

Words Matter?
Gender Disparities in Speeches,
Evaluation and Competitive
Performance

ISBN: 978 90 361 0670 2

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

©Huyền T.T. Nguyễn, 2021. All rights reserved. Save exceptions stated by the law, no part of this publication may be reproduced, stored in a retrieval system of any nature, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, included a complete or partial transcription, without the prior written permission of the author, application for which should be addressed to the author.

This book is No. **789** of the Tinbergen Institute Research Series, established through cooperation between Rozenberg Publishers and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Words Matter?
Gender Disparities in Speeches, Evaluation
and Competitive Performance

Woorden doen ertoe? Gendersverschillen in toespraken,
evaluatie en competitieve prestaties

Thesis

to obtain the degree of Doctor from the

Erasmus University Rotterdam

by command of the Rector Magnificus

Prof. dr. L. A. Bredenoord

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on December 1, 2021, at 15:30 hours

by

HUYỀN THỊ THANH NGUYỄN

born in Hà Nội, Việt Nam

Doctoral Committee:

Promotor: Prof. dr. O. H. Swank

Other members: Prof. dr. J. B. Hirschberg
Prof. dr. M. P. García-Gómez
Prof. dr. H. D. Webbink

Co-promotor: Dr. J. Delfgaauw

Acknowledgments

It feels surreal to type down these words, reflecting on my entire Ph.D. adventure, at this moment. Back in the days battling the first-year MPhil courses at the Tinbergen Institute, and during the toughest moments in this pandemic year, reaching the Ph.D. defense altar feels like such an impossible height for me. And it definitely would have, without the incredible humans and institutions that embraced me throughout these years.

First and foremost, this dissertation and the researcher I am today would not have come into existence without my advisor, Josse Delfgaauw. Your unwavering trust in my research capability, your sympathetic guidance, and your unflinching support, from the MPhil thesis stage to all these Ph.D. years, nurture my growth as a researcher. The unmatched intellectual freedom you enabled for my academic growth, even when it means venturing into other disciplines than economics, means the world to me. Stories about your *hundreds* of meticulous comments for my paper drafts (by now they collectively are in the thousands...), your never-ending enthusiasm to constructively discuss and critique research ideas, and your day-to-day open-door policy are often met with wide-eyed disbelief, if not downright jealousy. I learn how to become a supportive, inspiring, and effective academic from your dedicated attitude towards the work of colleagues and department visitors, as well as your efficient handling of numerous administrative tasks. You are crucial not only for my academic growth but also for my growth as a person. I am indebted to your sympathetic ear in my times of need, to your pragmatic advice about how to overcome the toughest of struggles, and to your incredible kindness. The fact that I could complete this dissertation against all odds, is (one of) the testaments to how blessed I am to have you as my supervisor.

Secondly, I would like to thank my promotor, Otto Swank, and all members of my defense committee for their valuable time and effort in assessing my dissertation. Julia Hirschberg,

thank you for your detailed comments to significantly improve my dissertation writing, as well as the countless ideas and practical tips on furthering my research agenda. Dinand Webbink and Pilar García-Gómez, thank you for your constructive feedback on the next steps for these chapters. Thomas Buser, I am honored to have you as part of my defense committee. Your expansive research work on gender and labor economics has been a major source of inspiration for me. On top of that, you have always been exceptionally generous in sharing your feedback and ideas on my work in earlier stages. Malia Mason, having my academic female role models, not just one, but two of them on my defense committee, makes this day exceptionally wonderful.

All these years would have been so dull without the supportive, approachable yet critical minds of ESE Economics Department. Otto, thank you for being the best head of department ever. The easy-to-mingle atmosphere in the seminars, the special wine tasting and department get-together events, are certainly the unique trademark of our department under your leadership. Vladimir, as much as I dread your intensive Micro I course, I hugely appreciate the opportunity to be your TA for Game Theory course, and learn from you a structured. Furthermore, your spot-on feedback on my paper presentation in its infancy stage definitely shapes it towards the right path. Dirk, Bas, Dinand and Sacha, thank you for your critical comments and warm encouragement for my work. Anne Gielen, thank you for your essential guidance on the academic paths, especially during pandemic times. Anne Boring, thank you for giving me necessary pointers to improve my paper in its infancy stages and organizing many meaningful network meeting opportunities for female academics at EUR. Anna Baiardi, Yao and Felix, thank you for giving me useful advice on my paper drafts, as well as the meaningful life conversations we had together. Ankimon and the wonderful secretariat ladies, thank you for always navigating the myriads of administrative and contractual arrangements for me. Judith and Caroline, thank you for all your timely help with any administrative and dissertation publication matters, as well as the wonderful moments during my MPhil years.

Next, I would like to thank the ESE Diversity & Inclusion Office, the Economics department and the Erasmus Trusfond for the generous financial support to collect the data and complete part of my research at Columbia University. I thank my research assistants: Fenna ten Haaf and Jan-Gunther Gosselke, for their concerted effort and adeptness in assisting me collecting and making sense of the data. Franka Boender, thank you for your initial

assistance in making sense of the argumentation system. Jan-Gunther, thank you especially for your timely coding expertise, which helps me through many complex data hurdles. Furthermore, I am extremely grateful to the ESE economics department for enabling me to participate in many useful research conferences, summer schools and workshops across Europe. From economics, political science to data science and machine learning, the intriguing discussions and presentations at these venues have definitely opened my minds to the ever-growing excitement and importance of interdisciplinary research.

It would be an unforgivable amiss to forget the highlight of my Ph.D. time: the research visit at Columbia Business School in New York. I am incredibly grateful to Malia Mason, my host, my mentor, and my role model female academic, for making me feel as inspired and homely as possible, during my entire stay. Ashley, Sam, Pete and Patrick, thank you for being so welcoming and inclusive towards me. Julia and Sarah Ita, thank you so much for enthusiastically embracing my debate research agenda; and I feel so blessed about our collaboration thus far. Furthermore, it was my honor to be part of the awe-inspiring I-House NY community. The deep exchanges at the bar, the fun ice cream socials, the Swedish Fit Mondays, and the talks of inspiring figures make me feel instantly at home in New York. Singing *Empire State of Mind* together at the Fall Fiesta after-party will always have a special place in my heart. This three-and-a-half month period at I-House and Columbia University is undoubtedly one of the most eye-opening experiences in my life.

Back to the days on the H-building *promovendus* 8th Floor, I am grateful to the joyful moments with my PhD friends. Nam, thank you for being the most cynical, yet indispensable office-mate, especially for our weekly restaurant exploration in Rotterdam. Yan, Benjamin, Jingni, Esmee and Mathijs, thank you for enliven the office with your optimistic and welcoming characters. All my MPhil friends, specifically, to name a few, Jenny, Dieter, Lingwei, Kostas and Johan, thank you for the numerous moments of fun get-together, of insightful conversations and of being the wonderful company you are, as always.

It is fair to say that, the vast majority of my enriching Ph.D life experiences is credited to the European debate community. Be it sparring with these critical minds over controversial topics, weighing up the merits and what-nots of their speeches, or simply just exasperating over bizarre motions with these "idea junkies" over drinks, debaters have always been the indispensable spices to enlighten the seemingly never-ending dark tunnels of research.

I thank my TU Delft Debate Club friends, Gerson, Bram, Tanya, Jorino and Albert, for embracing me into debating. Milla, Kat, Lucia, Sarah, and Gigi, I am in awe of your incredible wits and grit to navigate the game to the top, and even more inspired by your tireless effort to make this "mental" sport inclusive to everyone. Annabelle, my Corona buddy sister, I feel very lucky to share the ups and downs of this pandemic in Rotterdam with you. Milla and Migle, thank you for all the beautiful moments and soul-touching stories we share. To all debate friends whom I have been fortunate enough to cross paths with, this thesis is as much as what I hope to contribute to economics, as it is to our debate community.

