence from new survey data and students from disadvantaged backgrounds. *Journal of Public Economics 90*(8), 1435–1454.

Strøm-Erichsen, A. (2013). Kompetanse for en ny tid. *Government report*.

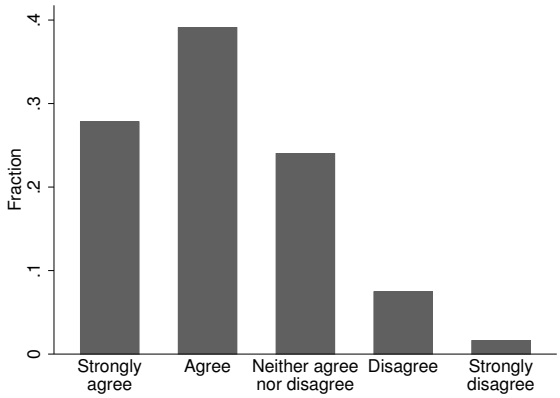
Van Laar, C., S. Levin, S. Sinclair, and J. Sidanius (2005). The effect of university roommate contact on ethnic attitudes and behavior. *Journal of Experimental Social Psychology 41*(4), 329–345.

Zimmerman, D. J. (2003). Peer effects in academic outcomes: Evidence from a natural experiment. *The Review of Economics and Statistics 85*(1), 9–23.

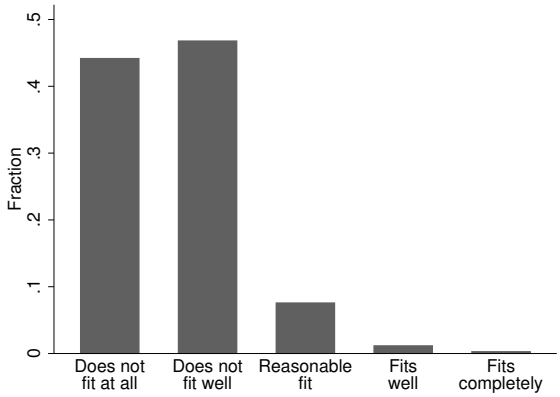
Figure I. Distribution of gender attitudes for male recruits at baseline



A. Mixed-gender productivity: "A team performs better when it is made up of the same gender"



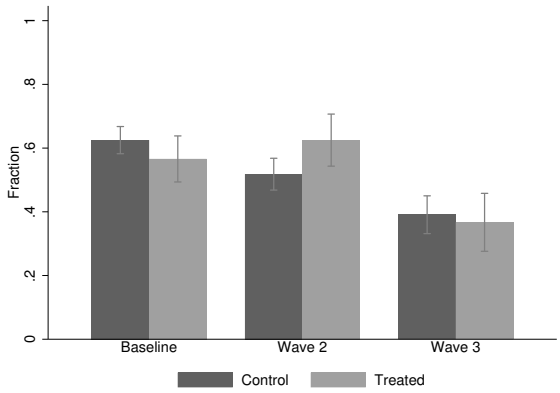
B. Gender roles: "It is important that men and women share household work equally"



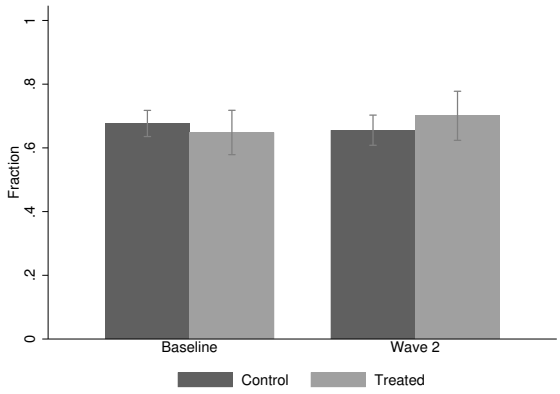
C. Gender identity: "How well does the following statement describe you: I am feminine"

Notes: Responses to baseline survey conducted prior to boot camp, excluding missings. There are 678, 683 and 686 observations in panels A, B and C, respectively.

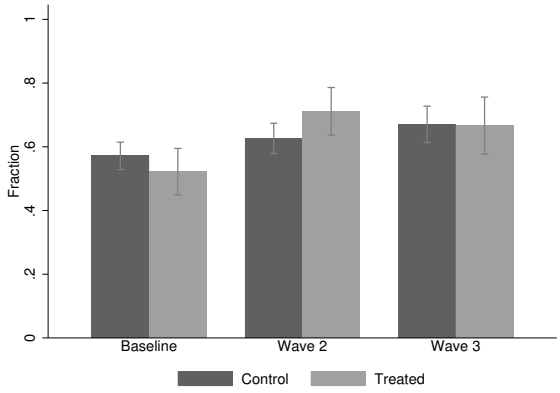
Figure II. Distribution of gender attitudes over time



A. Mixed-gender productivity: "A team performs better when it is made up of the same gender" *(Disagree=1)*



B. Gender roles: "It is important that men and women share household work equally" *(Agree=1)*



C. Gender identity: "How well does the following statement describe you: I am feminine" *(Statement does not fit me at all=0)*

Notes: Vertical lines denote 95% confidence intervals. There are 678, 683 and 686 baseline observations in panels A, B and C, respectively. There are likewise 522, 526 and 538 observations in wave 2 and 370 and 369 observations in wave 3.

