We are (not) anonymous

Essays on anonymity, discrimination and online hate

Joakim Jansson

Academic dissertation for the Degree of Doctor of Philosophy in Economics at Stockholm University to be publicly defended on Friday 30 November 2018 at 09.00 in William-Olssonsalen, Geovetenskapens hus, Svante Arrhenius väg 14.

Abstract

Haters gonna hate? - Anonymity, misogyny and hate against foreigners in online discussions on political topics. A crucial aspect of freedom of expression is anonymity, but anonymity is a contentious matter. It enables individuals to discuss without fear of repercussions, but anonymity can also lead to hateful writings threatening other's freedom. In this paper, we predict hateful content as well as estimate the causal link between anonymity and hateful content in civic discussions online. First, we make use of a supervised machine-learning model to predict hate in general, hate against foreigners and hate against females and feminists on a dominating Swedish Internet discussion forum. Second, using a difference-in-difference model we show that an exogenous decrease in anonymity leads to less hateful content in general hate and hate against foreigners, but an increase in hate against females and feminists. The mechanisms behind the changes is a combination of a decrease in writing hateful, as well as a decrease in writing in general and a substitution of hate against one group to another.

Gender grading bias at Stockholm University: quasi-experimental evidence from an anonymous grading reform. In this paper, we first present novel evidence of grading bias against women at the university level. This is in contrast to previous results at the secondary education level. Contrary to the gender composition at lower levels of education in Sweden, the teachers and graders at the university level are predominantly male. Thus, an in-group bias mechanism could consistently explain the evidence from both the university and secondary education level. However, we find that in-group bias can only explain approximately 20 percent of the total grading bias effect at the university level.

Anticipation Effects of a Board Room Gender Quota Law: Evidence from a Credible Threat in Sweden. Board room quota laws have recently received an increasing amount of attention. However, laws are typically anticipated and firms can react before the effective date. This paper provides new results on female board participation and firm performance in Sweden due to a credible threat of a quota law enacted by the Swedish deputy prime minister. The threat caused a substantial and rapid increase in the share of female board members in firms listed on the Stockholm stock exchange. This increase was accompanied by an increase in different measures of firm performance in the same years, which were related to higher sales and lower labor costs. The results highlight that anticipatory effects of a law could be detrimental to the analysis.

Differences in prison sentencing between the genders and immigration background in Sweden: discrepancies and possible explanations. I use data on punished drunk drivers to document differences in sentencing for the same crime between immigrants and native born and males and females respectively. Differences in past criminal activity or other individual observables can not explain the difference in sentencing. Instead, the difference between immigrants and native born seem to be due to statistical discrimination, while differences in recidivism rates might explain the gender difference. However, the higher incarceration rate for immigrants does not reduce their future number of crimes.

Keywords: *Anonymity, Discrimination, Hate, Applied Econometrics, Gender Economics, Labor Economics, Political Economics, Behavioral Economics.*

Stockholm 2018
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-160745

ISBN 978-91-7797-508-3
ISBN 978-91-7797-509-0
ISSN 1404-3491

Department of Economics

Stockholm University, 106 91 Stockholm

WE ARE (NOT) ANONYMOUS

Joakim Jansson

Stockholm University

We are (not) anonymous

Essays on anonymity, discrimination and online hate

Joakim Jansson

ISBN print 978-91-7797-508-3
ISBN PDF 978-91-7797-509-0
ISSN 1404-3491

Printed in Sweden by Universitetsservice US-AB, Stockholm 2018

To my grandparents, for
allowing me to live

To my parents, for
teaching me how to live

And to Inky and Lova, for
bringing joy to life

Contents

Acknowledgments

There are several people without whom this dissertation never would have been completed, or in some cases even started. I will present the ones I consider the most important here in no particular order. First and foremost, I would like to thank my main supervisor Björn Tyrefors for his support, feedback and general spirit. I have always felt that if I was stuck or didn't know where to go next, I always had the opportunity to ask and get some direction or encouragement along the way, and for this I am truly grateful. Furthermore, I have also learned a great deal about how do conduct applied empirical work by watching and working with Björn. Perhaps this is what I should thank him for the most. I would also like to thank my second advisor Jonas Vlachos for continued great feedback on my different projects. As with Björn, I have always felt welcome to ask for direction, however I must admit at times I have been to stubborn to seek out this support. Among other collaborators and colleagues over the past six years Emma von Essen stands out as the one I am the most indebted to. Always with a critical eye to whatever result I presented, she has taught me to always make sure I understand the underlying literature and assumptions behind what we find. Over time she has also become more than one of my co-authors, giving me advice on several additional topics about the inner workings of academia. And in addition, she along with Lina Eklund was willing to spend a good amount of research funds on a project involving web-scraping and machine learning, about which I knew nothing when we started.

I would also like to thank the rest of the staff at the department of economics, however a few stand out that I want to mention them by name. First and foremost, David Strömberg has given me a great deal of comments and feedback on both projects and applications for which I am truly grateful. I would also like to thank Peter Fredriksson for supervising me during my master thesis and by that giving me both the courage and interest to apply to the PhD-program. Per Petterson-Lidbom, Peter Skogman Thoursie,

Mahmood Arai, Andreas Madestam, Anders Åkerman and Anna Tompsett for helpful comments. Among my fellow PhD colleagues, a few stand out a little bit extra, among them my former office-mates Sirus Dehdari and Carl-Johan Rosenvinge as well as Daniel Almén, Daniel Knutsson, Nanna Fukushima, Erik Lindqvist, Wei Sei, Tamara Sobolevskaya, Roza Khoban, Mathias Ivanowsky and Richard Foltyn. In addition I have received many great comments and had great discussions with several folks from other places than the department of economics at SU. Among the foremost of these is Lena Hensvik, who provided me with a lot of great feedback during my final seminar, which have greatly improved all of my papers. I am also grateful to comments from Fredik Heyman, Joacim Tåg, Mathew Gentzkow and David Neumark. I would also like to thank Karin Blomqvist and Peter Langenius for providing me with data for one of the papers, as well as when I worked with both of them along with Anders Fjällström as a TA at the introductory courses back when I did my Master.

At last but definitely not least, I need to thank the people closest to me. In particular, I am truly grateful towards my family, extended family and friends for putting up with my odd working habits, other quirks and annoyances such as general lack of sense of time or other events taking place and overall lack of ability to take my mind off from work. Without all of you, I could never have done this, and I will forever be grateful for sharing my life with you. In particular, I would like to thank Inky and Lova, who have to not only put up with all of this to a greater degree than anyone else, but also provide compassion, support and general help with life. I am so happy to have you in my life, and I hope that one day I will be able to repay you in some way.

Joakim Jansson
Stockholm
October 2018

Introduction

Ever since my early teens, what I have considered unfair or unjust treatment has upset me. Although age has perhaps smoothed the pure anger I could once feel over such behavior, it is still a practice which I strongly dislike.

Even further back in time, in the days when I was just a kid, economics was hardly what I expected to do once I grew up. In fact at that point in time, I don't think I even could grasp the concept of economics. Instead, I devoted a lot of my time learning about dinosaurs, and perhaps at this time I rather looked destined to become a paleontologist. However, once I got into my mid teens my interest in societal development, politics and economics was growing by the day. As I grew older and started to study economics at the university, I was fascinated at first by the economic theories that postulated what could cause discrimination to occur in markets. Moving on through my higher education I eventually stumbled into applied econometrics and regressions. Again, I was amazed first by how regressions could filter out the effect of other characteristics out of a relationship between two variables, and later on the power of randomized controlled and natural experiments to provide evidence on how the world works compelled me. Perhaps it is this process that has led me to the topics and methods presented in this thesis.

This doctoral thesis consists of four independent essays in applied empirical microeconomics. Two of the chapters focus on anonymity as a policy tool and have been the main inspiration for the name of the thesis. Both also focus on discrimination in some sense, either through the direct effect on students grades or on peoples revealed attitudes toward foreigners and feminists. While discrimination in general can be seen as an equity problem, we as economists typically study it for efficiency reasons. The other two chapters are hence also related to discrimination, but with an increased focus on potential efficiency improvements. I will now briefly describe the main findings of the papers, before concluding.

Chapter 1: Haters gonna hate? – Anonymity, misogyny and hate against foreigners in online discussions on political topics. (Jointly written with Emma von Essen)

In this paper we study hate in discussions on political topics online under anonymity on a Swedish Internet forum called Flashback. First, we examine if it is possible to predict whether or not an entry in Swedish at Flashback as containing hateful content or not, based on the language used. This is followed by a more traditional difference-in-difference model where we investigate if a decreased perception of anonymity causes less hate in the discussions. We started by collecting entries on political topics regarding domestic politics, feminism and immigration from Flashback using a custom-built web-scraper in Python. We then draw a sub-sample of 100 randomly drawn discussion-threads from each of these areas and let a research assistant classify these entries as either hateful or not, and towards whom the hate was mainly directed. This gives us a classified sample of 4021 encoded entries, out of which we randomly draw 70 percent of the observations which we labeled the training data. Using this training data we then use machine learning algorithms to create prediction-models for hate in general and hate aimed at foreigners and misogyny. We then use the remaining 30 percent, called the test data, to test how well we can predict hate, hate against foreigners and misogyny. From this exercise we then find that the algorithms work well in predicting hate when it's aimed at a particular group, such as foreigners or feminists. Predicting hate in general is a lot harder.

We then use these prediction models in order to create dummy variables for if an entry contains hate, hate against foreigners or/and misogyny in all the data we scraped from Flashback. Using this data, we then use a difference-in-difference strategy to see to what degree the share of hate in the discussions is affected by the fact that the identity of the users was known by journalists. More specifically, the journalists had access to the identity of around 1/3 of all users registered at Flashback before March 2007. This became publicly known in September 2014. Thus, we label those registered before March 2007 the treatment group, and those registered after the control group. We then observe that the share of hateful entries and hateful entries against foreigners decrease for the users registered before March 2007 after September 2014, while it actually increases slightly for misogyny. These results seem to be driven in part by the fact that users that wrote a large share of hateful entries against foreigns decreased their activity in the post-period

and in part by them substituting hate against foreigners towards misogyny.

While previous research on hate online and anonymity has mainly been correlational (e.g. Moore et al. (2012); Suler (2004); Van Royen et al. (2017)) or studied hate in general (Cho et al., 2012), we contribute by using a natural experiment to study individual behavior and substitution as well as towards whom the hate is framed.

Chapter 2: Gender grading bias at Stockholm University: quasi-experimental evidence from an anonymous grading reform. (Jointly written with Björn Tyrefors)

This article is the second regarding anonymity, and thus one of the two from which the main inspiration for the name of the thesis is drawn. It is the first paper to look at gender bias in correction of exams at the university level. Using data from the entire Stockholm University between 2005 and 2014, we first employ a difference-in-difference-in-difference strategy using the fact that all written exams had to be anonymously graded from the start of the fall term 2009. As both all thesis' and oral- and laboratory exams where not affected by this reform, we can use these as a control group, and thus study the difference between the genders grades in the treated- and control-group before and after the reform. We find, contrary to previous evidence from lower educational levels[1], that women benefit from the anonymization of exams compared to men. A possible hypothesis raised by this reversed relationship is that it is caused by the male dominance at the university level, as compared to the female dominance among teachers at lower educational levels.

We then proceed to a sample consisting of the exams from the introductory course in macroeconomics at Stockholm University. Using this sample, we first replicate the main finding of females benefiting from the reform, though we this time use multiple choice questions as the control group, since these can't be graded with a bias. However, in this sample we also know the gender of the teachers assistant that corrected the question, thus allowing us to directly test the hypothesis of whether women correct women and men correct men more favorably. Furthermore, the allocation of these correctors are done through ballot, which should ensure randomization. We thus have an unintentional randomized experiment at hand. The finding from this experiment is that the randomization seemed to have worked, and that women

[1]See for example Lavy (2008), Hinnerich et al. (2011) and Kiss (2013).

indeed favor women in their grading and men favor men when exams are not anonymous. Once anonymity is introduced, this relationship goes away. This effect is however only large enough to explain about 20 percent at most of the total bias against women when the exams are not anonymous. We can thus conclude that although the graders gender matters for the bias in grading, other potential contributing factors, such as for instance organizational culture, seem in total more important.

Chapter 3: Anticipation Effects of a Board Room Gender Quota Law: Evidence from a Credible Threat in Sweden. (Joint with Björn Tyrefors)

The effect of board gender quotas have received an increasing amount of attention both from academia[2] and policymakers in countries such as Spain, Belgium, France, Germany, Iceland, Italy and the Netherlands (Eckbo et al., 2016). However, so far all papers have focused their attention on the Norwegian case. We instead use a credible threat in Sweden in the form of a law proposal by the Swedish deputy prime-minister Margareta Winberg in late 2002 and supported by prime minister Göran Persson. However, unlike in Norway a formal law mandating a certain share of women on the listed firms boards were never signed, primarily due to the fact that there was a change in political power in Sweden from a center-left government to a center-right. Yet we observe a sharp and rapid increase in the share of women on listed firms boards compared to non-listed firms.

We thus use the listed firms as a treatment group and the non-listed as controls in a reduced form difference-in-difference on firm performance. The underlying assumption of similar trends in absence of treatment seem reasonable both for the first stage (share of women on boards) and the reduced form (firm performance) based on pre-event graphical displays as suggested by Angrist and Pischke (2008). In contrast to previous research in the area, we find improved firm performance overall as the share of women on the listed firms boards increase, with a higher return-on-assets, lower labor costs and increased sales. This conclusion hold up even as we use a synthetic control-group approach (Abadie et al., 2010) and if we include linear separate time trends for treated and control groups or control for industry specific trends.

Chapter 4: Differences in prison sentencing between the genders and im-

[2]See for instance Ahern and Dittmar (2012), Bertrand et al. (2018), Matsa and Miller (2013) and Eckbo et al. (2016).

migration background in Sweden: discrepancies and possible explanations.

Previous research on discrimination in courts have typically focused on either judge/jury characteristics or the match between judge/jury and defendant characteristics and the outcome of the trial (Anwar et al., 2012, 2015, 2014; Abrams et al., 2012). In this paper I take a different approach and instead focus on how males and immigrants are sentenced compared to females and native born for the same committed crime. More specifically, I look at the difference in probability of being sentenced to prison between these groups for the same blood-alcohol content (BAC) when caught in a DUI (driving under influence).

I find that men are 5 and immigrants 10 percentage points relatively more likely to receive a prison sentence for the same committed crime compared to females and native born respectively. This difference can't be explained by differences in willingness to get treatment for substance abuse or observable underlying characteristics such as income or past criminal activity. However, there is a difference in crime recidivism rate between males and females that could explain the observed discrepancy. For immigrants, once one controls for underlying characteristics, there is no difference in crime recidivism compared to native born. In addition, reduced form estimates from a regression discontinuity strategy of immigrant and native born treatment effects of prison sentence on future crime does not suggest that the higher incarceration rate of immigrants reduces their future amount of criminal activity. Overall, the findings suggest that in the Swedish criminal justice system, there is room for improvement both in terms of efficiency and equity.

In sum, this dissertation shows the power anonymity can have as a policy tool. As economists we are used to think of trading off efficiency in order to obtain equity. But as the papers in this dissertation show, it is sometimes possible to obtain both, in particular in settings regarding discrimination. Furthermore, concealing the identity can in some cases improve both efficiency and equity. However, it needs to be used with caution and properly evaluated using research methods, as our results concerning misogyny online and anonymity show.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

Abrams, D. S., M. Bertrand, and S. Mullainathan (2012). Do judges vary in their treatment of race? *The Journal of Legal Studies 41*(2), 347–383.

Ahern, K. R. and A. K. Dittmar (2012). The changing of the boards: The impact on firm valuation of mandated female board representation. *The Quarterly Journal of Economics 127*(1), 137–197.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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.

Anwar, S., P. Bayer, and R. Hjalmarsson (2014). The role of age in jury selection and trial outcomes. *The Journal of Law and Economics 57*(4), 1001–1030.

Anwar, S., P. Bayer, and R. Hjalmarsson (2015). Politics in the courtroom: Political ideology and jury decision making. Technical report, National Bureau of Economic Research.

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (2018). Breaking the glass ceiling? the effect of board quotas on female labor market outcomes in norway. *The Review of Economic Studies*.

Cho, D., S. Kim, and A. Acquisti (2012). Empirical analysis of online anonymity and user behaviors: the impact of real name policy. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 3041–3050. IEEE.

Eckbo, B. E., K. Nygaard, and K. S. Thorburn (2016). Does gender-balancing the board reduce firm value?

Hinnerich, B. T., E. Höglin, and M. Johannesson (2011). Are boys discriminated in swedish high schools? *Economics of Education review 30*(4), 682–690.

Kiss, D. (2013). Are immigrants and girls graded worse? results of a matching approach. *Education Economics 21*(5), 447–463.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of public Economics 92*(10), 2083–2105.

Matsa, D. A. and A. R. Miller (2013). A female style in corporate leadership? evidence from quotas. *American Economic Journal: Applied Economics 5*(3), 136–69.

Moore, M. J., T. Nakano, A. Enomoto, and T. Suda (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior 28*(3), 861–867.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior 7*(3), 321–326.

Van Royen, K., K. Poels, H. Vandebosch, and P. Adam (2017). "thinking before posting?" reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior 66*, 345–352.

Haters gonna hate?
Anonymity, misogyny and hate against foreigners in online discussions on political topics*

Emma von Essen†, Joakim Jansson‡

Abstract

In this paper, we predict hateful content as well as estimate the causal link between anonymity and hateful content in civic discussions online. First, we make use of a supervised machine-learning model to predict hate in general, hate against foreigners and hate against females and feminists on a dominating Swedish Internet discussion forum. Second, using a difference-in-difference model, we show that an exogenous decrease in anonymity leads to a less hateful content in general and less hate against foreigners, but an increase in the hate against females and feminists. The mechanisms behind the changes constitute a combination of a decrease in writing in a hateful way, as well as a decrease in writing in general and a substitution of hate against one group with another.

Keywords: online hate, anonymity, discussion forum, machine learning, big data
JEL: C55, D00, D80, D90

*We want to thank Sara Ekman for excellent RA work. Furthermore, we thank Lina Eklund, Björn Tyrefors Hinnerich, Jonas Vlachos, David Strömberg, Lena Hensvik, Matthew Gentzkow, seminar participants at the seminars at the Department of Economics, the Swedish Institute for Social Research (SOFI) at Stockholm University and participants at the Association of Internet Reseaercher 2018 for providing fruitful comments. Financial support from the Swedish Research Council is also gratefully acknowledged.

†Institute for Social Research, Stockholm University, Stockholm Internet Research Group, Stockholm University, Department of Economics and Business, Aarhus University

‡Department of Economics, Stockholm University, Research Institute of Industrial Economics, Stockholm Internet Research Group, Stockholm University

1 Introduction

Freedom of speech is and has been a cornerstone in democratic societies, where speakers' corners served as symbols of spaces where members of society could express their opinions without fear of retaliation. Today, these spaces are more prominent online, such as Internet discussion forums. A crucial aspect of freedom of expression is anonymity (hiding identity-bearing information). When voting in free elections, anonymity is considered a fundamental right. The question of anonymity and free speech is not clear-cut, however. When partaking in a demonstration, some countries allow masks whereas other countries have anti-masking laws. Anonymity being a double-edged sword; on the one hand, more anonymity enables individuals to speak without any fear of social and governmental repercussions. But, on the other hand, due to lack of accountability, anonymity can lead individuals to involve in anti-social, untruthful, criminal and violent behaviors towards others. We investigate this in online discussions on political topics.

Today the Internet presents an important marketplace for information from a variety of sources, and access is largely free or less expensive as compared to traditional media. Individuals seek information, and debates are held on social media platforms, such as Facebook and Twitter.[1] A recent strand of literature in economics explores how social media and politics are interrelated. Allcott and Gentzkow (2017) highlight the importance of social media as a socurce of political information. Qin et al. (2017) show that social media discussions can predict protests and corruption charges. As online public discussions are becoming part of our everyday interactions, cyberhate - harassments, and threats - has turned into a growing concern (Cheng et al., 2017).[2] Communication where users are anonymous have been found to correlate positively with cyberbullying, cyberhate and aggressive speech, e.g. Moore et al. (2012); Suler (2004) and Van Royen et al.

[1] A recent survey presented in Swedish public radio indicates that traditional media focus on the game between politicians, whereas discussions in social media evolve around substantive political questions, http://sverigesradio.se/sverigesradio.se/sida/artikel.aspx?programid=83&artikel=6819036, Access November 23 2017 13:45.

[2] In a survey from 2014, 40 % of the U.S. population report having experienced online harassment (Duggan, 2014)

(2017). Not only does hate absorb time and cause serious harm to individuals, but factual content can also disappear in a sea of less constructive personal comments and opinions, leading to less informed decisions, for instance when voting in elections. Moreover, it can potentially silence specific targeted individuals or groups. For example, there is evidence of women and foreigners losing their jobs and shutting down their social media activity (Citron, 2014). Thus, cyberhate can limit the democratic participation in discussions of specific groups. The negative externalities of cyberhate have caused many countries to place cyberhate on the political agenda.[3] However, the anonymous discussions message boards are more than just cyberhate, they provide a space for political discussions and other informational content. In particular, discussions regarding sensitive topics might not be possible without anonymity, due to social stigma (Froomkin, 2017). Enforcing a decrease in anonymity to counter hate, by for example requiring users to disclose their names, might reduce the participation in the discussions (Cho et al., 2012). Understanding how anonymity affects cyberhate is vital to find efficient policies, such as regulation of anonymity and allocation of resources. This paper aims at investigating how anonymity affects hate in general and hate against females and foreigners in particular, in online political discussions. There is a lack of quantitative research joining anonymity and hate online. To our knowledge, the only study looking at this quantitatively is a conference paper by Cho et al. (2012) looking at changes in offensive language on two different forums caused by a real name policy in Korea. The key added value of our study is that our design allows us to control for self-selection and that we introduce analyses based on hate directed towards specific groups.

We study a large Swedish anonymous discussion forum named Flashback, similar to the U.S. based Reddit. Flashback is one of the most visited Internet pages in Sweden[4] with more than one million registered accounts. Anonymity is a requirement on Flashback, and their motto is: "Real freedom of expression". In

[3]The legal framework for hate speech and online harassment differs by country. In the U.S. hate speech is not a legal term, and freedom of speech is protected by the First Amendment, whereas in other countries such as Sweden hate speech is regulated by civil law or criminal law.

[4]At the beginning of 2017, Alexa ranks it as the 23rd most visited in Sweden, and 5214 in the world. Go to https://www.alexa.com/siteinfo/flashback.org to see the current ranking.

September 2014, the identity of approximately one-third of the accounts[5] registered before March 2007 was unexpectedly in the hands of journalists, and the journalists publicly exposed the legal identities of a handful of Flashback users. We consider users registered before March 2007 as the treatment group, since they ran a risk of having their identities exposed together with the content they had written. Users registered after this date will serve as the control group. We scraped text from the discussion threads and entries of three selected sub-forums containing political discussions: the domestic politics forum, the feminism forum and the immigration forum. A research assistant manually evaluated posts from a random subset of the scraped discussion threads, on whether each post contained hate or not as well as towards what group or individual the hate was directed. The groups were females and feminists, or foreigners, or others. We used classifications inspired by previous research on linguistic markers (Cohen et al., 2014).

The economic literature on transparency is diverse; theoretical and empirical research suggests that less privacy can give benefits or losses to welfare depending on the context (Acquisti et al., 2016).[6] If we think of cyberhate as pollution, the theoretical principal-agent model on public goods provision by Ali and Bénabou (2016) predicts a decrease in hateful content given a negative shock to transparency due to reputational concerns.[7] The empirical strategy we use combines a machine-learning approach with a difference-in-difference estimation, see (Mullainathan and Spiess, 2017) for an overview. Hansen et al. (2017) used a similar approach when studying the effects of transparancy on policy deliberations. They find both a positive and negative effect of transparency on monetary policy.

Using a Logistic Lasso (Least Absolute Shrinkage and Selection Operator), we

[5]According to the journalists that actually had access to the data, they had information on all the accounts. However when the owner of Flashback compared the list to registered users, he claimed that it was only one third of those registered before March 2007 as far as he could tell.

[6]Transparency can include elements of both anonymity and privacy. Privacy implies concealing and revealing an individual's personal information and actions, such as his or her medical records, whereas anonymity implies concealing the identity of the individual, such as the names on the medical records.

[7]In Appendix B we place our study in their theoretical framework to further understand this prediction.

can predict hateful content on the discussion forum. Furthermore, our empirical estimates from the difference-in-difference model show that less anonymity creates less hate in political discussions. Early users of the message board changed their share of hateful content more than did later registered users when there was a threat of being identity exposed. We also study hate against females and feminists as well as hate against foreigners as outcome variables. Hate against foreigners decreases as a result of decreased anonymity, whereas hate against females seems to increase. The mechanisms behind the decrease seem to be a combination of a decrease in writing hateful posts and a decrease in writing non-hateful posts. Interestingly, we find users to substitute hate against foreigners with hate against females and feminists.

The rest of the paper is structured as follows. Section 2 discusses related literature. Section 3 describes some background and the data collection, while section 4 depicts the prediction methodology and results. Section 5 concerns the empirical strategy and final data and section 6 discusses our main findings and potential mechanisms. In the final section, we conclude the paper.

2 Previous literature

This study speaks three strands of literature: the economic literature on transparency, economics of media and the literature in psychology and information science on the value of anonymity online.

The economic literature on transparency is sparse (Acquisti et al., 2016). Related to this paper is the economic literature of career concerns, where early papers state that more transparency and information about the agent improved the accountability and were never found to be detrimental to the principal (Holmström, 1979). However, later papers find that revealing more information about the agent can also be detrimental to the principal (Prat, 2005; Holmström, 1999). Another strand of studies looks at voting transparency and committee work. Concealing voting behavior by individual members of the committee involves both anonymity and privacy. The model by Nattrass (2007) shows that committees where votes are

transparent initiate more reforms. To our knowledge, there is no economic theory focusing on anonymity per se. However, a recent theory by Ali and Bénabou (2016) explicitly models transparency (privacy and anonymity) in a principal-agent model of public good provision. Here transparency can affect the aggregate provision of public good through agents' concern for reputation (social image).

Empirically there is a handful of economic articles looking at how anonymity can create a fair judgment of performance in labor and educational market contexts, such as anonymous procedures (Goldin and Rouse, 2000; Edin and Lagerström, 2006; Åslund and Skans, 2012; Bøg and Kranendonk, 2011; Hinnerich et al., 2015, 2011; Bengtsson et al., 2012; Jansson et al., 2018; Lavy, 2008). The use of anonymous application procedures seems to create a more fair judgment of performance. However, in an online auction setting, the anonymity strategy might backfire and produce less reputational feedback in the end (von Essen and Karlsson, 2013). In the current paper, we focus on how anonymity can change individuals' behavior and not how others view them. Closest to our study is a recent empirical article by Hansen et al. (2017) that investigates transparency – here a combination of privacy and anonymity - and monetary policy. At one point, the US Federal Reserve decided to make old as well as future taped records of the Federal Open Market Committee (FOMC) monetary policy discussions publicly available. The members of the FOMC were not aware that the tapes existed. Using a linguistic machine-learning model, Hansen et al. (2017) investigate how this change in transparency affected the deliberation of monetary policymakers. They combine the machine-learning approach with a difference-in-difference strategy and find both increased discipline in discussions, i.e. committee members refer to more facts and figures in the discussions, and a tendency to conform and agree with the chairman to a larger extent. However, the net effect suggests that transparency created a more informative monetary policy debate.

A recent strand of literature in economics explores the relationship between social media and politics. The paper by Qin et al. (2017) associate the amount of regime critical content in Chinese social media with offline outcomes. They study how social media discussions can predict both protests and corruption charges. All-

cott and Gentzkow (2017) look into the role of fake news in social media in the US presidential election 2016 and find among other things that fake news were heavily tilted in favor of the Republican presidential candidate. They also show that individual users look at fake news and to some extent recall fake news stories. Individual's interaction with fake news on Facebbok and Twitter seem to have increased during 2016 (Allcott et al., 2018).

The social value of anonymity online is a large research field in Human and Computer Interaction studies based on psychology and sociology. Anonymity online can have either positive or negative consequences. Related to the economic argument on social reputation, disinhibition or classical deindividuation, theory implies that lack of restraints and social control makes individuals disregard social conventions or have less self-awareness, which can lead to anti-normative behaviors (Suler, 2004; Postmes et al., 1998). Social Identity Theory of Deindividuation (SIDE), on the other hand, considers claims that anonymity accentuates the effect of the salient social identity and the reduction in constraint will lead to behavior they would not otherwise engage in (Postmes et al., 1998). Anonymity can thus make individuals more susceptible to group influence, creating in-groups along stereotypes (Reicher et al., 1995). According to this conceptual theory, anonymity on Flashback can create anonymous groups with norms where hateful content is a more or less integrated part of socializing and targeted towards specific groups. A recent paper confirms that previous exposure to hate can induce hateful writing in online groups (Cheng et al., 2017).

Reforms or changes in platform design and context can partly create contexts that motivate changes in behaviors (Kraut et al., 2012). For example, changes in transparency between end-users forced by platform design can nudge people into changing behaviors, such as sharing information (Chang et al., 2016) or socializing (Eklund and Johansson, 2013). The paper closest to ours in this respect is Cho et al. (2012). They evaluate the real name policy introduced on Korean discussion forums. The key difference between our study and theirs is that they compare different forums with possibly differing self-selection and that we introduce analyses based on hate directed towards specific groups and not just hateful content in

general.

3 Background, the event and data collection

According to Swedish law, hate speech is prohibited. The term is defined as publicly making statements that threaten or express disrespect for a group regarding its race, skin color, ethnic origin, faith or sexual orientation. However, it is not forbidden by law to write in a hateful way against a particular gender. During the 90's several members of the extreme right-wing movements were convicted of racial hate speech (Lööw and Nilsson, 2001).[8] Flashback, the discussion forum we study, started in the 90's as a small scale paper outlet promoting freedom of speech. In late 1990, Flashback went online and today it is one of the most visited Internet pages in Sweden and well known among users of other similar international forums and message boards, such as Reddit. According to a recent survey by Davidsson et al. (2018) 33% of the Swedish population state they use Flashback and the share is larger among men compared to women (40% vs 26 %). Flashback has more than one million registered accounts, and the site is (in)famous for its focus on the anonymity of the users as a way of promoting freedom of speech. According to Alexa, the average Flashback user spends approximately seven minutes per visit and on average seven pages per visit. Flashback hosts such a large variety of topics across and within forums, and there is no similar alternative forum in Swedish. If users leave Flashback, they have to migrate to more topic-specific message boards. The discussions are arranged in sub-forums that range from politics, sexual preferences and drug abuse to electronics and family relationships. Flashback categorizes each sub-forum into discussion threads, and within each thread, a member can contribute by posting a post. A user can never delete a posted message. Each sub-forum has users with moderator status supervising the discussions using Flashback's internal rules (Netiquette). If a user breaks any of the rules a moderator can give a warning and temporarily or permanently exlude the user from the forum. Moderators can also place discussion threads in the Recycle bin or

[8] In 1998 the law strengthened the responsibility of the publisher of online message boards to remove hate speech content. However, the publisher of Flashback is registered in the U.S. and is not affected by that part of the law.

lock threads for further posts. We find no indications on the forum of moderators changing their behavior during the period we study. Flashback did, for example, not change their internal rules for the moderators.

Previous literature finds the rate of abusive speech to differ depending on the topic of the forum (Cheng et al., 2017) where forums discussing general news and politics trigger more abusive language compared to forums discussing computer science.[9] We are interested in forums where individuals seek to discuss political and societal topics. In this respect, Flashback has become an important arena for testing opinions and collecting information, for example during periods before elections. We focus on the following three sub-forums: Immigration, feminism and domestic politics. In these forums, discussions about everything from party politics to everyday concerns of values and taxes take place.[10]

To study the impact of anonymity on hateful content we use an event. On September 10, 2014, Swedish and international media revealed that the identity of approximately one-third of the accounts registered at Flashback before March 2007 was unexpectedly in the hands of a group of journalists that called themselves Researchgruppen.[11] The group of journalists publicly exposed the identity of a few individuals with a history of writing hateful content, to indicate the increased risk of exposure for users registered before March 2007. For Flashback users this was considered a decrease in the anonymity of the users present at the site. A specific thread was initiated to discuss Researchgruppen, anonymity and the possible risk of identity exposure for some users. This event also received a great deal of publicity in national and international traditional media as well as social media. Figure 1a displays data from Google trends for the weekly relative search

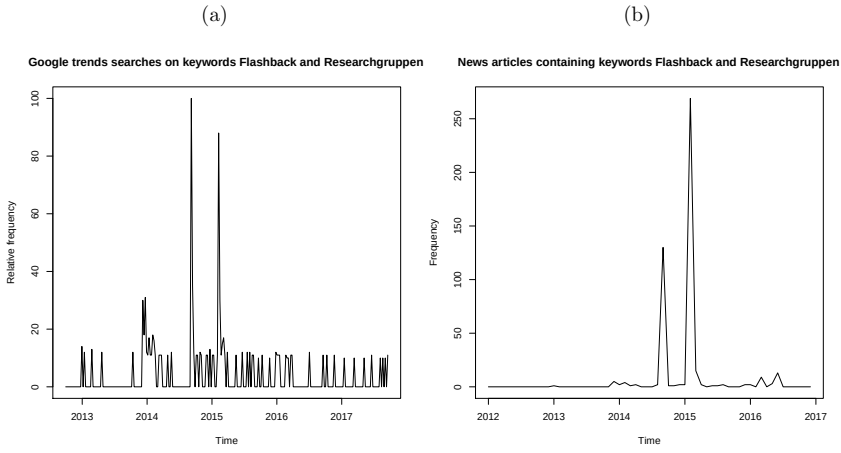
[9] Cheng et al. (2017) found that, while forums discussing general news triggered 21.4 % banned posts, computer science forums only triggered 2.3% banned posts.

[10] Examples of a thread in the domestic politics forum is "Keywords for a political alliance", in which participants discuss how all Swedish politicians have similar attitudes towards gender quotas and immigration. In another thread entitled "The number of sick days has dropped more than 50%" discussants focus on paid sick leave and people being forced to work while being ill. In the Feminism sub-forum the threads include discussions about how feminists affect Swedish politics, for example, the thread "Do feminists want gender equality?".