Undeniably, what keeps me going forward against all odds, is the boundless emotional support from my family. Mẹ, Tí, and Sunny, thank you for believing in my wildest of decisions, for better or worse, under any circumstances. I owe my ever-expanding life experiences and wisdom to your wholehearted trust and all-embracing love. Thank you to my extended family - chú Thành, cô Trang, Cindy, Nhật and Charly. Without your above-and-beyond generosity with your time, space, and practical assistance, I would have failed to keep my wits about me, especially in this past tumultuous year. Thank you for taking me in through the darkest hours, for making my pandemic relocation trips a reality, and for your loving care, ever since I came to Europe.

In these rare moments of gratitude, I would like to leave the last words to myself. As common wisdom goes, the hardest lesson one needs to learn is to love oneself, especially in academia. Hence, to my past, current and future self: Thank you for failing, wailing and trying over and over again, for never letting go of yourself even in the darkest hours, and for taking the time and space to care for yourself. Keep on being open to what makes you glow, inspired and fulfilled in life.

A handwritten signature in blue ink, appearing to read 'Huyen', with a long horizontal stroke extending to the right.

*"Không có việc gì khó
Chỉ sợ lòng không bền
Đào núi và lấp biển
Quyết chí ắt làm nên."*

Bác Hồ

CONTENTS

Acknowledgments	i
1 Introduction	1
2 The (Great) Persuasion Divide? Gender Disparities in Debate Speeches and Evaluations	7
2.1 Introduction	7
2.2 Institutional Setup	13
2.2.1 Tournament format & team allocation mechanism	13
2.2.2 Judging in debate tournaments	15
2.3 Descriptive Statistics & Linguistic Variables	15
2.3.1 Linguistic Variable Extraction	16
2.3.2 Linguistic Features of Speeches	18
2.3.3 Linguistic Features of Q&As	18
2.3.4 Speech Evaluation Scores	19
2.3.5 Relationship between score and demographic variables	19
2.4 Methodology & Hypotheses	20
2.4.1 Empirical Strategy	20
2.4.2 Hypotheses	22
2.5 Results	24
2.5.1 Do men and women persuade differently?	24
2.5.2 Are men and women evaluated differently?	28

2.6	Extensions	35
2.6.1	Do men and women receive different questions? Do they answer them differently?	35
2.6.2	Do female judges evaluate speeches differently?	37
2.6.3	Do speakers in preliminary rounds speak differently than those in elimination rounds?	40
2.7	Discussion & Conclusion	41
2.8	Appendix	44
2.8.1	Data Collection & Text Pre-processing Procedure	46
2.8.2	Figures	52
2.8.3	Tables	58
2.8.4	Word List	84
3	Gender Composition of Committees and Performance Evaluation: Evidence from Debate Tournaments	89
3.1	Introduction	89
3.2	Institutional Setup	94
3.2.1	Tournament format and allocation mechanism	94
3.2.2	Adjudication structure and deliberation rules	96
3.3	Data	97
3.3.1	Outcome Variable: Speech Scores	98
3.3.2	Judges & Evaluation Panels	99
3.3.3	Speakers	100
3.4	Descriptive Statistics	101
3.5	Methodology & Hypotheses	102
3.5.1	Empirical Strategies	102
3.5.2	Hypotheses	104
3.6	Results	106
3.6.1	Gender of Chair Judge and Speech Evaluation	106

3.6.2	Gender Composition of Committees and Speech Evaluation	108
3.6.3	Chair ft. Wing Gender Composition and Speech Evaluation	108
3.7	Extensions	111
3.7.1	Do accomplished chair judges evaluate speeches differently?	111
3.7.2	Do female judges evaluate speeches in higher vs. lower-ranked de- bates differently?	112
3.8	Conclusion	114
3.9	Appendix	116
3.9.1	Debater's Background, Institution Quota & Team Selection	116
3.9.2	Data collection: Debate topics, Language skill, Institution & ranking	116
3.9.3	Figures	118
3.9.4	Tables	129

4 Choking upon Facing (Fe)male Opponents? Evidence from Debate Tournaments 145

4.1	Introduction	145
4.2	Institutional Setup	149
4.3	Data & Descriptive Statistics	151
4.3.1	Data	151
4.3.2	Descriptive Statistics	153
4.4	Empirical Strategies	154
4.5	Results	156
4.5.1	Overall	156
4.5.2	Round 1s vs. Round 2s to 9s	159
4.6	Extensions	161
4.6.1	Higher vs. lower-ranked debates	161
4.6.2	Speaker gender ft. partner's gender	162
4.7	Conclusion	163
4.8	Appendix	165

4.8.1	Data collection: Judge panels, Debate topics, Language skill, Institution & ranking	165
4.8.2	Figures	167
4.8.3	Tables	174
	Summary	181
	Nederlandse Samenvatting (Summary in Dutch)	183
	Bibliography	185

1 Introduction

Humans speak to make ourselves understood, explain to get our ideas across, and debate to settle conflicts between one another. From convincing a friend to dine with you at your favorite restaurant to arguing for a case in front of the court, we engage with and persuade others in all aspects of our everyday lives. Mastering the art of persuasion is arguably the key to success, especially in highly competitive jobs with multi-dimensional tasks and complex organizational settings.

In the higher rungs of high-flying careers in business, academia, the law, or politics, one commonly observed fact is that women are persistently under-represented [Goldin et al., 2017; Blau and Kahn, 2017; Eckel et al., 2020]. In fact, recent research on gender disparities in willingness to negotiate salaries [Bohnet and Bowles, 2008; Leibbrandt and List, 2015], promote oneself [Exley and Kessler, 2019], perform a real effort task [Alan et al., 2020] or speak publicly [Buser and Yuan, 2020] has shed light on an important behavioral aspect: persuasion styles in high-stake contexts.

Notwithstanding the importance to understand how differences in persuasion styles across genders matter to gender representation and outcome gap, there exists very limited systematic evidence on gender disparities in speech patterns and evaluations. In most real-world settings, determining whether gender differences in outcomes are driven by differences in behavior or gender-specific evaluation patterns (i.e. discrimination) is inherently difficult due to two reasons. First, large-scale text data sets in a competitive setting where argumentation strength is unconfounded by *ad hominem* strategies or backdoor agreements are extremely scarce. In existing large-scale, textual communication data e.g. political debates [Gentzkow et al., 2019b], central bank communication strategies [Hansen et al., 2018] or judicial court opinion polarization [Ash et al., 2017], in addition to

an inordinate gender imbalance in the actors at hand, a transparent and rigorous evaluation procedure is absent. Consequently, linking specific speech patterns to the evaluation of persuasiveness across gender is infeasible. Second, persuasion, or communication in general, is context-dependent i.e. whether what one says is *more* persuasive, or deemed so, compared to others in the room, depends on who one faces and on who evaluates their speeches. Undeniably speaking, we listen through our own brain filters, biases of our lenses, and varying perceptions of the individuals across the tables. In order to comprehensively understand how and what matters for persuasion across genders, we need a well-defined, high-stake competitive setting with clear-cut and rigorous rules to value exclusively argumentation merits across contestants.