Table I. Examples of gender segregated occupations in Norway and the U.S. in 2014

	% Female			% Female	
	Norway	U.S.		Norway	U.S.
Male dominated	(1)	(2)	Female dominated	(3)	(4)
Plumbers	2	2	Dental assistants	97	97
Firefighters	6	6	Pre-K / K teachers	92	97
Aircraft pilots	6	7	Hair dressers	90	94
Truck drivers	10	6	Registered nurses	89	90
Military	**13**	**15**	Social workers	84	82
Computer programmers	20	20	Pharmacists	76	56
Civil engineers	22	17	Primary school teacher	74	81
Geoscientists	31	25	Physical therapists	73	70
Architects	46	25	Librarians	71	84

Notes: Fractions for Norway from authors' tabulations of register data. Fractions for the U.S. from the U.S. Bureau of Labor Statistics, BLS Reports: Women in the labor force: A databook (2015).

Table II. OLS estimates of gender attitudes

	"Same gender teams perform better" *Disagree=1* (1)	"Important to share HH work equally" *Agree=1* (2)	"I am feminine" *Statement does not fit me at all=0* (3)
Female friends	.0845**	.0185	.0310
	(.0263)	(.0258)	(.0271)
Has a sister	.0141	.0636**	-.0025
	(.0320)	(.0314)	(.0330)
Has a brother	-.0308	.0081	.0037
	(.0335)	(.0328)	(.0344)
High muscle strength	-.1088**	.0273	-.0529**
	(.0261)	(.0256)	(.0268)
High GAI test score	.0463	-.0844**	.0882**
	(.0259)	(.0254)	(.0267)
Mother higher education	.0000	.0167	.0347
	(.0271)	(.0266)	(.0279)
Father higher education	.0076	-.0364	.0219
	(.0275)	(.0271)	(.0284)
Mother works	-.0666	-.0205	.0215
	(.0455)	(.0448)	(.0472)
Parents divorced or separated	-.0122	-.0126	.0476*
	(.0277)	(.0272)	(.0285)
R-squared	.031	.019	.018
[p-value]	[.000]	[.049]	[.066]
N	1,430	1,439	1,442

Notes: Sample includes all male recruits from the baseline survey, including recruits in battalions which both did and did not participate in the experiment. Dummy variables for missing values of control variables are also included in the regressions. Female friends is a dummy for 40% or more of one's friends being female. High muscle strength is a dummy for above average or far above average muscle strength. GAI stands for general ability index, and is based on verbal comprehension and perceptual reasoning tests; high GAI test score is a dummy for scoring at or above the 6th stanine.

***p-value<.05, *p-value<.10*

Table III. Test of random assignment

	Female on Team
Female friends	-.0120
	(.0313)
Has a sister	-.0450
	(.0383)
Has a brother	.0464
	(.0400)
High muscle strength	-.0130
	.(0289)
High GAI test score	.0007
	(.0283)
Mother higher education	.0282
	(.0324)
Father higher education	-.0058
	(.0331)
Mother works	.0330
	(.0552)
Parents divorced or separated	.0048
	(.0337)
Joint F-statistic	.77
[p-value]	[.726]
Dependent mean	.282
N	781

Notes: Sample includes male recruits in battalions which participated in the experiment and who were assigned to rooms with between 5 and 7 members. The regression includes indicators for troop, since randomization of men and women to squads (i.e., rooms) occurs within troops. Dummy variables for missing values of the control variables are also included. The joint F-statistic has 17 degrees of freedom and is based on the controls shown in the table plus the dummies for missing values (there are 17 instead of 18 degrees of freedom, because a missing value for mother higher education is perfectly correlated with a missing value for father higher education).

***p-value<.05, *p-value<.10*

Table IV. Gender attitudes and occupation choices at the end of boot camp

	Attitudes after Boot Camp						**Military Occupation**	
	"Same gender teams perform better" *Disagree=1*		"Important to share HH work equally" *Agree=1*		"I am feminine" *Statement does not fit me at all=0*		Fraction women in chosen military occupation	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female on team	.1487**	.1333**	.0770*	.0821**	.1359**	.1382**	.0208**	.0221**
	(.0503)	(.0525)	(.0396)	(.0407)	(.0508)	(.0507)	(.0075)	(.0073)
Control variables	No	Yes	No	Yes	No	Yes	No	Yes
R-squared	.193	.221	.220	.260	.355	.386	.133	.161
Control group mean	.518	.518	.656	.656	.626	.626	.100	.100
N	522	522	526	526	538	538	657	657

Notes: Sample includes all male recruits in battalions which participated in the experiment. The attitude variables come from the wave 2 survey taken at the end of boot camp. The military occupation is chosen at the end of boot camp, with assignments made based on the choices of individuals combined with the needs of the military. Attitude and occupation data come from two different datasets which cannot be merged due to confidentiality laws in Norway. All regressions include indicators for troop. The attitude regressions additionally include the corresponding lagged attitude measured in the baseline wave 1 survey (as a fully saturated set of dummy variables for the possible answers). Controls for the attitude regressions are listed in Appendix Table A1; controls for the occupation regression are listed in Appendix Table A2. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Table V. Military preparedness/satisfaction attitudes and promotions at the end of boot camp