[11] https://www.technologyreview.com/s/533426/the-troll-hunters./

frequencies for the terms Flashback and Researchgruppen between 2013 and 2017. The relative search frequency has a spike if there are at least 50 searches. During this period, there are two major visible events, the first corresponding to the week during which the initial media revelation occurred. The second spike corresponds to the second week of February 2015, when the few identities were exposed. Panel 1b displays a similar graph using the monthly amount of news articles in Sweden containing the words "Flashback" and "Researchgruppen".[12] Again, the same two spikes are clearly visible, September 2014 and February 2015. In our study, we will exploit the users registered before March 2007 as the treatment group and label them early users. From September 2014 and onwards, this group ran an increased risk of having their identities publicly exposed in connection with the content they had written on Flashback. Users registered after March 2007 will serve as the control group, and we label them late users. The identities of members of this group did not run any risk of being identity exposed together with their written content.

Figure 1: Reactions to exposure



[12]The data on the number of news articles is obtained through searches in the database Mediearkivet.

To obtain data on the political discussions on the Flashback forum, we used scraping to retrieve the data from the Internet. Using a custom-built script in Python, we downloaded all posts (entries) in the three forums from the time the respective forum started until January 2017.[13] Next, we randomly selected 100 threads from each forum and let a research assistant classify the first twelve and last five posts from the threads. The randomization was implemented at the thread level because we wanted to classify whether initial hateful content was followed by more or less hateful posts and if a debate occurred criticizing previous posts. The research assistant was given instructions with definitions of the following main classifications; hateful content, threatening content, aggressive content, and towards whom the hate was directed. The assistant also classified the post according to whether the post confirmed or questioned the discussion in previous posts, whether the post expressed support for or against a specific political party, as well as whether the post contained a language of 'us and them'. Previous research on linguistic markers guided us in forming the classification definitions (Cohen et al., 2014). In this paper, we focus on hateful content. The final data set contains 4021 classified posts, about equally divided across the three forums.[14] In this paper, we focus on the classification of hateful content. Please see appendix section C for the instructions to the research assistant.

4 Prediction of hate

4.1 Methodology

This section describes the methodology we use to predict hateful content in new data using the classified data. To bring the text into something quantifiable, we use a so-called bag of words approach. First, we created a matrix consisting of the posts as rows and each word of the classified data as column names. Second, we

[13]Specifically, we downloaded all posts (entries) in these forums from the start of the respective forum until the day each script ended (they ended sequentially between January 2, 2017, and February 9, 2017), except for domestic politics from which we collected all posts from May 26, 2000. The feminism and integration forums started later, on May 25 2005 and July 4 2007.

[14]The data obtained from the RA contains 4043 observations; however 22 of these are not in the analysis because they contain only stopwords or numbers.

removed common stop-words. Stop-words are topic-neutral words such as articles and conjunctions. To reduce the dimensionality of the matrix, we also stemmed the data. Stemming is a common computer linguistic process removing some ending characters of a word and grouping similar words together. For example, words such as argues, arguing, argue are reduced to argu. The discussions are all in Swedish, and thus removing stop-words and stemming were adapted to the Swedish language.[15] To fill the cells in the matrix with a statistic that reflects the importance of a word for a post in the data set, we estimate a weighting factor for each word in each post. The type of weighting scheme we use is called *term frequency-inverse document frequency* (tf-idf). The value of the tf-idf increases proportionally to the number of times a word appears in a post but adjusted by the frequency of the word in the entire data set. Tf-idf is, for example, a common weighing scheme in recommender systems in digital libraries. The weights in each cell are estimated using the following procedure:

$$tfidf(t_k, d_j) = \#(t_k, d_j) * log\frac{|T_r|}{\#_{Tr}(t_k)} \tag{1}$$

where $\#(t_k, d_j)$ is the number of times the word t_k occurs in a post d_j and $\#_{Tr}(t_k)$ is the number of posts in the entire data set T_r in which t_k occurs.

The tf-idf matrix comprises our right-hand side variables in the prediction models. The left-hand side variable is a dummy for hateful content, a dummy for hateful content against immigrants or a dummy for misogynistic content. In line with methodological practice in machine learning, we split the classified data set into one training set of 2815 posts (observations) - roughly 70 percent - and one test set of 1206 posts. Moreover, we removed the words, which did not appear in the training data, but only in the test set. We started by describing the data of the full manually classified set. Then, we ran a Logistic Lasso as the machine learning model using only the training set. We compared the predictions from this model

[15]To create the matrix and to remove the stop-words as well as the stemming, we use the statistical software R. For the stemming, we use the package *SnowballC*.

with the actual (true) classifications by the RA in the test data set. Then, we evaluated the model by so-called confusion matrices; presenting probabilities of correct and incorrect classifications. The upcoming section 4.2 provides descriptive statistics for the data classified by the RA and briefly describes the Logistic Lasso model, while section 4 presents the main findings from the prediction model.

4.2 Description of the coded data set and Logistic Lasso

Table 1 presents summary statistics of the data that was manually classified by the research assistant. Besides basic descriptives such as the number of posts in total and by treatment- and control group, the number of users and threads, we also present the average share of the three main outcome variables we use in our analysis. Note that around one in five to one in four posts are hateful, with around every tenth being aimed at foreigners while every twentieth contains hate against females and feminist content. The outcome variable hateful entries comprises hate against foreigners, hate against females and feminists as well as hate against other groups, and we can see that the majority of hateful posts are concerned with either foreign or misogynistic attitudes. In the coded data set, 44 percent are considered aggressive, whereas only 1 percent of the entries is classified as threatful. Moreover, the share of disputing posts indicates that the users do not have the same opinion, but rather disagree with another user's standpoint in 27 percent of the posts.

Table 1: Summary statistics, data from research assistant

	Total	Early adopters	Late adopters	Pre-event registered
No. entries	4021.00	1153.00	2868.00	3794.00
No. users	2031.00	532.00	1499.00	1885.00
No. threads	295.00	214.00	256.00	294.00
Share hateful entries	0.23	0.21	0.25	0.23
Share hateful entries against foreigns	0.09	0.08	0.10	0.09
Share mysogynistic entries	0.06	0.03	0.07	0.05
Share of we/them entries	0.13	0.08	0.15	0.12
Share disputing entries	0.27	0.24	0.28	0.27
Share consenting entries	0.13	0.13	0.13	0.13
Share of threatful entries	0.01	0.01	0.01	0.01
Share aggressive entries	0.44	0.41	0.45	0.44

Across the forums, the largest share of hate is found in the immigration forum, where every third post contains hateful content, and the lowest share is found in the domestic policy forum. Hate against foreigners and hate against feminists and

females are mainly found on separate forums. The results are presented in Figures 6a-6c in the Appendix.

The machine learning tool we use is a Logistic Lasso.[16] Lasso is a regression analysis method, which does variable selection and regularization to increase prediction precision. Lasso reduces the coefficient estimates towards zero to balance the variance-bias trade-off, with some variable coefficients being reduced to zero. Formally, the Logistic Lasso computes a penalized maximization problem of the form given in equation 2.

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0 + \beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2}$$

Equation 2 is thus equivalent to the standard log-likelihood function for the logistic regression with the added penalty term $-\lambda \sum_{j=1}^{p} |\beta_j|$. The key parameter λ is chosen by tenfold cross-validation. Cross-validation is a method of evaluating models based on the idea of using the whole training data set when training. Using this method, we divide the training data into k-subsets, and when each of the subsets constitutes a test set, the other k-1 subsets become the training set. Running the algorithm k times, each observation will be in a test set once and in a training set k-1 times. Finally, the average error across all k trails is computed.[17] In this paper, we will focus on the variables measuring hate in general, hate against foreigners and misogyny as outcomes in three separate prediction models, and the tf-idf matrix comprises the regressors.

[16] We also ran a support vector machine model in the coded data. The Lasso made better predictions, with fewer incorrect and more correct classifications. The result of this exercise is presented in appendix section A.0.1.

[17] We use R as our statistical software along with the package *glmnet* for Logistic Lasso.

4.3 Prediction results of hate in the coded data

When evaluating the performance of the predictions from the Logistic Lasso model, we focus on maximizing the sum of the true positive rate or sensitivity;

$$\frac{Truepositives}{Truepositives + Falsenegatives}$$

and the true negative rate or specificity;

$$\frac{Truenegatives}{Truenegatives + Falsepositives}$$

Accuracy is another evaluation measure, defined as

$$\frac{True\ positives + True\ negatives}{Total\ cases}$$

Since all our outcomes are heavily skewed towards zero, focusing on maximizing the accuracy would not yield any fruitful predictions as the best accuracy will typically be predicting all posts as non-hateful. However, we report all three evaluation measures for each of the outcomes below. Figures 2a, 2b and 2c show how we have traded off the true positive rate against the true negative rate using ROC-curves (Receiver Operating Characteristic). Intuitively, this is a trade-off between type I and II errors, where we strive to minimize the sum of these two.

Table 2 presents the prediction results from the Logicstic Lasso model on hate split by the true classifications. Tables 3 and 4 do the same for hate against females and feminists and hate against foreigners. Starting with general hate in table 2 shows that out of the 1206 cases, the Lasso classifier predicted hateful content in 89 cases and no hateful content in 1117 cases. Posts that were classified by the research assistant as having hateful content constitute 290 cases, and 916 cases

had no hateful content. The true positive rate in table 2 is $\frac{53}{53+237} \approx 0.183$, while the true negative rate is $\frac{880}{880+36} \approx 0.961$, implying that our prediction for general hate still makes less than 5 percent of type I errors. The accuracy of the prediction is $\frac{53+880}{1206} \approx 0.774$. Proceeding to Table 3, the algorithm allows for a higher false positive rate, thus giving us a true negative rate of $\frac{969}{969+164} \approx 0.855$ and a true positive rate of $\frac{47}{16+47} \approx 0.644$. The accuracy rate is $\frac{47+969}{1206} \approx 0.843$, indicating that we are more successful in predicting hate against females as compared to general hate. However, as noted above, misogyny is a more skewed variable as compared to hate, and comparing the accuracy between the two will not be that meaningful since we can get a higher accuracy for misogyny simply by classifying all data as non-misogynistic. In the appendix, we compare the two classifications by the areas under the ROC-curves in Figures 2a and 2b. The Logistic Lasso predicts misogyny better than hate in general, the area under the curve for general hate is 0.57, while it is 0.75 for misogyny. The area under the ROC curve is called AUC, which is a fairly standard quality-of-prediction measure in Machine Learning applications. A value of 0.5 would imply that we are not doing any better than chance.

Table 4 provides the results from the prediction of the Logistic Lasso on hate against foreigners. For this variable, we have a true positive rate of $\frac{59}{68+59} \approx 0.465$ and a true negative rate of $\frac{973}{973+106} \approx 0.902$. The accuracy, in turn, is $\frac{973+59}{1206} \approx 0.856$. Thus, the accuracy indicates that we can predict hate against foreigners better than both general hate and misogyny. However, the area under the ROC-curve is 0.69, implying that we can predict hate against foreigners better than hate in general, but not better than misogyny. In sum, we can conclude that general hate will be our noisiest measure, and misogyny will be the most precise.

Table 2: Confusion matrix from Logistic Lasso on hate

	Truth	
Predict	No hate	Hate
No hate	880	237
Hate	36	53

Table 3: Confusion matrix from Logistic Lasso on misogyny

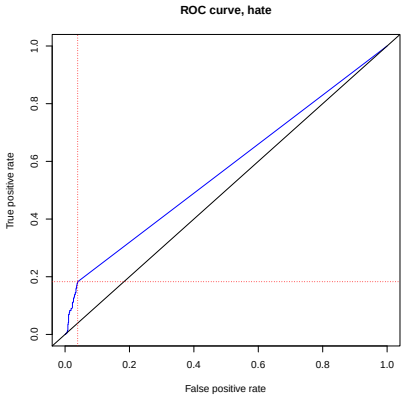
	Truth	
Predict	No misogyny	Misogyny
No misogyny	969	26
Misogyny	164	47

Table 4: Confusion matrix from Logistic Lasso on hate against foreigners

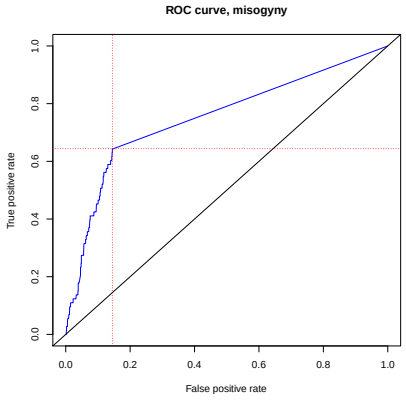
	Truth	
Predict	No hate foreign	Hate foreign
No hate foreign	973	68
Hate foreign	106	59

Figure 2: ROC curves

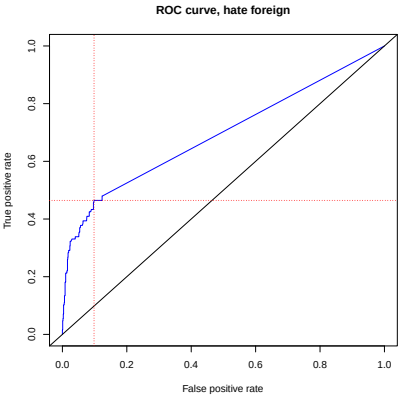
(a)



(b)



(c)



The Logistic Lasso performs a shrinkage to the coefficients for bias-variance trade-off, which implies that it will only include some variables, words in our case, in the final prediction model. We now turn our attention to which words that were included in the evaluated prediction models. Table 5 displays all words and

their associated coefficients from the Logistic Lasso model of general hate. The stemming has removed some of the ending characters. The first word is *arab*, and this is the same in Swedish as in English. The second word, *blatt*, comes from the Swedish racial slur word blatte, which is a derogatory word for someone with a dark skin tone. *Hor* probably comes from hora, which translates to whore, *lill* comes from lilla or lille, which are typically used to belittle someone. *Miljon* means a million and might refer to the cost of a political process, such as immigration or can also refer to the Million Programme, a Swedish public housing project from the 60's and 70's. Finally, *muslimsk* means muslim, *patetisk* translates to pathetic and *rån* means robbery. Overall, the words selected by the Logistic Lasso seem to conform with words connected to groups that are often targeted by cyberhate, women and foreigners (Citron, 2014). The coefficients are all positive, and this could indicate that the most common way of discussing in the coded data set is without any hateful content. The levels of the coefficients are not particularly useful to discuss since the Logistic Lasso produces biased estimates.

Table 5: Lasso coefficients for hate

name	coefficient
(Intercept)	-1.22
arab	0.22
blatt	0.95
hor	1.23
lill	1.53
miljon	1.27
muslimsk	3.58
patetisk	2.07
rån	2.81

Comparing the words picked for general hate and hate against foreigners and misogyny, there is some overlap between the words chosen by the Logistic Lasso. The words arab, muslimsk and blatt are also found in models predicting hateful

content against foreigners, whereas hor is the only overlapping word between the mysoginistic and the general hate model. For the full set of words picked out by the algorithm for these predictions, see Tables 23 and 24 in the appendix.

Using single words as primary features for classification can lead to misclassification since words can have different meanings in different contexts. To include some degree of context we tried using bi-grams, or word-pairs, occurring in a sequence. However, N-grams typically have issues with the distance between relevant words (Chen et al., 2012). The inputs into the Logistic Lasso are then weights for all pairs of words rather than for single words. However, it did not improve the classifier's prediction performance and therefore, we use the single word approach, with the results from the bi-gram approach available upon request.

5 Data and empirical strategy for exploring anonymity and hate

5.1 Descriptive results from the full data

The three models described above were used to predict hate in the full sample, including all posts that were not manually classified. Here, we restrict ourselves to the period from January 1, 2012, until December 31, 2016. December 31, 2016 is a natural ending point since we have data for all forums and all threads up to this date. The group of late adopters is decreasing as we move further back in time until it disappears completely in April 2007. Therefore, we do not to use any data before January 1, 2012. Table 6 shows the summary statistics for four different samples; the total sample of all users, early adopters, late adopters and users registered before the event took place. The statistics presented are the number of posts, the

number of users, the number of threads and the share of predicted hateful posts, the share of hateful posts against foreigners and hateful posts against females and feminists. The number of late adopters is larger compared to early adopters, and late adopters have written more posts than early adopters. Early adopters have written approximately 55 posts per user, whereas late adopters have written approximately 39 posts per user. About seven percent of all posts are predicted to have hateful content, while 16 percent are hateful against foreigners and 14 percent are misogynous. In the sample of manually coded posts displayed in Table 1, hate against foreigners and females and feminists will, by construction, be lower than general hate since both consist of shares of general hate. When we predict posts using the separate Logistic Lasso models in the full sample, this is not necessarily true, as illustrated in Table 6.

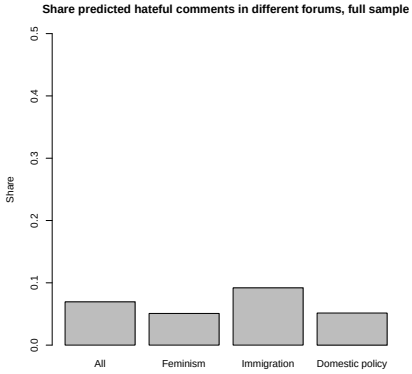
Table 6: Summary statistics, All data

	Total	Early adopters	Late adopters	Pre-event registered
No. entries	1984224.00	243604.00	1740620.00	1770474.00
No. users	48672.00	4474.00	44198.00	41738.00
No. threads	22360.00	17471.00	22246.00	22192.00
Share hateful entries	0.07	0.07	0.07	0.07
Share hateful entries against foreigns	0.16	0.17	0.16	0.16
Share mysogynistic entries	0.14	0.11	0.14	0.14

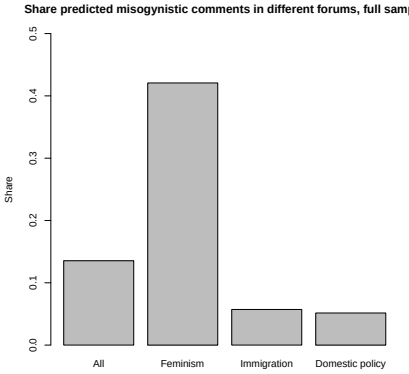
Splitting the outcomes by the forums, Figure 3a shows that most predicted hate seems to occur in the immigration forum, where every tenth post contains a hateful content, and the least predicted hate in the feminist and domestic policy forum, which resembles the manually coded data. As expected from the manually coded data, Figures 3b and 3c show that misogyny is most prominent in the feminist forum, and hate against foreigners is most prominent in the immigration forum.

Figure 3: Forum shares

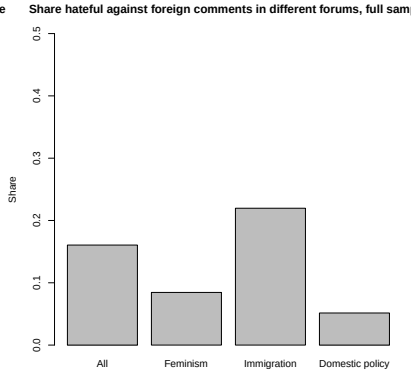
(a)



(b)



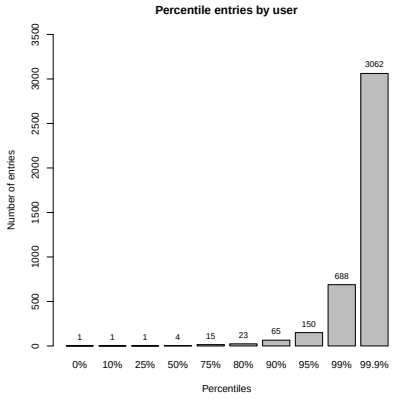
(c)



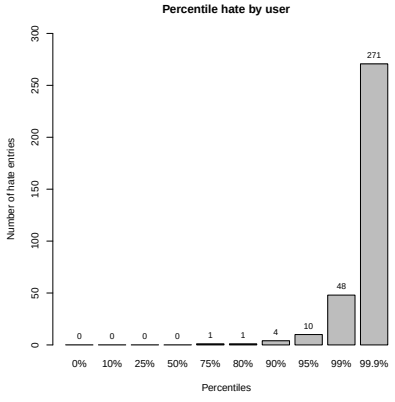
Even if the hateful posts against foreigners and hate against females and feminists occur in different sub-forums, it might be the same users that express these sentiments. To further understand users that write a hateful content in these online discussions, we take a glance at the distribution of the number of posts across percentiles of users. Figure 4a displays that there seems to be three types of Flashback users by means of activity. About 25 % of the users write only one post during the period of interest; a large middle group writes between 4 and 688 posts, and only 0.1 % of the users write a lot of posts (on average 3062 entries). Figure 4b shows that the number of hateful entries per user seems to be even more heavily skewed, where most users write zero hateful entries, the top 5 percentile around 10 hateful entries and the top 0.1 percentile writes 271 hateful posts. Finally, Figure 4c displays a similar distribution for the share of hateful entries. Once again, the bottom 50 percentiles have a share of hateful entries which is zero, while the 75 percentile and up to the 99 percentile write between 0.08 and 0.5 hateful entries. At the very top are people who only write hateful entries, but this group seems to consist of users with very few entries.

Figure 4: Distribution of entries and hate

(a)



(b)



(c)

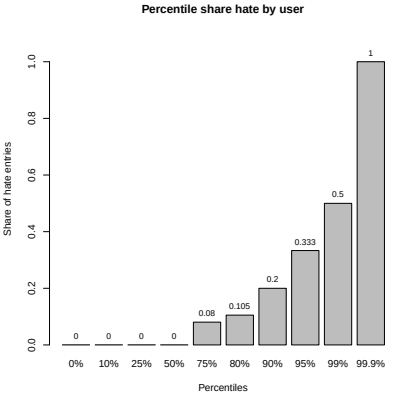


Table 18 in the appendix displays some simple correlations between our outcome measures as well as the number of entries. Interestingly, the relationship between the number of entries and the share of hateful entries a user writes seems to be quite weak. About one out of every twenty entries containing hate against

females and feminists is also classified as being generally hateful, while the same number for hate against foreigners and general hate is that around one out of every fourth entry is classified as both. Thus, the relationship seems a great deal stronger between hate in general and hate against foreigners, than hate in general and misogyny. The two types of hate show a surprisingly low correlation. Only 1.7 percent of the posts are classified as both being hateful against foreigners and hateful against females and feminists. The predictions indicate that the two types of hate do not overlap, but the same individual can compose hate against both foreigners and females separately. The same table displays no significant correlation between individuals' share of hate against foreigners and hate against females.

In sum, these results indicate that it is possible to find a prediction model of hateful messages in Swedish, and in this particular online discussion forum hate against foreigners is most prominent and seems easier to predict using this type of approach than hate against females and feminists. Moreover, hate against foreigners and hate against females seem to be conducted by separate users.

5.2 Empirical strategy for exploring anonymity and hate

To investigate whether a decrease in anonymity affects hate in discussions online, we use the event described in section 3 in a difference-in-difference strategy with the values of the predictions from the three Lasso models as outcomes. Users registered before March 2007 we call early adopters and those registered after we call late adopters. The first event on September 10, 2014 is the date that starts the post period in our difference-in-difference setting (the largest spike in Figure 1a and the smaller spike in Figure 1b). The strategy is formally wrapped up in

equation 3.

$$Y_{itg} = \alpha + \beta Early_{ig} * Post_t + \theta Post_t + \gamma Early_{ig} + \varepsilon_{itg} \tag{3}$$

Y_{itg} is the outcome variable, which will be a dummy for hateful content, hateful content against foreigners or misogyny, in a post by individual i at time t belonging to group g. $Early_{ig}$ is a dummy variable taking the value 1 if individual i belongs to the group g of early registered users, $Post_t$ is a dummy taking the value 1 in the post period and ε_{itg} is the error term. β thus measures the treatment effect of the change in anonymity, i.e. the increased probability of having one's identity publicly exposed.

Estimation of standard errors might be an issue in our setting since treatment primarily varies at the control-treatment group level (Bertrand et al., 2004; Donald and Lang, 2007; Conley and Taber, 2011). In our baseline estimates, we cluster the standard errors on the start date of the user since that is the treatment assignment variable. Moreover, we cluster the standard errors at the user level, the thread level, provide simple robust standard errors and use the Pettersson-Lidbom and Thoursie (2013) application of the results in Donald and Lang (2007). This final approach entails that we estimate the standard errors using a two-step approach, where we first aggregate the data to the group level by equation 4

$$\bar{Y}_{tg} = \alpha + \beta Early_g * Post_t + \theta Post_t + \gamma Early_g + \varepsilon_{tg} \tag{4}$$

where $\bar{Y}_{tg} = \sum_{i=1}^{N} Y_{itg}/N_g$. We can then note that we can rewrite equation 4 as the difference between the two groups, $g = 0, 1$:

$$\bar{Y}_{t1} - \bar{Y}_{t0} = \Delta Y_t = \pi + \beta Post_t + \Delta\varepsilon_t \tag{5}$$

Estimating equation 5 with the Newey-West estimator will give us standard errors adjusted for both correlation within treatment-control groups and serial correlation.

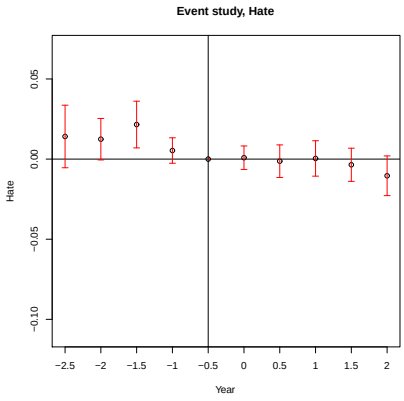
6 Difference-in-Difference results

Figure 5 shows the differences between early and late adopters along with 95 percent confidence intervals on six-month intervals before and after the event, as proposed in Angrist and Pischke (2008). Panel 5a provides us with the estimates for general hate. We can see that the estimates are slightly higher before the event, and somewhat smaller in the post-period, with no discernible trend over time in the coefficients. Panel 5b and panel 5c illustrate the same coefficients for hate against foreigners and misogyny. Anonymity seems to decrease not only general hateful content but also hate against foreigners. However, we find what seems to be a slight increase in misogyny. Again, neither of these figures appear to display any real trend in the pre-treatment period in the coefficient estimates, which thus implies that our identification strategy seems credible.

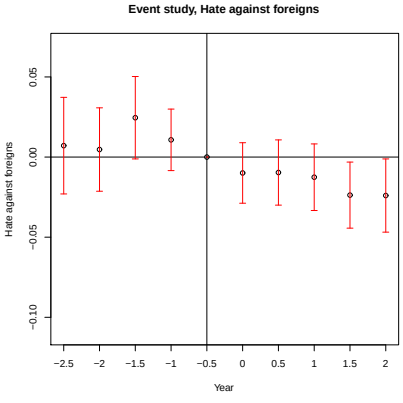
Tables 7-9 show the coefficients from the difference-in-difference estimations, corresponding to equation 3, with the various ways to treat the standard errors. The first column in each table clusters on user start date, the second column clusters on individual user, the third column clusters on thread, the fourth column collapses the data by weeks, and the fifth column uses no clustering of standard errors or collapsing of data. Overall, the regressions seem to confirm our conclusions from Figure 5. The main estimate in table 7 implies that the early adopters have a 1.5 percentage point lower probability of writing a hateful comment in the post period, with a baseline level of 7 percent. Taking this treatment effect literally, this means that instead of writing 7 out of 100 posts with hateful content,

Figure 5: Event study

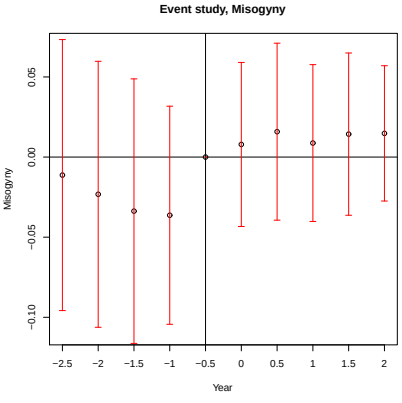
(a)



(b)



(c)



The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user startdate (172 clusters).

early adopters now only write hate in 5.5 out of 100 posts. The estimates in table 8 show us in turn that threat of exposure decreases the amount of hate against foreigners by 2.7 percentage points, while table 9 indicates that misogyny increases by 3.2 percentage points.[18]

It is not straightforward how we should treat the standard errors in the current setting with only one treatment and one control group. To accommodate this, we present several different approaches to the standard error estimation.[19] The first column clusters the standard errors at the user start date (172 clusters), as this is the variable determining treatment assignment, while the second column clusters at the individual level. Either approach gives similar standard error estimates. The third column in turn clusters the standard errors at the thread level, which decreases the standard errors quite substantially. Column four, in turn, collapses the data into the treatment (early) and control (late) groups on a weekly basis and then runs a regression on this collapsed data as demonstrated in equation 5, applying Newey-West standard errors estimated with four lags. This methodology should take care of all correlations within the two groups, as well as auto-correlation of the standard errors up to a month, thus providing us with the most reliable estimates. The standard errors remain at a similar level to when we cluster at the thread level, indicating that, if anything, we appear to overestimate the standard errors in our baseline specifications. Finally, we show the result with simple, robust standard errors. Doing this decreases the standard error even more. We can conclude that no matter how we treat the standard errors, we still find significant effects and, if anything, we seem to overestimate the standard errors in our baseline specifications.

[18]Changing specifications to logit regression instead of the linear probability model does not change our results. Table 22 in the appendix displays the marginal effects for such a specification.

[19]See, for instance, Bertrand et al. (2004), Donald and Lang (2007) and Conley and Taber (2011).

Table 7: DD results on hate

	Dependent variable:				
	Share hate				
	(1)	(2)	(3)	(4)	(5)
Post*Early	-0.015^{**}	-0.015^{**}	-0.015^{***}		-0.015^{***}
	(0.007)	(0.006)	(0.002)		(0.001)
Post reveal	-0.002	-0.002	-0.002	-0.013^{***}	-0.002^{***}
	(0.002)	(0.002)	(0.002)	(0.002)	(0.0004)
Early adopters	0.007	0.007	0.007^{***}		0.007^{***}
	(0.007)	(0.006)	(0.002)		(0.001)
Constant	0.070^{***}	0.070^{***}	0.070^{***}	0.005^{***}	0.070^{***}
	(0.002)	(0.002)	(0.001)	(0.002)	(0.0003)
Clustering	Start date	Individual	Thread	-	No clustering
Collapsed on weeks	No	No	No	Yes	No
Observations	1984224	1984224	1984224	262	1984224

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. The first column clusters the standard errors at the startdate, the second on the username, the third on thread, the fourth uses Newey-West standad errors with 4 lags on a collapsed time series and the final column uses normal robust standard errors.

Table 8: DD results on hate against foreigns

	Dependent variable:				
	Share hate foreigns				
	(1)	(2)	(3)	(4)	(5)
Post*Early	-0.027^{**}	-0.027^{***}	-0.027^{***}		-0.027^{***}
	(0.011)	(0.010)	(0.004)		(0.002)
Post reveal	-0.008^{**}	-0.008^{**}	-0.008^{**}	-0.023^{***}	-0.008^{***}
	(0.004)	(0.004)	(0.004)	(0.004)	(0.001)
Early adopters	0.016	0.016	0.016^{***}		0.016^{***}
	(0.011)	(0.011)	(0.003)		(0.001)
Constant	0.164^{***}	0.164^{***}	0.164^{***}	0.014^{***}	0.164^{***}
	(0.004)	(0.004)	(0.003)	(0.003)	(0.0004)
Clustering	Start date	Individual	Thread	-	No clustering
Collapsed on weeks	No	No	No	Yes	No
Observations	1984224	1984224	1984224	262	1984224

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. The first column clusters the standard errors at the startdate, the second on the username, the third on thread, the fourth uses Newey-West standad errors with 4 lags on a collapsed time series and the final column uses normal robust standard errors.

Table 9: DD results on misogyny

	Dependent variable:				
	Misogyny				
	(1)	(2)	(3)	(4)	(5)
Post*Early	0.032**	0.032**	0.032***		0.032***
	(0.014)	(0.014)	(0.005)		(0.001)
Post reveal	−0.021***	−0.021***	−0.021***	0.032***	−0.021***
	(0.006)	(0.007)	(0.005)	(0.007)	(0.001)
Early adopters	−0.044***	−0.044***	−0.044***		−0.044***
	(0.012)	(0.012)	(0.003)		(0.001)
Constant	0.149***	0.149***	0.149***	−0.042***	0.149***
	(0.007)	(0.006)	(0.004)	(0.004)	(0.0004)
Clustering	Start date	Individual	Thread	-	No clustering
Collapsed on weeks	No	No	No	Yes	No
Observations	1984224	1984224	1984224	262	1984224

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. The first column clusters the standard errors at the startdate, the second on the username, the third on thread, the fourth uses Newey-West standad errors with 4 lags on a collapsed time series and the final column uses normal robust standard errors.

A decrease in the share of hateful content can be caused by both changes in the number of hateful entries and changes in the number of entries in general (hateful and non-hateful entries). We ran the specification from the fourth column, collapsing the data on weeks using the number of entries and the number of hateful entries as outcome variables. Table 17 in the Appendix shows that the decrease in the share of hateful entries is a result of both a decrease in the number of hateful entries and a decrease in the number of total entries. Although the decrease in the number of non-hateful entries is larger as compared to the decrease in the number of hateful entries, the relative proportion of these changes causes a decrease in the share of hateful entries. This result indicates that one possible mechanism behind the decrease in the share of hateful entries is that the decrease in anonymity makes users leave the forum or decrease writing both hateful and non-hateful posts at the forum. Looking at the mere number of users pre and post the event, the number of early users decreases 4.5 percentage points more than the number of late users.[20]

[20]From the above results it would seem natural to study the probability that a user drops

6.1 Individual user behavior

A decrease in anonymity might also lead individual users to stop writing hateful content and proceed to writing non-hateful content. To this end, we take a closer look at the individual users and their behavior. We adjust the sample to comprise users that posted at least one entry before the event and remained active after the event by the same criteria. The sample thus consists of individual users that have at least one entry in both the pre- and the post-treatment periods, and we collapse the data into a pre- and post observation for each user. With this collapsed data we run difference-in-difference models for the sum as well as the share of hateful entries for the three outcomes; hateful entries in general and hateful entries against females and foreigners, respectively. The collapsed data consists of 11159 individual users, with two observations per user. Comparing with the number of users in the full data set, we see that the vast amount of users post entries only before the event. Table 10 shows the results of the collapsed data.

out of the discussion. The problem with this and our data is, however, that we only observe individuals if they are actually active both in the pre- and post-period. This implies that we cannot use the time dimension in the difference-in-difference setting if our outcome is a dummy on the user being active, since all users will, by definition, be active in the pre-treatment period. Thus, we can only observe if early users are more likely than late adopters to drop out after the event and not the change in the drop out rate between the two groups.