This dissertation contributes novel insights on the role of gender in persuasion tactics, competitive performance and evaluation patterns, in a unique setting of international university debate tournaments. These tournaments and their participants provide an attractive setting to systematically answer these questions. First, these competitions take place annually, at the European or worldwide scale, in a multi-round tournament setting following the widely used British Parliamentary Debate format across a variety of controversial topics. Second, participants are intrinsically motivated students representing various academic institutions worldwide, whose persuasion motives are similar to those of lawyers, politicians, and academics. In fact, many famous politicians, lawyers, and judges trained their persuasion skills in competitive debating, thus making this setup externally relevant to real-life competitive contexts. For each debate round, participants are randomly assigned debate topics, speaking positions (i.e. for or against the topic), opponents, and judge panels. Every participant gives a 7-minute speech to convince a panel of trained judges, who are incentivized to evaluate speeches fairly given past achievements and peer performance feedback. Such incentive architecture mirrors real-life committee decisions, where career concerns, authority play, and social pressures matter. Importantly, comparative argument strength is the yardstick to success in these tournaments. In other words, *ad hominem* argumentation strategies, which is the common confound with argumentation merits in political debates, are outlawed in debate tournament speeches. In the introduction section of each chapter, the reader will find the uniquely relevant debate tournament advantages to study the respective research questions.

By combining state-of-the-art, persuasion-relevant natural language processing with

econometric techniques, this dissertation is a collection of three empirical essays investigating persuasive communication performance and evaluations *across genders* in a relevant competitive context i.e. high-profile international debate tournaments. Chapter 2: [The \(Great\) Persuasion Divide? Gender Disparities in Debate Speeches and Evaluations](#) draws on recent advances in dictionary-based persuasion methods [Pennebaker et al., 2015] to extract spoken verbal tactics across genders in 1517 speech transcripts of the highest-profile inter-varsity debate tournaments to understand: (1) whether men and women persuade differently; and (2) how their persuasion patterns matter for competitive evaluations among committees. I find significant variation in speech patterns across genders. Female speakers use a more personal and disclosing speaking style, with more hedging phrases and non-fluencies in their speeches. In their answers to questions from opponents, women negate less, while having longer and more vague answers. On average, women receive lower evaluation scores than men. Across debates, having a less analytical speaking style and more positive sentiment is associated with higher scores for speeches by women, but not by men. Within debates, except for non-fluencies, there is no robust evidence of gender-specific evaluation standards. Noteworthily, within debates, even though evaluation patterns are similar for male and female speakers across judges and the judge panel gender compositions, committees with more female judges are significantly harsher towards female speakers. Overall, these insights suggest that the gender score gap arises because speeches of female speakers contain more score-reducing and fewer score-enhancing features, rather than discrimination.

Since evaluators play a critical role in determining persuasiveness among contestants, Chapter 3: [Gender Composition of Committees and Performance Evaluation: Evidence from Debate Tournaments](#) explores the causal impact of the gender composition of 4896 committees on 39 168 competitive speech performance scores across European and World Universities Debate Championships. Here I find that committees with a female chair judge give lower scores to both male and female speakers, particularly in higher-ranked debates. The gender of other committee members does not affect evaluations. While accomplished male chair judges are more generous in scoring, they are notably less so towards female speakers. These results demonstrate that gender quotas on evaluation committees do not necessarily eliminate the glass ceiling for women in high-stake, repeated competition contexts.

Last but not least, given that the gender of opponents has been hypothesized to impact the competitive performance of real-world contestants, Chapter 4: [Choking upon Facing \(Fe\)male Opponents? Evidence from Debate Tournaments](#) examines whether the gender of debate opponents causally affect the competitive performance of contestants, by exploiting the random assignment of 3153 participants to multiple rounds of debate matches. On average, I find that the performance of neither men nor women is affected by the gender composition of opponents. In higher-ranked debates, female speakers perform comparatively worse in rooms with more female opponents. These findings indicate that more inflow of women into competitions for high-profile careers does not necessarily reduce the thickness of the glass ceiling.

All in all, this dissertation serves to expand our understanding of how and to what extent oral persuasion patterns matter to performance evaluation, in a uniquely relevant competitive context. Findings from these chapters have three important implications. First, assuming that these results carry over to workplace settings, gender differences in outcomes of negotiations and job interviews would be attributable to differences in persuasion tactics, rather than how negotiation is evaluated. Since the lexical features investigated in this high-stake, competitive, male-dominated context correlate with confidence and charisma, the finding that female speakers have more features correlated with lower confidence and performance scores speaks to the exhibited gender gap in self-promotion, leadership tendency, and workplace authority. Second, the null finding of increasing female members or having a female chair in a committee raises doubts about the *direct* effectiveness of gender quota law, in and of its own, on smashing the glass ceiling for women to the top. Given the ever-growing implementation of such a law across the world, it is important to keep in mind other crucial institutional setups and mechanisms in truly creating an equitable competitive environment. Finally, since female speakers perform comparatively worse in debates with more female contestants, it is crucial that policymakers consider alternative setups of competitions into high-profile careers if their goal is to taper the barriers at the top.

"Persuasion is achieved by the speaker's personal character when the speech is so spoken as to make us think him credible. We believe good men more fully and more readily than others: this is true generally whatever the question is, and absolutely true where exact certainty is impossible and opinions are divided."

Aristotle

2 The (Great) Persuasion Divide? Gender Disparities in Debate Speeches and Evaluations

2.1 Introduction

Be it in business [Bertrand and Hallock, 2001; Matsa and Miller, 2011], academia [Weishaar, 2017], the law [Azmat and Ferrer, 2017] or politics [Matz and Bruschke, 2006], female under-representation in influential positions remains an empirical fact [Goldin et al., 2017; Blau and Kahn, 2017]. Recent work on gender disparities in willingness to negotiate salaries [Bohnet and Bowles, 2008; Leibbrandt and List, 2015], promote oneself [Exley and Kessler, 2019], perform a real effort task [Alan et al., 2020] or speak publicly [Buser and Yuan, 2020] has shed light on an important behavioral aspect: persuasion styles¹ in high-stake contexts.

Despite its criticality in addressing the gender representation and outcome gap,² very limited systematic evidence on gender disparities in speech patterns and evaluations exists. The budding descriptive evidence in political speeches points towards gendered persuasion style e.g. UK female politicians address different topics and use more emotional [Dietrich et al., 2019], less complex [Coates, 2015] and varied evidence types to support their arguments [Hargrave and Langengen, 2020]. Nonetheless, in political debates, argumentation merits are inevitably confounded with unobserved personal beliefs, backdoor agreements

¹*Persuasion* means using one's resources (e.g. spoken words, body language, etc) to change people's behaviors or attitudes toward an idea, both systematically and heuristically [Schacter et al., 2011].

²Public speaking preference has been shown to predict student's career expectations [Buser and Yuan, 2020]; bargaining skill plays an important role in the gender wage gap [Flinn et al., 2019], whereas negotiation training has been shown to have long-run positive economic outcomes for adolescent girls in low-income countries [Ashraf et al., 2020].

and the frequent usage of *ad hominem* personal attack strategies [Christine Banwart and McKinney, 2005; Wright and Holland, 2014]. On the evaluation side, the gender disparities found across teaching evaluations [Beg et al., 2021; Mengel et al., 2018; Boring, 2017], pitch contests [Brooks et al., 2014], grant proposals [Kolev et al., 2020] or academic articles [Hengel, 2020] are often qualified by potentially noisy evaluation metrics across topics, evaluators and participants. Therefore, determining whether gender differences in outcomes are driven by differences in behavior or gender-specific evaluation patterns (i.e. discrimination) in real-world settings is inherently difficult. Exploiting the well-defined speech evaluation rules in high-stake, male-dominated debate tournaments among future elites, I fill this gap with text analysis on the novel data set of 1517 debate speech transcripts and evaluation scores, to answer:

(1) Do men and women persuade differently?

(2) Given persuasion styles,³ does gender matter in speech evaluations?