	Attitudes after Boot Camp		**Military Evaluation**
	"Feel qualified for further service" *Strongly agree=1*	"Satisfaction with boot camp" *Good=1*	Promoted to Vice Corporal
	(1)	(2)	(3)
Female on team	-.0512	.0007	.0494
	(.0546)	(.0310)	(.0409)
Control variables	Yes	Yes	Yes
R-squared	.254	.078	.084
Control group mean	.521	.888	.161
N	528	534	657

Notes: Regressions mirror those in Table IV, except that the outcome in column 2 is not measured in wave 1, and so does not control for baseline attitudes. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Table VI. Female leadership attitudes

	Attitudes at the end of boot camp / end of service					
	"Which sex makes the best troop leaders?"		"Which sex makes the best leaders at the highest level?"		"Which sex makes the best leaders for foreign operations?"	
	Equally good or Females=1					
	wave 2	*wave 3*	*wave 2*	*wave 3*	*wave 2*	*wave 3*
	(1)	(2)	(3)	(4)	(5)	(6)
Female on team	-.0097	-.0450	-.0061	-.0036	-.0115	-.0690
	(.0493)	(.0582)	(.0479)	(.0610)	(.0434)	(.0590)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	.281	.255	.259	.250	.333	.280
Control group mean	.477	.402	.649	.577	.462	.395
N	528	369	524	369	519	370

*Notes: Regressions mirror those in Table IV with controls. Standard errors clustered by boot camp squad. **p-value<.05, *p-value<.10*

Table VII. Gender attitudes and education, occupation and workplace choices in the longer run

	Attitudes 6 Months into Service				**Civilian Education, Occupation and Workplace**		
	"Same gender teams perform better" *Disagree=1*		"I am feminine" *Statement does not fit me at all=0*		Fraction women in chosen:		
	wave 3	*wave 3-wave 2*	*wave 3*	*wave 3-wave 2*	education	occupation	workplace
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female on team	-.0120	-.1939**	-.0053	-.1895**	-.0028	.0250	.0255
	(.0578)	(.0940)	(.0670)	(.0608)	(.0226)	(.0231)	(.0226)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	.148	.112	.256	.142	.152	.161	.124
Control group mean	.391	-.093	.670	.091	.515	.450	.388
N	370	292	369	302	510	654	654

Notes: Sample includes male recruits in battalions which participated in the experiment. The attitude variables in wave 2 were collected at the end of boot camp, while wave 3 attitudes were collected approximately six months after boot camp. The construction of the civilian education, occupation and workplace variables are explained in the text. The data used in columns 1-4 and columns 5-7 come from two different datasets which cannot be merged due to confidentiality laws in Norway. All regressions include indicators for troop. The attitude regressions additionally include the corresponding lagged attitude measured in the baseline survey (as a fully saturated set of dummy variables for the possible answers). Controls for columns 1-4 are listed in Appendix Table A1; controls for the columns 5-7 are listed in Appendix Table A2. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Table VIII. Preparedness/satisfaction attitudes and military evaluations at the end of service

	Attitudes at the end of service		**Performance evaluations**				
	"Feel qualified for further service"	"Satisfaction with service"	Mean (4)-(7)	Conduct	Cooperation	Independence	Overall
	Strongly agree=1	*Good=1*		*Exceeds expectations or Excellent=1*			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female on team	-.0360	.0313	.0090	-.0185	.0090	-.0196	-.0011
	(.0542)	(.0488)	(.0434)	(.0470)	(.0453)	(.0434)	(.0474)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	.208	.129	.117	.090	.093	.118	.143
Control group mean	.635	.785	.472	.502	.468	.484	.407
N	369	368	657	657	657	657	657

Notes: Regressions mirror those in Table VII with controls. Outcomes in columns 3-7 are only measured at the end of service. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Online Appendix

Does Integration Change Gender Attitudes?
The Effect of Randomly Assigning Women to Traditionally Male Teams

Gordon B. Dahl, Andreas Kotsadam and Dan-Olof Rooth

Appendix A: Details on Robustness Checks

This appendix provides details on the robustness checks reported in Appendix Table A6. For ease of comparison, the first specification reports our baseline estimates from Table 4.

Panel B uses lasso regression, following the method outlined in Belloni, Chernozhukov, and Hansen (2014), to chose which pre-determined variables to include. We allow lasso to choose from the baseline attitudes, the full list of demographic controls and their interactions with each other. This yields remarkably similar estimates compared to baseline, but with somewhat smaller standard errors. In panel C, we omit the baseline attitudes and other controls, all of which were measured prior to randomization. The estimate becomes insignificant for the household work question, but remains significant for the other two outcomes. We favor controlling for baseline attitudes, both because it is standard practice in experimental designs like this and also because it reduces the standard errors somewhat. We also favor keeping the set of control variables constant across outcomes and specifications, rather than having lasso choose a different set of controls for each regression.