Table 10: Individual behavior

	Dependent variable:						
	No. hate	Share hate	No. hate foreign	Share hate foreign	No. misogyny	Share misogyny	No. entries
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Post*Early	-1.825^{**}	-0.002	-3.658^{**}	-0.006	0.111	0.0003	-14.711^{**}
	(0.778)	(0.004)	(1.499)	(0.006)	(1.019)	(0.005)	(5.977)
Post reveal	-0.289	-0.010^{***}	-1.037^{**}	-0.015^{***}	-1.434^{**}	-0.012^{***}	-3.060
	(0.204)	(0.002)	(0.497)	(0.002)	(0.687)	(0.003)	(2.933)
Early adopters	1.339	-0.003	2.996^{*}	-0.013^{***}	-1.072	-0.021^{***}	11.623
	(0.837)	(0.003)	(1.753)	(0.004)	(1.408)	(0.004)	(7.135)
Constant	3.997^{***}	0.061^{***}	9.449^{***}	0.144^{***}	8.667^{***}	0.103^{***}	59.355^{***}
	(0.229)	(0.001)	(0.517)	(0.002)	(0.855)	(0.002)	(3.085)
Observations	22318	22318	22318	22318	22318	22318	22318

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016.
Standard errors clustered at the user startdate.

The first column shows that early adopters decrease their amount of hateful entries by 1.8 from the pre- to the post period as compared to the control group. However, column two shows that the same is not true for the share of individual hateful comments. Column three, in turn, highlights that the treated substantially decrease their amount of hateful content against foreigners, writing on average 3.7 fewer hateful entries against foreigners as compared to the control group. Once again, however, the share of the individual users remains unaltered, as demonstrated in column four. In line with the results, column five in tables 7-9 indicates that users do not decrease their amount of misogyny, the coefficient is rather positive but insignificant. Furthermore, the share of misogyny remains at a similar level as in the pre-period. Finally, column seven shows the estimated effect on the total amount of user entries. We can note that those that are at the risk of being exposed decrease their level of overall activity quite drastically, writing 14.7 fewer entries in the post period as compared to the change in the control group. However, this is a decrease in the number of entries which is driven primarily by users with a higher share of hateful entries in the pre-treatment period. This is illustrated in table 11. The first column is the same regression as column 7 in table 10, but restricting the sample to only users in the bottom 2/3 of the distribution of the share of hate at the individual level in the pre-period. We can see that the coefficient from the simple DD-model has now shrunk to being in essence zero and being insignificant. The second column in turn gives the same estimate but only using the individuals in the bottom 2/3 when it comes to foreign hate in the pre-period. Again, the coefficient is insignificant. Finally, the third column performs the same exercise but using pre-treatment misogyny as the cut-off variable. Interestingly, we find an almost identical significant coefficient as in column 7 in table 10. Thus, it looks like it is the individuals with a high rate of hate in general and hate against foreigners that alter their behavior, but those with a high proportion

of misogyny do not appear to change. A threat of exposure seems to primarily affect users that have misbehaved in the past, and they alter their behavior by reducing their level of activity.

Table 11: Individual behavior, low haters

	Dependent variable:		
	No. entries		
	(1)	(2)	(3)
Post*Early	0.701	−6.865	−14.698**
	(4.301)	(5.610)	(5.793)
Post reveal	0.115	−0.898	2.956
	(2.601)	(3.153)	(2.234)
Early adopters	−0.463	3.630	14.524**
	(5.093)	(7.231)	(6.263)
Constant	34.290***	50.866***	38.039***
	(3.133)	(3.546)	(1.821)
Cut off	Hate< 0.039	Hate foreign< 0.143	Misogyny< 0.063
Observations	14236	13992	14244
Note:	*p<0.1; **p<0.05; ***p<0.01 The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user startdate. Each column uses the data in the bottom 2/3 of the distribution in individual share hate, share hate foreign and share misogyny in the pre-treatment period, respectively.		

In sum, a decrease in anonymity, by a threat to expose users' identities, can lead to a decrease in hateful content in online discussions, but it can also lead to a decrease in non-hateful entries. Individuals leave the forum or decrease their activity level.

6.2 Subtitution of hate

The main result displays an increase in the share of hate against females and feminists, but no decrease in the number of misogynous entries. A possible explanation of this result, is that a decrease in anonymity can make users substitute hate, i.e. users might decrease their hate against foreigners but increase their hate against females and feminists. To estimate this, we use the simple framework outlined in equations 6-8, where X_{ig0} is the number of entries with hate against foreigners in the pre-period by user i, in group g in period $t = 0$ and Y_{ig0} is the number of misogynous entries.

$$Y_{ig0} = \rho Early_{ig} * X_{i0} + \pi X_{i0} + \delta Early_{ig} + \kappa_{ig0} \tag{6}$$

$$Y_{ig1} = \rho Early_{ig} * X_{i1} + \pi X_{i1} + \delta Early_{ig} + \kappa_{ig1} \tag{7}$$

Taking the difference between equation 6 and 7, i.e. the first difference over time, thus yields

$$\Delta Y = \rho Early_{ig} * \Delta X_i + \pi \Delta X_i + \kappa_{ig1} - \kappa_{ig0} \tag{8}$$

Thus, under the assumption of similar trends between the treated and non-treated individuals, ρ will measure the degree to which treated individuals substitute hate against foreigners with misogyny.

Table 12 investigate the substitution effect using a regression as demonstrated in equation 8.

Table 12: Individual substitution

	Dependent variable:			
	ΔNo. misogyny	ΔNo. hate	ΔNo. entries	ΔNo. entries
	(1)	(2)	(3)	(4)
Early*ΔNo. hate foreign	−0.574**	0.078	−1.396***	
	(0.239)	(0.080)	(0.516)	
ΔNo. hate foreign	0.831***	0.397***	4.634***	
	(0.238)	(0.038)	(0.365)	
Early*ΔShare hate foreign				16.698**
				(8.434)
ΔShare hate foreign				7.040**
				(2.952)
Constant	−0.506	0.122	1.108	−5.075*
	(0.530)	(0.092)	(1.205)	(2.642)
Observations	11159	11159	11159	11159
Note:	*p<0.1; **p<0.05; ***p<0.01 The time period used in the regressions is from January 1st 2012 until December 31st 2016.			

Column 1 in Table 12 shows a negative coefficient for the interaction term in the first column, which implies that users in the treatment group who decrease their amount of hateful entries against foreigners by one betwee pre- and post-period also increase their amount of misogyny by 0.57 entries. Thus, the treated seem to substitute one hateful entry against foreigners with one-half entry containing misogyny. As expected from the main results, the second column indicate that there is no substitution between hate against foreigners and hate in general.

Column 3 and 4 probes into how users in the treatment group who decrease their amount of hateful entries against foreigners change their general posting activities. The thrid column show that a decrease by one hateful entry against foreigners is associated with an increase in the overall entries by 1.4 among the treated. The fourth column indicate that a reduction of hateful content against foreigners from 10 % to 0 % decreases the number of general posts with 1.7. There are many possible explanations for these seemingly conflicting results. The results

in this table differ from the tables above by probing into a specific group of users. Also, the results highlight that changes in number of hateful posts and in share of hateful posts might have different effects on the general posting activity.

6.3 Limitations and robustness

In Tables 13-15 we proceed to investigate the robustness of our main results. The first column simply replicates our baseline estimate from 7-9. During the same week as the event took place elections where held in Sweden. We therefore add a control variable for registration date in the second column to linearly control for the registration time of all individuals. Controlling for the user start date is equivalent to looking at the difference in discontinuity before and after the revelation date between early and late adopters, using the same slope before and after the revelation date and all data on both sides of the cutoff. In this specification, the coefficient falls slightly as compared to our main estimate, down to -0.011 for hate to -0.021 for hate against foreigners and increases to 0.038 for misogyny. Moreover, both general hate and hate against foreigners fall in significance to the ten percent level.

The results could be an effect of Flashback users creating multiple accounts, and then becoming passive on one account and continuing to discuss and produce hate using another account. First, there are strict forum rules to guide the discussions and accounts. Moderators on Flashback can suspend users from all forums if they, for example, use multiple accounts to support their arguments in the same sub-forum. A suspended user cannot use his or her current account, and cannot create a new Flashback account. Thus, it is unlikely that we have individuals with multiple accounts producing hateful content in the same sub-forum. Users that worry about their identity being in the hands of journalists could create a new account and continue to write in a hateful way using the new account. However, this is not

without its cost to regular users, as they might have built up a certain reputation for their pseudonym. To take this possibility into account, column two restricts the sample to everyone registered before the event, that is September 2014. Again, the coefficient for general hate falls to -0.011 and loses all significance at this point. Hate against foreigners decreases, but is significant at the ten-percent level. Hate against females and feminists remains stable at 0.032 and significant. In column four, we use the sample from column three, but include the linear control in start date from column two. By construction, there will be entries from a user with a late start date that has not had the time to write so many entries. A few outliers could influence the slope of the linear control function. Controlling for this in column four shows no quantitative or qualitative effect of this kind, for any of the three outcomes. In column five, we interact the linear control in start date with the post dummy variable, to let the slope of the start date differ between the pre- and post period. This specification is a global version of the difference-in-discontinuity approach as pioneered by Grembi et al. (2016), and is an even more robust control to the issue with the Swedish election being in the same week as the event. Here the point estimates for hate in general and hate against foreigners are still in the same direction, but the effect becomes non-significant. The effect on hate against females and feminists remains significant at the ten-percent level.

Table 13: DD robustness, hate

	Dependent variable:				
	Share hate				
	(1)	(2)	(3)	(4)	(5)
Post*Early	−0.015**	−0.011*	−0.011	−0.011	−0.013
	(0.007)	(0.007)	(0.007)	(0.007)	(0.009)
Post reveal	−0.002	−0.004**	−0.005**	−0.005***	0.817
	(0.002)	(0.002)	(0.002)	(0.002)	(1.622)
Early adopters	0.007	0.019**	0.007	0.017*	0.018*
	(0.007)	(0.008)	(0.007)	(0.009)	(0.009)
Registration date		0.002***		0.002**	0.002**
		(0.001)		(0.001)	(0.001)
Post*Registration date					−0.0004
					(0.001)
Constant	0.070***	−4.302***	0.070***	−3.344**	−3.724**
	(0.002)	(1.056)	(0.002)	(1.434)	(1.640)
Last registered	-	-	Sep 2014	Sep 2014	Sep 2014
Observations	1984224	1984224	1770474	1770474	1770474

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user startdate (172 clusters).

Table 14: DD robustness, hate foreign

	Dependent variable:				
	Share hate foreign				
	(1)	(2)	(3)	(4)	(5)
Post*Early	−0.027**	−0.021*	−0.020*	−0.019*	−0.016
	(0.011)	(0.011)	(0.011)	(0.011)	(0.018)
Post reveal	−0.008**	−0.013***	−0.015***	−0.016***	−0.925
	(0.004)	(0.004)	(0.004)	(0.004)	(4.166)
Early adopters	0.016	0.038***	0.016	0.030**	0.028*
	(0.011)	(0.014)	(0.011)	(0.015)	(0.017)
Registration date		0.004***		0.002*	0.002
		(0.001)		(0.001)	(0.002)
Post*Registration date					0.0005
					(0.002)
Constant	0.164***	−7.482***	0.164***	−4.637	−4.217
	(0.004)	(2.288)	(0.004)	(2.881)	(4.017)
Last registered	-	-	Sep 2014	Sep 2014	Sep 2014
Observations	1984224	1984224	1770474	1770474	1770474

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user startdate (172 clusters).

Table 15: DD robustness, misogyny

	Dependent variable:				
	Misogyny				
	(1)	(2)	(3)	(4)	(5)
Post*Early	0.032**	0.038***	0.032**	0.035**	0.045*
	(0.014)	(0.014)	(0.015)	(0.014)	(0.027)
Post reveal	−0.021***	−0.026***	−0.021***	−0.023***	−3.538
	(0.006)	(0.006)	(0.008)	(0.007)	(7.796)
Early adopters	−0.044***	−0.021	−0.044***	−0.014	−0.018
	(0.012)	(0.018)	(0.012)	(0.024)	(0.027)
Registration date		0.004		0.005	0.004
		(0.002)		(0.004)	(0.004)
Post*Registration date					0.002
					(0.004)
Constant	0.149***	−7.717	0.149***	−10.496	−8.872
	(0.007)	(4.896)	(0.007)	(7.570)	(8.335)
Last registered	-	-	Sep 2014	Sep 2014	Sep 2014
Observations	1984224	1984224	1770474	1770474	1770474

Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user startdate (172 clusters).

Going from a standard DD-approach to something that is essentially a difference-in-discontinuity methodology does not seem to change our main estimates to any considerable extent.

7 Concluding remarks

Our study contributes to the current policy debates on how to combat online hate. We estimate the link between anonymity and hateful content and find that reduced anonymity leads to reduced hate in online discussions about politics. First, we predict hateful content in political discussion forums using a Swedish social media website. Here we let a research assistant classify a random part of the entries into hateful content or

no hateful content. Using a supervised machine learning model, Logistic Lasso, we then predict hateful content in the full data set. Second, we use these predictions to quantify the causal link between anonymity and hateful content using a difference-in-difference design. The exogenous variation comes from an event where anonymity unexpectedly decreased (a threat of possibly being exposed). We find that reduced anonymity leads to less hateful content in political discussions online. The effect seems to be driven by a combination of a decline in writing hateful posts and a decrease in the number of non-hateful posts. We also find suggestive evidence of individuals substituting hate against foreigners with hate against females. One possible explanation for the substitution is that hate against foreigners is associated with previous convictions of hate speech in Sweden. Our results open up for an exciting avenue of research on how to understand different types of hate online.

Our results suggest that a policy combating online hate is not as simple as to require users to expose their names. For example, many news sites have blocked the possibility to provide anonymous comments to their articles and Facebook now requires users to display their real names. An effect of less hateful content can imply both changes in how individuals write their entries and that individuals stop discussing. In this paper, we see both effects. Discussions on social media seem to affect political outcomes (Qin et al., 2017; Allcott and Gentzkow, 2017). Fewer individuals that discuss politics online can have adverse consequences such as lower political accountability or less informed decisions (Strömberg, 2015).

8 References

Acquisti, A., C. R. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 2*(54), 442–92.

Ali, S. N. and R. Bénabou (2016). Image versus information: Changing societal norms and optimal privacy. Technical report, National Bureau of Economic Research.

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*(2), 211–36.

Allcott, H., M. Gentzkow, and C. Yu (2018). Trends in the diffusion of misinformation on social media. *arXiv preprint arXiv:1809.05901*.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Åslund, O. and O. N. Skans (2012). Do anonymous job application procedures level the playing field? *Industrial & labor relations review 65*(1), 82–107.

Bengtsson, R., E. Iverman, and B. T. Hinnerich (2012). Gender and ethnic discrimination in the rental housing market. *Applied Economics Letters 19*(1), 1–5.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *The Quarterly journal of economics 119*(1), 249–275.

Bøg, M. and E. Kranendonk (2011). Labor market discrimination of minorities? yes, but not in job offers.

Chang, D., E. L. Krupka, E. Adar, and A. Acquisti (2016). Engineering information disclosure: Norm shaping designs. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 587–597. ACM.

Chen, Y., Y. Zhou, S. Zhu, and H. Xu (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pp. 71–80. IEEE.

Cheng, J., M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. *arXiv preprint arXiv:1702.01119*.

Cho, D., S. Kim, and A. Acquisti (2012). Empirical analysis of online anonymity and user behaviors: the impact of real name policy. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 3041–3050. IEEE.

Citron, D. K. (2014). *Hate crimes in cyberspace*. Harvard University Press.

Cohen, K., F. Johansson, L. Kaati, and J. C. Mork (2014). Detecting linguistic markers for radical violence in social media. *Terrorism and Political Violence 26*(1), 246–256.

Conley, T. G. and C. R. Taber (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics 93*(1), 113–125.

Davidsson, P., M. Palm, and A. Melin Mandre (2018). *Svenskarna och Internet 2018*. IIS (Internetstiftelsen i Sverige).

Donald, S. G. and K. Lang (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics 89*(2), 221–233.

Duggan, M. (2014). *Online harassment*. Pew Research Center.

Edin, P.-A. and J. Lagerström (2006). Blind dates: quasi-experimental evidence on discrimination. Technical report, Working Paper, IFAU-Institute for Labour Market Policy Evaluation.

Eklund, L. and M. Johansson (2013). Played and designed sociality in a massive multiplayer online game. *Eludamos. Journal for Computer Game Culture 7*(1), 35–54.

Froomkin, A. M. (2017). Lessons learned too well: Anonymity in a time of surveillance. *Ariz. L. Rev. 59*, 95.

Goldin, C. and C. Rouse (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *The American Economic Review 90*(4), 715–41.

Grembi, V., T. Nannicini, and U. Troiano (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics 8*(3), 1–30.

Hansen, S., M. McMahon, and A. Prat (2017). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*.

Hinnerich, B. T., E. Höglin, and M. Johannesson (2011). Are boys discriminated in swedish high schools? *Economics of Education review 30*(4), 682–690.

Hinnerich, B. T., E. Höglin, and M. Johannesson (2015). Discrimination against students with foreign backgrounds: evidence from grading in swedish public high schools. *Education Economics 23*(6), 660–676.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74–91.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The review of Economic studies 66*(1), 169–182.

Jansson, J., B. Tyrefors, et al. (2018). Gender grading bias at stockholm university: Quasi-experimental evidence from an anonymous grading reform. Technical report.

Kraut, R. E., P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl (2012). *Building successful online communities: Evidence-based social design*. Mit Press.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of public Economics 92*(10), 2083–2105.

Lessig, L. (2006). Code: Version 2.0 nueva york.

Lööw, H. and L. Nilsson (2001). Hets mot folkgrupp. *BRÅ Rapport 17*.

Moore, M. J., T. Nakano, A. Enomoto, and T. Suda (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior 28*(3), 861–867.

Mullainathan, S. and J. Spiess (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

Nattrass, B. (2007). Decision making in committees: Transparency, reputation, and voting rules. *The American Economic Review 97*(1), 150–168.

Pettersson-Lidbom, P. and P. S. Thoursie (2013). Temporary disability insurance and labor supply: evidence from a natural experiment. *The Scandinavian Journal of Economics 115*(2), 485–507.

Postmes, T., R. Spears, and M. Lea (1998). Breaching or building social boundaries? side-effects of computer-mediated communication. *Communication research 25*(6), 689–715.

Prat, A. (2005). The wrong kind of transparency. *The American Economic Review 95*(3), 862–877.

Qin, B., D. Strömberg, and Y. Wu (2017). Why does china allow freer social media? protests versus surveillance and propaganda. *Journal of Economic Perspectives 31*(1), 117–40.

Reicher, S. D., R. Spears, and T. Postmes (1995). A social identity model of deindividuation phenomena. *European review of social psychology 6*(1), 161–198.

Strömberg, D. (2015). Media and politics. *economics* 7(1), 173–205.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior* 7(3), 321–326.

Van Royen, K., K. Poels, H. Vandebosch, and P. Adam (2017). "thinking before posting?" reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior 66*, 345–352.

von Essen, E. and J. Karlsson (2013). A matter of transient anonymity: Discrimination by gender and foreignness in online auctions.

A Appendix

A.0.1 Support vector machine

Table 16: Confusion matrix from SVM on hate

	Truth	
Predict	No hate	Hate
No hate	804	234
Hate	112	56

Table 17: Difference in number of entries between treated and control

	Dependent variable:				
	No. entries	No. hateful entries	No. hateful entries foreign	No. misogyny entries	No. entries by low-haters
	(1)	(2)	(3)	(4)	(5)
Post reveal	−772.596*	−57.231*	−95.249	50.596	−245.630
	(417.926)	(32.121)	(75.970)	(64.930)	(222.582)
Constant	−5,354.043***	−368.343***	−863.743***	−844.957***	−2,734.600***
	(198.811)	(12.722)	(34.765)	(35.453)	(116.587)
Observations	262	262	262	262	262
Collapsed on weeks	Yes	Yes	Yes	Yes	Yes

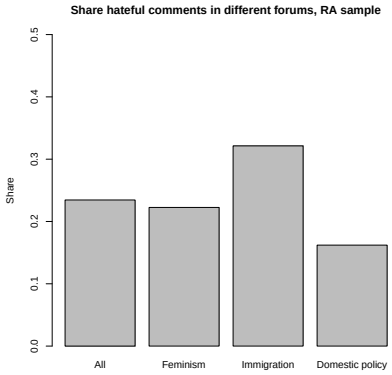
Note: *p<0.1; **p<0.05; ***p<0.01
The time period used in the regressions is from January 1st 2012 until December 31st 2016. Data is collapsed on a weekly level and the standard errors are computed using the Newey-West estimator with 4 lags. The outcome variables are each a time series of the difference between the early and late adopters.

Table 18: Simple correlations

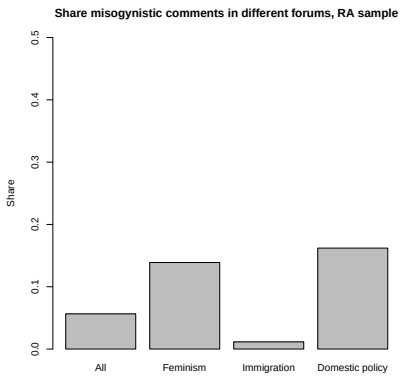
	Dependent variable:						
	Individual share hate	Hate		Misogyny	Individual share hate		Individual share misogyny
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No. individual entries	0.00001*						
	(0.00000)						
Misogyny		0.057***					
		(0.001)					
Hate foreign			0.234***	0.017***			
			(0.0005)	(0.001)			
Individual misogyny					0.062***		
					(0.003)		
Individual hate foreign						0.245***	−0.0002
						(0.003)	(0.004)
Constant	0.062***	0.062***	0.032***	0.133***	0.055***	0.026***	0.105***
	(0.001)	(0.0002)	(0.0002)	(0.0003)	(0.001)	(0.001)	(0.001)
Observations	48672	1984224	1984224	1984224	48672	48672	48672
Observations	48,672	1,984,224	1,984,224	1,984,224	48,672	48,672	48,672
Note:	*p<0.1; **p<0.05; ***p<0.01						

Figure 6: Forum shares, RA sample

(a)



(b)



(c)

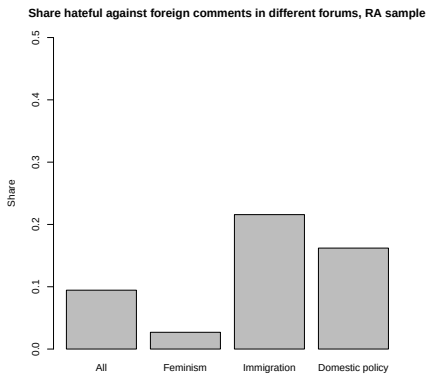


Table 19: Summary statistics, Domestic forum

	Total	Early adopters	Late adopters	Pre-event registered
No. entries	639768.00	101709.00	538059.00	592712.00
No. users	28765.00	2993.00	25772.00	26138.00
No. threads	10153.00	7335.00	10044.00	10083.00
Share hateful entries	0.05	0.05	0.05	0.05
Share hateful entries against foreigns	0.13	0.13	0.13	0.13
Share mysogynistic entries	0.04	0.04	0.04	0.04

Table 20: Summary statistics, Immigration forum

	Total	Early adopters	Late adopters	Pre-event registered
No. entries	887928.00	104897.00	783031.00	760274.00
No. users	35036.00	3200.00	31836.00	29801.00
No. threads	16179.00	11059.00	16040.00	15904.00
Share hateful entries	0.09	0.10	0.09	0.09
Share hateful entries against foreigns	0.22	0.23	0.22	0.22
Share mysogynistic entries	0.06	0.06	0.06	0.06

Table 21: Summary statistics, Feminism forum

	Total	Early adopters	Late adopters	Pre-event registered
No. entries	456528.00	36998.00	419530.00	417488.00
No. users	20435.00	1782.00	18653.00	19905.00
No. threads	5810.00	3817.00	5792.00	5785.00
Share hateful entries	0.05	0.04	0.05	0.05
Share hateful entries against foreigns	0.08	0.07	0.09	0.08
Share mysogynistic entries	0.42	0.44	0.42	0.42

Table 22: Marginal effects from logistic regressions

	Share hate	Share hate foreign	Misogyny
Post*Early	-0.013	-0.025	0.037
	(0.006)	(0.009)	(0.018)
Post reveal	-0.002	-0.008	-0.02
	(0.002)	(0.004)	(0.006)
Early adopters	0.007	0.016	-0.042
	(0.007)	(0.011)	(0.011)

Note: The time period used in the regressions is from January 1st 2012 until December 31st 2016. Standard errors clustered at the user startdate.

Table 23: Lasso coefficients for hate against gender

name	coefficient
(Intercept)	-2.95
alltm	2.96
avundsjukan	1.22
beatric	5.95
betrak	1.27
bottn	0.03
egotripp	0.06
feminism	0.14
feminist	1.33
fjortis	2.35
hor	8.81
klubb	3.29
kvinn	2.18
mental	1.66
oskyd	2.25

Table 24: Lasso coefficients for hate against foreign

name	coefficient
(Intercept)	-2.39
arab	3.08
blatt	2.05
fruar	4.27
förankr	0.54
gruppvåldtäk	2.87
intel	0.45
kameran	0.59
komplet	0.24
koranskol	2.19
krasch	2.08
käk	1.82
landskron	1.86
muslimsk	6.21
neg	1.48
negr	0.22
parasit	3.68
rån	7.87
separat	0.32
serb	2.04
svennehor	2.28
svensk	0.12

B Theoretical framework

To structure our empirical analysis of the relationship between anonymity and hateful content in social media, we use the logic of the principal-agent model by Ali and Bénabou (2016). It explicitly models anonymity (transparency) and reputation in a public goods setting. Contribution to the public good is writing in an honest and respectful manner, free from the pollution of hateful speech. For the purpose of our study, we will use the agents' equilibrium behavior to form empirical expectations. Below we describe in brief relevant parts of the model.

A single state or a general authority is concerned with all citizens having access to

honest and civil political discussions in social media, i.e. free from hateful content. A continuum of forum users (agents) $i \in [0,1]$ take part in the political discussions on social media, and can contribute to the public good by not subverting to hateful comments. Anonymity guides how much the other agents as well as the principal can view individual contributions.

A user's contributions depend on 1) an intrinsic preference of political debates free from hateful comments, v_i, 2) an individual signal/perception of the common value of a hate-free debate, θ_i, and 3) a concern for reputation μ_i (social image). Users care about other's beliefs about them, i.e. they wish to appear prosocial. The strength of the reputational concern varies across individuals, communities and time periods. User $\jmath$ estimates other users' reputation by using his own signal and reputational concern as well as the aggregate contribution $\bar{a}$. User $\imath$ incorporates how he will be judged by others, and makes contribution decisions thereafter. The user's contribution decision a_i at a cost $C(a_i)$ depends on his non-reputational pay-off and his reputational payoff[21]

$$\max_{a_i \in \mathbb{R}} \{E[U_i(v_i\ \theta_i\ w; a_i\ \bar{a}\ a_p)|\theta_i] + x\mu_i[R(a_i\ \theta_i\ \mu_i) - \bar{v}]\} \tag{11}$$

Anonymity, x, can change through an exogenous shock or the principal can choose the degree of anonymity. However, we will only investigate an exogenous change in anonymity.[22] Anonymity affects utility only through the reputational concerns, the risk of being exposed as producing hateful content. A user $\jmath$ observes another user $\imath$'s increase in contribution relative to the aggregate contribution, and knows it was motivated by a

[21]

$$U_i(v_i\ \theta_i\ w; a_i\ \bar{a}\ a_p) \equiv (v_i + \theta)a_i + (w + \theta)(\bar{a} + a_p) - C(a_i) \tag{9}$$

$$R(a_i\ \theta_i\ \mu_i) - \bar{v} \equiv E_{\bar{a},\theta\_i,\mu\_i}\Big[\int_0^1 E[v_i|a, \bar{a}, \theta_j, \mu_j]\mathrm{d}j\ |\theta_i, \mu_i\Big] \tag{10}$$

(a_p is the contribution by the principal.)

[22] A current debate concerns state regulation of anonymity on the Internet by, for example, requiring retention of information (Froomkin, 2017).

high intrinsic motivation, a high social signal of hate-free debates or a high concern for reputation. With linear strategies, there is a unique equilibrium ($x \geq 0$), the expected returns to social image are the same for all users, despite them having different signals of the common value of hate-free discussions in social media and a different strength of reputational concern. When there is no variation in reputational concerns across individuals ($s_\mu^2 = 0$), the marginal returns to reputation become a value, implying that a decrease in anonymity increases aggregate contribution one-to-one. If individuals vary in their reputational concern ($s_\mu^2 > 0$), the signal of an individual writing in a less hateful way becomes less informative. The behavior could be due to reputational concerns. Then, the marginal return to image concerns $\xi(x)$ decreases when anonymity decreases. This, in turn, will have less than a one-to-one impact on aggregate behavior. See the original article for the derivation of the marginal return to reputation.

Changes in anonymity thus have opposing effects. On the one hand, less anonymity will make users write less hateful comments, on the other hand, it affects how users judge each other's changes in production of hateful content, thus affecting the returns to reputation. If users suddenly get a better view of how much hateful content each user produces (decreased anonymity), it is also less clear whether a change in hateful writing is due to a high reputational concern or a high value of hate-free debate. When anonymity increases, agents' decisions become more driven by the variation in reputational concern than the signals about the common social value of the public good. The signal about the common value of the public good becomes noisy, which depresses the marginal return to reputation.

C Instructions to the RA

Instructions to Research Assistant Spring 2017.

The population we investigate is from the Internet forum Flashback. We have scraped text from the following three sub-forums; immigration, feminism and domestic policy. We have drawn a random sample of 100 threads from each forum. Each thread and each post has an id-number. We want you to code 12 posts starting from the beginning of the thread and 5 posts starting from the end of the thread.

The unit of coding is the post. Please read the full post. You will receive the threads and posts in an Excel sheet, where we want you to insert your classifications. Below you can find descriptions of how we want you to classify the posts.

Start with 2 threads and after this we can meet to discuss the progress before you proceed.

	Responds to Svarar på	
	0 = Doesn't seem to respond to any particular post 999 = Response to several posts from several authors [tomt] = Responds to a post that's not in the sample	*The value noted here is the id-number of the post to which the writer seems to respond.*
1	Questioning Ifrågasättande	
	0 = Neither nor 1 = Affirmative 2 = Nuancing 3 = Questioning	*- If the post **quotes another post** the coding relates to the quoted post.* *- If the post **doesn't contain a quote** the coding relates to the first post.* *- The **first post is always coded** as "neither-nor"*
2	Understanding Förståelse	
	0 = No 1 = Yes	If the writer shows understanding of the thoughts and intentions expressed in an earlier post, the coding should be "yes". **Regardless of whether the writer agrees or not.**
3	Party politics positive. *Does the post express a opinion in favor of any party or coalition of parties or Feminist Initiative?*	
	0 = No, not positive to any coalition of parties 1 = Yes, the red-green coalition 2 = Yes, the liberal-conservative coalition 3 = Yes, the Sweden Democrats 4 = Feminist Initiative 5 = Feminist initiative and the Left Party 6 = The seven traditional parties 7 = Sweden Democrats and the Right.	***Only to be coded "yes" if it is obvious,** e.g. when the parties or their representatives are mentioned, either explicitly or through paraphrases*
4	Party politics negative. *Does the post express a **negative** opinion of any party or coalition of parties or Feminist Initiative?*	
	0 = No, not positive to any coalition of parties 1 = Yes, the red-green coalition 2 = Yes, the liberal-conservative coalition 3 = Yes, the Sweden Democrats 4 = Feminist Initiative 5 = Feminist Initiative and the Left Party 6 = The seven traditional parties 7 = Sweden Democrats and the Right.	***Only to be coded "yes" if it is obvious,** e.g. when the parties or their representatives are mentioned, either explicitly or through paraphrases*

	Aggressiveness (tone)		
	0 = Not at all aggressive 1 = Partly aggressive 2 = Predominantly aggressive.	*In what **tone** is the post as a whole written?* *- If **some part** of the post is read as aggressive, it should be coded **partly aggressive**.* *- If the post contains **mostly aggressive** text, it should be coded **predominantly** aggressive.*	
	HATRED HAT		
6	Another Flashback user	0 = No 1 = Yes	*If the post contains **words or statements** that indicate persecution (in the broad sense of the term) of **a group or an individual,** it should be coded "yes", Possible examples are:* *- threat* *- expressions of disrespect* *- insults* *- verbal violations* ***Use the coding "yes" also for isolated hateful statements. It doesn't have to be blatant.***
7	Specific public person		
8	Persons with specific sex/gender		
9	Persons who were born abroad, or whose parents are born abroad		
10	Persons with a specific ethnicity		
11	Persons with a specific sexual inclination		
12	Persons with specific skin color		
13	Something else		
14	If the hatred is pointed toward something else, please specify	Text	
	HOT		
15	Another Flashback user	0 = No 1 = Yes	*Does the post contain words that are explicit threats or assault? Assault means that someone threatens to **harm an individual or his or her property.** The assault **can be directed to other persons, animals or objects** that are important to the individual.*
16	Specific public person		
17	Persons with specific sex/gender		
18	Persons who were born abroad, or whose parents are born abroad		
19	Persons with a specific ethnicity		
20	Persons with a specific sexual inclination		
21	Persons with specific skin color		
22	Something else		
23	If the threat is pointed toward something else, please specify	Text	
24	Male preference		
	0 = No 1 = Yes	If the post contains **words that in any way state the superiority of men** over women, it should be coded "yes"	
25	Female preference		
	0 = No 1 = Yes	If the post contains **words that in any way state the superiority of women** over men, it should be coded "yes"	
26	Gender equality preference		
	0 = Nej 1 = Ja	If the post contains **words of men and women being equal**, it should be coded "yes"	
27	Attitudes towards foreigners		
	0 = No opinion 1 = Positive attitude 2 = Neutral attitude 3 = Negative attitude	By foreigners is meant people who are **born abroad or whose parents were born** abroad.	

28	Gender - disadvantaged	
	0 = No opinion 1 = Men are disadvantaged 2 = Women are disadvantaged	If the post contains words that express women as disadvantaged or men as disadvantaged, it shall be coded as 1 or 2 respectively.
29	Ethnicity - disadvantaged	
	0 = No opinion 1 = Swedes are disadvantaged 2 = Immigrants are disadvantaged	If the post contains words that express Swedes as disadvantaged or immigrants as disadvantaged it shall be coded as 1 or 2 respectively.
30	Us and Them	
	0 = No 1 = Yes	If the post explicitly contains a language of ”us and them” or clearly expresses an in-group out-group view the variable should be coded yes.
31	Sarcasm or irony	
	0 = No 1 = Yes partly 2 = Yes, fully	If the post contains sarcasm or irony in part or in full it should be coded yes.