Debate tournaments and their participants provide an attractive setting to systematically answer these questions. First, intrinsically motivated students who represent a diverse range of academic institutions, compete in a multi-round tournament setting across a variety of controversial topics.⁴ This diverse and ambitious subject pool with similar persuasion incentives to lawyers, politicians and academics makes this setup externally relevant to real-life competitive contexts. In fact, many famous politicians, lawyers and judges trained their persuasion skills in competitive debating.⁵ Second, given *exogenously assigned* debate topics, speaking positions (i.e. for or against the topic), opponents and judge panels, every participant gives a 7-minute speech to convince a panel of trained judges. This randomization feature rules out confounds on persuasion merits such as personal beliefs and backdoor agreements. Noteworthy, judges⁶ are incentivized to evaluate speeches fairly with a *tournament-for-judges* system based on peer performance

³i.e. speech length, complexity of language usage, composition of function words, sentiment polarity, analytical/authentic/positive tone, proportion of non-fluencies and hedges.

⁴i.e. 5 rounds Hobart William Smith (HWS) Round Robin Invitational Champion League, and 9 rounds in the World and European Universities Debate Championship.

⁵Examples include the current UK Prime Minister Boris Johnson, US House Speaker Nancy Pelosi and Senator Elizabeth Warren, economist John-Maynard Keynes, former Pakistani prime minister Benazi Bhutto, US Circuit Judge Stephanos Bibas, 29th Australian Prime Minister Malcom Turnbull.

⁶Details on judge background and selection criteria is in Section 3.2.2.

feedback.⁷ Such incentive architecture mirrors real-life committee decisions, where career concerns, authority play and social pressures matter. Finally, the *primary* production function for speech success, defined in the transparent judge guide book and score scale, is comparative argument strength. In other words, *ad hominem* personal attack strategy, which is commonly confounded with argumentation merits in political debates, is outlawed in debate tournament speeches.

The data set consists of 1517 speech transcripts and evaluation scores across 189 debates from 22 World and European Universities Debate Championship and 9 HWS Debate Champion League tournaments, collected from YouTube video recordings from 2008 to 2018.⁸ To extract relevant linguistic variables from these transcripts, I take a dictionary-based approach⁹ drawing from three validated sources: (i) the speech elements at the word and phrase level in persuasiveness studies by [Petukhova et al., 2017b] and [Dubiel et al., 2020]; (ii) Linguistic Inquiry and Word Count (LIWC) [Pennebaker et al., 2015]; and (iii) the politeness studies of [Danescu-Niculescu-Mizil et al., 2013] and [Yeomans et al., 2018]. Dictionary-based methods drawn from these studies are suitable to study gender disparities in debate speeches because of two reasons. First, since the dictionary categories of persuasion elements and styles are trained and validated on debate speeches [Petukhova et al., 2017b; Dubiel et al., 2020] and policy discussions [Yeomans et al., 2018], they provide the most applicable prior specifications for this competitive debate speech data set. Second, given that the 1517 speech transcripts of 748 debaters across 128 topics is sufficiently small and sparse, using a prior dictionary of speech features provides a reliable method to extract interpretable linguistic variables [Gentzkow et al., 2019a]. In this paper, to understand descriptive differences between speeches given by men and women, I conduct logistic regression of gender as the dependent variable against these linguistic variables. To check whether men and women are evaluated differently given speech content, I run linear and fixed effects regression of score against these variables, interacting with the gender dummy while controlling for debate and individual characteristics.

⁷Round-by-round assessments by judges who adjudicated with them and debaters who were judged by them is instrumental to the room allocation and promotion/demotion in subsequent rounds.

⁸See Table 2.8 for a detailed overview of speeches by year and competition.

⁹For any given text file, this method calculates the number of words that match each of the categories in the dictionary, and then expresses these frequencies as a percentage of the total number of words in the text.

I find notable differences in the language usage of men and women. On average, speeches given by men are more analytical¹⁰ and less authentic.¹¹ Women use slightly more hedges,¹² fillers¹³ and personal pronouns, but fewer nouns and adjectives than men in their speeches. In the 1828 pairs of strategic questioning & answering (Q&A) during these speeches, I find no difference in the numbers of requested questions towards both sexes. Yet, there are more hedges in the accepted questions posed to female debaters. Answers of female speakers are significantly longer, with significantly less negation and more hedges.

In terms of evaluation, speeches given by women receive 0.16 standard deviation (SD) lower scores compared to those given by men. Across debates, controlling for relevant factors,¹⁴ the use of analytical style correlates negatively with score only for women; whereas the use of personal pronouns and positive emotional tone correlates more positively for women than for men. These correlations are sizable: compared to men, a one standard deviation increase in analytical style relates to a 0.13 SD larger reduction in score for women. On the other hand, a one standard deviation increase in the use of personal pronouns and positive emotional tone relates to 0.09 and 0.13 SD larger increase in score for women, respectively. Within debates, except for fillers, most effects become insignificant. This is potentially due to the large restriction on degrees of freedom¹⁵ of debate fixed effects, coupled with power-matching team allocation mechanism,¹⁶ which leads to insufficient variation across genders in speeches within a debate. Noteworthy, further analysis given gender of the judges shows that, even at the debate room level, controlling for speech patterns, committees headed by female judges give female speakers significantly lower scores than they give male speakers. Altogether, linguistic-wise, except for fillers, I found no robust evidence for discrimination. The overall performance gap is attributable to more score-reducing speech patterns in speeches of women. Obviously, these results cannot exclude the possibility of differing "optimal" persuasion styles between men and women, i.e. conditional on argumentation quality, the current evaluation standards

¹⁰i.e. uses more distant, guarded discourse style.

¹¹i.e. uses more personal, disclosing styles.

¹²e.g. "in a way", "technically", "somewhat", "a little bit". A detailed list is in Appendix 2.8.4.2.

¹³"huh", "uh", "erm", "um", "well", "so", "like", "hmm". A detailed list is in Appendix 2.8.4.3.

¹⁴i.e. native English speaker status, speaking position, competition & year, debate topics, room and judge gender composition.

¹⁵123 clusters of six to eight observations per debate, for total observations of 984 speeches.

¹⁶i.e. Teams of comparable ability are matched to debate against each other. See Section 3.2.1 for tournament setup explanation.

might have evolved to reward more the speech features that men are, on average, better at.

These findings have three important implications. First, assuming that these results carry over to workplace settings, gender differences in outcomes of negotiations and job interviews would be attributable to differences in persuasion tactics, rather than how a negotiation is evaluated. As the lexical features investigated in this high-stake, competitive, male-dominated context correlate with confidence and charisma,¹⁷ the finding that female speakers have more features correlated with lower confidence and performance scores¹⁸ speaks to the exhibited gender gap in self-promotion [Exley and Kessler, 2019], leadership tendency [Born et al., 2020] and workplace authority [Wright et al., 1995]. The result on how analytical speaking style is not necessarily working for women also speaks against "lean in" advice [Exley et al., 2020], thus highlighting the double-bind dilemma in persuasive communication styles for women [Jamieson et al., 1995]. Second, the robust finding of harsher evaluation results for female speakers from female judges¹⁹ calls for further investigation into the "black box" of discussion dynamics in deliberation decisions in committees. These findings speak to the ambiguous or low effect of gender quota in committee decisions [Bagues and Esteve-Volart, 2010; Bagues et al., 2017; Bagues, 2017]. Coupled with the evidence on gendered behavior and stereotypes in groups [Coffman, 2014; Sarsons, 2017b; Coffman et al., 2019], these results caution against blindly imposing gender quotas on evaluation committees. As successful women at the top might very well have internalized the male-skewed assessment system during their career paths, such policies do not necessarily guarantee gender-neutral evaluation standards. Finally, given the parliamentary debate format and the diverse pool of policy-oriented participants and debate topics, these findings provide the first step towards understanding how communication styles matter in deliberative democracy practices in political debates, public interviews and briefings [Chambers, 2004; Karpowitz et al., 2012].