Panel D estimates a probit instead of a linear model. The marginal effects are similar. In panel E, we instead estimate the model in first differences.[1] The estimates from this first difference model are positive and significant, and somewhat larger for the first two outcomes. In panel F, we estimate the baseline model, but instead of using binary outcomes, we use the full scale of responses which have been converted to a numerical scale as needed. These coefficients have a different interpretation, but reveal a similar pattern. In panel G, we likewise alter the first difference model to use the full linear scale of possible responses for both wave 2 and wave 1 attitudes instead of dichotomous variables. The estimates are all statistically significant.[2]

The final set of robustness checks applies to both the attitude questions and the occupational choice question. It would be interesting to explore the differential impacts of having 1 versus 2 or more females in the room (i.e., squad). However, the experiment was not set up to do this, as the military had a preference for 2 females in a room. There are only 11 rooms with 1 women in them, and even fewer with 3 or more women. With this caveat in mind, in specification H, we use two treatment variables: one female in the room and 2 or

[1] This changes the baseline specification in two ways. First, it dichotomizes lagged attitudes, instead of allowing for a set of dummy variables for the possible categorical responses in wave 1 as control variables. Second, it constrains the coefficient on lagged attitudes to be 1.

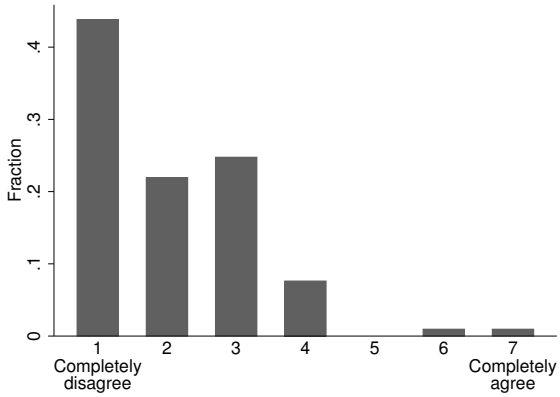
[2] We also explored alternative codings for the binary outcomes used in the baseline and first difference specifications, recoding the neutral category as a 1 instead of a zero for the first two attitude questions (for the gender identity question, there is no meaningful recode, as most individuals give just one of two answers). The magnitudes of the estimates fall and are no longer statistically significant. This suggests that at least some of the attitude changes are for individuals moving from a neutral attitude to a pro-gender equality attitude.

more females in a room. While somewhat imprecise, the interacted estimates are similar and not statistically different from each other for any of the outcomes. Specification I includes all room sizes, rather than our baseline restriction of rooms with 5-7 recruits in them. This makes little difference. Specification J restricts the sample to rooms with exactly 6 recruits and either 0 females or 2 females in the room. While this reduces the number of observations in the baseline sample by almost a quarter, the estimates are similar to baseline and remain statistically significant.

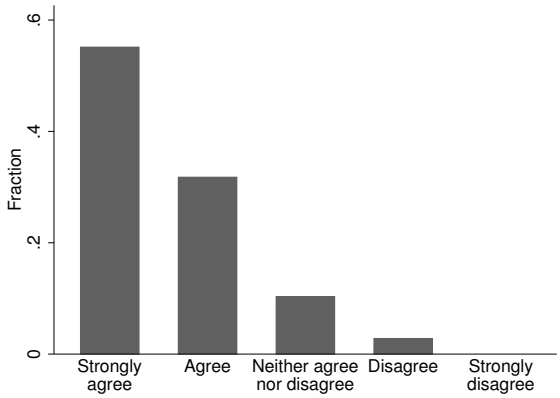
References

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

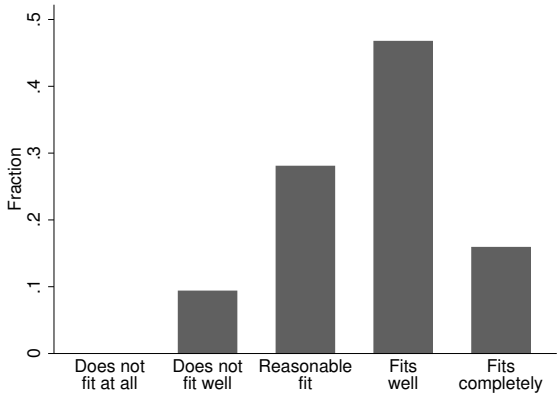
Appendix Figure A1. Distribution of gender attitudes for female recruits at baseline



A. Mixed-gender productivity: "A team performs better when it is made up of the same gender"



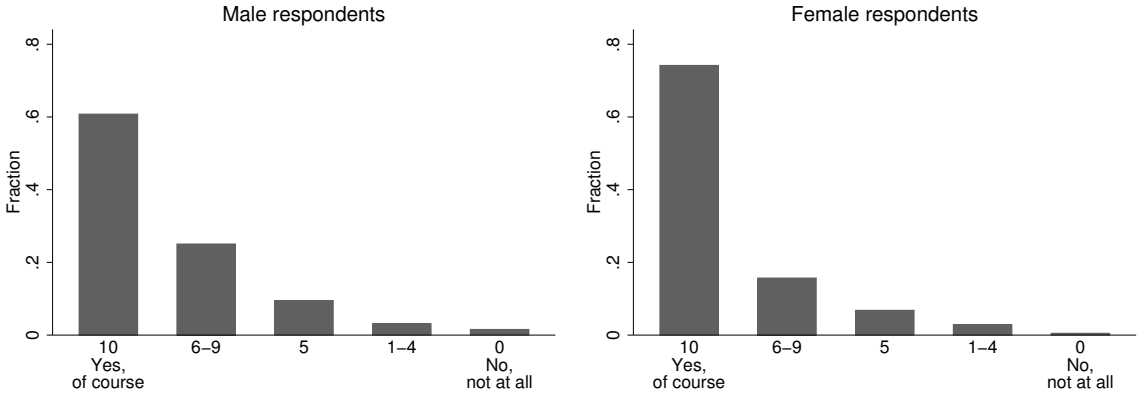
B. Gender roles: "It is important that men and women share household work equally"