Gender grading bias at Stockholm University: Quasi-experimental evidence from an anonymous grading reform*

By Joakim Jansson[†] and Björn Tyrefors[‡]

Abstract:

In this paper, we first present novel evidence of a grading bias against women at the university level. This is in contrast to previous results at the secondary education level. Contrary to the gender composition at lower levels of education in Sweden, the teachers and graders at the university level are predominantly male. Thus, an in-group bias mechanism could consistently explain the evidence from both the university and the secondary education level. However, we find that in-group bias can only explain approximately 20 percent of the total grading bias effect at the university level.

Keywords: grading bias, university, discrimination, education, anonymous grading
JEL: I23, I24, J16, J71, D91

* We thank Handelsbanken's Research Foundations for generous financial support. We thank Karin Blomqvist and Peter Langenius for supplying us with parts of the datamaterial. We thank Per Pettersson-Lidbom, Mahmood Arai, Jonas Vlachos, Peter Skogman Thoursie, Fredrik Heyman, Joachim Tåg, David Neumark and Lena Hensvik, seminar participants at Stockholm University and The Research Institute of Industrial Economics, participants at SUDSWEC 2015 and at the 2nd Conference on Discrimination and Labour Market Research for fruitful comments.

[†] Department of Economics, Stockholm University, Research Institute of Industrial Economics, Stockholm Internet Research Group, Stockholm University

[‡] Department of Economics, Stockholm University, Research Institute of Industrial Economics

1 Introduction

Biased grading has recently received increasing attention in economics. This literature is generally motivated by the growing gender gap in educational attainment and the sorting of males and females into specific fields.[1] However, previous studies on grading bias have focused on pre-tertiary education levels and have typically found a bias against males or no effect.[2] The teaching profession has been increasingly staffed by women, which has been proposed as one mechanism explaining the grading bias against boys, through, for instance, so-called in-group bias.[3] A related strand of literature has thus focused its attention on how having a teacher of the same gender or ethnicity affects students' grades and performance.[4] In contrast to the lower education levels, a large majority of university teachers are male. Therefore, a study of grading bias at the university level could inform us about the role of both institutional culture and in-group bias as mechanisms. Furthermore, there are not, to our knowledge, any large-scale studies based on quasi-experimental methods evaluating gender grading bias at universities.[5]

This study aims at filling this gap by making two main contributions: documenting the effect of an anonymous grading reform at the university level and then credibly estimating and quantifying how much of the effect that can be explained by having your examination corrected by someone of the same gender as yourself. For this purpose, we combine two unique data sets with two related experimental designs. In both cases, we make use of an exam reform at Stockholm University, where all standard exams had to be graded with no information about the exam-taker's identity. This reform was put in place at the beginning of the fall term of 2009. Using a difference-in-difference-in-difference design, we first find a positive effect of the anonymous grading reform on the test results of female students. Thus, consistent with the findings of the work by Goldin & Rouse (2000), for example, being evaluated anonymously causes

[1] See, for instance, Lavy and Sand (2015), Kugler et al. (2017) and Terrier (2015) for evidence on educational sorting and grading bias.
[2] See, for instance, Lavy (2008) or Hinnerich et al. (2011). The exception is Breda and Ly (2015) who, however, focus on how the effect varies with the male domination of a field and not the general effect.
[3] The phenomenon that people tend to favor other people of their own group is usually referred to as in-group bias effects. See, for instance, Sandberg (2018).
[4] See, for instance, Dee (2005, 2007), Lee et al. (2014), Lusher et al. (2015), Feld et al. (2016) and Lim and Meer (2017a).
[5] A pilot study on parts of the sample was undertaken by Eriksson and Nolgren (2013) under the supervision of Björn Tyrefors Hinnerich.

improved evaluations for females. In fact, the pre-reform gender gap in grades appears to be closed by the reform. We argue that this is likely explained by a gender bias in grading.[6] These findings are consistent with the fact that there are more male graders at the university level in contrast to lower academic levels, accounting for the reversed sign compared to what is found in the studies at lower levels. To test for this explanation, we make use of a second experiment. By using a particular exam, namely, the introductory exam in macroeconomics, we can collect more detailed information on grader gender and, more importantly, we can utilize a nonintentional randomized experiment setting for this exam, in which graders of different genders were randomly assigned to correct different questions. First, we also confirm in the subsample a negative bias effect against females, similar to the specification used in the full sample. Then, by random assignment of the gender of the grader, we can estimate the causal effect of same-sex bias among correctors and quantify how much of the total effect it constitutes. We find strong evidence of same-sex bias in the TAs' corrections of exams. Furthermore, this bias disappears once anonymous exams are introduced at the university, showing the efficiency of the policy of name removal on the exam. However, in-group bias accounts for only approximately 20 percent of the total effect, indicating that the bias is mainly determined by factors other than graders simply favoring their own gender.

There is an increasing number of studies investigating the different dimensions of grading bias at the pre-tertiary levels. As a whole, there are two strands in this literature. First and foremost, there are studies investigating the general gender grading bias of teachers, where test scores are compared across anonymous and non-anonymous exams. Lavy (2008) looks at the gender bias in Israeli matriculation exams in nine subjects among high school students. Using a difference-in-difference approach, he finds evidence of a bias against male students. The size of the effect varies to some degree between different subjects and depends on teacher characteristics. A similar approach is taken by Hinnerich et al. (2011,

[6] Even though we are estimating the causal effect of the anonymous grading reform, we can never be certain that the outcome is *only* due to grading bias. In fact, we can think of a situation where the behavior of the students changes, where they could, for example, start to exert more effort as a consequence of the reform. However, we share this drawback with many prominent studies in the field based on anonymous and non-anonymous observations of the outcome.

2015), where the bias of both gender and foreignness is studied. Related to this is also Sprietsma (2013) who compared the grades given for the same essay with either a German- or Turkish-sounding first name. She finds that essays believed to be written by Turkish students receive significantly worse grades. Kiss (2013) studies the grading of immigrants and girls once test scores have been taken into account and finds a negative impact on immigrants' grades in primary education. Furthermore, girls are graded better in upper-secondary school. Lindahl (2007), on the other hand, finds that male test scores increase with the share of male teachers, whereas grades decrease at the same rate. Second, there are studies looking at more reduced form effects of having a male or female teacher depending on your own gender. Most notable is probably Dee (2005) who looks at the effect of having a teacher of the same gender or ethnicity as yourself in eighth grade and finds a positive effect. A similar approach is taken in Dee (2007); however, more long-run and behavioral responses were instead considered. It is worth noting that none of these studies are at the university level.

However, there are a few studies using the university as a testing ground. Closely related is Breda and Ly (2015). They use oral (non-blind) and written (blind) entry-level exams at elite universities in France and find that females' oral performance is graded better than that of males in more male-dominated subjects. Our setting differs in many ways in addition to scale. We make use of a change in policy over time and the examiners in our study are typically the teachers of the students and not external examiners as in Breda and Ly (2015). Thus, we study standard examination at a large (approximately 30 000 students per year) state-financed non-selective university. Lastly, and most importantly, they are estimating a non-linear model (an interaction model) and the overall average effect may still be consistent with our finding.[7] Additionally, the effect of teachers as role models at the university level is investigated in Hoffmann and Oreopoulos (2009). Still other papers look at how the classroom gender composition affects student performance (Lee et al., 2014), how the matching of TA/teacher and student ethnicity/gender affects their performance (Lusher et al., 2015; Lim and Meer, 2017a; Lim and Meer,

[7] Breda and Ly (2012) show an overall grading bias effect of the same sign and size as ours. See also Breda, and Hillion (2016).

2017b; Coenen and Van Klaveren, 2016) and whether biased grading seems to be driven by favoring your own type (endophilia) or by discriminating against other types (exophobia) (Feld et al., 2016).

The rest of the paper is organized as follows. Section 2 describes the two empirical strategies and data sets that we use, section 3 presents the results and section 4 concludes the paper.

2 Data and empirical designs

2.1 Data

Both of our designs are based on a reform that forced a removal of the test-taker's identity on standard exams from the start of the fall term of 2009. For our first design, we use the fact that other graded activities, such as thesis, oral and home assignments, were not anonymized for practical reasons. All departments except the Department of Law were affected, but only because the Department of Law already had a long-standing practice of anonymous grading.[8] Thus, all examinations at the Department of Law and activities such as thesis work, oral and home assignments at other departments serve as a control group in a difference-in-difference-in-difference design. This design uses the universe of grades at Stockholm University from the fall of 2005 to the spring of 2014.

Our second design makes use of a particular exam where we instead hand-collected more detailed information. This approach creates an opportunity to evaluate the importance of in-group bias, as the graders were randomly allocated to questions by ballot. We employ data from the macroeconomics exam for the introductory course at Stockholm University from the spring of 2008 to the fall of 2014. In addition to the random assignment of teachers, the design is again based on the reform that forced a removal of the test-taker's identity on standard exams from the start of the fall term of 2009. However, the control group differs in this case. The introductory exam consists of two multiple-choice questions

[8] Table A2 shows that our results are robust when excluding the Department of Law.

and seven essay-style questions, each worth ten points.[9] As multiple-choice questions only have one correct answer, it is impossible or at least very costly for the grader to grade with a bias. Thus, for this design, the multiple-choice questions serve as the control group and the essay questions as the treatment group.

For brevity, we will, in the empirical specifications, define exam and essay questions as *treated* and thesis, oral, home assignments, exams at the Department of Law and multiple-choice tests as *control*. We are fully aware that our different assessment types may measure different skills and hence the difference between treated and control in a given cross section will not be informative with respect to grading bias as we would compare apples and oranges. Fortunately, we can make use of the time dimension and the policy intervention in a difference-in-difference-in-difference setting. Consequently, the control and treatment test types may well be measuring different skills without posing a threat to internal validity. The important assumption will be that for each assessment type, treatment and control, the difference in test scores between sexes should move in parallel over time in the absence of anonymization, a regularity that can be partly tested by estimating pre-trends across series.

2.1.1 Data from all graded activities at Stockholm University

In the relevant time period, there were three main grading systems in place: the original, consisting of G (pass), VG (pass with distinction) and U (fail); a special grading scheme implemented for most of the courses at the Department of Law, consisting of AB (highest), BA (middle), B (lowest) and U (fail); and finally the system imposed by the Bologna process in the European Union. The Bologna scheme had to be implemented from the fall of 2008 at the latest, although it was used at certain departments and courses before that deadline. However, the Department of Law still has an exception to this rule. The Bologna

[9] This is, however, only true up until the fall term of 2013, after which the multiple choice questions need to be answered to take the exam and the essay questions are each worth twelve points.

system uses the letters A through F, where A is the highest grade and F (along with Fx) is a fail.[10] The numeration of these different systems is given in Table A1, while Figures A2-A4 provide the histograms for each of them. The histogram plots in all look quite normally distributed, except for the grades in the Department of Law.[11] To make the different grading systems comparable, we standardized each of them separately by subtracting the mean and dividing by the standard deviation. [12]

We collected data on all grades at Stockholm University in the period from the fall of 2005 up to the spring of 2014, recorded in the administrative system Ladok.[13] Our data contain information on the date of the exam, the course, the course credits, and the responsible department, as well as basic information on the individual taking the exam. Summary statistics are provided in Table 1. Table 1, Panel A shows the data for all graded activities, and we find that there is a majority of female students for these activities (63 percent) and that students are on average 28 years old.

The data do not explicitly document whether it was a written exam (graded anonymously after the fall of 2009) or not. To identify examination forms that were still not anonymous after the introduction of the reform, we made use of the fact that graded activities come with a text-based-name indicating the type of examination. For example, a bachelor's thesis grade comes with a text stating "thesis." Since theses and term papers are never anonymously graded, as the name is written on the front page, we coded them as being non-anonymous. Other examination forms that can never truly be anonymously graded are lab assignments and different types of presentations requiring physical attendance. We define non-anonymously graded activities by searching through the column of text indicating examinations of these

[10] Table A2 shows that our results are robust when excluding different grading schemes.

[11] Anecdotal evidence suggests that the Department of Law strives for normally distributed grading on the main exam as well as for retakes, which could explain why the distribution does not look normal. However, it could also be because being accepted as a law student requires quite high grades starting in high school. Other anecdotal evidence suggests that students always received the highest grade on their final thesis up until recently (dropping all observations classified as thesis at the Department of Law does not change our results).

[12] It is important to note that although all departments had to adopt the new grading scheme by the start of the fall 2008, some students still received grades from the old system (i.e. VG-U) after that point. This is for two reasons, the first being that certain parts of courses are still either awarded a pass (G) or a fail (U), typically seminars requiring attendance or hand-in assignments. However, if a student first got registered in a course when the old grades were still in use at that department, failed first and then passed it later on when the new A-F grades had been introduced, that student would still be awarded a grade from the VG-U scale.

[13] We should note that we drop the department "Lärarutbildningskansliet" since it was not a formal department over the full period and was affected by massive reforms.

types. For example, if the word “thesis” or “home assignment” is found, that activity is coded as non-anonymous. We thus obtain a dummy indicating whether we know that tests are always non-anonymous even from the fall of 2009 and onwards. Then, we combine this with all examinations from the Department of Law, which were either anonymous or non-anonymous throughout the entire period.[14] Table 1, Panel A shows that in fact 77 percent of the activities are classified as affected (treated) by the reform. Thus, our treatment group of interest will be residually determined and hence will have a potential measurement error by misclassification.[15] However, this would imply that we, if anything, are underestimating the true effect.[16]

[14] For the entire coding, contact us for the code-file (Stata).

[15] The misclassification problem when using the full population is also one motivation for why we subsequently focus on the data set from the Department of Economics since treated and non-treated are clearly categorized in that setting. Furthermore, we can use a more precise outcome since we observe the students’ score on each question, which varies between 0 and 10.

[16] The logic behind this is simple: since we determine treatment status residually, we will likely classify some of the in fact not treated as being treated. Hence, our treatment indicator will capture some of the effect of the non-treated, thus biasing our estimates towards zero. This is usually referred to as classical measurement error and attenuation bias.

Table 1: Summary statistics

	(1) Mean	(2) S.D.	(3) Min.	(4) Max.
Panel A: Full sample				
Female student	.6276207	.4834388	0	1
Age	28.22348	8.983703	16	88
Thesis and hand-ins	.169309	.3750247	0	1
Department of Law	.0652571	.2469791	0	1
Treated	.7678218	.4222222	0	1
Autumn 09	.5714777	.4948647	0	1
Observations	1856027			
Panel B: Introductory macroeconomics sample				
Female student	.4882662	.4998672	0	1
Female teacher	.3250679	.4684048	0	1
Same sex	.4992282	.5000043	0	1
Fall 09	.7982883	.401282	0	1
Retake	.2091369	.4066963	0	1
Age of student	23.23196	4.156114	18	71
Observations	51177			

2.1.2 The introductory macroeconomics sample

The data on student performance were collected from the course administrator and the course coordinator. The main benefit of the introductory exam is that it consists of two multiple-choice questions as well as seven essay questions, each worth ten points—that is, up until the fall term of 2013, after which the essay questions were worth 12 points and the multiple-choice questions were a prerequisite for eligibility to take the exam.[17] Since we know which questions are multiple choice, we have no measurement error in this sample. One additional benefit of this setting is that each of the 7 essay questions was corrected by a separate TA. Furthermore, the TAs were assigned to the specific questions by ballot, thus creating a nonintentional experiment.[18] The first names of the TAs were collected from the course coordinator's correction templates and then typed into a spreadsheet by hand. These numbers were then merged together. Table 1, Panel B provides some key characteristics of the collected data. As can be seen, both

[17] Details regarding the exam and the process that underlies the correction are described in the Appendix.

[18] The exam also contains an 8th essay-like question that the typical student do not have to answer due to a credit system. Hence, these questions are excluded from the analysis.

same-sex and female students correspond to around half of the sample, while most exams are from the anonymous period and most TAs are male. Hence, if male students performed better than female students on average, we would overestimate a positive in-group bias effect simply because the majority of TAs are male. Thus, it is necessary that we condition on the female students' average score in both the pre- and post-anonymization periods when we estimate the in-group bias effect.

Since we have collected both the gender of the students and the name (and thus the gender) of the TAs assigned to each question, these exams provide an optimal setting for studying possible same-sex bias effects. More specifically, the randomization of TAs to questions ensured that there is no selection by gender or ability into questions of different difficulty levels. It is thus possible to compare one student's score on each question depending on whether the corrector is of the same gender or not, as long as we condition on the average performance of each gender in order to avoid including general gender discrimination in our estimates.[19]

2.2 Empirical designs

2.2.2 The effect of anonymization on gender differences

Our two designs are based on the same reform that forced a removal of the test-taker's identity on standard exams from the fall of 2009. Thus, we can formulate an empirical model similar to a difference-in-difference-in-difference (Katz (1996), Yelowitz (1995)):

(1) $$testscore_{ijt} = \delta_0 + \delta_1 female_{i,j,t} * fall\ 09_t * treated_{j,t} + \delta_2 fall\ 09_t * female_{i,j.t} + \delta_3 female_{i,j.t} * treated_{i,t} + \delta_4 fall\ 09_t * treated_{i,t} + \delta_5 fall\ 09_t + \delta_6 female_{i,j.t} + \delta_7 treated_{i,t} + \varepsilon_{i,j,t}$$

[19] It is important to note here that the gender of the corrector is unknown to the student at the time the exam is taken.

On test/question-type *j*, for individual *i*, in time period *t*, *treated* is an indicator taking the value one if it is an exam/essay question and zero if it is a thesis/multiple-choice question, *fall 09* is a dummy for the time periods taking the value one when anonymization was implemented in the fall of 2009, and *female* is a gender dummy. The coefficient of interest is δ_1, which measures the effect of anonymization on female grades compared to male grades. However, as our treatment is varying on test type level, there are typically small gains from using disaggregated individual data and (in the absence of compositional effects) we could equivalently use aggregated data (Angrist and Piscke, 2009) and the identity of δ_1 by estimating:

(2) $Y_{jt} = \xi + \varsigma treated_j + \omega fall\ 09_{\ t} + \delta_1 treated * fall\ 09_{jt} + \kappa_{ijt}$,

where $Y_{jt} = \overline{testcscore}_{jt}^{women} - \overline{testcscore}_{jt}^{men}$, the difference in group means. From this it becomes clear that the identifying assumption is now that the difference in test score between sexes should move in parallel in the absence of anonymization across the two test types, as this formulation can be viewed as a standard difference-in-difference specification. Under the identifying assumption of parallel trends of the difference of group means in the absence of anonymization, we will estimate δ_1 with no bias, and it will be the causal effect of anonymization on female grades compared to male grades. To test this identifying assumption, we will estimate time separate treatment effects over time in accordance with Angrist and Pischke (2009).

Moreover, we acknowledge that the estimations of the standard errors are problematic in our study since treatment only changes once for one group (standard written exams), as discussed by Bertrand, Duflo and Mullainathan (2004), Donald and Lang (2007) and Conley and Taber (2011). We begin by clustering them at the student level. However, since treatment only varies once at the control-treatment group level,

this might not be conservative enough. Here, we follow the Pettersson-Lidbom and Thoursie (2013) application of the results in Donald and Lang (2007) and aggregate data to the group level and estimate a time series model with a structural break. For this purpose, we note that we can write $\Delta Y_t = \bar{Y}_{jt}^{treated} - \bar{Y}_{jt}^{control}$, giving us:

(3) $\Delta Y_t = \gamma + \delta_1 fall\,09_t + \pi_t$

Thus, we aggregate equation (2) once more and estimate a time series model on the differences in test scores of men and women across test types and use standard errors robust to heteroscedasticity and serial correlation by applying the Newey-West estimator with one lag.

2.2.3 In-group bias

In-group bias in our context is the inclination of teachers to give superior grades to those who belong to the same group with which they identify. Part of a gender bias in grading could be culturally determined irrespective of the gender of the grader, but another mechanism could be in-group bias. At Stockholm University, a majority of the teachers are men and hence, in-group bias could potentially explain all grading bias. The main benefit of the data from the introductory macro course is that we have a randomization of graders and hence of the gender of the grader on each question. However, unfortunately, we do not observe any gender for the corrector of the multiple-choice questions, leaving us with no control contrast in that test type dimension. Hence, to consistently separate the in-group bias effect from the total effect of anonymization, we need to be able to estimate the total effect relying on a before-and-after design. For our two designs to be comparable, we need the difference in gender ability in exam performance to be constant from the control to the treatment period. This corresponds to $\delta_2 = 0$ in

equation (1) or $\omega = 0$ in equation (2). Under this condition, we can consistently estimate the gender difference of the effect of anonymity by a regression corresponding to equation (4).[20]

(4) $testscore_{ijt} = \delta_0 + \delta_1 female_{i,j,t} * fall\ 09_t + \theta_2 female_{i,j.t} + \theta_3 fall\ 09_t + \varepsilon_{i,j,t}$

Moreover, we can then separate the in-group bias effect from the total effect by using the following regression equation:

(5) $testscore_{i,t} = \lambda_0 + \lambda_1 female_{i,t} * fall\ 09_t + \lambda_2 fall\ 09_t + \lambda_3 female_{i,t} + \lambda_4 I(same\ sex_{i,j,t}) + \lambda_5 I(same\ sex_{i,j,t}) * fall\ 09_t + \varepsilon_{i,j,t}$

where $I(same\ sex_{i,q,t})$ is an indicator function for cases in which the student answering the question and the TA correcting it have the same gender. Thus, we can also observe if any potential in-group bias disappears after the introduction of anonymous grading. With regard to the standard errors in this specification, we make use of a two-way cluster at the individual and TA level.

3 Results

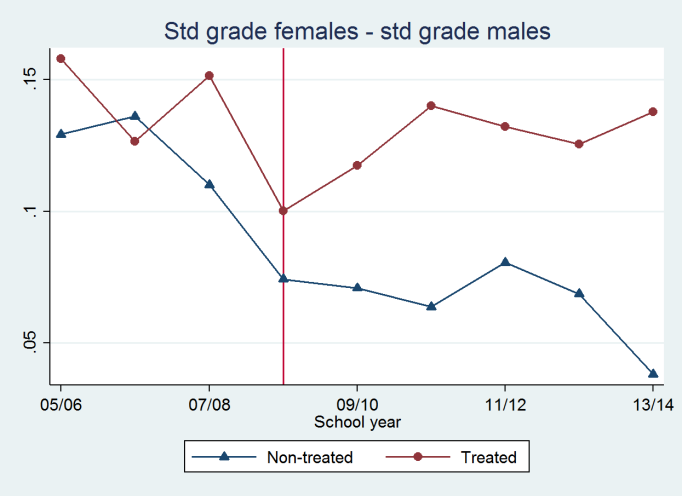
3.1 Results from the full sample

In this section, we present our results for gender bias at the entire Stockholm University. Since the underlying assumption is the parallel trends assumption, we begin by plotting the difference in standardized grade between genders as two time series (Figure 1, Panel A). Moreover, as proposed by Angrist and Piscke (2009), we also plot annual "treatment" effect estimates from a regression both before

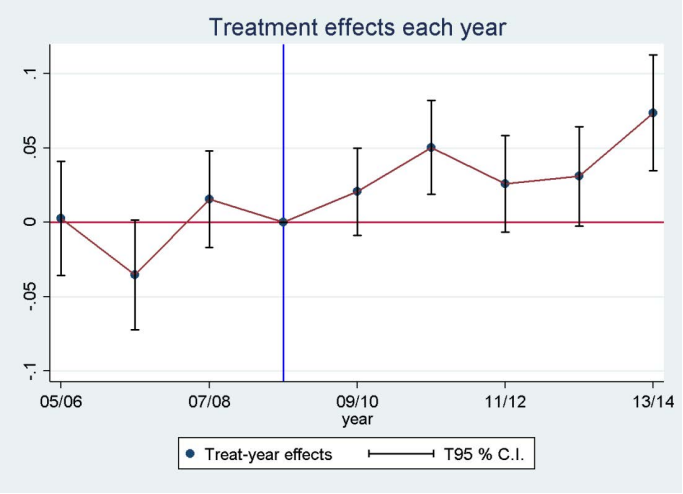
[20] This is verified in a simple simulation exercise in Stata in the file generating the main results and formally shown in the Appendix.

and after the implementation of the reform showing "placebo" effects before the reform and dynamic causal effects after the reform (Figure 1, Panel B). The results are presented in Figure 1. Panel A displays the difference between men and women for the control and treatment test types over time. We can see that prior to the reform, both series exhibit a similar negative trend which continues after the reform for the non-treated group. However, for the treated group, we see that in the post-reform period, the decline is halted or perhaps even reversed. Panel B then in turn plots the coefficient for the treatment effect over time. The estimates are fairly stable around zero in the pre-treatment period, and then increase in the post-treatment periods, with estimates being consistently positive in contrast to the pre-treatment period. Hence, we conclude that even if the test types may be measuring different skills, we show strong evidence that this does not pose a threat to interval validity as the parallel trends assumption seems likely to hold.

Fig. 1:
Panel A: Difference in levels between treated and non-treated groups.



Panel B: Annual differential effects across the treated and non-treated groups.



Note: Standard errors clustered at the individual level.

We continue with our regression results, which are presented in Table 2. Column 1 corresponds to a regression equivalent to equation (1). We can see that the anonymous examination raises female grades

relative to male grades by approximately 0.043 of a standard deviation. We further note that for males, there is a decrease in grades for exams by 0.0426 (Column 1, row 2) while for female students there is an additional effect of an increase discussed above of 0.043. Thus, female student test scores on exams remain about the same, while those of male students decrease (relative to the development of the control test type). Overall, this points at an average decrease of the test score of exams of about 0.02 of a standard deviation due to anonymous grading.[21]

Column 2, in turn, presents the results from a regression on the collapsed time series data. We can note that the standard error is essentially unchanged as compared to the first column. Moreover, aggregation leaves the estimate unchanged, which makes it likely that any compositional bias is of little importance. In column 3, we then include nonparametric gender and exam specific trends.[22] This is possible thanks to the DDD-like identification design. The estimate decreases slightly, though it is still close to the coefficient in column one. Hence, the results in this column further back up the credibility of our design, as the estimated effect does not seem to be driven by unobserved trends. Finally, column 4 once more runs a regression corresponding to equation (1) but this time it uses the number of course credits as weights, thus giving more weight to more important examinations. This increases the coefficient slightly, indicating that the effect is bigger for more important examination forms.

Additional robustness tests are performed in Table A2 in the Appendix. The first column replicates column 1 in Table 2, column 2 entirely excludes the Department of Law from the analysis and columns 3 and 4 restrict the analysis to A-F grades during their mandatory period. All these restrictions increase the coefficient slightly. Finally, column 5 alternates the numbers from Table A1 such that B for law students is 1, BA is 2 and AB 3, while G is 1 and VG is 2. Reassuringly, this does not change any estimate at all, since we standardize each grading scheme.

[21] That anonymous grading is more conservative is often found in the literature (e.g. Hinnerich et al., 2011). The calculation of 0.02 is based on the assumption of 50 percent females students. In fact, a DID-estimate on the effect of anonymization yields 0.018.

[22] In other words, we include gender*month fixed effects and treatment group (exams or papers)*month fixed effects.

Table 2: Gender grading bias effects. Full sample.

	(1) Stand. score	(2) Stand. score	(3) Stand. score	(4) Stand. score
female*treated*fall 09	0.0430***		0.0318***	0.0676***
	(0.0110)		(0.0110)	(0.0107)
treated*fall 09	-0.0426***		0.245***	-0.0319***
	(0.00853)		(0.0539)	(0.00839)
female*treated	0.0222**		0.0280***	-0.0186**
	(0.00935)		(0.00930)	(0.00868)
female*fall 09	-0.0488***		-0.0496	-0.0677***
	(0.00918)		(0.0582)	(0.00858)
treated	-0.123***			-0.154***
	(0.00725)			(0.00687)
female student	0.114***			0.155***
	(0.00793)			(0.00705)
fall 09	-0.0671***	0.0447***	0.147***	-0.0646***
	(0.00665)	(0.0114)	(0.0491)	(0.00608)
constant	0.0744***	0.0215**	-0.263***	0.0981***
	(0.00573)	(0.00733)	(0.0223)	(0.00498)
Month*gender FEs	No	No	Yes	No
Month*treated FEs	No	No	Yes	No
Course credits weights	No	No	No	Yes
Collapsed	No	Yes	No	No
Time period	Fall 2005-spring 2014	Fall 2005-spring 2014	Fall 2005-spring 2014	Fall 2005-spring 2014
N	1856027	9	1856027	1856027

Note: Standard errors clustered at the student level except in column 2. In column 2, Newey-West standard errors are used with one lag. The dependent variable is standardized score. Column 3 includes both month*gender and month*treated FEs as well as the interactions between gender*fall 2009 and treated*fall 2009 in order to saturate the model. Since we have data on the exact date of the exams, and since we define treatment to start on the official start date of the fall term in 2009, which is the 22nd of August, we need these additional interactions to saturate the model. Excluding them increases the coefficient of interest. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.2 How much of the aggregate effect can be attributed to in-group-bias? Results for the introductory macroeconomics sample.

Table 3 contains the main results from the sample from the introductory macroeconomics exam. Column one gives the result from a regression corresponding to equation (1), with the first row presenting the treatment effect. We can observe a slightly higher coefficient compared to the full sample of approximately 0.09 standard deviations. However, this is consistent with the fact that we have no measurement errors in the dependent variable and hence, no attenuation bias in contrast to the previous design. It is also worth noting that the "placebo coefficient", δ_2 in equation (1) and the second row in the table, is very close to zero and far away from significant at any level. This enables us to use a before-and-after design and still obtain an unbiased estimate of δ_1 in this setting. The result from such a regression is presented in column two for the same time period as in the first column. We can note that the coefficient is essentially unchanged at approximately 0.09 standard deviations and is still highly significant. The coefficients imply that before the anonymization reform, females performed approximately $1/10^{th}$ of a standard deviation worse than male students (the fourth row, δ_2 in equation 4), while after the reform, the scores of females increased by $1/10^{th}$ of a standard deviation (the second row, δ_1 in equation 4). The sum of these two coefficients is presented at the bottom of the table along with the *p*-value from a Wald test on whether their sum is equal to zero. One can note that the sum of the coefficients is close to zero and not significantly different, indicating that the gender difference in grades falls to zero once anonymous exams are introduced. Finally, the third column runs the same regression as column two but uses the entire available data for the macroeconomics sample, with a largely unchanged coefficient.[23]

[23] Figure A1 in the Appendix provides a similar graph as Figure 1 but uses the DDD-setting in the macroeconomics example.

Table 3: Gender grading bias effects. Introductory macroeconomics sample.

	(1) Stand. score	(2) Stand. score	(3) Stand. score
female*treated*fall 09	0.0849**		
	(0.0379)		
fall 09*female student	0.00883	0.0910**	0.103**
	(0.0426)	(0.0410)	(0.0402)
fall 09*treated	-0.149***		
	(0.0271)		
female*treated	-0.0708**		
	(0.0324)		
female student	-0.0410	-0.109***	-0.109***
	(0.0383)	(0.0366)	(0.0366)
treated	-0.421***		
	(0.0230)		
fall 09	0.0879***	-0.0598**	-0.0646**
	(0.0307)	(0.0278)	(0.0272)
constant	0.381***	0.0645***	0.0645***
	(0.0277)	(0.0246)	(0.0246)
Time period	Spring08-Spring13	Spring08-Spring13	Spring08-Autumn14
Sum treatments	0.0141	-0.0176	-0.00544
P-value	0.481	0.412	0.776
N	49700	39684	51177

Note: Standard errors clustered at the student level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4 then proceeds to investigate the importance of same-sex bias in the aggregate effect. The first column simply replicates the third column in Table 3 in order to make the comparison easier. The second column in Table 4 shows the estimation results when including an in-group bias variable corresponding to equation (5). We conclude that having a teacher of the same gender as yourself raises your points on that question by 0.04 standard deviations from the mean. Once more, similar to the main gender difference effect, this effect also goes back to zero as soon as anonymous exams are introduced. At the bottom of the table, the row "Sum treatments" gives the sum of the coefficients λ_4 and λ_5, i.e., the sum of the same-sex coefficients before and after anonymization, respectively. It can be seen that this estimate is close to zero. The row below this one then provides the p-value from a Wald test testing the hypothesis that $\lambda_4 + \lambda_5 =$

0, which cannot be rejected. Thus, a removal of the name from the exam seems to be sufficient to prevent both a general gender bias and a same-sex bias in correctional behavior. Since many suspect that content and handwriting style may also signal gender after the anonymization reform, this is indeed an interesting finding.[24] It is also of interest to analyze what happens to the aggregate gender bias when including the in-group bias variable. As can be seen, both the pre- and post-anonymization coefficients are altered by approximately 0.02. Thus, it seems as if part (approximately 20 percent) of the gender difference is due to in-group bias but not the entire effect. Column three then adds a dummy for retakes, while column four in turn adds question-specific fixed effects. The fact that the coefficients in essence are unchanged is reassuring in the sense that the randomization of TAs to questions seems to have worked.[25] Column five then adds gender-specific nonparametric trends, in other words, female student multiplied by the date of the exam fixed effects. This is to ensure that the estimated same-sex effects are not driven by any underlying trends in gender performance, at the cost of not being able to estimate the female student coefficients from the first two columns. Since the coefficients are essentially unchanged, we conclude that this does not seem to be a concern. This robustness to all controls should not be surprising, however, given the randomization of TAs to the questions.

[24] However, Breda and Ly (2015) demonstrate that female handwriting is not easily distinguishable from male handwriting.
[25] It is important to note here that the question-specific fixed effects are even more flexible and reliable than controlling for TA fixed effects.

Table 4: Results in-group bias

	(1) Stand. Score	(2) Stand. score	(3) Stand. score	(4) Stand. score	(5) Stand. score
fall 09*female student	0.103**	0.0863**	0.0815**	0.0889**	
	(0.0428)	(0.0388)	(0.0370)	(0.0359)	
female student	-0.109***	-0.0874**	-0.0879***	-0.0939***	
	(0.0380)	(0.0340)	(0.0322)	(0.0317)	
fall 09	-0.0646	-0.0410	-0.0276		
	(0.0928)	(0.0980)	(0.0944)		
same sex		0.0439***	0.0439***	0.0415***	0.0368***
		(0.0101)	(0.00971)	(0.0113)	(0.0128)
fall 09*same sex		-0.0302**	-0.0343**	-0.0330**	-0.0295**
		(0.0138)	(0.0143)	(0.0149)	(0.0143)
retake			-0.307***		
			(0.0499)		
Sum treatments $\lambda_4 + \lambda_5$		0.0137	0.00965	0.00850	0.00732
P-value $\lambda_4 + \lambda_5 = 0$		0.164	0.380	0.413	0.0839
Question FEs	No	No	No	Yes	Yes
Genderspecific trends	No	No	No	No	Yes
N	51177	51177	51177	51177	51177

Note: Standard errors clustered at the TA (49 clusters) and student (6 521 clusters) level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

To further investigate if the randomization was carried out properly, Table 5 uses the few background characteristics we have as outcome variables in a regression on TA gender. If the TAs are successfully randomly assigned to the questions, then the characteristics of the question as well as the student answering the question should be the same for both male and female TAs.[26] Thus, column 1 starts by comparing the question number between male and female correctors. If randomization did work, we should expect male and female correctors to answer questions with the same number to the same degree. Indeed, there is no significant difference between genders, with a very small coefficient. The second column then proceeds to look at the probability that a female TA corrects a female student's exam. If female TAs corrected questions that female students found easier to answer, we might see that female TAs were more likely to correct answers by female students, due to the fact that fewer females simply answered the questions corrected by male TAs. However, if anything, the reverse seems to be true as we find a small negative coefficient significant at the 10 percent level. Since the coefficient is so small, around 1 percent with a baseline of 49 percent, we argue that this is to be interpreted as a rather precisely estimated zero and will not cause any concern.