Contribution-wise, this paper is the first study to systematically relate *spoken linguistic* features to persuasive speech evaluations in naturally occurring high-stake, male-dominated,

¹⁷e.g: fillers [Dinkar et al., 2020], hedges [Mihatsch, 2012; Holmes, 1990], speaking tone and language complexity [Yang et al., 2020; Hirschberg and Rosenberg, 2005]

¹⁸fillers, hedges

¹⁹To conclusively pinpoint the driving mechanism of such evaluation patterns against female speakers, deliberation discussions among judges and further information on demographic background are necessary. We also need significantly larger data points for power detection, both of which we leave for future work.

competitive contexts. Thus far, the text-as-data literature on both the gender differences in behavior and evaluation primarily focuses on written or cleanly curated texts. In research grant proposals, [Kolev et al., 2020] show that female applicants use narrower words than male applicants to describe their contribution potential. Comparatively, this paper takes into account also speech length, word complexity, lexical components and speech styles, thus giving a more comprehensive picture of persuasive speeches. On evaluations of academic articles, [Hengel, 2020] found that female economists are held to higher peer review standards using readability scores on paper abstracts. In comparison, I analyze spoken linguistic features in the whole speeches, and find no robust evidence of speech style discrimination. The under-performance of women is mostly due to women using more score-reducing tactics in their speeches. On questioning behavior, [Bohren et al., 2018] documented more opinion words and wider range of sentiments in questions posted by female accounts in the online Math forum. In contrast, I found no sentiment differences in questions posed to women in the high-stake, face-to-face competitive context. Instead, such questions contain more hedging language, and are responded to with less negation and more hedges by women. Furthermore, this work contributes a comprehensive dictionary-based automatic extraction of persuasive speech features [Pennebaker et al., 2015; Yeomans et al., 2018] to the literature linking gender differences in communication patterns and attitudes to outcomes, i.e. the work of [Ash et al., 2021] on how gendered attitudes impact judicial decisions.

In hiring and promotion literature, this paper speaks to the observed gender biases in evaluations associated with speech behavior. In student teaching evaluations (SETs), studies by [Mengel et al., 2018] and [Boring, 2017] on large-scale university evaluations corroborated the seminal finding of [MacNell et al., 2015], whereby students disproportionately vote in favor of male professors, conditional on quality of instructors. In business, female-founded startups are less likely to raise capital in male-dominated sectors [Hebert, 2018]; their pitches are also less preferred by investors [Brooks et al., 2014] and whenever female entrepreneurs present more masculine behaviors, their speeches are more likely to be positively evaluated [Balachandra et al., 2013]. On committee evaluation behavior, recent lab experiments on opinion aggregation in group deliberations, [Mengel, 2020] found strong and significant gender biases under open communication among committee members. By quantifying how people use and evaluate linguistic tactics in natural competitive settings with rigorously tested computational linguistic techniques

[Pennebaker et al., 2015; Yeomans et al., 2018], this paper enhances our understanding of behavioral differences across genders beyond entry decisions and outcomes.

Finally, the investigation of persuasion strategies relates to the linguistic and social psychological discourse on gender differences in linguistic styles (powerless vs. powerful language). Powerless language style,²⁰ which strongly correlates with lower perceived likeability and competence [Bradac and Mulac, 1984], has been found more in speeches of women [Lakoff, 1973]. However, findings regarding gender differences were mixed,²¹ where, essentially, variation in women’s linguistic styles are more associated with *social powerlessness* than to gender per se, which exhibits itself most in male-dominated environment. In politics, [Hargrave and Langengen, 2020] recently found from 200 UK House of Commons debates that women use less adversarial debate style and cite evidence more from personal experience. Nonetheless, these studies have limited explanatory power, due to either limited sample sizes or the lack of integrated and systematic assessment mechanisms without confounding motives (e.g. career concerns, asymmetric preferences for various leadership styles, varying institutional setups). Comparatively, this larger-scale debate data set in democratically relevant, competitive format enables us to investigate systematic gender differences in linguistic tactics.

This chapter proceeds as follows. Section 2.2 highlights the institutional setup of British Parliamentary debates, tournament format and judging criteria. Section 2.3 provides the descriptive statistics and the linguistic variables. Section 2.4 explains the empirical strategies and hypotheses. Sections 2.5 and 2.6 summarize the main results and extension analyses respectively. Section 2.7 concludes with discussions on avenues for future research.

2.2 Institutional Setup

2.2.1 Tournament format & team allocation mechanism

Tournament format. Every year, around 200+/- teams across Europe attend the European Universities Debate Championship (EUDC); 450+/- teams across the world participate in

²⁰i.e. using hedges e.g: *You know, sort of, I think*, tag questions, modal verbs, etc

²¹For instance, [Crosby and Nyquist, 1977] on the use of tag questions [Dubois and Crouch, 1975], on conversational directness [Holtgraves, 1997] [Rundquist, 1992].

the World Universities Debate Championship; whereas 16 teams who have won prestigious continent-wide tournaments across the globe are invited to join the HWS Round Robin Champion League. Fixed teams of two speakers go through nine preliminary rounds (i.e. *in-rounds*) in EUDC and WUDC, while only five rounds for HWS Round Robin tournaments. All debates are conducted in British Parliamentary (BP) Debate style, with four teams per debate, opening half and lower half, speaking for or against the motion.²² For each round, teams are exogenously assigned team speaking positions, opponents and judges. All speakers have 15 minutes to prepare with the debate partner for the given motion and speaking position. Each speaker gives a 7-min speech in a debate, whereby during their speeches, they can accept or reject questions from opponents. After each round, individuals receive two scores: (1) team score²³ and (2) individual speaker scores,²⁴ where the sum of individual scores of higher ranked teams must be higher than that of next-ranked team. The accumulated team points and speaker points²⁵ determine the top performing teams proceeding to elimination rounds (i.e. *out-rounds*). The top 10 – 15% ranked teams advance to the out-rounds in each language category. In these out-rounds, only team ranking decisions are made, i.e. teams that are ranked 1st and 2nd advanced into further rounds, whereas those on 3rd and 4th place are eliminated. In the final debate, the best team is crowned the champion. The best speaker is an individual with highest cumulative individual speech scores across all preliminary rounds.

Team allocation mechanism. In Round 1, debate teams are randomly matched; whereas in the remaining in-rounds, teams are power-paired. Specifically, teams debate against those who have gained the same (aggregate) number of team ranking points from previous rounds.²⁶ Due to this power-pairing mechanism, the universal 50-to-100 individual speech score scale ensures consistent evaluation of speech quality across all debate rooms i.e. winning from a lower-ranked room does not necessarily mean higher individual speaker scores than, for instance, taking a 2nd or 3rd in a higher ranked room. Moreover, the further the rounds proceed, the more "sorted" teams are in terms of speech performances, i.e. consistently higher-performing teams debate against one another in a room, and vice versa.

²²More details of BP debate style and format is in Appendix 2.8.0.1.

²³Team ranking 1st receives 3 points, 2nd receives 2 points, 3rd receives 1 points and 4th receives no point.

²⁴50-to-100 score scale, with 50 as the lowest. See Appendix 2.8.2.1 for a speaker score scale example of European Universities Debate Championship 2017.

²⁵Speaker points are used for : (i) award best performing speakers in the form of top 10 speaker awards; and (ii) determine the advancing team into elimination rounds in case of ties.