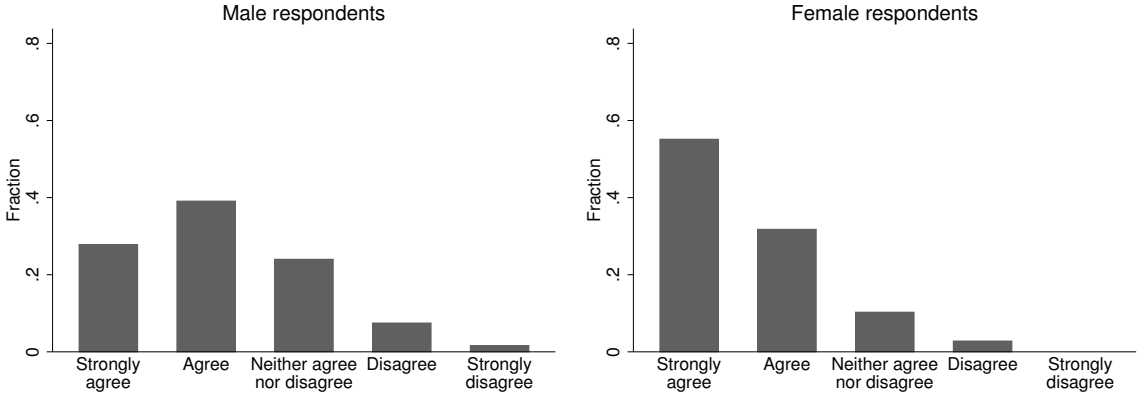
C. Gender identity: "How well does the following statement describe you: I am feminine"

Notes: Responses to baseline survey conducted prior to boot camp, excluding missings. There are 105, 107 and 107 observations in panels A, B and C, respectively.

Appendix Figure A2. Comparison of attitudes in the general population versus in the military on the importance of sharing household work



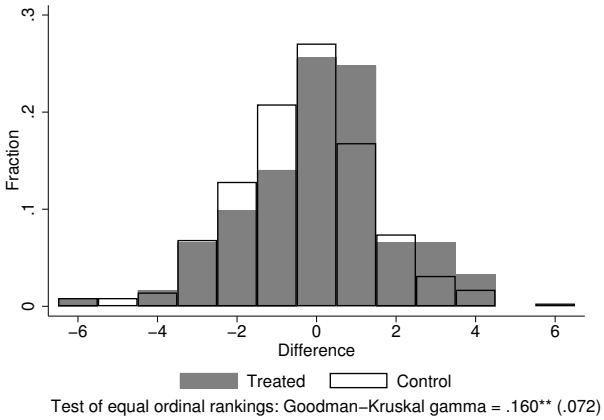
A. Gallup question on gender roles: "It is important that women and men share responsibility for the household"



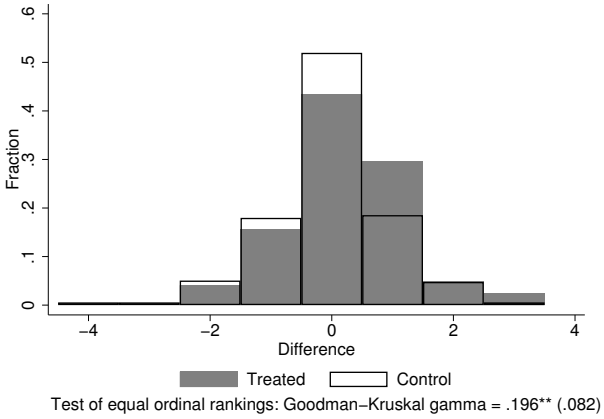
B. Military question on gender roles: "It is important that men and women share household work equally"

Notes: Panel A presents aggregated results from nationwide surveys conducted by Jakobsson and Kotsadam in conjunction with Gallup in 2008, 2009 and 2010. Panel B presents results from our sample of all military recruits, based on their attitudes at the start of boot camp in 2014. There are 2,889 male and 3,211 female respondents in the Gallup data and 1,439 male and 234 female respondents in our military data.

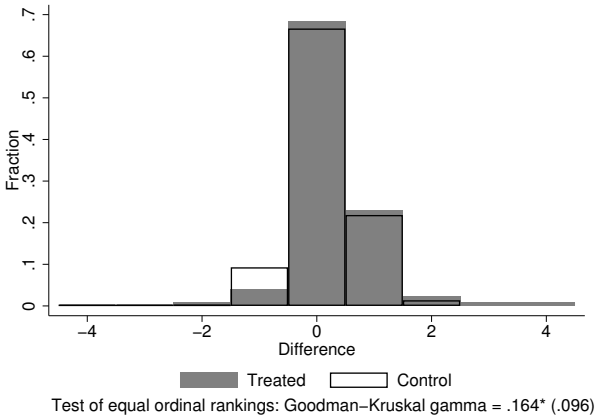
Appendix Figure A3. Differences the full scale of responses between survey wave 2 and baseline attitudes.