Next, the third column looks at the age of the answering student, following a similar reasoning as column 2. Once more, the coefficient is very small and indicates that female TAs correct questions by students who are 0.08 years younger, though the estimate is insignificant. Finally, column 4 looks at the probability that females are more likely to correct questions on retake exams. Since randomization takes place within exams, it could be the case that there is still sorting in gender across exams, though the question fixed effects in Table 4 should take care of any such bias. It is still reassuring to see an insignificant coefficient. We can thus conclude that the TAs for the introduction course in macroeconomics indeed seem to favor students of their own gender and that this effect seems to disappear once the exams are anonymous.

[26] The latter is an indication that certain students do not avoid answering questions corrected by, for instance, females.

However, this can only explain approximately 20 percent of the total effect of the reform on the gender difference.

Table 5: Randomization of TAs to questions

	(1) Question number	(2) Female student	(3) Age of student	(4) Retake
female teacher	0.0833	-0.0111*	-0.0802	-0.0161
	(0.439)	(0.00593)	(0.0514)	(0.0683)
Constant	6.222***	0.492***	23.26***	0.214***
	(0.297)	(0.00718)	(0.0636)	(0.0282)
N	51177	51177	51177	51177

Note: Standard errors clustered at the TA (49 clusters) and student (6 521 clusters) level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Conclusion

There are few studies investigating biased grading at the university level. Bias at the university level is important since it is typically not enough to make it "in" to get a job in your field—you also have to make it "out." Furthermore, your choice of courses, and in the end the degree you end up with, might depend on the signal you get in terms of grades in that area, as suggested by the model presented in Mechtenberg (2009). We find a sizable bias against female students. This is in sharp contrast to most of the literature studying bias prior to entering university studies, which typically has found a bias against males or no effects.

A major difference comparing the university level to lower levels of education is that male teachers are in the majority. Thus, one determinant could be same-sex bias, rationalizing the sign shift when studying grading bias at the university in contrast to lower levels. Previous studies on in-group bias have generally either been on noneducational data or have suffered from possible problems with teacher or student sorting. In this paper, we furthermore use an unintended randomized experiment to provide evidence that TAs correcting exams at the university favor students of their own gender. However, the size of the in-group bias is only approximately 20 percent of the total effect. Interestingly, both the in-group bias and the

general bias disappear when exams are graded anonymously, indicating the effectiveness of removing identity from exams, even though handwriting and content are otherwise left unchanged. This is a finding that could potentially be applied to many other evaluation settings as well and hence increases the policy relevance of our findings. More research is needed in order to truly get to the core of the underlying mechanisms, however, as we cannot explain all of the difference with our estimates.

5 References

Breda, T., & Hillion, M. (2016) "Teaching accreditation exams reveal grading biases favor women in male-dominated disciplines in France", *Science*, 29 Jul 2016: Vol. 353, Issue 6298, pp. 474-478.

Breda, T., & Ly, S. T. (2012). Do Professors Really Perpetuate the Gender Gap in Science? Evidence from a Natural Experiment in a French Higher Education Institution. CEE DP 138. *Centre for the Economics of Education (NJ1)*.

Breda, T., & Ly, S. T. (2015). Professors in core science fields are not always biased against women: Evidence from France. *American Economic Journal: Applied Economics, 7(4)*, 53-75.

Coenen, J., & Van Klaveren, C. (2016). Better test scores with a same-gender teacher?. *European Sociological Review, 32*(3), 452-464.

Conley, T. G., & Taber, C. R. (2011). Inference with “difference in differences” with a small number of policy changes. *The Review of Economics and Statistics, 93(1),* 113-125.

Dee, T. S. (2005). A teacher like me: Does race, ethnicity, or gender matter?. *American Economic Review, 95(2)*, 158-165.

Dee, T. S. (2007). Teachers and the gender gaps in student achievement. *Journal of Human Resources, 42*(3), 528-554.

Donald, S. G., & Lang, K. (2007). Inference with difference-in-differences and other panel data. *The review of Economics and Statistics, 89(2),* 221-233.

Eriksson, A. & Nolgren, J. (2013) Effekter av anonym rättning på tentamensbetyg vid Stockholms universitet – En empirisk studie i hur kvinnors och mäns betyg påverkas av anonym rättning. Mimeo Stockholm University

Feld, J., Salamanca, N., & Hamermesh, D. S. (2016). Endophilia or exophobia: beyond discrimination. *The Economic Journal, 126(594)*, 1503-1527.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review, 90(4)*, 715-741.

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2011). Are boys discriminated in Swedish high schools?. *Economics of Education Review*, *30*(4), 682-690.

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2015). Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools. *Education Economics*, *23*(6), 660-676.

Hoffmann, F., & Oreopoulos, P. (2009). A professor like me the influence of instructor gender on college achievement. *Journal of Human Resources, 44*(2), 479-494.

Katz, L. F. (1996). *Wage subsidies for the disadvantaged* (No. w5679). National bureau of economic research.

Kugler, A. D., Tinsley, C. H., & Ukhaneva, O. (2017). *Choice of Majors: Are Women Really Different from Men?* (No. w23735). National Bureau of Economic Research.

Kiss, D. (2013). Are immigrants and girls graded worse? Results of a matching approach. *Education Economics*, *21*(5), 447-463.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of public Economics*, *92*(10-11), 2083-2105.

Lavy, V., & Sand, E. (2015). *On the origins of gender human capital gaps: Short and long term consequences of teachers' stereotypical biases* (No. w20909). National bureau of economic research.

Lee, S., Turner, L. J., Woo, S., & Kim, K. (2014). *All or Nothing? The Impact of School and Classroom Gender Composition on Effort and Academic Achievement* (No. w20722). National Bureau of Economic Research.

Lindahl, E. (2007). Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden. *Institute for Labour Market Policy Evaluation (IFAU) Working Paper, 25.*

Lim, J., & Meer, J. (2017a). The impact of teacher-student gender matches: Random assignment evidence from South Korea. *Journal of Human Resources*, 1215-7585R1.

Lim, J., & Meer, J. (2017b). *Persistent effects of teacher-student gender matches* (No. w24128). National Bureau of Economic Research.

Lusher, L., Campbell, D., & Carrell, S. (2015). *TAs like me: Racial interactions between graduate teaching assistants and undergraduates* (No. w21568). National Bureau of Economic Research.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The Review of Economic Studies, 76*(4), 1431-1459.

Pettersson-Lidbom, P., & Thoursie, P. S. (2013). Temporary disability insurance and labor supply: evidence from a natural experiment. *The Scandinavian Journal of Economics, 115(2),* 485-507.

Sandberg, A. (2018). Competing identities: a field study of in-group bias among professional evaluators. *Forthcoming in The Economic Journal.*

Sprietsma, M. (2013). Bias in grading: experimental evidence from primary school teachers. *Empirical Economics, 45*(1), 523-538.

Terrier, C. (2015). *Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement.* London School of Economics. CEP Discussion Paper 1341, March 2015, London.

Yelowitz, A. S. (1995). The Medicaid notch, labor supply, and welfare participation: Evidence from eligibility expansions. *The Quarterly Journal of Economics*, *110*(4), 909-939.

6 Appendix

6.1 The procedure underlying the correction of exams at the introductory macroeconomics course

Each of the 7 questions is corrected by a TA, usually a separate one for each question, although there are some exceptions, in particular for the retakes. Before the correcting process starts, all TAs, the lecturer and the course coordinator assemble and discuss in broad terms how many points that should be given for different answers. At the end of this meeting, the allocation of TAs to questions 4-10 is determined by lottery.

Once this process has been completed, each TA receives his/her approximately 500 answers to his/her question (approximately 100 if it is a retake) and are then left with the daunting task of correcting each answer as fair as possible. By Swedish law, it is required that the students should know the results within 3 weeks after the exam at the latest and thus, one has less time than this to actually complete the correction. Hence, after approximately 2-2.5 weeks, the TAs and the course coordinator gather once more to look at students 1-2 points below a higher grade and then try to move those above the threshold. It is important to note that they are still anonymous at this stage since the fall of 2009. After this, the results are posted, and a session is announced, during which the template that everyone agreed upon during the first meeting is presented to the students. At the end of this session, students are allowed to make complaints directly in person to the TAs, which usually leads to a 1-2 point increase for 1-2 students at the most. It is important to note that, in general, we have data on the students' points right after they have been determined by the

TAs only and thus, they are not subject to bias from anyone other than the TA. The exceptions are one exam from the fall of 2009 and one question on another exam.

6.2 Reduction of DD to before-and-after

It is stated in section 2.2.3 that if $\omega = 0$, we can consistently estimate the DD-effect using a simple before-and-after framework. Equations (6)-(8) illustrate how this works in our simple regression framework, where Y_{11} is the gender difference in testscore for the treated group in the post treatment period and Y_{10} is the gender difference in testscore for the treated group in the pre-treatment period. Formally, we can write this as:

(6) $Y_{j=1,t=1} = \xi + \varsigma treated_{j=1} + \omega fall\ 09_{t=1} + \delta_1 treated_{j=1} * fall\ 09_{t=1} + \kappa_{i11} = \xi + \varsigma + \omega + \delta_1 + \kappa_{i11}$

(7) $Y_{j=1,t=0} = \xi + \varsigma treated_{j=1} + \omega fall\ 09_{t=0} + \delta_1 treated_{j=1} * fall\ 09_{t=0} + \kappa_{i10} = \xi + \varsigma + \kappa_{i10}$

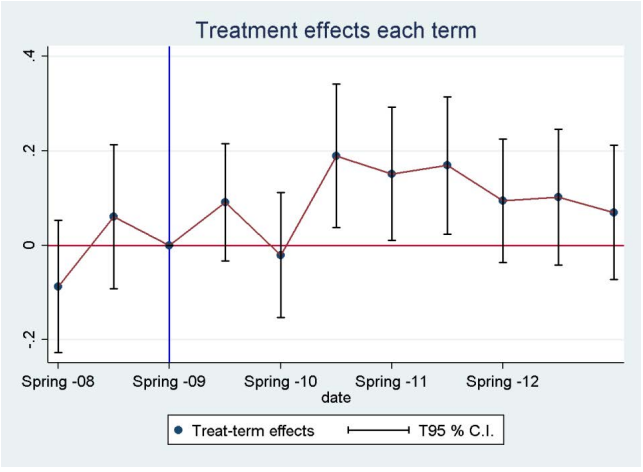
Thus, since $fall\ 09_t = 0$ when $t = 0$, the difference before and after for the treated group $j = 1$ is reduced to:

(8) $Y_{11} - Y_{10} = \omega + \delta_1 + \kappa_{i11} - \kappa_{i10}$

Thus, if $\omega = 0$ we can estimate the true treatment effect δ_1 using equation (8), that is the simple before and after in gender difference in the treatment group, by a regression corresponding to equation (4).

6.3 Figures and tables

Fig. A1:



Note: Standard errors clustered at the student level.

Fig. A2:

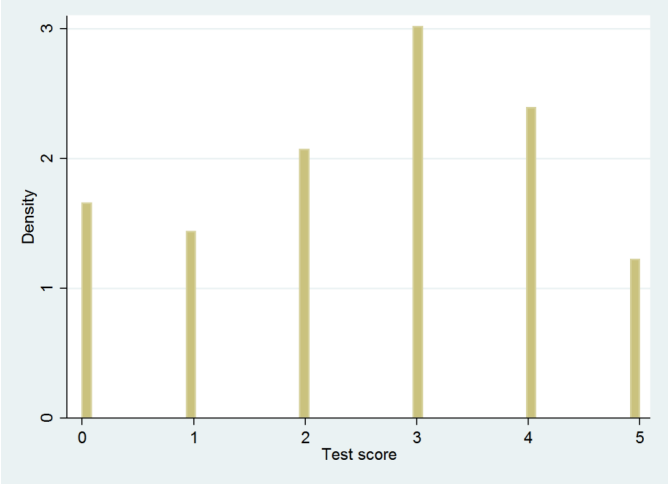


Fig. A3:

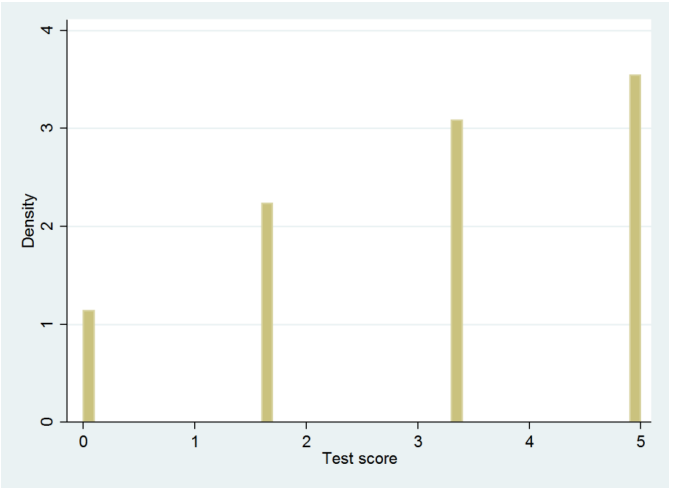
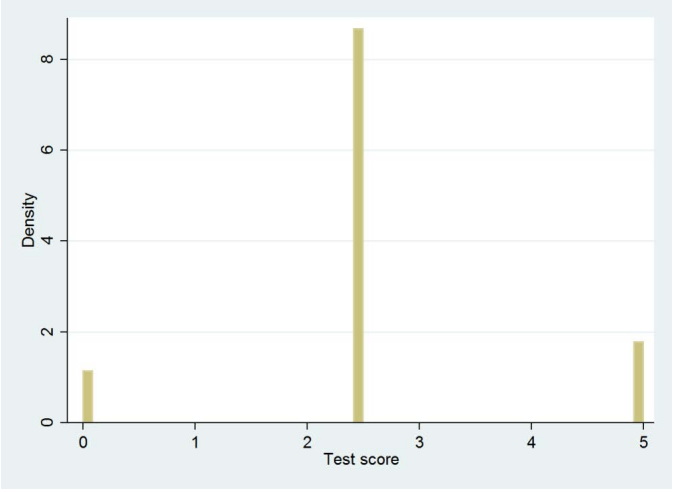

Density
0
1
2
3
4
0
1
2
3
4
5
Test score

Fig. A4:


Density
0
2
4
6
8
0
1
2
3
4
5
Test score

6.3 Tables

Table A1: Grades and their values

Grades A-F	Values	Grades AB-U	Values	Grades VG-U	Values
A	5	**AB**	5	**VG**	5
B	4	**BA**	3.33	**G**	2.5
C	3	**B**	1.67	**U**	0
D	2	**U**	0	-	-
E	1	-	-	-	-
F/Fx	0	-	-	-	-

Table A2: Additional Robustness

	(1) Stand. score	(2) Stand. score	(3) Stand. score	(4) Std. score	(5) Std. score
female*treated*fall 09	0.0430***	0.0554***	0.0541***	0.0527***	0.0430***
	(0.0110)	(0.00924)	(0.0177)	(0.0204)	(0.0110)
treated*fall 09	-0.0426***	-0.0316***	-0.0922***	-0.143***	-0.0426***
	(0.00853)	(0.00725)	(0.0143)	(0.0166)	(0.00853)
female*rreated	0.0222**	-0.00305	0.0814***	0.0828***	0.0222**
	(0.00935)	(0.00771)	(0.0158)	(0.0188)	(0.00935)
female*fall 09	-0.0488***	-0.0613***	-0.0243	-0.0177	-0.0488***
	(0.00918)	(0.00741)	(0.0171)	(0.0196)	(0.00918)
Treated	-0.123***	-0.171***	-0.273***	-0.222***	-0.123***
	(0.00725)	(0.00612)	(0.0128)	(0.0153)	(0.00725)
female student	0.114***	0.139***	0.0327**	0.0260	0.114***
	(0.00793)	(0.00631)	(0.0157)	(0.0184)	(0.00793)
fall 09	-0.0671***	-0.0781***	0.0720***	0.114***	-0.0671***
	(0.00665)	(0.00524)	(0.0138)	(0.0158)	(0.00665)
Exclude Dep. of Law	No	Yes	No	No	No
Only A-F grades	No	No	Yes	Yes	No
A-F grades are mandatory	No	No	No	Yes	No
Alternative numbers	No	No	No	No	Yes
N	1856027	1734908	973477	901927	1856027

Note: Standard errors clustered at the student level. The dependent variable is standardized score.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Anticipation Effects of a Board Room Gender Quota Law: Evidence from a Credible Threat in Sweden*

By JOAKIM JANSSON[†] AND BJÖRN TYREFORS[†]

Board room quota laws have recently received an increasing amount of attention. However, laws are typically anticipated and firms can react before the effective date. This paper provides new results on female board participation and firm performance in Sweden due to a credible threat of a quota law enacted by the Swedish Deputy Prime Minister. The threat caused a substantial and rapid increase in the share of female board members in firms listed on the Stockholm Stock Exchange. This increase was accompanied by an increase in different measures of firm performance in the same years, which were related to higher sales and lower labor costs. The results highlight that anticipatory effects of a law could be detrimental to the analysis.

Keywords: board behavior, gender equality, antidiscrimination policy

JEL: G34, J16, J78, D22, K29, M12

* We thank Handelsbanken's Research Foundations for generous financial support. We thank Per Pettersson-Lidbom, Joacim Tåg, Karin Thorburn, Peter Skogman Thoursie, Jenny Säve-Söderberg, Johanna Rickne, Alexander Ljungqvist, Robert Östling, Dan-Olof Rooth, Matthew Lindquist, Camille Hebert, and seminar participants at IFN, Stockholm University, Linneus University, SOFI, Århus University, and 2018 FMA European Conference for helpful comments.

[†] Department of Economics, Stockholm University, Research Institute of Industrial Economics, Stockholm Internet Research Group, Stockholm University

[†] Department of Economics, Stockholm University, Research Institute of Industrial Economics

I. Introduction

Policymakers in Europe have recently begun to focus on the relative underrepresentation of women on corporate boards, and numerous countries consider or have implemented gender quotas. The first quota law, adopted in Norway in December 2005, required public limited liability companies (ASA) to increase the female representation on their boards of directors to 40 percent within two years. The law increased female representation by approximately 20 percentage points for the typical firm (Matsa and Miller 2013). Other countries, including Spain, Belgium, France, Germany, Iceland, Italy and the Netherlands, have subsequently implemented quotas (Eckbo, Nygaard and Thorburn 2016). In Sweden, the policy debate has been intense as well. In 2002, Swedish Deputy Prime Minister Margareta Winberg, supported by Prime Minister Göran Persson, threatened to impose a mandatory law if considerable improvements in board room representation were not achieved in the listed companies within two years. Specifically, the listed companies were asked to increase their share of female directors to 25 percent, an increase of approximately 20 percentage points.

Our main contribution is that we estimate a pure anticipation effect of a gender quota law and that the effects are large in magnitude. We use a difference-in-difference-design where listed companies, the treatment group, saw a direct threat of a quota law where comparable non-listed firms, the control group, did not. Interestingly, the threat increased firm performance, a result which differs from other quasi-experimental studies evaluating gender quota laws. Our main results specifically show that the threat caused a substantial and rapid increase in the female board share in firms listed on the Stockholm Stock Exchange; the short-term effect size was approximately 5-10 percentage points or an approximately 100-200 percent increase. Interestingly, this increase was accompanied by an increase in the measures of firm performance in the exact same years. On average,

profits over assets (ROA) increased by approximately 2-4 percent among listed firms after the threat, relative to the change in ROA in unlisted firms in the same time period. However, increased female representation on boards did not lead to the higher recruitment of females as CEOs, either in the short or the long run. In fact, our results indicate the opposite, suggesting that certain female CEOs were recruited to the boards and not always replaced by female CEOs. One way of explaining the magnitude of the estimated firms' performance effects is to acknowledge that we are estimating an anticipation effect of a law. The net effect of a law may still be small.

The results still seem hard to rationalize theoretically from a classic economics perspective, where agents are profit maximizing and have perfect information. Then, it is reasonable to conjecture that a quota law or a credible threat of a law should reduce profits, in particular in the corporate sector where the competition pressure is high which should limit suboptimal board composition.

However, recently, Besley et al. (2017) study quotas in party politics. They show theoretically that, in a setting where the competence of new candidates of a party ballot is positively related to success for the party but, at the same time, is threatening the power of the incumbent, the incumbent will trade off party success against survival of power. Thus, a gender quota could lead to better candidates as mediocre men are replaced by both better men and women. They also find strong empirical support for the model and explicitly point out that this model "could be applied, for example to private organizations such as corporate boards".

Correspondingly, when evaluating the quota law in Norway, Bertrand et al. (2014) find that "the average observable qualifications of the women appointed to the boards of publicly limited companies significantly improved after the reform".

A similar conclusion is reached in Ferreira et al. (2017) when finding a greater stability of post-quota female appointments.[1]

Moreover, as has been frequently noted in the literature, we also acknowledge that if the male directors have a distaste for women and/or a taste for homogeneity, then diversity and independence could increase firm performance (e.g. Adams and Ferreira, 2009, Smith 2014 and Ferreira 2015).[2] A credible threat could push the board to be more gender neutral and firms could perform better. In the models proposing potentially positive effects of a quota, there must be a supply of competent women or women with different characteristics than men to recruit.[3] The diversity could be manifested in less permanent characteristics such as level of formal training and experience.[4] But gender differences could also be more stable. Related to decision making are differences in preferences and attitudes such as differences in risk attitudes,[5] attitudes towards competition and negations.[6]

Thus, theoretically we cannot determine which of the effects that prevail and we would ideally like to randomize gender quotas on corporate boards in order to evaluate the causal effects. The Norwegian law of quotas in 2005 has been used as such an exogenous shock (Ahern and Dittmar, 2012 and Matsa and Miller,

[1] Ferrari et al. (2016) also find that the quota in Italy led to overall higher levels of education of board members. See also Comi et al. (2016).

[2] A related approach would be to assume that shareholders or directors have a bias when evaluating female competence. A quota may then ex ante reduce the bias, analogous to the findings in Beaman (2009).

[3] Women have been more highly educated than men for many years in most OECD countries. Related to the supply argument is the literature on compensation, in particular at the top level of organizations. See, for example, Bertrand and Hallock (2001) or Keloharju et al. (2017) for evidence on Swedish data.

[4] As discussed in Adams (2016), diversity could be either temporary or more of a permanent type. Differences such that female directors are likely to be younger (see e.g. Adams and Ferreira, 2009 and Adams and Funk, 2012) or being an outsider of the "old boys club" could to change over time.

[5] For example, it has been suggested that the Lehman Brothers' crisis would never have occurred if it had been Lehman Sisters (Adams and Ragunathan, 2014). However, this argument misses out on the selection into boards as pointed out and documented by Adams and Funk (2012) where they find that the selected female directors are less risk adverse, invalidating the Lehman Sisters "hypothesis" with respect to risk aversion differences.

[6] See e.g. the survey of the literature and empirical evidence in Bertrand (2011).

2013).[7] Ahern and Dittmar (2012) use the pre-reform share of women on the board of listed firms and the fact that early adopters are not affected by the law to the same extent. Using this strategy, they find a large negative effect on firms' Tobin's Q ratio. However, as discussed by Ferreira (2015), early adopters are unlikely to be similar in trends to their counterpart. When we replicate their first stage in our setting, the parallel trend assumption is violated due to mean reversion. This finding is illustrated in Figure A1 in the Appendix. Turning to the most similar study, Matsa and Miller (2013) also use a difference-in-difference design, in which a sample of non-listed limited liability firms act as the control group to the listed firms. Again, the effect found in Matsa and Miller (2013) on firm performance is negative.[8] Conversely, Nygaard (2011) finds a positive effect of quotas on firm performance when evaluating the Norwegian reform. However, the robustness of the results from these papers has been questioned (Ferreira 2015; Eckbo, Nygaard and Thorburn 2016). When critically assessing the empirical design used in previous papers, Eckbo, Nygaard and Thorburn (2016) find a zero effect of the quota reform on firm performance measures. One major point made in Eckbo, Nygaard and Thorburn (2016) is that firms could anticipate the law after the change in the political debate in February 2002. Anticipatory effects are a direct threat to validity in a difference-in-difference setting if they are not properly accounted for (Angrist and Pischke 2009). For example, if a law was anticipated, but not acknowledged by the econometrician, the estimated effect may well have the wrong sign. One way of understanding the bias is to picture a quota law with heterogeneous treatment effects. Some firms will see an increase

[7] The Norwegian reform was implemented sequentially in practice. The first discussions began in 1999, and the first proposal was released in 2001 by the then center-left government. In 2002 the newly elected center-right government made statements both in support of and in defiance against a quota law, which in the end resulted in a law being passed in late 2005. The law in turn gave the affected companies two years to comply.

[8] The authors pick the treatment period as post-2006. As demonstrated by Figure 1 in Bertrand et al. (2014), the increase in the share of females on boards began back in 2002 and continued until 2008. Thus, their first stage does not seem to exhibit parallel trends prior to their treatment period.

in firm performance due to more female board members and some will be hurt. Under the reasonable assumption that firms with positive treatment effects are more likely to start the process of recruiting female directors, we would estimate a positive firm anticipation effect and a negative effect of the effective law. The net of the law, the anticipation and the effective law effects, may be zero, positive or negative. Thus, a credible difference-in-difference strategy uses the first date when the law was anticipated as the treatment date.

Given the large degree of disagreement regarding the effects of the Norwegian reform and the debate regarding the suitability of using the Norwegian setting for causal interpretation, we propose another testing ground, where we use a credible threat by the Swedish Deputy Prime Minister as the exogenous variation, to provide evidence of the effects of gender quotas.

The remainder of the paper is outlined as follows: In section 2, we document the background of the threat. In section 3, we describe the methodology, data, and sampling. In section 4, we provide the results, and in section 5 we conclude the paper.

II. Background

Sweden has a long history of male-dominated board rooms in listed companies. In the 1990s, the female share was steady at just below 5 percent. In 2003, the female share began to increase, tripling within 3 years. Anecdotally, the increase has been attributed to threats of a gender quota law made by the minister of gender equality, Margareta Winberg, during the second half of 2002. Winberg, a prominent feminist figure with a long history in the Social Democratic Party and the government, took office in 1998 as a minister of gender equality. In our study, identification is linked to the timing of the threat, and therefore it is crucial to describe the threats carried out over time. Figure 1 shows the number of printed

articles in newspapers in Sweden, a major channel used by policy makers to propose new policy ideas. The number of articles is based on a search that includes the minister's name, quota, women and board.[9] In 1999, as depicted in Figure 1, Winberg began to discuss, although rarely, the role of board room quotas for women in listed companies. Previously, she had acknowledged that a female quota in the business world could be problematic since competencies might be scarce. In three articles in leading Swedish newspapers in 1999, Winberg stated that she was not hostile to a law but instead hoped to see voluntary improvements within 5 years. In the following years, gender quotas in the board rooms were absent from the debate, as depicted in Figure 1.

In 2002, the temperature rose. During that year, the number of printed articles mentioning Winberg's name in combination with quotas, women and boards exploded. In July, in the leading business daily *Dagens Industri*, Winberg indicated that she was contemplating a quota law to increase the pressure on listed firms (*Dagens Industri* June 17 2002). As a result, the debate became heated. Following Winberg's appointment as Deputy Prime Minister in October, a series of articles intensified the tone and outlined the quota threat in detail. In an article in *Dagens Industri,* she stated that "the threat is real", noting that if the listed companies were not making significant progress, "there will be a law" (*Dagens Industri* October 22 2002). In another article in the leading daily paper *Svenska Dagbladet*, Winberg defined significant progress: the share of female directors must increase to 25 percent within two years. She noted that she had full support from Prime Minister Göran Persson and that a formal "Investigation Directive" was under way and would be ready by the spring. After that, a formal investigation could proceed. Winberg estimated that the law would be ready in

[9] Source: Mediaarkivet, a digital archive containing more than 700 printed newspapers. See http://www.retriever-info.com/sv/category/news-archive/. The search was "margareta winberg kvotering kvinnor styrelse".

2004 or 2005. Thus, the magnitude of articles significantly increased and the tone concerning a quota was sharpened at the end of 2002. Winberg's new political appointment, her well-known feminist ideology, combined with the backing of the Prime Minister, strengthened the credibility of the quota threat. For the first time in history, the representation of women on the boards of listed companies began to rise consistently.

The dotted line in Figure 1 denotes 2002. In this study, we set 2002 as the baseline year since we observe data annually. This choice is reasonable for two reasons: the explicit threats were laid out at the end of 2002, and shareholders appoint new directors at an annual meeting. Since the annual meeting typically occurs in the late spring, 2003 will be the first year of treatment.[10]

The time series of the articles ends in 2003, the year when Winberg resigned. However, the investigation of the law was established by the Minister of Justice, Thomas Bodström, in the summer of 2005, and in June 2006 a law proposal was finished. The proposal stated that listed firms (and government-controlled limited liability companies) should have at least 40 percent women on their boards by 2008; otherwise, a fine would have to be paid every time a new board was elected. The investigator argued that other limited liability companies should also not be subject to the law.[11] Thus, the law proposal was consistent with the content in the previous threats made to listed limited liability firms.

In September 2006, the Social Democratic Party lost the election and a new conservative-liberal government was formed. The new government was against the gender quota law proposal and, as depicted in Figure 1, the share of female representation was halted for several years. In February 2010, both Anders Borg, the Finance Minister, and Per Schlingmann, the spin doctor and secretary of the

[10] In the Appendix, Table A4 depicts the results if 2001 is set as the baseline year. The results do not differ substantially.

[11] See the investigation proposal "Könsfördelningeni bolagsstyrelser" (2006) for a full description.

leading party in the government "Nya Moderaterna", complained that progress toward female representation was too slow (it had been steady since the Social Democrats lost the election and the law proposal was rejected), again opening up the discussion of a law (*Dagens Industri*, February 2 2010). However, at Nya Moderaterna's annual convention a year and a half later, party members reacted strongly and rejected any quota law (*Dagens Industri*, October 22 2011).

Generally speaking, the development of female representation on corporate boards responds to different threat levels. However, in this paper, we will focus on the first major threats at the end of 2002 and study their effects. From a causal point of view, everything else may be an endogenous response.

III. Methodology, Data and Sampling

A. Methodology

A naïve regression population function could be written as follows:

$$Y_{ct} = a + \beta Share\_female_{ct} + e_{ct} \quad (1)$$

where Y_{ct} is firm *c´s* performance outcome such as operating profits/assets (ROA) at time *t*. It is clear that unobserved firm characteristics can determine the variable of interest, the share of female directors on a firm's board, as well as the outcome. Thus, to estimate β with no bias, we would need an instrument for the variable of interest. In addition to being strong, an instrument must be: (i) "as good as" randomly assigned and (ii) excludable, i.e., the only channel through which it operates is the endogenous variable (exclusion restriction). The "as good as" randomly assigned condition ensures a causal interpretation of the reduced form. In our setting, we could under (i) estimate the causal effect of the threat of a quota

law. In a DID-setting (i) translates into parallel trends of the outcome across treatment and control groups. Thus, the reduced form in our setting becomes

(2) $$Y_{clt} = \alpha + \gamma Listed_l + \lambda\, Post_t + \delta(Listed_l * Post_t) + \varepsilon_{clt}$$

where *Post* is a dummy taking the value one for the period after 2002 and otherwise taking the value zero. *Listed* is equivalently a dummy for listed firms in 2002. Under the assumption of parallel trends, δ, the parameter of the interaction, will measure the causal effect of the threat of a quota law on, for example, the share of female directors or the ROA. The subscript l=1,2 denotes treatment or control group.

If we also assume the exclusion restriction to hold, we could also write the first-stage equation as the following:

(3) $$Share\ female_{clt} = b + \tau Listed_l + \phi\, Post_t + \xi(Listed_l * Post_t) + \omega_{clt},$$

and we could estimate the causal effect (a LATE) of increasing the share of women from 0 to 1 on firm performance by OLS with $\frac{\widehat{\delta}}{\widehat{\xi}}$.

In this paper, we suggest that it is unlikely to assume that the exclusion restriction would hold both in the setting of a law and in the setting of the threat of a law. First, imposing quotas could affect firms' recruitment procedure in numerous ways. Having to recruit women will most likely include using new expertise, networks and recruitment firms, which could have a direct effect on the outcome as evidenced in Ferreira et al. (2018). Moreover, the threat of a law might signal future government interventions in general, which could influence firm actions. Further, the presence of more women on corporate boards might increase the size of the board; research suggests that board size may be important

for performance through monitoring and advising (Jensen 1993; Yermack 1996). Lastly, having additional women on the board is correlated with other factors that have been found to be of importance for firm performance, such as director independence (see the survey in Adams, Hermalin and Weisbach 2010) and the size of the board. Thus, director independence could affect firm performance, and any outsider group, not just females, would affect independence and potentially firm performance. Consequently, we view equation (3) as an interesting reduced form and one potential channel. Thus, this paper focuses on estimating the causal effect of the threat of imposing gender quotas for listed firms and hence, parallel trends will be the major identifying assumption.

Given the large amount of disagreement in the evaluations of the Norwegian reform, we provide a battery of specification tests in this paper. First, we address compositional bias by adding industry fixed effects and thus, non-parametrically control for the industry-level specific factors.[12] An even more flexible specification could include firm-specific effects instead of the dummy *Listed* and year fixed effects instead of the dummy *Post*. However, in the absence of compositional effects, this should not affect the coefficient of interest.