²⁶More information about power-pairing mechanism is given in [Monash Debate Review](#)

2.2.2 Judging in debate tournaments

Adjudication panel structure. A debate room is adjudicated by a *chair (C)* judge (one chair/room) and several *wing (W)* judges. Typically two to four wing judges are allocated in preliminary rounds, whereas four to eight are allocated in elimination rounds. All judges are responsible for keeping track of the key arguments and determining the team ranking, speaker scores and justifications thereof. The chair (C) judge has the *ultimate power* and responsibility to assign the definitive ranking²⁷ and speaker scores, as well as delivering verbal ranking explanations to debaters after panel discussion.

A judge's role. Judges need to act as an *informed global citizen*, who evaluates the argumentative cases *holistically*, given their relevance and plausibility. Judging is done comparatively, i.e. decide which team, when weighed against another team, gave the most persuasive case for their side, given one's *impartial* reading of the entire debate.²⁸ The standard is only on general knowledge, found in the front pages of major articles in the national or international newspapers.²⁹ Qualified judges must accurately weigh what was *actually said* by teams in the debate, without inserting their preconceptions or expert knowledge into their decisions. Importantly, judges arguably have strong incentives to exert their best adjudication effort, as making it to judge in elimination rounds means being recognized as the best judges of Europe or the world by the community.³⁰

2.3 Descriptive Statistics & Linguistic Variables

This section provides the key linguistic variables used in the analysis and descriptive statistics of the data obtained in Figure 4.1. Detailed data collection procedure on (i) speech transcripts; (ii) questions and answers (Q&A) during speeches; and (iii) metadata information about speech evaluations and individuals can be found in Appendix 2.8.1.

²⁷In case of ranking conflict, the vote of the chair judge will be the tie-breaker vote.

²⁸For a detailed description of judge's role, please refer to page 4 - 10 of [Novi Sad EUDC 2018 Judge Briefing](#).

²⁹For instance, discussing the reparations for WWII, the Iraq conflict or AI ethics would be a fair game, not on the technical or esoteric knowledge about these issues.

³⁰For detailed information on the judge allocation mechanism, see Appendix 2.8.0.2.

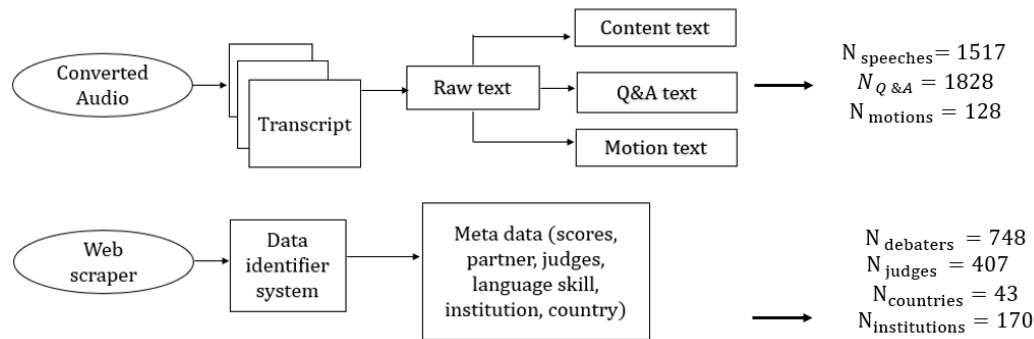


Figure 2.1: Overview: Data collection and construction procedure

2.3.1 Linguistic Variable Extraction

Following the standardized procedure on text stemming and tokenization in text analysis of [Gentzkow et al., 2019a,b; Ash et al., 2017], I deleted hyphens and apostrophes, replaced all other punctuation with spaces; and removed non-spoken parenthetical insertions in the cleaned transcripts. Next, I extracted a subset of LIWC linguistic and psychometric variables suited to study persuasiveness [Petukhova et al., 2017b] and conversational receptiveness [Yeomans et al., 2018, 2020], serving as independent variables in regression analysis. LIWC analyzes a text file on a word-by-word basis, comparing each word in the file to almost 5000 words and word stems in an internal dictionary [Pennebaker et al., 2015]. The words in this dictionary have been rated by judges as representing a variety of different linguistic (i.e. word count, pronouns, articles, etc) or psychological categories (i.e. emotional, cognitive, and sensory processes). Empirical studies using LIWC in computational linguistics, social psychology and communication research have demonstrated its ability to detect meaning in a wide variety of settings [Tausczik and Pennebaker, 2010]. For debate speech transcripts, the chosen variables can be categorized into six groups:

1. Basic speech features:³¹ word count, sentence count, words per sentence, character per word, words with more than 6 letters.
2. Parts of Speech (POS) composition:³² proportion of verbs, nouns, adjectives, adverbs and personal pronouns.

³¹These variables measures speaking rate in persuasive contexts, whereby faster speaking rate have been noted to correlate with higher credibility in linguistics [Miller et al., 1976; Smith and Shaffer, 1995]

³²Parts of speech tags have been found useful as cues for credibility in court trials [Pérez-Rosas et al., 2015] and interview dialogues [Levitan et al., 2018].

3. Basic linguistic group:³³ Argument indicators, fillers, hedges.
4. BP procedural words and phrases and POI rejects: Procedural BP debate style words and phrases³⁴ and percentage point of rejecting questions during one's speech³⁵
5. Sentiment analysis:³⁶ i.e. Net sentiment³⁷ = Positive sentiment - Negative sentiment. Valence-based sentiment analysis is based on lexicons of sentiment-related words, where each word is rated as: (1) *whether* it is positive or negative; and (2) *how* positive or negative. The intensity scale ranges from -1 (strongly negative) to + 1 (strongly positive). To avoid multi-collinearity issues in regression analysis and capture both the proportion of positive vs. negative emotion in the speech, I use net sentiment, which is effectively a linear combination between positive and negative emotion words.
6. LIWC variables: Lists of Certainty and Uncertainty³⁸ words and scaled score of *Analytic* [Pennebaker et al., 2014], *Authentic* [Newman et al., 2003], and (*Emotional*) *Tone* [Cohn et al., 2004] from LIWC 2015 dictionary of [Pennebaker et al., 2015].

These psychometric variables include words and phrases that have been consolidated from previous lab findings and converted to percentiles using Categorical Dynamic Index (CDI) Score. CDI Score is a bipolar continuum derived from factor loadings of its component word categories, based on large comparison samples. That makes it highly applicable to a wide variety of text samples, from studying application essays of students [Pennebaker et al., 2014], speeches of politicians [Abe, 2011] to emotional experience [Sun et al., 2020]. Appendix 2.8.1.1 gives detailed information on the linear combination of these variables and what the respective scales refer to.

³³Argument words indicate, at the beginning of the sentence or clause, whether a speaker gives a claim/premise. Fillers [Dinkar et al., 2020] and hedges [Holmes, 1990] have been robustly shown to be strong correlates of confidence and credibility. A detailed list of these words are available in Appendix 2.8.4. Note that word categories used from the LIWC and Politeness study can be found in their standard dictionary package.

³⁴See Appendix 2.8.4.

³⁵i.e. "No/No thank you/Sit down."

³⁶Emotional load has been shown to enhance or decrease message effectiveness [Gatti et al., 2014] in various contexts, e.g: political speeches [Charteris-Black, 2011, 2018], judgement [Marcus et al., 2000], or public service announcement [Dillard and Peck, 2000]

³⁷The commonly used sentiment analysis package VADER [Hutto and Gilbert, 2014] yields similar result as psychometric variable Tone in LIWC dictionary [Pennebaker et al., 2015]

³⁸i.e. "Tentative" word list in LIWC 2015 dictionary.