A. "Same gender teams perform better"



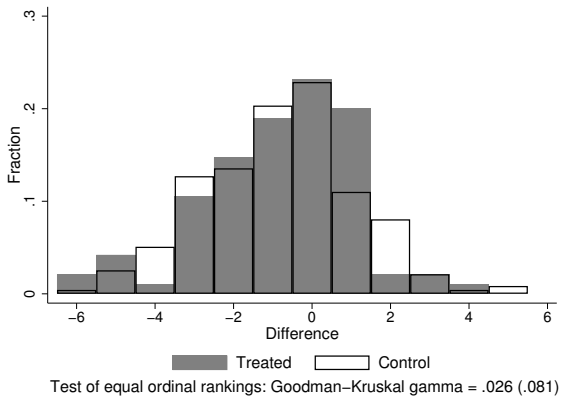
B. "Important to share household work equally"



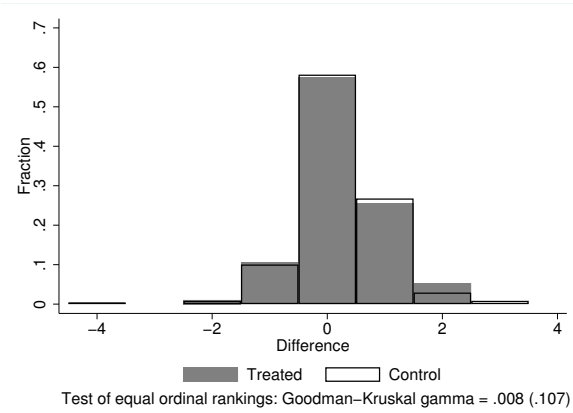
C. "I am feminine"

Notes: Baseline survey was conducted prior to boot camp, and survey wave 2 at the end of boot camp. There are 472, 478 and 492 observations in panels A, B and C, respectively. See footnote 20 for details on the test using the Goodman-Kruskal gamma.
***p-value<.05, *p-value<.10*

Appendix Figure A4. Differences the full scale of responses between survey wave 3 and baseline attitudes.



A. "Same gender teams perform better"



B. "I am feminine"

Notes: Baseline survey was conducted prior to boot camp, and survey wave 3 six months into military service. The question on sharing household work was not asked in wave 3. There are 331 and 333 observations in panels A and B, respectively. See footnote 20 for details on the test using the Goodman-Kruskal gamma.
***p-value<.05, *p-value<.10*

Appendix Table A1. Summary statistics for men and women at baseline

	Full sample		Experimental sample	
	Men	Women	Men	Women
	(1)	(2)	(3)	(4)
Opposite sex friends	.321	.408	.330	.370
Share missing	*.158*	*.111*	*.128*	*.101*
Has a sister	.505	.531	.521	.571
Share missing	*.299*	*.252*	*.279*	*.252*
Has a brother	.533	.573	.561	.538
Share missing	*.296*	*.263*	*.270*	*.311*
High muscle strength	.496	.485	.540	.496
Share missing	*.042*	*.080*	*.045*	*.084*
High GAI test score	.484	.290	.510	.252
Share missing	*.010*	*.031*	*.012*	*.042*
Mother higher education	.450	.447	.515	.412
Share missing	*.159*	*.115*	*.129*	*.101*
Father higher education	.527	.546	.575	.521
Share missing	*.159*	*.115*	*.129*	*.101*
Mother works	.767	.813	.796	.815
Share missing	*.157*	*.111*	*.129*	*.101*
Parents divorced or separated	.276	.286	.273	.286
Share missing	*.160*	*.118*	*.129*	*.109*
N	1,697	262	781	119

Notes: The samples in columns 1 and 2 include recruits in battalions which both did an did not participate in the experiment. The samples in columns 3 and 4 only include recruits in battalions which participated in the experiment. Opposite sex friends is a dummy for 40% or more of one's friends being of the opposite sex. High muscle strength is a dummy for above average or far above average muscle strength. GAI stands for general ability index, and is based on verbal comprehension and perceptual reasoning tests; high GAI test score is a dummy for scoring at or above the 6th stanine.

Appendix Table A2. Summary statistics for longer-run control variables

	Mean	SD	N
Share of women in high school education track	0.50	0.16	656
Final grades in junior high school	4.34	0.61	649
Immigrant	0.15	0.36	657
Has a sister	0.66	0.48	657
Has a brother	0.71	0.46	657
Mother's share of women in their occupation	0.67	0.21	585
Father's share of women in their occupation	0.32	0.24	569
Mother has higher education	0.55	0.50	654
Father has higher education	0.47	0.50	649
Mother's share of women in field of study	0.61	0.25	358
Father's share of women in field of study	0.35	0.24	303

Notes: These are the control variables contained in the merged Military and Statistics Norway register datasets. The pre-determined share variables in this table are defined similarly to the share variables appearing in Table VII. Final grades are the average of all grades as of the end of 9th grade. Higher education is defined as any type of post-secondary schooling.

Appendix Table A3. Placebo tests: Are prior attitudes affected by future exposure?