Second, we acknowledge that the estimations of the standard errors are problematic in our study since treatment only changes once for one group, as discussed by Bertrand, Duflo and Mullainathan (2004), Donald and Lang (2007) and Conley and Taber (2011). Regarding the standard errors, we begin by clustering them at the industry level, thus acknowledging not only firm correlated shocks but also industry shocks. Compared to the related literature, this is a conservative treatment of the standard errors. However, since treatment only varies once at the control-treatment group level, this might not be conservative

[12] In the Appendix, Table A5, we also estimate our main model in which we leave out one industry at a time. This model is motivated by the fact that potentially 2003, the first year of treatment, is three years after the burst of the dot-com bubble and one could worry that certain industries, for example IT or telecom, would drive our results. Fortunately, our results are robust when leaving out one industry at a time.

enough. Here, we follow the Pettersson-Lidbom and Thoursie (2013) application of the results in Donald and Lang (2007). The problem is that treatment only varies one time at the group level *l*, listed and non-listed, and not at the firm, *c*, or industry level. The error term could contain both a firm error r_{clt} and a group time-error j_{lt}; therefore, $\varepsilon_{clt} = r_{clt} + j_{lt}$. In the presence of a group time error, standard errors are biased; clustering at the firm or industry level will not help, and clustering on *l* cannot be done due to the low size of 2.

We address the clustering problem as discussed in Moulton (1986) by aggregation. Thus, we calculate the mean for every time period for the groups listed and non-listed and estimate equation (2) at the group level (listed and non-listed). Although this addresses the Moulton (1986) problem, the error could still be serial correlated. Taking the difference between the two groups, however, we represent one time series as:

$$\Delta Y_t = \gamma + \delta Post_t + \Delta\mu_t, \tag{4}$$

where $\Delta Y_t = Y_{listed,t} - Y_{non-listed,t}$, $\gamma = \gamma_{listed} - \gamma_{non-listed}$ and $\Delta\mu_t = \mu_{listed,t} - \mu_{non-listed,t}$. With this transformation, the estimate of δ will be identical to an estimate from a fixed-effect model (where N=2 and T=15 when using annual data). When estimating equation (4), we make the standard errors robust to heteroscedasticity and serial correlation by applying the Newey-West estimator.

It is straightforward to introduce two specification tests for parallel trends, as discussed by Angrist and Pischke (2009). First, we could add the leads of the independent variable $Post$. If the parallel trends assumption holds, the coefficient should come out both close to zero and statistically insignificant. We show these results graphically. Furthermore, we could add a linear trend to the specification, and if the parallel trend assumption holds true and there are no dynamic effects,

then the effect should remain stable. However, since the election of board members often occurs at the annual meeting in the late spring, we could expect the effects to be smaller in 2003. We could also match on the pre-trends according to the method of synthetic control, developed in Abadie et al. (2010).

Importantly, there could be other major factors affecting listed companies differently than non-listed companies around 2002-2003. In any DID-setting with one policy change and two groups, and in particular with annual data, this is the major concern. In the end it is not testable. However, there are some sanity checks that could be made. Firstly, we have identified two other potential drivers. Ferreira (2015) notes the changed Norwegian Code of Practice for Corporate Governance and changed accounting rules (Norway adopted IFRS accounting rules in 2005). Since Sweden also implemented both of these practices in 2005, we provide estimation results from a shorter window, namely, 1998-2004, which can be found in Table A6, Columns (2) and (3). Our results are similar for this shorter period, which makes it less likely that these two changes are drivers.

Lastly, in our main specifications, we make a few restrictions on data, as discussed below. For the sake of transparency, the sensitiveness of the results for these restrictions can also be found in the Appendix.

B. Data

Our data consist of two data sets that have been merged. The first is composed of all, except financial, limited liability firms' final accounts and key figures over the time period 1998-2012[13] To these data we add information on all individual board members in limited liability firms and the years during which they were on

[13] Some firms do, however, produce two or even three accounts during one calendar year. To avoid weighting these firms more heavily, we identify their final accounts by the observation with the highest turnover in each year. Since the turnover only (weakly) increases over the fiscal year, this should leave us with the final accounts only. Notably, not all variables and measures exist for all firms in our sample.

the board. These data contain information for the time period 1998-2012[14] Specifically, we take all board members who are on the board at some point during the given year and then compute the average share of women on the board based on these members. All data come from the Swedish Companies Registration Office (but in two mergable data sets). The office keeps track of all companies and their CEOs and directors. The firm data are available for the universe of limited liability companies, excluding financial firms. For example, the office keeps track of the financial statement items and the number of employees. Each firm must by Swedish law file this information within 6 months after the end of a fiscal year.

From a causal point of view, anything occurring after the threat and onwards could be endogenous, including delisting. Any restriction on data before the threat is non-problematic since it is based on pre-treatment characteristics. All restrictions made below will therefore be based on characteristics in 2002. In the Appendix, we will relax our restrictions, one by one, to verify and disclose the robustness of our results. The results are found in the Appendix, Table A2.

We begin with the sampling restriction wherein we limit our analysis to all firms that are active in 2002. A non-active firm is a firm in which there is no intent to operate a normal business. Furthermore, we define treatment status based on whether a firm is listed or not in 2002. This means that we can use the number of firms as an indicator of compositional bias due to delisting.

Since non-listed firms may have a board size of 1, we limit our analysis to firms with a board size of at least 5 directors for the firms to be comparable.

[14] The data on boards contain information for more years than 1998-2012; however, it is censored from both the top and the bottom outside the range of 1998-2012. There are no dates assigned for those that start on a board prior to 1993 or who quit after 2012. Likewise, those quitting a board prior to 1988 or after 2012 have no date recorded. Since the data on the final accounts begin in 1998, the censoring prior to 1993 does not matter. Similarly, since both the board and final accounts data end in 2012, any censoring after that point is irrelevant to this study.

Furthermore, we only consider ordinary board members as part of the board and thus, we exclude labor union representatives, deputy directors and the likes, although our results are not very sensitive when also including these.

While a number of other reasonable restrictions could be made, our main analysis will hinge on these restrictions. However, in Appendix Table A3, we show results for other plausible restrictions, including restrictions on the share of capital that differs across groups or public or private limited liability firms and number of employees.[15] These different restrictions are not driving the results.

Finally, we determine the gender of the board members through their personal identification number for all Swedish residents. Using personal numbers, we obtain exact gender information for 95.72 percent of the data.[16] For non-Swedish residents, however, we rely on board members' first name only. We obtain our results by using the list of all names given to more than 10 born boys or girls in the previous year (2014) from Statistics Sweden, dropping all duplicates between the genders, and then defining the gender of the board member by checking her first name against this list. This process increased the hit rate to 98.15 percent. If we could not determine the gender of a board member after this process, the board member's gender was coded as missing. Thus, we end up with final account data for the universe of limited liability firms in 2002 (except financial firms) for the time period 1998-2012, along with information on the boards' gender composition.

Moreover, since a firm can belong to a group of firms, we focus our analysis on the parent firm if it belongs to a group. If the firm is not part of a group then we study this sole firm. The definition of a parent firm is one that controls other firms

[15] A public firm might have more than 200 stock owners and should have at least 500 000 SEK (approximately 60 000 USD) in share capital, whereas private limited liability firms may have as little as 50 000 SEK. Before 2005, this amount was twice as high at 100 000 SEK. Moreover, public firms need a board size of 3, whereas for private firms, it suffices with 1 member.

[16] A regression using only those in which the gender is identified from the personal number can be found in Table A6, column 1. The results are again robust.

in the group (the subsidiaries). Policies affecting a parent company thus have spillover effects on other companies in the group. Since listed companies are commonly the parent of non-listed subsidiaries, including the subsidiaries would mean a violation of SUTVA (Rubin, 1980). Thus, we focus on the parent companies as the unit of observation if there exists a group and subsidiaries are not part of the main analysis. Since the parent company board is in charge of the subsidiaries, this poses no problem with respect to measuring the female director share, which is simply the share in the board of the parent. However, regarding firm performance measures such as operating profits/assets, we could either use the parent company financial statements or the group financial statements. Using the parent financial statements would generally underestimate the firm performance. However, the DID estimation hinges on a parallel trends assumption, and thus we need not only this underestimation to be different across groups but also to evolve differently over time across groups to cause a methodological problem. Therefore, using the financial statement of the parent company should not automatically pose a threat to internal validity. To verify this, we also use the financial variables from the group financial statement; our coefficient of interest is indeed unchanged. Lastly, we also redo the analysis only using parent firms that are part of a group, i.e. also excluding single firms (with no subsidiaries). Lastly, In the Appendix, Table A2, Column (6) also shows the results when all individual firms are treated as independent, whether they are parent firms or subsidiaries.

As is standard in the previous literature, we winsorize all financial variables at the 1 percent and 99 percent level. Thus, we cap all values above the 99th percentile and below the 1st percentile to the value at the 99th and 1st percentile, respectively. This procedure is conducted separately for listed and non-listed firms. The results after alternative levels of winsorizing can be found in Table A4

and it is reassuring that point estimates are unaffected by winsorizing levels as only the precision changes

The summary statistics for listed and non-listed firms after the process of winsorizing are presented in Table 1. Panel A shows the statistics for all independent firms, that is parent firms or firms that belong to no group, i.e., firms that are independent with no subsidiaries. First, the share of female directors is approximately 14 percent for the period. Second, one can note that the mean of the operating profits/assets is negative for the period on average, although the median remains positive. Turning to Panel B, where we have instead used the group financial statement for the parent firms belonging to a group, we see no major differences, although both the balance sheets and the results are larger in absolute terms to some extent. Mostly, we observe approximately 170 000 observations, where one observation represents a parent firm or an independent firm for a given year.

IV. Main Results

A. Graphical Evidence

We begin by inspecting the number of firms in the treatment group over time. Since we condition based on the firms being listed in 2002, it must follow that there are (weakly) fewer firms before and after 2002. Cleary, attrition in the treatment group after 2002 might be an outcome causing a survival bias when examining firm performance measures. If we find that the quota threat caused listed firms to perform better, we are worried that the worst-performing listed firms have exited. Figure 2 below shows the number of listed firms conditioned on their existence in 2002. We first notice that there is no substantial attrition in

the listed group until the financial crisis in 2009. Thus, the threat does not seem to have caused a large outflow of firms from the listed group.

Turning to the share of female directors as an outcome, we begin by graphically inspecting the time series in Figure 3. Column 1 shows the share for independent firms and Column 2 shows the share for independent firms but for the matched sample, where the group financial statement has been used for the firms with subsidiaries. Since the match rate is high, the time series should be similar, which is shown in Figure 3. Interestingly, in the years before the quota threat, we can see a slightly upward and parallel trend in both listed and non-listed firms, although non-listed firms have a higher share of female directors. After the threat, there is an extraordinary increase for listed firms, whereas non-listed firms remain in the same approximate trend. After 2006, when the law was rejected, parallel trends emerge once again. The first year's reactions are the mildest, showing some dynamic effects before stabilizing around 2006. Panel B shows the estimates as annual treatment effects, as discussed by Angrist and Pischke (2009). The estimates suggest small and mostly non-significant effects before the threat, with sharply increasing effects in the first few years after the threat, which then appear to flatten out around 2006. Although the estimates show small effects before the threat, there may be weak evidence of an increase in the share of female board members before the threat, i.e., testing whether the effect survives when including linear treatment and control group trends will be of interest. However, the overall pattern is consistent with a causal interpretation of the effects. The effects size seems to be approximately 8 percentage points.

We now turn to our main firm performance measure, operating profits divided by total assets (ROA), as used in Matsa and Miller (2013).[17] Figure 4 of Panel A

[17] Ahern and Dittmar (2012) use Tobin's Q as their measure of firm performance. To compute this metric, however, one needs the market value of the firm, which we cannot observe for non-listed firms. We thus focus on the other commonly used firm performance measures that are available both for our treatment and control groups.

shows a rather similar downward trend until 2002. The sharp decrease in ROA due to the burst of the dot-com bubble in 2000 is visible for both groups. The dot-come bubble decline pedagogically shows the point of having a control group. Interestingly, listed parent companies have a negative ROA for the entire period, not only in the crisis following 2000. Clearly, negative ROA for such a long period can hardly resemble real firm performance. Thus, it is of interest to instead use the operating profits/assets from the group financial statement if the parent belongs to a group. Column 2 of Panel A shows that using the group financial statement instead of only the financial parent statement yields a more reliable measure of firm performance. However, there is also a slight tendency for profits to decline more for the listed groups between 2000 and 2001, potentially indicating a mild Ashenfelter's dip. When analyzing the annual treatment effects in Panel B, the dip does not seem to significantly influence the results. We also note that the Lehman Brothers crisis in 2008 yielded a sharp decline in profits as well and that the decrease is again somewhat larger for listed firms. It is reassuring that we do not see a pattern that the listed firms after the Lehman Brothers crisis are seeing some years of faster growth rates of profits/assets. Thus, the estimated effects for the threats in the period from 2003 and onwards are unlikely to merely be a convergence effect driven by the dot-com bubble in 2000. Profits increased by approximately 2-4 percent of the assets among listed firms after the threat, relative to the change in profits in unlisted firms in the same time period.

Moreover, there is an interesting correspondence between Figures 3 and 4. Both outcomes appear to be parallel before the threat. Then, there is a large reaction for the listed group until 2005-2006, both for the share of females and profits over assets, before stabilization occurs.

Lastly, to address any concerns about linear trends in the reduced form regarding the share of female directors and concerns that the effect might be driven by an Ashenfelter's dip, we perform a robustness check using a synthetic control group approach. Following the advice in Abadie et al. (2010), we match the dependent variable in 1998, 2000 and 2002. Both graphs show a good correspondence before 2002 and a sharp divergence afterwards. The effect sizes are 8 percentage points for the share of female directors and approximately 3 percentage points for profits.[18] Thus, concerns about pre-trends or dips are not critical for our results. Notably, Figure 5 also suggests that our results are not driven by functional form assumptions.

B. Main Regression Result

In Table 2 we present our main results, beginning with estimating the model outlined in equation (2), in Column 1. In Panel A we show the results when the share of female directors is the outcome. The threat of quotas caused the share of females to increase by approximately 8 percentage points, an increase of approximately 150 percent. Adding industry flexible time trends in column 2 does not alter the results, thereby strengthening the indication that attrition does not cause a compositional bias. In column 3, linear trends are added. Thus, our identification strategy no longer hinges on a parallel trend assumption; instead, if the trend differs, it may only differ linearly. Since Figure 3 indicates a slightly upward trend, it is not surprising that the estimate is changed. However, it remains significant and large at approximately 4 percentage points. Notably, if the first

[18] To implement Abadie et al. (2010), we collapse the data into the treatment group (in other words, all listed firms) and the remaining companies into industries. This leaves us with 57 time series, where one is the treatment group and the other 56 are the remaining companies in their respective industries. To these data we then apply the synthetic control method as in Abadie et al. (2010), where the control group is a weighted combination of the industries without the listed firms. As matching variables, we simply use the values of the dependent variable in 1998, 2000 and 2002. The exact resulting estimates of the effect can be found in Table A1 in the Appendix.

year reaction is the mildest due to dynamic effects, which has been suggested since directors are appointed in late spring, then part of the "true" effect is controlled away when adding linear trends. Lastly, in Column 4 we present the results from estimating equation (4), i.e., using collapsed data and a time series of 15 observations to address the Moulton and serial correlation problem when estimating the standard errors. Although the standard errors double in size, the effect remains significant.

Turning to firm performance and profits, we see in general that using the financial statements from the parent firm (Panel B) yields somewhat smaller estimates compared to using the group financial statements if the firms is the parent of a group (Panel C). However, in relation to the size of the standard errors, the effects are roughly the same. In summary, profits increased by approximately 2 – 5 percent of the assets among listed firms after the threat relative to the change in profits in unlisted firms in the same time period.

Lastly, in Table 3, we restrict the sample by only using parent firms belonging to groups; this means using approximately 30 000 observations (groups) compared to approximately 170 000 observations in Table 2. In general, the results depicted in Table 2 remain.

C. Additional Results

In Tables 4 and 5, we use the group's financials to construct other outcomes. We use our basic DID model, as presented in equation (2). In Column 1, Table 4, the basic estimate in which the outcome is operating profits over assets is re-tabulated. Since operating profits include depreciation and amortization, we also show the effect for the outcome EBITDA/assets in Column (2). Again, our estimate is a statistically significant EBITDA/assets increase of approximately 4

percent among listed firms after the threat, relative to the change in profits in unlisted firms in the same time period. When only considering total revenue/assets, we again obtain a positive estimate, although less precisely estimated. Interestingly, labor costs/assets decrease by approximately 2 percent of the assets among listed firms after the threat, relative to the change in profits in unlisted firms in the same time period. Again, this finding contrasts with that of Matsa and Miller (2013). Due to the accounting identity, an increase in profits must reflect some mixture of an increase in revenues and/or a decrease in costs. Although estimated with low precision, revenues seem to increase and labor costs to decrease. Two alternative outcomes, operating profits per employee and value added per employee, are presented in Columns (5) and (6). The results show the same sign as our other firm performance measure but are imprecisely estimated.

Turning to Table 5, Column (1), we confirm that the numerator of our major outcome, operating profits /assets, is positively and significantly affected by the threat. Thus, our effect is not driven by decreasing the denominator. Columns (2) and (3) show an increase in the number employed, although the figures are somewhat functionally specific because the effect becomes insignificant when using the logs instead of the levels. Columns (4)-(6) speak directly to our concern about using a gender quota law or a threat as an instrument with respect to the validity of the exclusion restriction. Column (4) shows that the female proportion of CEOs decreases by 2.5 percentage points. This result is consistent with female CEOs being recruited to corporate boards and not solely replaced by women. Columns (5) and (6) suggest that the board is also increasing in size. A back of the envelope calculation suggests that boards are expanded by one woman due to the quota threat. Thus, this finding clearly illustrates how the gender quota threat is affecting numerous potential channels that affect firm performance.

V. Conclusion

Gender quotas on corporate boards have recently received increased attention. The first quota law was adopted in Norway in December 2005. Other European countries have subsequently implemented quotas. Empirically, we know little about the effects of quotas in the board rooms on firm performance. This paper uses a credible threat of gender quotas aimed at listed firms. We find that the threat caused a substantial and rapid increase in the female board share in firms listed on the Stockholm Stock Exchange. The effect size was approximately 5-10 percentage points or a 100-200 percent increase. Thus, the anticipation effects of the quota law were large, consistent with a credible threat. Interestingly, this increase was accompanied by an increase in measures of firm performance in the same years. We can generally reject effect sizes that are smaller than 0.005 measured as operating profits/total assets; on average, profits increased by approximately 2-4 percent of the assets among listed firms after the threat, relative to the change in profits in unlisted firms. However, increased female representation on boards did not lead to a more frequent recruitment of females as CEOs, either in the short or the long run. In fact, our results indicate the opposite, which suggests that some of the female CEOs were recruited to the boards and were not always replaced by female CEOs. Moreover, labor costs decreased and sales increased, although these figures were imprecisely estimated. Our results indicate that parallel trends are a reasonable assumption, and our result is highly robust.

Although we attempt to make substantial progress with respect to the implementation of the method, we cannot rule out the possibility that, in comparison to the Norwegian studies, our conflicting results are due to differences across countries and reforms. In particular, although the Swedish quota threat was converted to a law proposal, it was never implemented due to a

new government. Second, the threat increased female representation from approximately 5 to approximately 15 percent. This result was far from the level of 40 percent that was the intended goal in Norway. Clearly, the effects of gender quotas on firm performance might be a nonlinear function of female representation.

In the future, we plan to collect additional information regarding how organizational structures are affected by more female directors, in line with the questions posed by Bertrand et al. (2014). For example, will there be more females positioned in middle and top management? Will male workers and managers utilize the generous parental leave system in Sweden to a larger extent?

REFERENCES

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American statistical Association*, 105, 490, 493-505.

Adams, Renee B., 2016. "Women on boards: The superheroes of tomorrow? *The Leadership Quarterly*, 27, 3, 371-386

Adams, Renee B., Benjamin E. Hermalin, and Michael S. Weisbach. 2010. "The Role of Boards of Directors in Corporate Governance: A Conceptual Framework and Survey." *Journal of Economic Literature*, 48, 58–107.

Adams, Renée B. and Daniel Ferreira. 2009. "Women in the Boardroom and their Impact on Governance and Performance," *Journal of Financial Economics*, 94, 291–309.

Adams, Renée B. and Patricia C. Funk. 2012. "Beyond the Glass Ceiling: Does Gender Matter?" *Management Science*, 58, 219–235.

Adams, Renee B. and Ragunathan, Vanitha. 2015 "Lehman Sisters" FIRN Research Paper.

Ahern, Kenneth R., and Amy K. Dittmar. 2012. "The Changing of the Boards: The Impact on Firm Valuation of Mandated Female Board Representation." *The Quarterly Journal of Economics*, 127, 137–197.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Beaman, Lori, Raghabendra Chattopadhyay, Esther Duflo, Rohini Pande, and Petia Topalova. 2009. "Powerful Women: Does Exposure Reduce Bias"*Quarterly Journal of Economics*, 124, 1497–1540.

Bertrand, Marianne & Hallock, Kevin F. 2001. "The gender gap in top corporate jobs" *Industrial and Labor Relations Review, 55*, 3-21.

Keloharju, Matti and Knüpfer, Samuli and Tåg, Joacim. "What Prevents Female Executives from Reaching the Top?" 2017. *IFN Working Paper* No. 1111; *Harvard Business School Research Paper Series* No. 16-092.

Bertrand, Marianne. 2011. "New Perspectives on Gender." *Handbook of Labor Economics*, vol. 4B, edited by D. Card and O. Ashenfelter, 1545– 92. Amsterdam: Elsevier B.V.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics*, 119, 249–275.

Bertrand, Marianne, Sandra E. Black, Sissel Jensen, and Adriana 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 20245.

Besley, Timothy, Olle Folke, Torsten Persson, and Johanna Rickne. 2017. "Gender Quotas and the Crisis of the Mediocre Man: Theory and Evidence from Sweden." *American Economic Review,* 107(8): 2204-42.

Comi, Simona, Mara Grasseni, Federica Origo, and Laura Pagani, 2016. "Where Women Make the Difference. The Effects of Corporate Board Gender - Quotas on Firms' Performance Across Europe." Unpublished mimeo.

Conley, T.G. and Taber, C.R. 2011. "Inference with 'difference in differences' with a small number of policy changes", *Review of Economics and Statistics*, vol. 93(1), pp. 113–25.

Donald, Stephen G., and Kevin Lang. 2007. "Inference with Difference-in-Differences and Other Panel Data." *The Review of Economics and Statistics*, 89, 221–233.

Eckbo, B. Espen, Knut Nygaard and Karin S. Thorburn. 2016. "Does Gender-

Balancing the Board Reduce Firm Value?" CEPR Discussion Paper Series, DP11176.

Harald Dale-Olsen, Pål Schøne & Mette Verner. 2013. "Diversity among Norwegian Boards of Directors: Does a Quota for Women Improve Firm Performance?", *Feminist Economics*, 19:4, 110-135

Ferrari, Giulia, Valeria Ferraro, Paola Profeta, and Chiara Pronzato, 2016.

"Gender Quotas: Challenging the Boards, Performance, and the Stock Market," *IZA Discussion Paper* No. 10239,.

Ferreira, Daniel. 2015. "Board Diversity: Should We Trust Research to Inform Policy?" *Corporate Governance: An International Review*, 23, 108–111.

Ferreira, Daniel, Edith Ginglinger, Marie-Aude Laguna and Yasmine Skalli. 2017. "Board Quotas and Director-Firm Matching" C*EPR Discussion Paper* No. DP12117.

Jensen, Michael C. 1993. "The Modern Industrial Revolution, Exit, and the Failure of Internal Control Systems." *The Journal of Finance*, 48, 831–880.

Justitiedepartementet. 2006. *Könsfördelningen i bolagsstyrelser Ds 2006:11*. Stockholm.

Matsa, David A., and Amalia R. Miller. 2013. "A Female Style in Corporate Leadership? Evidence from Quotas." *American Economic Journal: Applied Economics*, 5, 136–169.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics*, 32, 385–397.

Nygaard, Knut. 2011. "Forced Board Changes: Evidence from Norway." Norwegian School of Economics and Business Administration Discussion paper.

Pettersson-Lidbom, Per, and Peter S. Thoursie. 2013. "Temporary Disability Insurance and Labor Supply: Evidence from a Natural Experiment." *The Scandinavian Journal of Economics*, 115, 485–507.

Rubin, Donald. **B.** 1980. “Randomization analysis of experimental data: The Fisher randomization test comment.” *Journal of the American Statistical Association*, 75(371), 591-593.

Smith, Nina 2014. “Gender quotas on boards of directors.” *IZA World of Labor* 2014: 7.

Yermack, David. 1996. "Higher Market Valuation of Companies with a Small Board of Directors." *Journal of Financial Economics*, 40, 185–211.

FIGURES

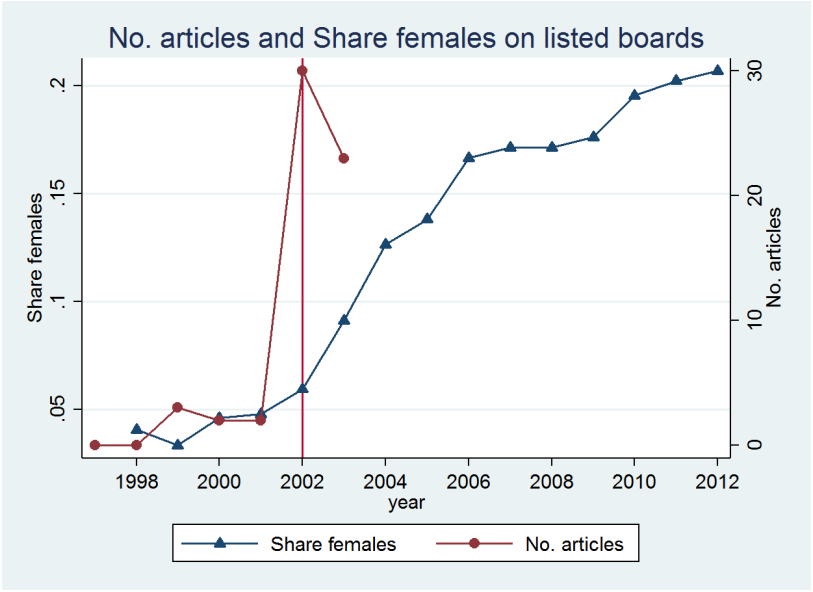


FIGURE 1. SHARE OF FEMALE REPRESENTATION ON THE BOARDS OF LISTED FIRMS AND ANNUAL NUMBER OF PRINTED ARTICLES IN THE SWEDISH PRESS 1998-2003

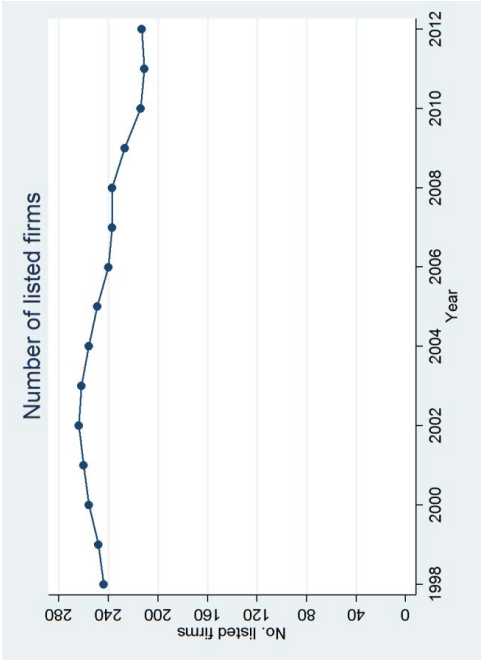


FIGURE 2. NUMBER OF LISTED FIRMS OVER TIME ON THE STOCKHOLM STOCK EXCHANGE

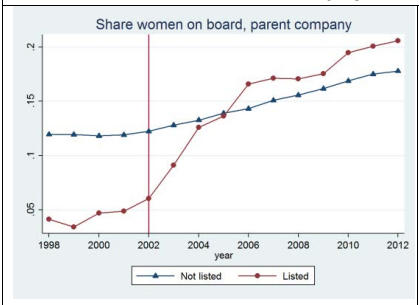
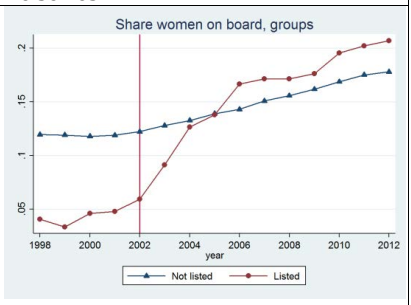
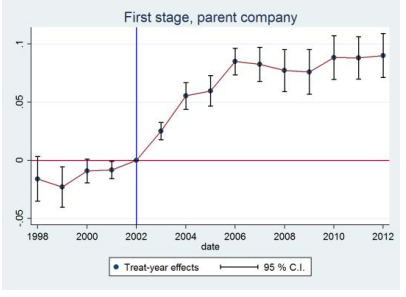
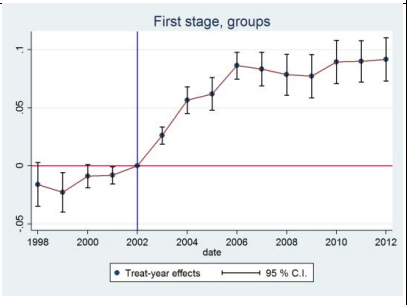
Independent firms (Parent or firms with no subsidiaries)	Independent firms but the sample where parent firms are matched with the group financial report
Panel A: Time series	
	
Panel B: Annual treatment effects	
	

FIGURE 3. SHARE OF FEMALE DIRECTORS ON BOARDS, 1998-2012

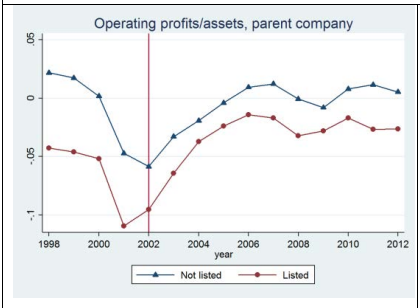
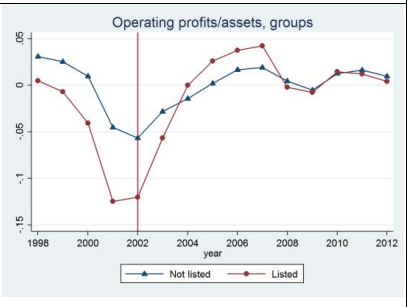
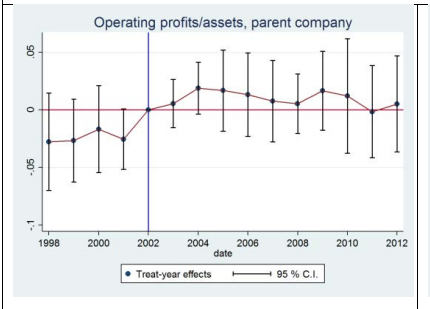
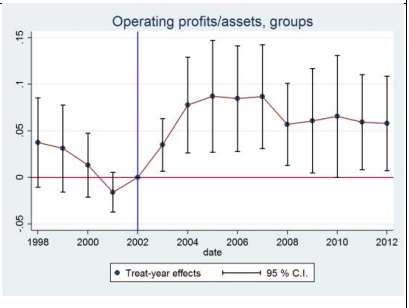
Independent firms (Parent or firms with no subsidiaries)	Independent firms but the sample where parent firms are matched with the group financial report
Panel A: Time series	
	
Panel B: Annual treatment effects	
	

FIGURE 4. PROFITS/ASSETS, 1998-2012

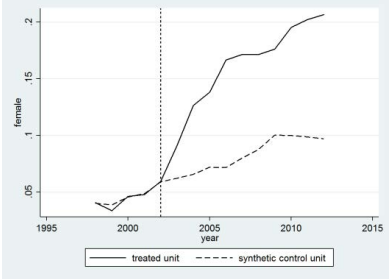
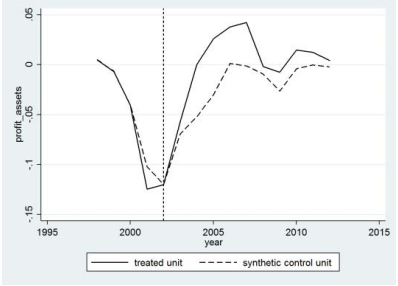
Female share of directors Independent firms but the sample where parent firms are matched with the group financial report	**Operating profits over assets** Independent firms but the sample where parent firms are matched with the group financial report
	

FIGURE 5. SYNTHETIC CONTROL (ABADIE ET AL. 2010), GROUP FINANCIAL STATEMENTS USED

TABLES

Table 1—Summary Statistics, 1998-2012

	Mean	p50	sd	Min	Max	Count
Panel A: Parent firm financial statements						
Female board share	.1390776	0	.200584	0	1	168643
Operating profits	5490755	67000	6.89e+07	-7.60e+08	3.38e+09	168534
Total assets	3.00e+08	9546000	2.57e+09	66000	1.09e+11	168563
Profits/assets	-.0094904	.0177392	.2813509	-1.743985	.5924171	168290
Total revenue	1.03e+08	6516000	4.57e+08	0	1.59e+10	168587
No. on board	5.636735	5	2.728714	1	63	169130
Labor cost/assets	.5686929	.3574007	.6485608	.0006615	3.393548	114408
Labor cost	2.71e+07	4923000	1.06e+08	9000	3.93e+09	114445
R&D costs/assets	-.0147597	0	.0576047	-.4634188	0	22609
Selling costs/assets	-.1233812	-.0051427	.2473822	-1.446863	0	22588
Performance pay board	23.39417	0	277.3693	0	12000	165156
No. employed	51.52356	5	392.4991	0	26379	162678
EBITDA	9970088	224000	9.17e+07	-5.74e+08	4.76e+09	166351
Average board age	51.56429	52	7.128098	19	97	169130
Observations	170019					
Panel B: Group financial statements						
Female board share	.1390241	0	.2005795	0	1	169079
Operating profits	2.42e+07	172000	3.78e+08	-1.09e+09	1.76e+10	168681
Total assets	4.33e+08	1.10e+07	4.38e+09	66000	2.01e+11	168706
Profits/assets	-.0036505	.0308635	.2888744	-1.775194	.5973451	168405
Total revenue	3.08e+08	1.09e+07	3.26e+09	0	1.29e+11	168752
No. on board	5.635953	5	2.728101	1	63	169566
Labor cost/assets	.4628314	.2175555	.6405655	2.03e-06	3.25526	122963
Labor cost	1.36e+07	1765000	4.13e+07	549	3.12e+08	123029
R&D costs/assets	-.0183407	0	.0653342	-.5386208	0	22869
Selling costs/assets	-.161734	-.0557467	.2627354	-1.490032	0	22847
Performance pay board	71333.71	0	1461417	0	6.60e+07	165570
No. employed	201.6156	7	2881.23	0	279641	163390
EBITDA	3.81e+07	450000	5.09e+08	-4.86e+08	2.36e+10	167317
Average board age	51.56599	52	7.131224	19	97	169566
Observations	170460					

Table 2—Effect of the Threat of a Quota Law

Outcome	(1) Basic	(2) Compositional bias test	(3) Linear Trends	(4) Collapsed
		Panel A: Effect on share of female directors		
Share Female	0.0838***	0.0816***	0.0409***	0.0840***
	(0.00505)	(0.00460)	(0.00761)	(0.00959)
		Panel B: Effect on firm performance. Parent company financial statement used		
Profits /assets	0.0260***	0.0227**	0.0273***	0.0292***
	(0.00777)	(0.00919)	(0.00660)	(0.00529)
		Panel C: Effect on firm performance. Group financial statement used		
Profits /assets	0.0516***	0.0488***	0.0658***	0.0540***
	(0.0158)	(0.0151)	(0.0186)	(0.0124)
Industry trends	No	Yes	No	No
Standard errors	Clustered at industry	Clustered at industry	Clustered at industry	Newey-West

The standard errors are clustered at the industry level (57 clusters) errors in Columns 1-4. Column 5 presents Newey-West standard errors.*p< 0.10, **p< 0.05, ***p< 0.01. The number of observations is168 643; 168 290; and 168 405 in panels A, B and C, respectively. In Column 4, the number of observations is always 15 across all panels.