For the Q&As during speeches, due to their short and dialogue nature, the chosen variables are: (1) Basic Features: Word Count, Fillers, Hedges and Negation; (2) Parts of Speech: Impersonal pronouns, first person pronouns, second- and third-person pronouns; and (3) LIWC: Certainty and Uncertainty Indicators, Analytic, Authentic and Emotional Tone. Section 2.3.3 summarizes key descriptive statistics for Q& A between male and female speakers.

2.3.2 Linguistic Features of Speeches

Table 2.19 describes the independent variables by group and provides summary statistics, whereas Table 2.20 gives two-sample t-test of unequal variances results of these variables between speeches of men vs. women. Noteworthy, while speeches given by both sexes have roughly the same number of words, men use shorter sentences and more complex words, as well as having notably fewer fillers and hedges in their speeches. With respect to LIWC variables, men give more analytical and less authentic speeches, with slightly more negative sentiment, compared to women. On debate-strategic strategies, there does not seem to be a difference in terms of numbers of questions offered to them (i.e. POI rejects), though men use slightly more debate procedural words & phrases in their speeches.

2.3.3 Linguistic Features of Q&As

For this data set of 1828 pairs of Q&As, given their short dialogue nature, I extracted six linguistic features: (i) word count; (ii) hedges; (iii) fillers and pauses; (iv) personal (first and second) and impersonal pronouns; (v) negations; and (vi) sentiments. Table 2.21 and 2.22 provide the summary statistics of linguistic and psychometric variables for accepted questions and associated answers during speeches; whereas Table 2.23 and 2.24 give the two-sample t-tests of unequal variance and p-value of these variables. There appears to be no difference across these features in the questions, the answers given by female debaters in their speeches are longer, contain significantly more second and third-person pronouns³⁹ slightly more hedges, and less negation,⁴⁰ compared to those given by male debaters.

³⁹i.e you, he, she, they and their abbreviated versions e.g: you've, they'd, etc.

⁴⁰e.g: can't, shouldn't, wouldn't, etc

2.3.4 Speech Evaluation Scores

The kernel density distribution of score between men and women of Figure 2.5 shows that, although scores are normally distributed scores for both men and women, women receive lower scores than men. Table 2.17 provides the detailed score descriptive statistics for both genders across three competition groups. Comparing the scores of Table 2.17 to those of all tournaments from 2008 to 2018 ($N = 107405$) in Appendix 2.8.3.2 gives us the percentile of these speeches in the overall speech evaluation population. For WUDC speeches, the selected sample is among the upper percentile. This is less pronounced for EUDC speeches. The recorded HWS speeches are representative for this tournament.

Overall, these speeches represent the upper end of the tournament participants. This confirms the prevailing recording practice across tournaments to film higher-level debate rooms in the later preliminary rounds. Noteworthy, the higher scores are largely driven by male debaters, thus complying with the existing research showing persistent under-performance of females in intercollegiate debating [Pierson, 2013]. Appendix 2.8.2.5 provides further box plots on score differentials across teams and competitions.

2.3.5 Relationship between score and demographic variables

Table 2.25 provides the single regression of room control variables against scores, i.e: speaker gender, language status, competition, speaking position, institution rank and motion type. Gender-wise, there is a significant difference of 0.16 standard deviation (SD) in score. Noteworthy, debate rooms with ≥ 4 female speakers is associated with -0.40 SD lower score, and this variable remains significant even after pooled together with other room and judge control variables. Non-native speakers, on average, receive 0.92 SD lower score, compared to native speakers. This coefficient remains significant at 0.46 SD, even after controlling for all demographic variables.

2.4 Methodology & Hypotheses

2.4.1 Empirical Strategy

2.4.1.1 Do men and women persuade differently?

To analyze whether elements of speeches given by women differ from those given by men, I run logistic regression of above-mentioned groups of linguistic and psychometric markers⁴¹ against the speaker's gender, controlling for relevant factors. All continuous variables are standardized. As a robustness check exercise, I also run a linear probability and probit model with these elements, documented in Appendix 2.8.3.8.

To avoid multi-collinearity issues due to overlapping word and phrase categories across linguistic variables, I run identical regression analysis separately for three groups: (i) Basic Features⁴² (ii) Parts of Speech⁴³ and (iii) LIWC psychometric measures.⁴⁴ For Q&A data, I run the same analysis on three groups: (i) Word Count, Fillers, Hedges and Negation; (ii) Impersonal, First-person, Second and third-person Pronouns; and (iii) LIWC psychometric measures. This analysis is summarized below:

$$\log_{ik} \left(\frac{Pr_{Female}}{1 - Pr_{Female}} \right) = \sum_{j=1}^n \beta_j \mathbf{X}_{jik} + \sum_{j=1}^n \gamma_j \mathbf{Y}_{ik} + \varepsilon_{ik}$$

where the dependent variable is $\log_{ik} \left(\frac{Pr_{Female}}{1 - Pr_{Female}} \right)$ is the log odds ratio of the probability that the speech i in debate k is given by a female speaker Pr_{Female} . Throughout all analysis, I cluster standard errors at the debate level. The independent variables are:

- \mathbf{X}_{ijk} , which is the frequency matrix of the total count of words and phrases (standardized) in a category j in speech i of debate k , listed in Table 4 of section IV.1.
- \mathbf{Y}_{ik} are the following control categorical variables:

1. $\mathbb{I}_L = 1$ if a debater is a non-native English speaker

⁴¹Table 4, section IV.1

⁴²Word Count, Words per Sentence, \geq 6-letter words, Argument Indicators, Fillers, Hedges, BP Words and Phrases, POI rejects

⁴³Noun, Verb, Adjective, Adverb, Personal Pronoun

⁴⁴Certainty Words, Uncertain Words, Analytic, Authentic, Tone

2. $\mathbb{I}_R = 1$ if the institution of the debater is ranked in the top 50 worldwide⁴⁵
3. \mathbb{I}_C is the group competition type, with $\mathbb{I}_C = 0$ as WUDC, $\mathbb{I}_C = 1$ as EUDC, $\mathbb{I}_C = 2$ as HWS.
4. Speaking position (1st to 8th in a given debate)
5. Motion topic type (17 groups) in a given debate⁴⁶
6. $\mathbb{I}_D = 1$ if a room has ≥ 4 female speakers
7. $\mathbb{I}_J = 1$ if the chair judge is a female
8. $\mathbb{I}_P = 1$ if the judge panel contains same or higher number of female judges compared to male judges

2.4.1.2 Does gender relate to speech content evaluation scores?

The evaluation score given to each speech based on their argumentation persuasiveness is used as the dependent variable in this regression. Independent variables consist of above-mentioned groups of linguistic and psychometric features, in addition to interaction variable with the gender indicator variable $\mathbb{I}_{Female=1}$ of the speaker, as shown:

$$S_{ik} = \theta_{ik}\mathbb{I}_{Female} + \sum_{j=1}^n \beta_j \mathbf{X}_{jik}\mathbb{I}_{Female} + \sum_{j=1}^n \alpha_j \mathbf{X}_{jik} + \sum_{j=1}^n \gamma_j \mathbf{Y}_{ik} + \eta_k + \varepsilon_{ik}$$

where S_{ik} is the speech score i in debate k , \mathbf{Y}_{ik} are the same control categorical variables mentioned in the previous section, $\sum_{j=1}^n \beta_j$ are the coefficients of interest. To account for all unobserved heterogeneity at the debate room level, debate-room level fixed effect η_k is imposed. Note though, this fixed effect imposes a large restriction on degrees of freedom.⁴⁷ Along with the power-matching team allocation mechanism,⁴⁸ it leaves very limited room for variation within debates. Therefore, the effect size interpretation takes into account findings from model specification with and without this fixed effect.