	"Same gender teams perform better" *Disagree=1* (1)	**"Important to share HH work equally"** *Agree=1* (2)	**"I am feminine"** *Statement does not fit me at all=0* (3)
Female on team	-.0621	-.0056	-.0497
	(.0525)	(.0460)	(.0532)
Control variables	Yes	Yes	Yes
R-squared	.064	.087	.058
Dependent mean	.609	.669	.558
Control group mean	.625	.677	.571
N	678	683	686

Notes: Regressions mirror those in Table IV with controls, but use attitudes measured in the baseline survey (before boot camp began). Since attitudes are not measured prior to the baseline, these regressions do not control for lagged attitudes. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Appendix Table A4. Testing for differential attrition by treatment status

	"Same gender teams perform better"		**"Important to share HH work equally"**	**"I am feminine"**	
	Missing outcome in survey wave = 1				
	wave 2	*wave 3*	*wave 2*	*wave 2*	*wave 3*
	(1)	(2)	(3)	(4)	(5)
Female on team	-.0420	-.0366	-.0360	-.0518	-.0284
	(.0405)	(.0449)	(.0412)	(.0417)	(.0437)
Control variables	Yes	Yes	Yes	Yes	Yes
R-squared	.143	.109	.138	.134	.114
Control group mean	.312	.535	.307	.294	.535
N	781	781	781	781	781

*Notes: Regressions mirror those in Table IV with controls. Wave 2 timing corresponds to the end of boot camp while wave 3 corresponds to 6 months into military service. Standard errors clustered by boot camp squad. **p-value<.05, *p-value<.10*

Appendix Table A5. Testing for differential attrition by baseline answers

	"Same gender teams perform better"		**"Important to share HH work equally"**	**"I am feminine"**	
	Missing outcome in survey wave = 1				
	wave 2	*wave 3*	*wave 2*	*wave 2*	*wave 3*
	(1)	(2)	(3)	(4)	(5)
Baseline answer = 1	-.0347	.0147	.0024	.0001	-.0035
	(.0382)	(.0407)	(.0393)	(.0369)	(.0387)
Control variables	Yes	Yes	Yes	Yes	Yes
R-squared	.139	.099	.136	.126	.098
Control group mean	.292	.524	.289	.274	.526
N	678	678	683	686	686

*Notes: Regressions mirror those in Table IV with controls. Wave 2 timing corresponds to the end of boot camp while wave 3 corresponds to 6 months into military service. Standard errors clustered by boot camp squad. **p-value<.05, *p-value<.10*

Appendix Table A6. Additional military attitudes

	"Want to do military service"	**"Satisfaction with room"**		**"Plan to continue in the military"**	
	Strongly agree=1	*Good=1*		*Yes=1*	
	wave 2 only	*wave 2*	*wave 3*	*wave 2*	*wave 3*
	(1)	(2)	(3)	(4)	(5)
Female on team	.0028	.0004	-.0206	.0213	-.0154
	(.0425)	(.0439)	(.0588)	(.0356)	(.0460)
Control variables	Yes	Yes	Yes	Yes	Yes
R-squared	.273	.070	.106	.515	.309
Control group mean	.585	.847	.782	.203	.215
N	521	533	370	534	370

Notes: Regressions mirror those in Table IV, except that the outcome in columns 2 and 3 is not measured in wave 1, and so does not control for baseline attitudes. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Appendix Table A7. Robustness checks

	Attitudes after Boot Camp			Military Occupation
	"Same gender teams perform better" *Disagree=1*	"Important to share HH work equally" *Agree=1*	"I am feminine" *Statement does not fit me at all=0*	Frac. women in chosen military occupation
	(1)	(2)	(3)	(4)
A. Baseline model	.1333**	.0821**	.1382**	.0221**
	(.0525)	(.0407)	(.0507)	(.0073)
N	522	526	538	657
B. Lasso choosing among	.1380**	.0797**	.1359**	.0201**
baseline attitudes and controls	(.0502)	(.0389)	(.0495)	(.0071)
N	522	526	538	657
C. Excluding baseline	.1353**	.0474	.1070*	–
attitudes and controls	(.0523)	(.0459)	(.0628)	
N	522	526	538	
D. Probit, marginal effect	.1581**	.0968*	.1638**	
	(.0600)	(.0474)	(.0543)	
N	522	520	538	
E. First difference model	.1953**	.1231**	.1209**	–
	(.0715)	(.0491)	(.0527)	
N	472	478	492	
F. Baseline model,	.3949**	.1357	.1971**	–
using full scale	(.1983)	(.0836)	(.0716)	
N	522	526	538	
G. First difference model,	.5012**	.2090**	.1594*	–
using full scale	(.2385)	(.0882)	(.0862)	
N	472	478	492	
H. One female	.1340	.1296	.1621**	.0240*
	(.0914)	(.1262)	(.0799)	(.0134)
Two or more females	.1390**	.0606	.1155**	.0210**
	(.0561)	(.0426)	(.0535)	(.0080)
N	522	526	538	657
I. All room sizes	.1144**	.0706*	.1304**	.0239**
	(.0515)	.0405	(.0485)	(.0073)
N	539	543	556	685
J. Room size = 6 and	.1530**	.1029*	.1619**	.0186**
females=0 or 2	(.0776)	(.0533)	(.0627)	(.0087)
N	396	399	406	476