Table 3—Effect of the Threat of a Quota Law, Only Groups

Outcome	(1) Basic	(2) Compositional bias test	(3) Linear Trends	(4) Collapsed
		Panel A: Effect on share of female directors, only groups		
Share Female	0.0869***	0.0781***	0.0534***	0.0833***
	(0.00736)	(0.00679)	(0.00924)	(0.00913)
		Panel B: Effect on firm performance. Group financial statement used, only groups		
Profits /assets	0.0344**	0.0303*	0.0386**	0.0354**
	(0.0165)	(0.0171)	(0.0169)	(0.0149)
Industry trends	No	Yes	No	No
Standard errors	Clustered at industry	Clustered at industry	Clustered at industry	Newey-West

The standard errors are clustered at the industry level (57 clusters) errors in Columns 1-4. Column 4 presents Newey-West standard errors.*p< 0.10, **p< 0.05, ***p< 0.01. The number of observations is 31 270 in panel A and 31 325 in panel B. In Column 4, the number of observations is always 15 across both panels.

Table 4—Other Outcomes of the Effect

	(1) Profits/assets	(2) EBITDA/assets	(3) Total revenue/assets	(4) Labor cost/assets	(5) Operating profits/employee	(6) Value added/employee
Estimate	0.0516***	0.0375**	0.0329	-0.0225*	167.6	199.8
	(0.0158)	(0.0152)	(0.0379)	(0.0134)	(332.7)	(331.4)

Standard errors in parentheses, Clustered at industry*$p<0.10$, **$p<0.05$, ***$p<0.01$.

Table 5—Additional Outcomes

	(1) Operating profits	(2) No. employed	(3) log(No. employed)	(4) Female as CEO	(5) No. on board	(6) log(No. on board)	(7) Average board age
Estimate	444628782.6***	1004.2**	0.0586	-0.0253***	0.722***	0.213***	-0.630
	(85660024.3)	(386.0)	(0.0908)	(0.00757)	(0.110)	(0.0230)	(0.496)

Standard errors in parentheses, Clustered at industry $^{*}p< 0.10$, $^{**}p< 0.05$, $^{***}p< 0.01$.

APPENDIX

Table A1—Synthetic Control Difference Estimates

	(1) Difference female	(2) Difference profits/assets
Post 2002	0.0818	0.0313
Constant	-0.000857	-0.00473
Synthetic control difference	Yes	Yes
N	15	15

Table A2—Remove Restrictions

	(1) Non-active used	(2) Board>2	(3) All board sizes	(4) 2001 as base	(5) 2 lags in NW	(6) All individual firms
			Panel A: Share females			
Estimate	0.0832***	0.0976***	0.109***	0.0795***	0.0852***	0.0838***
	(0.00490)	(0.00520)	(0.00555)	(0.00437)	(0.0107)	(0.00505)
			Panel B: Operating profits/assets			
Estimate	0.0511***	0.0661***	0.0856***	0.0337**	0.0540***	0.0260***
	(0.0160)	(0.0156)	(0.0167)	(0.0143)	(0.0115)	(0.00777)
Standard errors	Clustered at industry	Clustered at industry	Clustered at industry	Clustered at industry	Newey-West	Clustered at industry

Standard errors in parentheses, $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A3—Add Restrictions

	(1) At least 5 employees	(2) At least 10 employees	(3) At least 20 employees	(4) At least 500k SEK in share capital	(5) At least 1000k SEK in share capital
			Panel A: Share female		
Estimate	0.0821***	0.0797***	0.0755***	0.0843***	0.0841***
	(0.00561)	(0.00568)	(0.00645)	(0.00498)	(0.00541)
			Panel B: Operating profits/assets		
Estimate	0.0577***	0.0576***	0.0487***	0.0516***	0.0490***
	(0.0162)	(0.0166)	(0.0174)	(0.0160)	(0.0170)

Standard errors in parentheses, Clustered at industry $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A4— Winsorizing at Different Levels. Outcome is Profits/Assets

	(1) 1 percent	(2) 2 percent	(3) 0.5 percent	(4) No winsorizing
Estimate	0.0516***	0.0474***	0.0565***	0.0412
	(0.0158)	(0.0155)	(0.0156)	(0.0394)

Standard errors in parentheses, Clustered at industry $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A5, Panel A—Leaving One Industry out

Profits/assets	0.0514***	0.0509***	0.0517***	0.0516***	0.0498***	0.0517***	0.0504***	0.0521***	0.0515***	0.0514***
	(0.0158)	(0.0156)	(0.0158)	(0.0158)	(0.0154)	(0.0159)	(0.0155)	(0.0160)	(0.0158)	(0.0158)
Industry code	01	02	05	10	100	13	14	15	16	17
N	166774	167666	168321	168308	152283	168351	168110	166631	168384	168033

Table A5, Panel B—Leaving One Industry out

Profits/assets	0.0516***	0.0515***	0.0515***	0.0521***	0.0513***	0.0516***	0.0504***	0.0512***	0.0514***	0.0519***
	(0.0158)	(0.0158)	(0.0159)	(0.0161)	(0.0160)	(0.0158)	(0.0156)	(0.0158)	(0.0158)	(0.0160)
Industry code	18	19	20	21	22	23	24	25	26	27
N	168250	168317	166852	167648	164311	168347	167026	167585	167729	167934

Table A5, Panel C—Leaving One Industry out

Profits/assets	0.0516***	0.0520***	0.0499***	0.0532***	0.0488***	0.0515***	0.0515***	0.0515***	0.0522***	0.0516***
	(0.0160)	(0.0161)	(0.0156)	(0.0164)	(0.0149)	(0.0159)	(0.0158)	(0.0159)	(0.0160)	(0.0158)
Industry code	28	29	30	31	32	33	34	35	36	37
N	166041	165737	168169	167644	167884	167383	167751	167957	167504	168198

Table A5, Panel D—Leaving One Industry out

Profits/assets	0.0515***	0.0516***	0.0516***	0.0514***	0.0479***	0.0521***	0.0511***	0.0513***	0.0518***	0.0516***
	(0.0159)	(0.0158)	(0.0161)	(0.0158)	(0.0155)	(0.0162)	(0.0157)	(0.0157)	(0.0162)	(0.0158)
Industry code	40	41	45	50	51	52	55	60	61	62
N	163137	168294	163857	166589	152471	163709	165027	163543	167732	168232

Table A5, Panel E—Leaving One Industry out

Profits/assets	0.0522***	0.0498***	0.0558***	0.0515***	0.0554***	0.0508***	0.0513***	0.0443***	0.0528***	0.0712***
	(0.0161)	(0.0152)	(0.0176)	(0.0158)	(0.0167)	(0.0164)	(0.0159)	(0.0125)	(0.0166)	(0.0180)
Industry code	63	64	65	66	67	70	71	72	73	74
N	164001	167469	165716	168343	165579	151045	167032	160375	166061	137007

Table A5, Panel F—Leaving One Industry out

Treated	0.0516***	0.0491***	0.0509***	0.0514***	0.0515***	0.0514***	0.0517***
	(0.0158)	(0.0151)	(0.0157)	(0.0158)	(0.0158)	(0.0159)	(0.0158)
Industry code	75	80	85	90	91	92	93
N	168231	165678	165258	167531	166568	163004	168063

Standard errors in parentheses, Clustered at industry. $^{*}p<0.10$, $^{**}p<0.05$, $^{***}p<0.01$.

Table A6—Window Size and Alternative Female Measure

	(1) Share females, only known ID	(2) Share females	(3) Profits/assets
Treated	0.0904***	0.0520***	0.0456***
	(0.00548)	(0.00652)	(0.0116)
Window	1998-2012	1998-2004	1998-2004
N	164311	88231	87239

Standard errors in parentheses. Note: The standard errors are clustered at the industry level (57 clusters).

*$p<0.10$, **$p<0.05$, ***$p<0.01$.

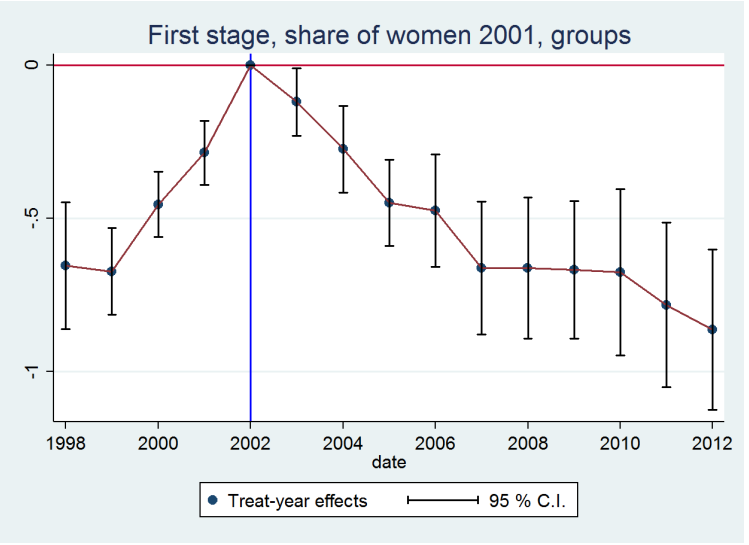


FIGURE A1. ALTERNATIVE FIRST STAGE ANNUAL EFFECTS

Differences in prison sentencing between genders and immigration background in Sweden: Discrepancies and possible explanations*

Joakim Jansson†

Abstract

I use data on punished drunk drivers to document differences in sentencing for the same crime between immigrants and native born and males and females, respectively. Differences in past criminal activity or other individual observables cannot explain the difference in sentencing. Instead, the difference between immigrants and native born seems to be due to statistical discrimination, while differences in recidivism rates might explain the gender difference. However, the higher incarceration rate for immigrants does not reduce their future number of crimes. Overall, the findings suggest that there is room for both equity and efficiency improvements in incarceration rates.

Keywords: crime, discrimination, conviction, punishment severity, drunk driving
JEL: J71, K0, K14, K40

*I thank Björn Tyrefors, Jonas Vlachos, David Strömberg, Lena Hensvik, Mahmood Arai and seminar participants at the department of economics at Stockholm University.

†Department of Economics, Stockholm University, Research Institute of Industrial Economics, Stockholm Internet Research Group, Stockholm University

1 Introduction

A shared view among democratic countries over the last centuries has been that everyone should be equal before the law. Yet research in recent years has shown that the matching of juror, defendant and victim characteristics influences the probability of conviction (Anwar et al., 2012, 2015), that jurors' age increases the likelihood to convict (Anwar et al., 2014) and that judges differ substantially in conviction rates and the harshness of the punishment across races (Abrams et al., 2012). This literature is generally motivated by the right to a fair and impartial trial, and suggests that, for instance, people of different races are not treated equally under the current rule of law. However, the findings in the above mentioned literature mainly focus on the effect of judge and jury characteristics and/or the matching of jury and defendant characteristics and not the overall level of discrimination in the legal system. For one, none of the papers present evidence on whether or not immigrants are sentenced more severely for the same crime. In this paper, I hence try to shed some light on whether males or immigrants are sentenced differently than females and native born in courts in Sweden for the same committed crime.

I show that immigrants and males are more likely to receive a prison sentence for the same crime as compared to native born and females. More specifically, I look at the probability of being sentenced to prison for a given level of blood-alcohol content (BAC henceforth) when caught in a DUI (driving under influence) control and document a 0.1 higher probability for immigrants and a 0.05 higher probability for males to receive a prison sentence. I then proceed by investigating possible explanations for the observed differences. Controlling for other individual characteristics leaves the estimates largely unchanged, making unobserved differences an unlikely explanation. However, there is a difference in the crime recidivism rate between males and females that can potentially explain the difference in sentencing. Between immigrants and native born, the difference in recidivism disappears once individual characteristics are taken into account, suggesting that statistical discrimination in perceived recidivism rates might cause the difference in sentencing. Yet, using the found discontinuity in the probability of receiving a prison sentence around BAC=1 from Hinnerich et al. (2016) and using it in a

difference-in-discontinuity (Grembi et al., 2016) like setting between immigrants and native born reveals that immigrants are not less likely to reduce their future criminal activities due to the extra imprisonment they receive. Unfortunately, the same exercise cannot be done for the gender difference.

The previous economic literature on discrimination in the legal system has in general focused on either judge and jury characteristics or how the matching of judge/jury and defendant characteristics affects the trial. Few deal with the potential overall discrimination in courts. An exception is Steffensmeier et al. (1993) who use guidelines sentencing data from Pennsylvania from 1985-1987 in a regression control framework. Their data allows them to control for severity of the crime as well as individual characteristics. Using this approach, they find a small effect of lesser imprisonment of females. Nordström (1998) also employs a regression control framework using Swedish data and DUI offenses. He finds that prior to the reform of the drunk driving regulations on the 1st of July in 1990, immigrants fared no worse than the native born. However, after the reform, which explicitly stated that other factors besides the BAC-level should be taken into account, immigrants received relatively more prison sentences. Anwar et al. (2015) take a different approach and instead investigate the bias of politically affiliated lay jurors in Sweden. They find that lay jurors typically vote to a larger degree in accordance with their political affiliation. Anwar et al. (2012) in turn show that jury pools consisting of white-only jurors are 16 percent more likely to convict black than white defendants, while Abrams et al. (2012) instead show that judges differ significantly in their incarceration rates between black and white defendants. McConnell and Rasul (2017) compare Hispanics to whites that committed crimes prior to 9-11 but were either sentenced before or after that date and find that the difference increases after the attack. Finally, Goncalves and Mello (2017) show that police officers are less likely to reduce the charged speed of blacks than whites around discontinuities in speed for fines.

I contribute to this literature in two ways. First, since I only use people that have committed DUI offenses, I have a simple measure which I can use to hold the crime constant within a very small neighborhood. This allows me to rule out any uncertainties regarding whether one group on average commits worse crimes. The detailed register data of Sweden further allows me to control for

a wide arrangement of other control variables, including past criminal activity, and I let a Lasso algorithm from Belloni et al. (2014a) pick which of all these variables to include as controls. Second, the data allows me to rule out different explanations and mechanisms behind the found relationship, such as immigrants refusing treatment for alcohol abuse, higher recidivism rates or differing responses to imprisonment.

My paper is also related to other strands of literature. It lies close to Hansen (2015) and Hinnerich et al. (2016), both of which estimate the deterrence effect of receiving a prison sentence using an RD-design on convictions from DUI offenses. Methodologically, this paper is also reminiscent of Schwartz et al. (2016), which documents a retention heterogeneity effect in New York schools, using a similar RD-design as parts of this paper. Kuziemko (2012) looks into parole boards and finds that they typically decrease societal costs by granting parole to inmates that are less likely to fall back into criminal behavior. This paper, in turn, displays what seems to be an inefficient use of legal resources, since imprisonment is a costly punishment in terms of expenses.

The rest of the paper is outlined as follows. Section 2 provides some background and describes the data and the econometric strategy. Section 3 then presents the main results, while section 4 concludes the paper.

2 Data, econometric strategy and some background

2.1 Background

The Swedish court system differs somewhat from its Anglo-Saxon counterpart. There is no jury. Instead, when prison is a potential sentence, a professional judge along with three lay jurors known as *nämndemän* decide the outcome of the criminal case. If the harshest punishment is a fine, then a single professional judge decides the verdict.[1] The lay jurors are politically affiliated, appointed officials

[1]More specifically, it is stated in The Swedish Code of Judicial Procedure, first chapter, section 3b, second paragraph "During the main hearing of a criminal case for which the punishment is not harsher than fines or a six-month prison sentence the district court is able to rule without any *nämndeman* present, if there is no reason for the sentence to be any other than fines..." (my translation). In Swedish the paragraph reads "Vid huvudförhandling i mål om brott för vilket

that serve four-year terms and are more or less randomly assigned to cases. In order to become elected as a *nämndeman*, you need to be nominated by a political party and then be appointed by either the municipality or the county council. When it comes to the court's ruling, the verdict does not have to be unanimous, it is enough that a majority of the *nämndemen* and the judge find the defendant guilty.

In 1990, some major changes were made to the Swedish law on driving under influence. First, the threshold for being eligible for fines was decreased from 0.5 to 0.2 BAC. Secondly, where the BAC-level had previously been the sole determinant of the punishment, the law was changed in order for the court to take other circumstances into account as well. These included how reckless the driver was, if other substances were used as well, the age of the perpetrator, previous crimes and the overall life situation. Four years later, in 1994, the threshold for being subject to imprisonment was lowered from 1.5 to 1.[2] However, a convention still remained that BAC-levels above 1.5 should receive an even harsher punishment. These two cut-off points remain to this very day, which is documented in Hinnerich et al. (2016) and can also be seen in figures 1A and 1B.

2.2 Data

In this paper I use the same data as Hinnerich et al. (2016).[3] It consists of a combined data set on the universe of recorded BAC-levels from breathalyzer tests and blood tests between 2008 and 2012 along with the sentences related to these crimes from the courts. Breathalyzer- and blood-tests were obtained from the National Forensics Centre (NFC), while the data on the verdicts comes from the Swedish National Council for Crime Prevention. In addition, I have access to some of the variables from the Longitudinal integration database for health insurance and labour market studies (LISA), which is based on Swedish registers. Regarding

inte är föreskrivet svårare straff än böter eller fängelse i högst sex månader är tingsrätten domför utan nämndemän, om det inte finns anledning att döma till annan påföljd än böter och det i målet inte är fråga om företagsbot."

[2]Thus, in the data we should expect *nämndeman* to be present more frequently on verdicts with a BAC equal to or above 1 and 1.5, respectively, and to a lesser degree at lower level cases.

[3]I thank Björn Tyrefors Hinnerich, Mikael Priks and Per Pettersson-Lidbom for gaining access to this data.

the procedure performed by the police in suspicion of drunk driving, when caught in a police control, the individual first performs a breathalyzer test. If it comes out positive, two new tests are administered and in court the average of these two, with 0.15 deducted, is the proof material. If the individual refuses the test or there is suspicion that other substances have been used as well, a blood test is performed at the police station. Furthermore, I restrict the sample to the verdicts where drunk driving is the only crime in order to ensure that those within the same BAC-interval have committed the same offense. This leaves me with 29788 observations. However, I also restrict my main attention to the BAC-levels for which prison is a viable sentence, implying that only those caught in a DUI with at least a BAC above 1 will be considered.[4] This leaves me with 10520 observations, the summary statistics of which are provided in table 7 along with a subset of covariates.[5]

2.3 Econometric strategy

I first start by using local linear regressions to construct graphs of the difference in prison sentences across BAC values between native born and immigrants and males and females. Specifically, I estimate the probability of being sentenced to prison for all observations belonging to group k with $1 \leq BAC \leq 3$ with local linear regressions, taking the cut-offs at $BAC = 1$ and $BAC = 1.5$ into account. Formally, I estimate

$$\{\hat{\alpha}, \hat{\beta}\} = \underset{\alpha,\beta}{\arg\min} \sum_{i=1}^{N} w_i(y_i - \alpha - \beta(x_i - x_0))^2, \quad \forall k \in \{1, 2\} \tag{1}$$

within the rule-of-thumb selected bandwidth, were w_i is the kernel weight, x_i is the BAC-level, y_i is a dummy taking the value 1 if the individual i is sentenced to prison and k is to which group the individual belongs; female/male or native born/immigrant.[6] I then proceed to compute the main estimates for the difference

[4] I also restrict the sample from above, including only those with a BAC less than 3. However, there are exceedingly few with that level of blood alcohol content.

[5] A full list of all available control variables is provided in table 8.

[6] In practice I use the lpoly command in Stata for the local linear regressions, using it separately for each group.

in sentencing by using BAC fixed effects, where each fixed effect is computed on a 0.1 BAC-interval, thus giving me 20 BAC fixed effects. This is wrapped up in equation 2, where β will measure the effect of either being immigrant or female as compared to being native born or male, controlling for 20 BAC fixed effects (BAC_j).

$$y_i = \beta\, group_i + \sum_{j=1}^{20} \phi_j BAC_j + \varepsilon \tag{2}$$

I also estimate the difference in cut-off probability around $BAC = 1$ and $BAC = 1.5$ for being sentenced to prison between both immigrants against native born and females against males. These specifications will in essence be difference-in-discontinuity estimates, as first outlined in Grembi et al. (2016). I will thus estimate regressions of the form:

$$y_i = \alpha + \beta\, group_i * \mathbf{1}[BAC_i > cutoff] + \mathbf{1}[BAC_i > cutoff]$$

$$+\mathbf{1}[BAC_i > cutoff] * f(BAC_i) + group_i * \mathbf{1}[BAC_i > cutoff] * f(BAC_i)$$

$$+group_i * f(BAC_i) + f(BAC_i) + \pi group_i + \varepsilon_i \tag{3}$$

where β is the coefficient of interest, y_i is a dummy indicating if individual i was sentenced to prison, $group_i$ is a dummy taking the value 1 if the individual is an immigrant or a female, depending on the specification, $\mathbf{1}[BAC_i > cutoff]$ is an indicator function for if the BAC value is above the cutoff and $f(BAC_i)$ is the control function in the running variable. I estimate this specification separately for the upper ($BAC = 1.5$) and lower ($BAC = 1$) cut-offs.

The benefit of this specification is that it allows me to compare immigrants or females just above and below the cutoff as well as the difference to their comparison group (native born or males). While the comparison group (born in Sweden or males) might not have the same characteristics as the group under review (immigrants or females) for any given BAC-level, the group's individuals just above and below the cutoff should be comparable in all dimensions except the probability of receiving a prison sentence. This implies that any potential difference in discontinuity is driven by the increased amount of prison sentences just above the cut-off. It thus leaves us with two potential explanations for any observed difference in

cut-off probability; either discrimination only occurs when prison is a potential outcome, or some other discontinuity that affects the difference is present at the same BAC-level. For instance, as is stated in section 2.1, the three lay jurors or *nämndemän* are only present when prison is a viable outcome of the trial. Thus, as *nämndemän* are politically affiliated, any difference in discontinuity might be due to the introduction of *nämndemän*. Unfortunately, this is a potential mechanism that I cannot yet look into; however, I hope to gather some data in the near future to investigate this.

Unlike standard regression discontinuity approaches, the discontinuity parameters will not be non-parametrically identified in equation 3. Rather do they rely on additive separability in the difference in discontinuity.

3 Results

Figure 1, panels A and B sum up the main attention of this paper. The graph shows the difference in the probability of being sentenced to prison between immigrants and native born and males and females, respectively. It is evident that immigrants and males are substantially more likely to be sentenced to prison given that they have committed the same crime. This is also summarized numerically in table 1, panels A and B.[7] The first column of the table simply computes the average difference in probability of being sentenced to prison, not taking the severity of the DUI offense into account. The second column introduces the BAC fixed effects, as showed in equation 2. As can be seen, this does not substantially change the results. Columns 3 and 4 include past crimes, age and kids at home as control variables, with little to no change in the coefficients. Finally, column 5 includes control variables picked by a Lasso model out of all available background variables in my data.[8] More specifically, I run the Lasso using all available background

[7]The standard errors in the table are clustered at the individual level. This is due to the fact that a few individuals appear more than once in the data.

[8]The Lasso (abbreviation for Least Absolute Shrinkage and Selection Operator) is a regression analysis method which both does variable selection and regularization in order to increase prediction precision and simplify the interpretation. In essence, it reduces the coefficient estimates towards zero in order to balance the variance-bias trade-off, with some variable coefficients being reduced to zero.

characteristics in my data on first the outcome variable, in other words prison sentence, and then the treatment variable (immigrant and female, respectively) controlling for BAC fixed effects as well as age and past number of crimes, as suggested in Belloni et al. (2014b) and Belloni et al. (2014a). Then, I use the union of the picked covariates as the control variables in column 5.[9] If anything, the inclusion of these controls increases the magnitude of the coefficient. Overall, the estimates imply that immigrants are about 10 percent more likely to be sentenced to prison for the same crime, while females are between 5 and 7 percent less likely to be sentenced to prison for the same crime.

[9]In practice, I use the Stata ado-file lassoShooting from the supplementary data to Belloni et al. (2014a).

Figure 1

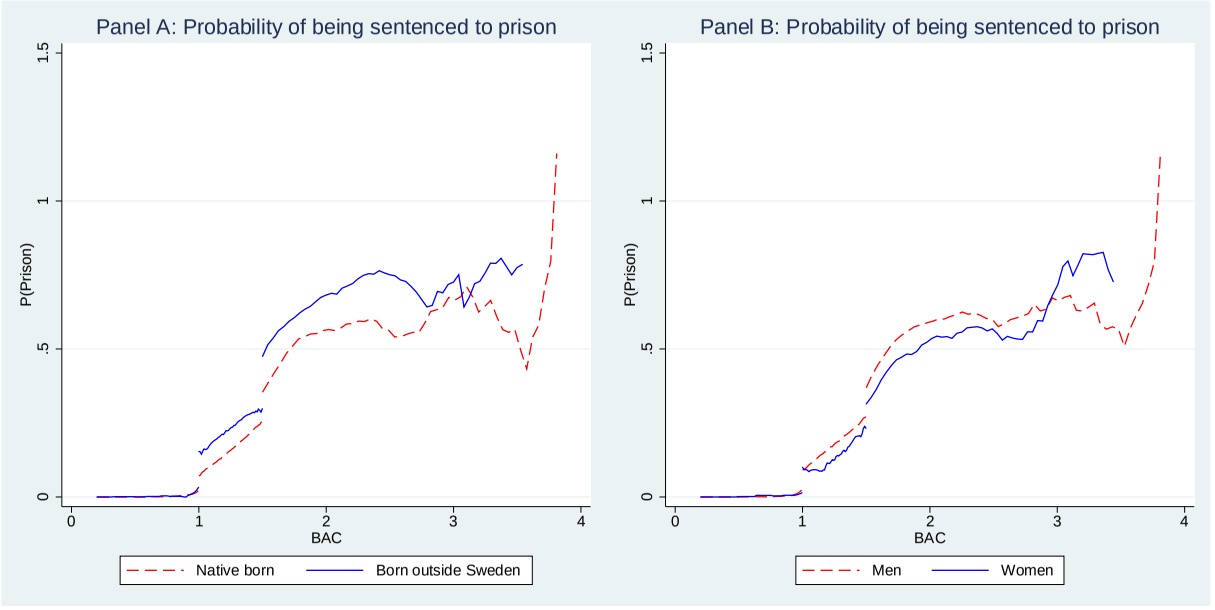

Panel A: Probability of being sentenced to prison
P(Prison)
0
.5
1
1.5
0
1
2
3
4
BAC
Native born
Born outside Sweden
Panel B: Probability of being sentenced to prison
P(Prison)
0
.5
1
1.5
0
1
2
3
4
BAC
Men
Women

Table 1: Panel A, Immigrant difference in sentencing

	(1) Prison	(2) Prison	(3) Prison	(4) Prison	(5) Prison
Immigrant	0.0866***	0.0982***	0.0975***	0.0914***	0.110***
	(0.0132)	(0.0119)	(0.0119)	(0.0124)	(0.0241)
Past No. crimes			-0.00242***	-0.00271***	-0.00301***
			(0.000701)	(0.000724)	(0.000727)
Age				0.00168***	0.00250***
				(0.000321)	(0.000387)
No. of kids at home				-0.0154***	
				(0.00558)	
BAC intervall	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3
BAC FEs	No	Yes	Yes	Yes	Yes
Lasso controls	No	No	No	No	Yes
N	10520	10520	10520	10007	10007

Table 1: Panel B, Gender difference in sentencing

	(1) Prison	(2) Prison	(3) Prison	(4) Prison	(5) Prison
Female	-0.0472***	-0.0520***	-0.0562***	-0.0595***	-0.0701***
	(0.0128)	(0.0117)	(0.0117)	(0.0119)	(0.0158)
Past No. crimes			-0.00284***	-0.00313***	-0.00323***
			(0.000708)	(0.000731)	(0.000745)
Age				0.00184***	0.00234***
				(0.000322)	(0.000400)
No. of kids at home				-0.0112**	
				(0.00563)	
BAC intervall	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3
BAC FEs	No	Yes	Yes	Yes	Yes
Lasso controls	No	No	No	No	Yes
N	10520	10520	10520	10007	10007

Standard errors in parentheses
Note: The standard errors are clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Thus, we can conclude that it seems highly unlikely that any unobserved individual characteristics will be able to explain the difference in sentencing. Below I hence investigate other potential explanations for the observed discrepancy. First, table 2 investigates what sentences immigrants and females receive instead of prison, again controlling for BAC fixed effects. The first column looks at the difference in probability of being sentenced to rehabilitation for alcohol addiction. This is of particular interest since Nordström (1998) notes that his found difference in prison sentencing might be due to the fact that immigrants reject rehabilitation

to a greater degree than native born. We can note that immigrants are less likely to be sentenced to rehabilitation than native born; however, this difference is relatively minor compared to the difference in the probability of being sentenced to prison. But when it comes to the gender difference, the coefficient is of the opposite sign and of equal magnitude as the results in table 1. In Sweden, there are two probation sentences, one of them being more severe and automatically turning into a harder prison sentence if the offender misbehaves during the probation period, while the other is somewhat milder. I have labeled these two sentences harsh and mild probation,[10] and columns 2 and 4 look into the difference in probability for these two sentences. We can note that both immigrants and males are relatively more likely to be sentenced to the harsh probation sentence, while the opposite is true for the milder one. However, the difference is considerably larger for the milder sentences. Finally, column 3 reports the difference in probability of instead receiving a fine, which should be viewed as the least harsh punishment overall. We can see that immigrants and females are less likely to receive fines as punishment. In sum, we can conclude that the largest differences are among the mild probation sentences. Furthermore, immigrants are more likely to receive the harsher punishment and less likely to receive the milder ones as compared to native born, while the same is not obviously true for females and males, seeing that males are more likely to be sentenced to fines. We can also note that denying rehabilitation treatment for alcohol addiction cannot explain the immigrant difference, though potentially the gender one.

[10]For those more familiar with the Swedish legal system, harsh probation corresponds to *villkorlig dom*, while the milder one is *skyddstillsyn*.

Table 2: Panel A, Immigrants' other sentences

	(1) Rehab	(2) Harsh probation	(3) Sentenced to fine	(4) Mild probation
Immigrant	-0.0224***	0.0423***	-0.0340***	-0.0814***
	(0.00655)	(0.0113)	(0.00719)	(0.00774)
BAC intervall	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3
BAC FEs	Yes	Yes	Yes	Yes
N	10520	10520	10520	10520

Table 2: Panel B, Females' other sentences

	(1) Rehab	(2) Harsh probation	(3) Sentenced to fine	(4) Mild probation
Female	0.0484***	-0.0259**	-0.0558***	0.0942***
	(0.00868)	(0.0110)	(0.00647)	(0.0112)
BAC intervall	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3	1≤ BAC <3
BAC FEs	Yes	Yes	Yes	Yes
N	10520	10520	10520	10520

Standard errors in parentheses
Note: The standard errors are clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3 looks into the difference in recidivism rate for men compared to women and immigrants compared to native born. The purpose of the table is to demonstrate how good predictors being immigrant and female really are for future criminal behavior, as this might influence a rational judge to give harder sentences to the group more likely to relapse back into criminality. However, Hinnerich et al. (2016) demonstrate that mild prison sentences deter future criminal activity, implying that the outcome is endogenous when prison is available as a punishment. I thus only focus on individuals with a $BAC < 1$, as 98.6 percent of all in this interval are sentenced to fines. Since almost everyone receives fines as punishment in this interval, any potential endogeneity issues between punishment and future criminal activity should be out of the question, since everyone is punished in the same way in any case.

The first column simply shows the difference between immigrants and native born and males and females in the future number of crimes. The coefficients imply that immigrants commit 0.08 more future crimes and females 0.14 less, compared to a mean of 0.31 and a standard deviation of 1.15 in this interval. The second column includes BAC fixed effects. We can note that not much happens to neither the immigrant nor the gender difference in crime recidivism. The third and fourth column then introduce the number of fines as well as the past number of crimes as

control variables. While the introduction of the number of fines does not change anything to any considerable extent, the past number of crimes cuts the estimate for women in half. Finally, column five once again uses the Lasso to select the optimal control variables out of the available ones, controlling for the number of fines, the past number of crimes and BAC fixed effects. The introduction of these causes the coefficient for immigrants to fall to close to zero, although the standard error is too large to rule out sizable effects in either direction. For the gender difference, however, the estimate has returned to -0.11. We can thus conclude that once individual characteristics are taken into account, there seems to be no significant difference between immigrants and native born in crime recidivism. If the judges deciding the verdict do not have access to the same background information as us, or fail to take it into account, the results regarding immigrants above are in line with a statistically discriminatory behavior. However, this is not true for the gender difference.

One other possibility is that the difference in sentencing between immigrants and native born is due to the fact that prison as a punishment is more effective against immigrants. Next, I will thus use differences in cut-off probabilities to see if immigrants and native born respond in the same way to being sentenced to prison.