Notes: Regressions mirror those in Table IV with controls. The military occupation regressions do not include baseline attitudes due to confidentiality issues with merging datasets. See text for details on different specifications. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Appendix Table A8. Heterogeneity by characteristics of males

FOT = female on team	**"Same gender teams perform better"** *Disagree=1* (1)	**"Important to share HH work equally"** *Agree=1* (2)	**"I am feminine"** *Statement does not fit me at all=0* (3)	**Frac. women in chosen military occupation** (4)
A. Female friends				
Female on team	.0708	.0724	.1363**	–
	(.0671)	(.0495)	(.0644)	
FOT×Female friends	.1284	.0062	-.0991	
	(.1055)	(.0907)	(.0828)	
B. Has a sister				
Female on team	.0129	.0611	.1185	.0173
	(.1026)	(.0851)	(.0925)	(0.0158)
FOT×Has a sister	.1499	.0163	-.0138	.0071
	(.1294)	(.1019)	(.0995)	(.0187)
C. Mother higher educ.				
Female on team	.2014**	.0567	.1333*	.0230**
	(.0869)	(.0739)	(.0683)	(.0106)
FOT×Mother higher educ.	-.1321	.0296	-.0553	-.0011
	(.1077)	(.0941)	(.0788)	(0.0147)
D. Muscle strength				
Female on team	.1611**	.0997*	.1637**	–
	(.0695)	(.0531)	(.0643)	
FOT×High strength	-.0263	-.0268	-.0580	
	(.0996)	(.0969)	(.0936)	
E. GAI test score				
Female on team	.1811**	.0801	.1756**	–
	(.0644)	(.0663)	(.0682)	
FOT×High test score	-.0751	.0117	-.0699	
	(.0897)	(.0898)	(.0792)	
F. Baseline gender stereotype				
Female on team	.0939	.0080	.1137**	–
	(.0693)	(.0452)	(.0491)	
FOT×Neg. stereotype	.0641	.1733	-.0305	
	(.1163)	(.0964)	(.0811)	
N	522	526	538	657

Notes: Regressions mirror those in Table IV, with the addition of interacting the treatment variable (female on team) with various dummy variables. In Panel F, negative baseline stereotype is defined using the relevant lagged gender attitude variable (i.e., the baseline survey responses). Standard errors clustered by boot camp squad.

***p-value<.05, *p-value<.10*

Appendix Table A9. Gender attitudes and education, occupation and workplace choices at the end of service, without control variables

	Attitudes 6 Months into Service				**Civilian Education, Occupation and Workplace**		
	"Same gender teams perform better" *Disagree=1*		"I am feminine" *Statement does not fit me at all=0*		Fraction women in chosen:		
	wave 3	*wave 3-wave 2*	*wave 3*	*wave 3-wave 2*	education	occupation	workplace
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female on team	-.0175	-.2108**	.0021	-.1659**	-.0067	.0065	-.0008
	(.0584)	(.0841)	(.0635)	(.0600)	(.0226)	(.0251)	(.0240)
Control variables	No	No	No	No	No	No	No
R-squared	.120	.082	.221	.119	.052	.038	.041
Control group mean	.391	-.093	.670	.091	.515	.450	.388
N	370	292	369	302	510	654	654

*Notes: Regressions parallel those in Table VII, except that they do not include control variables. Standard errors clustered by boot camp squad. **p-value<.05, *p-value<.10*

Appendix Table A10. Lasso regressions for gender attitudes and education, occupation and workplace choices in the longer run

	Attitudes 6 Months into Service				**Civilian Education, Occupation and Workplace**		
	"Same gender teams perform better" *Disagree=1*		"I am feminine" *Statement does not fit me at all=0*		Fraction women in chosen:		
	wave 3	*wave 3-wave 2*	*wave 3*	*wave 3-wave 2*	education	occupation	workplace
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female on team	-.0299	-.2091**	-.0015	-.1560**	.0039	.0194	.0134
	(.0554)	(.0792)	(.0612)	(.0535)	(.0214)	(.0231)	(.0232)
Control variables	lasso	lasso	lasso	lasso	lasso	lasso	lasso
N	370	292	369	302	510	654	654

Notes: See notes to Table VII. All regressions include indicators for troop, since that is the level of randomization. The other controls are chosen optimally from the total list of controls, including the baseline attitude variables for columns 1-4, and their interaction with each other using lasso, following the method outlined in Belloni et al. (2014). Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*

Appendix Table A11. Commendations received during service

	ln(total awards)	**Sharp-shooting bronze**	**Sharp-shooting silver**	**Ski march bronze**	**Field sport bronze**	**Loaded march bronze**	**Infantry run bronze**
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female on team	-.0929**	.0128	-.0491	-.0081	-.0060	-.1164**	-.0132
	(.0413)	(.0526)	(.0398)	(.0119)	(.0151)	(.0422)	(.0379)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-squared	.305	.227	.173	.085	.342	.271	.300
Control group mean	.894	.456	.331	.030	.028	.448	.284
N	657	657	657	657	657	657	657

Notes: Regressions mirror those in Table VII. Standard errors clustered by boot camp squad.
***p-value<.05, *p-value<.10*