Table 3: Panel A, Immigrant as predictor of recidivism

	(1) Future No. crimes	(2) Future No. crimes	(3) Future No. crimes	(4) Future No. crimes	(5) Future No. crimes
Immigrant	0.0798***	0.0748***	0.0740***	0.0815***	0.00501
	(0.0225)	(0.0226)	(0.0189)	(0.0182)	(0.0410)
No. fines			-0.00682***	-0.00325**	0.000572***
			(0.00168)	(0.00147)	(0.000204)
Past No. crimes				0.0625***	0.0624***
				(0.00766)	(0.00809)
BAC intervall	0.2≤ BAC <1	0.2≤ BAC <1	0.2≤ BAC <1	0.2≤ BAC <1	0.2≤ BAC <1
BAC FEs	No	Yes	Yes	Yes	Yes
Lasso controls	No	No	No	No	Yes
N	19172	19172	18908	18908	18370

Table 3: Panel B, Female as predictor of recidivism

	(1) Future No. crimes	(2) Future No. crimes	(3) Future No. crimes	(4) Future No. crimes	(5) Future No. crimes
Female	-0.141***	-0.143***	-0.101***	-0.0397**	-0.113***
	(0.0187)	(0.0188)	(0.0174)	(0.0177)	(0.0229)
No. fines			-0.00663***	-0.00312**	0.000562***
			(0.00167)	(0.00147)	(0.000206)
Past No. crimes				0.0622***	0.0618***
				(0.00770)	(0.00814)
BAC intervall	0.2≤ BAC <1	0.2≤ BAC <1	0.2≤ BAC <1	0.2≤ BAC <1	0.2≤ BAC <1
BAC FEs	No	Yes	Yes	Yes	Yes
Lasso controls	No	No	No	No	Yes
N	19172	19172	18908	18908	18370

Standard errors in parentheses
Note: The standard errors are clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.1 Differences in cut-off probability

If immigrants are less likely to commit future crimes after being imprisoned, it may make sense to sentence them to prison to a greater degree if we want to reduce crimes as much as possible for a given prison budget. In this section, I follow Hinnerich et al. (2016) and look at how imprisonment affects immigrants and native born differently. Figure 2, panel A and B, illustrates the difference in cut-off probability of being sentenced to prison around the cut-offs at $BAC = 1$ and $BAC = 1.5$ between immigrants and native born. Especially at the lower discontinuity, it is evident that there is a larger increase for immigrants than for native born at the discontinuity, though a similar pattern exists at the upper cut-off as well. Figure 3, in turn, shows the same discontinuities but for the gender difference. Once more, one can see what appears to be a difference in discontinuity, in particular at the lower cut-off.

Figure 2

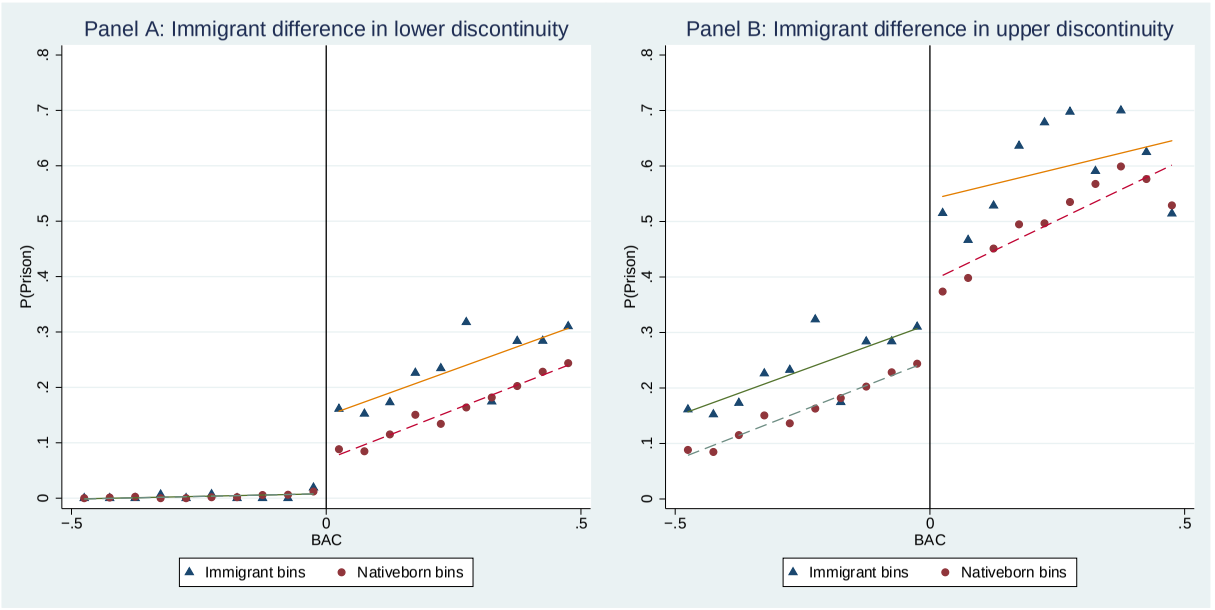

Panel A: Immigrant difference in lower discontinuity
Panel B: Immigrant difference in upper discontinuity
P(Prison)
0
.1
.2
.3
.4
.5
.6
.7
.8
−.5
0
.5
BAC
Immigrant bins
Nativeborn bins

Figure 3

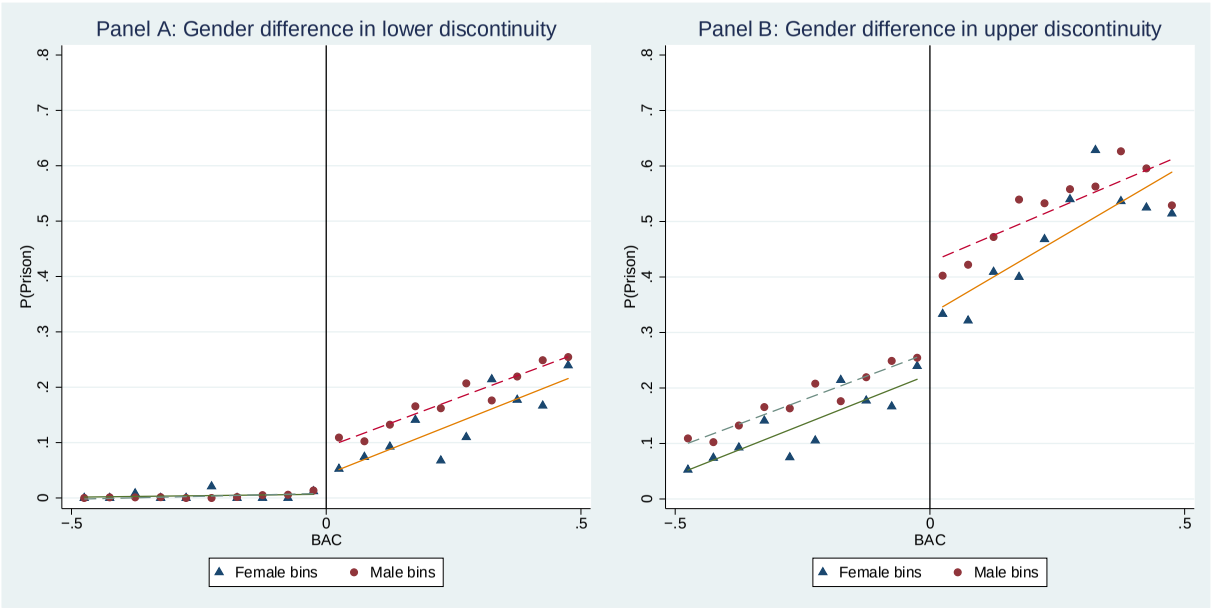

Panel A: Gender difference in lower discontinuity
Panel B: Gender difference in upper discontinuity
P(Prison)
0
.1
.2
.3
.4
.5
.6
.7
.8
−.5
0
.5
BAC
Female bins
Male bins

Table 4, panels A and B, presents the coefficients from the regressions corresponding to figure 2, panels A and B. First, one can note that the coefficient is quite stable in panel A around the lower cut-off, and on average seems to fluctuate slightly below the estimated difference from table 1. Second, the difference in panel B around the upper discontinuity is in general somewhat smaller, and not significantly different from zero. However, as far as the numeric estimate goes, the coefficients are still in the same ball park as those from panel A. Table 5 in turn displays the same estimates but for the gender difference. In general, the coefficients are quite a lot smaller than in table 4 both at the upper and lower discontinuity, and sway back and forth between negative and positive. There thus seems to be no gender difference in the discontinuity probability of being sentenced to prison. All in all, we might thus use the relationship found in table 4, panel A as a first stage in order to look at the difference in treatment effect between immigrants and native born.

Table 4: Panel A, Immigrant RD estimates

	(1) Prison sentence	(2) Prison sentence	(3) Prison sentence	(4) Prison sentence	(5) Prison sentence
Immigrant*Above	0.0791***	0.0795**	0.0691*	0.0963*	0.119
	(0.0285)	(0.0316)	(0.0376)	(0.0494)	(0.0898)
Above	0.0634***	0.0637***	0.0642***	0.0638***	0.0232
	(0.0101)	(0.0111)	(0.0127)	(0.0162)	(0.0281)
N	11713	8679	6140	3586	1181
Bandwidth	0.45	0.35	0.25	0.15	0.05
Linear spline	Yes	Yes	Yes	Yes	Yes
Immigrant * Linear spline	Yes	Yes	Yes	Yes	Yes
Cut-off	Lower	Lower	Lower	Lower	Lower

Table 4: Panel B, Immigrant RD estimates

	(1) Prison sentence	(2) Prison sentence	(3) Prison sentence	(4) Prison sentence	(5) Prison sentence
Immigrant*Above	0.0666	0.0730	0.0507	0.0619	0.0669
	(0.0589)	(0.0675)	(0.0794)	(0.104)	(0.188)
Above	0.115***	0.108***	0.0981***	0.0973***	0.0237
	(0.0216)	(0.0245)	(0.0286)	(0.0368)	(0.0636)
N	7988	6349	4677	2872	966
Bandwidth	0.45	0.35	0.25	0.15	0.05
Linear spline	Yes	Yes	Yes	Yes	Yes
Immigrant * Linear spline	Yes	Yes	Yes	Yes	Yes
Cut-off	Upper	Upper	Upper	Upper	Upper

Standard errors in parentheses
Note: Standard errors clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Panel A, Gender RD estimates

	(1) Prison sentence	(2) Prison sentence	(3) Prison sentence	(4) Prison sentence	(5) Prison sentence
Female*Above	-0.0267	-0.0265	0.00899	0.00674	0.0965
	(0.0250)	(0.0290)	(0.0331)	(0.0451)	(0.0924)
Above	0.0823***	0.0826***	0.0766***	0.0795***	0.0282
	(0.0105)	(0.0116)	(0.0134)	(0.0170)	(0.0287)
N	11713	8679	6140	3586	1181
Bandwidth	0.45	0.35	0.25	0.15	0.05
Linear spline	Yes	Yes	Yes	Yes	Yes
Female * Linear spline	Yes	Yes	Yes	Yes	Yes
Cut-off	Lower	Lower	Lower	Lower	Lower

Table 5: Panel B, Gender RD estimates

	(1) Prison sentence	(2) Prison sentence	(3) Prison sentence	(4) Prison sentence	(5) Prison sentence
Female*Above	-0.0304	-0.0620	-0.0398	-0.0133	0.0847
	(0.0562)	(0.0639)	(0.0740)	(0.0945)	(0.172)
Above	0.130***	0.127***	0.111***	0.106***	0.0178
	(0.0219)	(0.0248)	(0.0291)	(0.0376)	(0.0652)
N	7988	6349	4677	2872	966
Bandwidth	0.45	0.35	0.25	0.15	0.05
Linear spline	Yes	Yes	Yes	Yes	Yes
Female * Linear spline	Yes	Yes	Yes	Yes	Yes
Cut-off	Upper	Upper	Upper	Upper	Upper

Standard errors in parentheses
Note: Standard errors clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6 hence looks at the difference in reduced form at the lower cut-off between immigrants and native born on crime recidivism. However, it is important to note that all first-stage relationships in table 4, panel A are week, and thus the estimates in table 6 should be treated with some caution.[11] Nevertheless, there is no sign of prison having a more deterrent effect on immigrants' future criminal behavior. If anything, it appears to be a less efficient policy against immigrants, although as said, the statistical relationships found for the smaller bandwidths should be interpreted with extra caution. Unfortunately, we cannot do the same exercise for the gender difference, since we have no first-stage relationship.

[11] For instance, the F-statistic for the first column in table 4, panel A is 7.68.

Table 6: Crime recidivism, reduced form

	(1) Future No. crimes	(2) Future No. crimes	(3) Future No. crimes	(4) Future No. crimes	(5) Future No. crimes
Immigrant*Above	0.0725	0.110	0.165	0.356*	0.632**
	(0.109)	(0.120)	(0.148)	(0.183)	(0.278)
Above	-0.128***	-0.148***	-0.223***	-0.149*	-0.451***
	(0.0485)	(0.0574)	(0.0714)	(0.0828)	(0.172)
N	11713	8679	6140	3586	1181
Bandwidth	0.45	0.35	0.25	0.15	0.05
Linear spline	Yes	Yes	Yes	Yes	Yes
Immigrant * Linear spline	Yes	Yes	Yes	Yes	Yes
Cut-off	Lower	Lower	Lower	Lower	Lower

Standard errors in parentheses
Note: Standard errors clustered at the individual level.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Conclusion

In this paper I document a difference in the probability of being sentenced to prison between immigrants and native born and males and females for the same committed DUI offense. While the evidence presented here is not enough to conclude discriminatory behavior in either case, the potential explanations behind the two differences seem to differ. The gender difference in prison sentences seems to be in line with equalizing future crime recidivism between genders. However, the same cannot be said about the difference between immigrants and native born. This instead seems to be driven by statistical discrimination in crime recidivism. The results suggest that everyone is indeed not equal before the law in Sweden, and that there seems to be room for both efficiency and equity improvements when it comes to prison utilization and immigrants.

It is worth noting at this point as well that there really should not be any difference in discontinuity around the cut-offs in BAC unless immigrants are truly more harshly sentenced even as we hold their background characteristic constant. The intuition for this is simple; though a simple comparison between two groups (born in Sweden and immigrants for instance) can tell us that everyone is not punished in the same way for the same crime, we cannot yet conclude that this is due to immigration background. This is due to the fact that although they are being punished for the same crime, the two compared groups might not have the same (un)observed characteristics for any given BAC-level. Above, I have tried to take this into account using the very detailed register data in Sweden. However, when one uses the difference-in-discontinuity approach, as in section 3.1, the group individuals just above and below the cutoff should be comparable in all dimensions except the probability of receiving a prison sentence. This implies that any potential difference in discontinuity is driven by the increased amount of prison sentences just above the cut-off. But since it is stated in Swedish law that politically affiliated lay jurors called *nämndemän* should be present in rulings where prison is a likely sentence, a possible explanation for the observed difference in discontinuity is the fact that lay jurors are present in the rulings. Furthermore, previous research has shown that lay jurors both vote in accordance with their political affiliation and may sway their fellow jurors their way, thus potentially

altering the sentences (Anwar et al. (2015)). In addition, the difference observed at the lower cut-off might simply be due to the fact that so few are sentenced to prison below the cut-off. In upcoming work, I will thus try to obtain more data around the upper cut-off ($BAC = 1.5$) in order to rule out the difference being due to no prison sentences being handed out below $BAC = 1$. I will also try to gather some data on the political lay jurors in order to evaluate whether or not they are a viable explanation for the observed difference in discontinuity.

5 References

Abrams, D. S., M. Bertrand, and S. Mullainathan (2012). Do judges vary in their treatment of race? *The Journal of Legal Studies 41*(2), 347–383.

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.

Anwar, S., P. Bayer, and R. Hjalmarsson (2014). The role of age in jury selection and trial outcomes. *The Journal of Law and Economics 57*(4), 1001–1030.

Anwar, S., P. Bayer, and R. Hjalmarsson (2015). Politics in the courtroom: Political ideology and jury decision making. Technical report, National Bureau of Economic Research.

Belloni, A., V. Chernozhukov, and C. Hansen (2014a). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives 28*(2), 29–50.

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

Goncalves, F. and S. Mello (2017). A few bad apples? racial bias in policing. *Industrial Relations Section Working Paper 608*, 1–39.

Grembi, V., T. Nannicini, and U. Troiano (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics 8*(3), 1–30.

Hansen, B. (2015). Punishment and deterrence: Evidence from drunk driving. *The American Economic Review 105*(4), 1581–1617.

Hinnerich, B. T., P. Pettersson-Lidbom, and M. Priks (2016). Do mild sentences deter crime? evidence using a regression-discontinuity design.

Kuziemko, I. (2012). How should inmates be released from prison? an assessment of parole versus fixed-sentence regimes. *The Quarterly Journal of Economics 128*(1), 371–424.

McConnell, B. and I. Rasul (2017). Hispanic-white sentencing differentials in the federal criminal justice system.

Nordström, T. (1998). Påföljdsval vid rattfylleri: Effekter av 1990 års reform av trafikbrottslagen. *Nordisk Tidsskrift for Kriminalvidenskab 85*(2).

Schwartz, A. E., D. Almond, A. Lee, et al. (2016). Retention heterogeneity in new york city schools. Technical report, Center for Policy Research, Maxwell School, Syracuse University.

Steffensmeier, D., J. Kramer, and C. Streifel (1993). Gender and imprisonment decisions. *Criminology 31*(3), 411–446.

6 Appendix

Table 7: Summary statistics

	est1					
	mean	p50	sd	min	max	count
BAC	1.571561	1.49	.4225663	1	2.996	10520
Female	.1473384	0	.3544598	0	1	10520
Immigrant	.1576046	0	.3643871	0	1	10520
Past No. crimes	2.173004	1	5.862797	0	139	10520
Age	42.31274	44	13.58741	18	68	10520
Have income related to studies	.0544619	0	.2269382	0	1	10007
Income from labor	1584.982	1213	1821.06	0	65784	10007
Have income related to sickness	.1996602	0	.3997649	0	1	10007
No. days unemployed	37.61457	0	82.12341	0	366	10007
No. of kids at home	.3351654	0	.7602886	0	6	10007
Prison	.3480038	0	.4763599	0	1	10520
Days of prison sentence	12.51768	0	18.80141	0	120	10520
Sentenced to fine	.0981939	0	.2975908	0	1	10520
No. fines	93.24782	90	21.60759	30	180	1033
Harsh probation	.2996198	0	.4581132	0	1	10520
Mild probation	.1473384	0	.3544598	0	1	10520
Rehab	.0797529	0	.2709231	0	1	10520
Observations	10520					

Table 8: Available variables for Lasso selection

Variables:
Total number of past crimes
Age
Birth year
Number of children, age 0-3
Number of children, age 4-6
Number of children, age 7-10
Number of children, age 11-15
Number of children, age 16-17
Number of children, age 18-19
Number of children, age 20+
Wage earnings
Earnings business
Declared wage earnings
Acquisition income and work-related compensation
Acquisition income and work-related compensation, incl. deficit of business activity
Acquisition income and work-related compensation, incl. deficit of active business activity
Income related to studies
Incidence of student income
Income due to parental leave
Incidence of income due to parental leave
Parental benefit number of gross days
Parental benefit number of net days
Parental benefit, remuneration
Maternity allowance, remuneration
Occurrence of sickness or occupational injury compensation
Occurrence of rehabilitation allowance
Total income due to illness / occupational injury / rehabilitation
Marking if sickness has been reported since the previous year
Total gross days for remuneration included in reported sickness
Total net days for remuneration included in reported sickness
Sums paid amount of compensation paid in reported sickness
Income due to unemployment
Occurrence of unemployment benefits
Number of days in unemployment
Number of days in part-time unemployment
Total income due to labor market policy program
Number of days in arranged employment
Occurrence of arranged employment
Total income due to early retirement compensation / sickness benefit / activity allowance
Occurrence of early retirement compensation / sickness benefit / activity allowance
Compensation for months with non fixed-term and / or fixed-term sickness compensation
Compensation for months with activity allowance
Income from capital
Total retirement pension
Total occupational pension
Occurrence of occupational/retirement pension
Private pension insurance
Total income from pensions
Social security benefits for the family
Housing benefits for the family
Housing supplementary allowance for the family
Disposable income for the family
Number of days in new start job
Activity support for education, compensation amount
Number of kids
Marital status dummies
Position in family dummies
Dummies for immigration decade
Family type dummies
Type of employment dummies
Dummies for different combinations of unemployment and unemployment benefits
Labor market program participation dummies

Sammanfattning

Ända sedan mina tidiga tonår har jag lätt blivit upprörd av vad jag upplevt som orättvis behandling av någon. Även om åldern kanske har stillat den värsta ilskan jag kunde känna en gång över sådant betéende är det ändå fortfarande något jag starkt ogillar.

Än längre tillbakas, till den tiden då jag bara var ett barn, så hade jag knappast några direkta planer på att ha nationalekonom som yrke. Faktum är att jag nog knappast kunde greppa nationalekonomi som koncept då. Istället ägnade jag en stor del av min tid åt att lära mig om dinosaurier och såg kanske mer ut som att jag var ämnad att bli paleontolog. Men när jag väl gått in i tonåren så ökade mitt intresse för samhällsutvecklingen, politik och ekonomi för varje dag. När jag sedan började läsa nationalekonomi på universitetet så fascinerades jag av teorierna för hur diskriminering kunde uppstå på marknader. När jag sedan på högre nivåer stötte på ekonometri och regressioner upplevde jag det som fantastiskt hur man genom regressioner kunde filtrera ut effekten av andra variabler ur relationen mellan två variabler av intresse. Ytterligare senare kom jag att känna samma sak för hur randomiserade och naturliga experiment kan erbjuda bevis för hur världen fungerar. Kanske är det denna utveckling som lett mig fram till de ämnen och metoder som finns med i denna avhandling.

Denna avhandling består utav fyra av varandra obereoende uppsatser där samtliga på något sätt använder tillämpade mikroekonometriska metoder. Två utav kapitlen fokuserar på anonymitet som policy verktyg och det är ifrån dessa uppsatser som inspiration till avhandlingens namn primärt har hämtats. Bägge fokuserar dessutom på diskriminering i någon mening, antingen genom den direkta effekten på studenters betyg eller folks betéende mot främlingar och feminister när dom tror att dom är anonyma. Diskriminering i allmänhet kan ses som ett rättviseproblem så studerar vi som nationaleknomer det typiskt av effektivitetsskäl. Dom två andra kapitlen i avhanlingen är därför också relaterade till diskriminering, men fokuserar i högre

grad på potentiella effektivitetsvinster. Jag kommer nu kortfattat beskriva huvudslutsatserna från samtliga uppsatser, innan jag gör en kort sammanfattning.

Kapitel 1: Haters gonna hate? – Anonymity, misogyny and hate against foreigners in online discussions on political topics. (skrivet tillsammans med Emma von Essen)

I den här uppsatsen så studerar vi hat i diskussioner om rörande politik under anonymitet på ett Svenskt Internet-forum kallat Flashback. Först så undersöker vi om det är möjligt att förutsäga ifall ett inlägg på forumet innehåller hatiskt innehåll eller inte, baserat på språket i inlägget. Detta följs av en mer traditionell s.k. difference-in-difference modell, där vi undersöker hur hat i diskussionerna påverkas av en upplevd minskning av anonymitet. Vi börjar med att samla in inlägg om inrikespolitik, feminism och integration och invandring från Flashback med hjälp av en egenhändigt byggd web-scraper i Python. Vi drar sedan ett slumpmässigt urvall på 100 så kallade diskussions-trådar från varje underforum och låter sedan en forskningsassistent klassa dessa som antingen hatiska eller inte och mot vem hatet i huvudsak riktar sig. Detta ger oss att slutgiltigt klassat urval av 4021 kodade inlägg, ur vilka vi tar ett slumpmässigt urval om 70 procent som så kallat träningsdata. Med hjälp av denna träningsdata använder vi oss sedan av en maskinlärningsalgoritm för att skapa en prediktionsmodell för hat i allmänhet, hat riktat mot främlingar och misogyni. Vi använder sedan dom återstående 30 procenten som så kallat testdata för att testa hur väl vi kan förutsäga hat, hat mot främlingar och misogyni. Vi här att algoritmen fungerar som bäst när den får klassa hat riktat mot en specifik grupp, såsom invandrare eller feminister. Att förutsäga hat i allmänhet visar sig vara mycket svårare.

Vi använder oss sedan av dessa modeller för att skapa dummy-variabler för ifall ett inlägg är hatiskt, ifall det är hatiskt mot främlingar och/eller innehåller misogynt innehåll i all data som vi hämtat hem från Flashback. Med hjälp av detta stora dataset med samtliga inlägg använder vi oss sedan av en difference-in-difference strategi för att se hur hat i diskussionerna påverkas av faktumet att identiteterna bakom konotona som skulle vara anonyma hade hamnat i händerna på journalister. Mera specifikt så hade journalisterna tillgång identiteten för cirka en tredjedel av alla användare registrerade på Flashback innan mars 2007. Detta blev allmänt känt i september 2014. Därför klassar vi alla som registrerade sig innan mars 2007 som tillhörande

behandlingsgruppen och alla registrerade efter som tillhörande kontrollgruppen. Vi ser sedan att andelen hatiska inlägg och hatiska inlägg mot främlingar minskar för användarna registrerade innan mars 2007 efter september 2014, medan andelen misogyni faktiskt ökar en aning. Dessa resultat verkar delvis drivas av faktumet att användare som skrev en stor andel hatiska inlägg mot främlingar minskade sin aktivitet i efter-perioden och delvis av att dom övergår till att skriva misogynt innehåll istället.

Tidigare forskning om hat på Internet och anonymitet har i huvudsak varit baserat på korrelationer (se till exempel Moore et al. (2012); Suler (2004); Van Royen et al. (2017)) eller bara studerat hat i allmänhet (Cho et al., 2012). Vi bidrar därför till litteraturen genom att använda ett naturligt experiment, studerar individuellt betéende och vem hatet i första hand riktar sig mot.

Kapitel 2: Gender grading bias at Stockholm University: quasi-experimental evidence from an anonymous grading reform. (skrivet tillsammans med Björn Tyrefors)

Denna uppsats är den andra som rör anonymitet, och därmed en av dom två från vilka huvudinspirationen till namnet av avhandlingen tagits. Det är även den första uppsatsen som undersöker köns-bias i rättningen av tentor på universitetsnivå. Med hjälp av data från hela Stockholms universitet mellan åren 2005 och 2014 så använder vi oss först av en difference-in-difference-in-difference strategi genom att uttnyttja faktumet att alla skrivna tentamina var tvugna att vara anonyma från och med höstterminen 2009. Eftersom både uppsatser, muntliga övningar och laborationer inte påverkades av denna reform, kan vi använda dessa som kontrollgrupp och därmed studera könsskillnaden mellan behandlings- och kontrollgruppen innan och efter reformen. Vi finner, i motsats till tidigare forskning på lägre utbildningsnivåer[1], att kvinnor gynnas av införandet av anonyma tentamina jämfört med män. En möjlig hypotes till det omvända sambandet är att universitetet fortfarande är en mansdominerad miljö, medan lägre utbildningsnivåer sedan länge är kvinnodominerade.

Vi fortsätter därför med att använda ett mindre urval bestående av tentamina från grundkursen i makroekonomi vid Stockholms universitet. Med hjälp av dessa data så replikerar vi först resultatet att kvinnor gynnas jämfört med män vid anonyma tentamina, dock med flervalfrågor som kontrollgrupp

[1]Se till exempel Lavy (2008), Hinnerich et al. (2011) och Kiss (2013).

denna gång eftersom dessa omöjligen kan rättas med bias. Fördelen med denna data är dock att vi vet könet på den gruppövningslärare som har rättat en specifik fråga, vilket därigenom tillåter oss att direkt testa hypotesen ifall kvinnor som rättar kvinnor och män som rättar män gör så mer fördelaktigt. Vidare så har dessa gruppövningslärare allokerats till sina frågor genom lottdragning, vilket därför ska säkra randomisering. Alltså har vi i praktiken tillgång till ett randomiserat experiment. Vi finner att randomiseringen verkar ha lyckats och att kvinnor och män gynnar folk av sitt eget kön vid rättning när tentorna inte är anonyma. Så snart anonymitet införts så försvinner dock detta samband. Därutöver så är denna effekt enbart tillräckligt stor för att förklara cirka 20 procent av den totala biasen mot kvinnor som mest. Vi drar därför slutsatsen att rättarens kön spelar roll för vilket håll biasen går åt, men att andra faktorer totalt sett verkar viktigare.

Kapitel 3: Anticipation Effects of a Board Room Gender Quota Law: Evidence from a Credible Threat in Sweden. (Skrivet tillsammans med Björn Tyrefors)

Effekten av kvottering av kvinnor i bolagsstyrelser har fått stor uppmärksamhet senaste tiden både från akademiker[2] och politiker i länder såsom Spanien, Belgien, Frankrike, Tyskland, Island, Italien och Holland (Eckbo et al., 2016). Så här långt har dock samtliga akademiska uppsatser fokuserat på införandet av lagstiftning i Norge. Vi använder oss istället av ett trovärdigt hot om lagstiftning i Sverige föreslaget av vice-statsminister Margareta Winberg i slutet av 2002 med stöd av statsminister Göran Persson. Tillskillnad från i Norge så realiserades dock aldrig lagförslaget om en viss andel kvinnor i listade bolagsstyrelser, i första hand som en konsekvens av att Socialdemokraterna förlorade makten till Alliansen i valet 2006. Trots detta så ser vi en skarp ökning av andelen kvinnor i listade bolagsstyrelser i jämförelse med olistade bolag.

Vi använder därför listade bolag som behandlingsgrupp medan olistade bolag blir kontrollgruppen i en difference-in-difference modell med bolagsprestation som utfall. Antagandet för att vår skattning av effekten ska vara tillförlitlig bygger på antagandet om lika trender vid avsaknad av behandling, vilket verkar trovärdigt baserat på jämförelsen av trender mellan behandlings- och kontrollgruppen i perioden innan 2002, d.v.s. innan hotet formuler-

[2]Se bland annat Ahern and Dittmar (2012), Bertrand et al. (2018), Matsa and Miller (2013) och Eckbo et al. (2016).

ades, såsom föreslås av Angrist and Pischke (2008). Tillskillnad från tidigare forskning på området så finner vi att listade bolag börjar prestera bättre när andelen kvinnor i deras bolagsstyrelser ökar, med en högre vinst som andel av tillgångar, lägre arbetskraftskostnader och ökade intäkter. Dessa resultat är robusta för användandet av syntetisk kontrollgrupp (Abadie et al., 2010) och ifall vi inkluderar separata linjära tidstrender för behandlings- och kontrollgruppen.

Kapitel 4: Differences in prison sentencing between the genders and immigration background in Sweden: discrepancies and possible explanations.

Tidigare forskning gällande diskriminering inom rättsväsendet har typiskt sett fokuserat på antingen hur domar eller jury egenskaper alternativt matchningen mellan domar och jury egenskaper och den åtalades egenskaper påverkar utfallet av en rättegång (Anwar et al., 2012, 2015, 2014; Abrams et al., 2012). I den här uppsatsen anlägger jag istället en annan ansats och fokuserar på hur män och invandrare döms jämfört med kvinnor och inrikesfödda när dom begått samma brott. Mera specifikt så tittar jag på skillnaden i sannolikhet att få fängelsestraff mellan dessa grupper för samma nivå av promille i blodet vid en rattfyllekontroll.

Jag finner att det är 5 respektive 10 procentenheter mer troligt att en man får en fängelsedom än en kvinna och att en invandrare får en fängelsedom än en inrikesfödd när dom begått samma brott. Denna skillnad kan inte heller förklaras av skillnader i vilja att genomgå behandling för missbruk eller underliggande observerbara skillnader mellan grupperna såsom inkomst eller begågna brott. Det finns dock en skillnad i återfallsbenägenhet mellan män och kvinnor som skulle kunna förklara varför män oftare får fängelse. För invandrare finns dock ingen skillnad i återfallsbenägenhet när hänsyn tas till observerbara egenskaper jämfört med inrikesfödda. Därutöver så använder jag en så kallad regression discontinuity metod för att bedöma skillnaden i återfallsbenägenhet som ett resultat av den högre benägenheten att få fängelse mellan invandrare och inrikesfödda. Jag finner att invandrare inte blir mindre benägna att återfalla i brott som ett reslutat av den högre benägenheten att få fängelse, vilket antyder att det finns utrymme inom det svenka rättsväsendet för att förbättra både effektiviteten och rättvisan.

Sammantaget så visar denna avhandling på styrkan anonymitet kan ha som ett policyverktyg. Som nationaleknomer är vi vana att se en avvägning mellan effektivitet och rättvisa. Men som en del av uppsatserna visar i

denna avhandling kan man i vissa fall uppnå bägge målen samtidigt, särskilt i fall där diskriminering förekommer. Detta kan i vissa fall uppnås enbart genom att man döljer någons identitet. Det är dock viktigt att vara försiktig med när anonymitet bör tillämpas, och införanden bör alltid utvärderas med forskningsmetodik, vilket inte minst våra resultat gällande misogynitet och anonymitet på internet visar.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

Abrams, D. S., M. Bertrand, and S. Mullainathan (2012). Do judges vary in their treatment of race? *The Journal of Legal Studies 41*(2), 347–383.

Ahern, K. R. and A. K. Dittmar (2012). The changing of the boards: The impact on firm valuation of mandated female board representation. *The Quarterly Journal of Economics 127*(1), 137–197.

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

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.

Anwar, S., P. Bayer, and R. Hjalmarsson (2014). The role of age in jury selection and trial outcomes. *The Journal of Law and Economics 57*(4), 1001–1030.

Anwar, S., P. Bayer, and R. Hjalmarsson (2015). Politics in the courtroom: Political ideology and jury decision making. Technical report, National Bureau of Economic Research.

Bertrand, M., S. E. Black, S. Jensen, and A. Lleras-Muney (2018). Breaking the glass ceiling? the effect of board quotas on female labor market outcomes in norway. *The Review of Economic Studies*.

Cho, D., S. Kim, and A. Acquisti (2012). Empirical analysis of online anonymity and user behaviors: the impact of real name policy. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pp. 3041–3050. IEEE.

Eckbo, B. E., K. Nygaard, and K. S. Thorburn (2016). Does gender-balancing the board reduce firm value?

Hinnerich, B. T., E. Höglin, and M. Johannesson (2011). Are boys discriminated in swedish high schools? *Economics of Education review 30*(4), 682–690.

Kiss, D. (2013). Are immigrants and girls graded worse? results of a matching approach. *Education Economics 21*(5), 447–463.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? evidence from a natural experiment. *Journal of public Economics 92*(10), 2083–2105.

Matsa, D. A. and A. R. Miller (2013). A female style in corporate leadership? evidence from quotas. *American Economic Journal: Applied Economics 5*(3), 136–69.

Moore, M. J., T. Nakano, A. Enomoto, and T. Suda (2012). Anonymity and roles associated with aggressive posts in an online forum. *Computers in Human Behavior 28*(3), 861–867.

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior 7*(3), 321–326.

Van Royen, K., K. Poels, H. Vandebosch, and P. Adam (2017). "thinking before posting?" reducing cyber harassment on social networking sites through a reflective message. *Computers in Human Behavior 66*, 345–352.