⁴⁵Appendix 2.8.1.4 provides summary information on how institutions & academic ranking is collected.

⁴⁶See Figure 3.4 in Appendix 2.8.2.9 for the distribution of speeches across topics.

⁴⁷123 clusters of six to eight observations per debate, for total observations of 984 speeches

⁴⁸In Round 1, participants are randomly matched. From Round 2 onward, clusters of four teams who earn cumulatively same team scores are matched to compete against one another.

2.4.2 Hypotheses

In terms of spoken speech patterns, the use of fillers in speeches have often been associated with negative evaluations on various dimensions [Bradac and Mulac, 1984], including impression in the courtrooms [Hosman and Wright, 1987], attitude change and source credibility [McCroskey and Mehrley, 1969] [Miller and Hewgill, 1964]. These earlier studies are confirmed by [Bortfeld et al., 2001] in disfluencies in conversations and [Dinkar et al., 2020]'s work on fillers and confidence. Hedges,⁴⁹ which are words and phrases that convey a sense of uncertainty [Lakoff, 1973], deference or politeness [Haas, 1979; Holmes, 1984, 1986, 1990], are consistently confirmed to be more characteristic in speeches of women. Regarding uncertainty and certainty indicators, women have been shown to use more tentative language, [Leaper and Robnett, 2011] as well as intensifiers in their face-to-face communication [Samar and Alibakhshi, 2007; Hanafiyeh and Afghari, 2014] and problem-solving interactions in [Bradac et al., 1995]. Given that a recent larger-scale study of [Newman et al., 2008] confirms these patterns, the first hypothesis is as follows.

H1.1 (Speech Patterns): *Speeches given by women use less complex words, contain more hedges, fillers and uncertainty indicators compared to those given by men.*

On speech articles and styles, according to [Pennebaker et al., 2015], pronouns track the relationship between a speaker and a listener/audience. Women have been found to use more intensive adverbs [Biber et al., 1998], and overall more personal pronouns [Lenard, 2017]. Even though earlier research showed that women tend to adopt masculine strategies when in predominantly male fields, [Wessel et al., 2015], recent work on political speeches have found that speeches of female politicians are more emotional [Dietrich et al., 2019], less aggressive [Grey et al., 2002], use simpler language [Lin and Osnabrügge, 2018] and more personal styles to support their arguments [Hargrave and Langengen, 2020]. Since these speeches follow a parliamentary debate format, similar patterns are expected.

H1.2 (Speech Patterns): *Speeches given by women have more adverbs and personal pronouns. Their speeches have wider range of sentiments, more personal, disclosing style*

⁴⁹E.g. *I think, You know, probably.* See Appendix 2.8.4.2 for the full list.

compared to those given by men.

On the evaluation side, in countries with strong debate and forensics institutions like the US, female speakers under-represented and under-performed in debate tournaments [Bruschke and Johnson, 1994; Stepp and Gardner, 2001]. This pattern persists to modern days, where the latest analysis on debate performance by [Pierson, 2013] is from 2,225 teams with 35,062 speaker scores across 14 EUDC and one WUDC in 2001 to 2013 and found a gender gap: on average, male speaker scores are 1.2 points⁵⁰ higher than female speaker score for every debate round. Given the similar gender evaluation gap found in other contexts that involve speech evaluation e.g. pitch contests [Brooks et al., 2014], teaching evaluations [Boring, 2017; Mengel et al., 2018], I expect similar gender gap patterns in the 984 speeches of this data set, as captured in the following hypothesis.

H2.1 (Speech Evaluations): *On average, speeches given by female speakers receive lower scores than male speakers.*

The mere presence of a gender gap in speaker scores does not inherently imply discrimination, but could be attributable to any other factor such as speech quality, debate topic, institutional and/or personal reputation. Regarding attitude towards speech behavior across genders, [Bradac et al., 1981; Holmes, 1984; Wright et al., 1995] showed that when women use hedges and tag questions, they are viewed negatively in terms of their persuasiveness and intelligence. Given the recent finding of [Phillimore, 2017] in 53 WUDC speeches of 2018 about the negative correlations of hedges and speech scores of women, I expect harsher punishment for female speeches compared to male speeches.

H2.2 (Speech Evaluations): *Conditional on speech features, speeches given by women are associated with lower scores than speeches by men. Women are less rewarded/ more heavily punished for specific speech features, such as hedges and fillers, than men.*

⁵⁰This is the difference in the 50-to-100 score scale, because [Pierson, 2013] uses raw scores to run regression, instead of standardized scores.

2.5 Results

2.5.1 Do men and women persuade differently?

This section reports the average marginal effect results of logistic regression of linguistic variables against gender from five models, varying in terms of control variables taken into account on all 1517 speeches.

All variables are standardized to consistently interpret effect magnitude with respect to evaluation scores. To avoid multi-collinearity issues due to overlapping word and phrase categories across variable groups, I ran identical regression analyses for each group: (i) Basic Features,⁵¹ (ii) Parts of Speech⁵² and (iii) LIWC psychometric measures.⁵³ Each result table for these groups reports five regression models: (1) only linguistic variables; (2) control for room variables;⁵⁴ (3) control for room variables and chair judge gender; (4) control for room variables and panel gender composition;⁵⁵ and (5) room controls, chair judge gender and panel gender composition. Analyses using linear probability and probit models in Table 2.42 and 2.43 of Appendix 2.8.3.8 yield similar results.

2.5.1.1 Basic features

Table 2.1 summarizes the average marginal effects from logistic regression of basic features against the probability that the speech is given by a woman. Unconditionally, Column (1) shows that, speeches with slightly fewer words, longer sentences, less complex words and contain more fillers and hedges are more likely to be given by female debaters. In Column (2) and (3) where room control variables and chair judge gender are accounted for, the difference in word count between men and women becomes insignificant. However, one SD increase in sentence length and the amount of hedges increases the probability that the speech is given by a woman by 4.3 p.p. and 3.8 p.p., respectively. Once all variables are controlled for in Column (5), one SD increase in sentence length and proportion of hedges

⁵¹Word Count, Words per Sentence, \geq 6-letter words, Argument Indicators, Fillers, Hedges, BP Words and Phrases, POI rejects

⁵²Noun, Verb, Adjective, Adverb, Personal Pronoun

⁵³Certainty Words, Uncertain Words, Analytic, Authentic, Tone

⁵⁴i.e. (i) language status, (ii) speaking position, (iii) competition, (iv) whether the room has \geq 4 female speakers, (v) institution ranking, and (vi) motion type.

⁵⁵i.e. whether the judge panel is female-dominated.

in speeches increase such probability by 4.7 and 3.2 p.p, respectively. Conversely, one SD of BP procedural words & phrases decreases the probability that the speech is given by a woman by 3 p.p.

Table 2.1: Average Marginal Effects from Logistic regression of basic features

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Word Count	-0.025** (0.01)	-0.018 (0.01)	-0.018 (0.01)	-0.016 (0.01)	-0.015 (0.01)
Words per Sentence	0.051*** (0.01)	0.043*** (0.01)	0.043*** (0.01)	0.047*** (0.01)	0.047*** (0.01)
Complex Words	-0.028** (0.01)	-0.015 (0.01)	-0.015 (0.01)	-0.018 (0.01)	-0.017 (0.01)
Argument Words	-0.008 (0.01)	-0.002 (0.01)	-0.002 (0.01)	-0.001 (0.01)	-0.001 (0.01)
Fillers	0.024** (0.01)	0.012 (0.01)	0.012 (0.01)	0.011 (0.01)	0.011 (0.01)
Hedges	0.053*** (0.01)	0.038*** (0.01)	0.038*** (0.01)	0.032*** (0.01)	0.032** (0.01)
BP Words and Phrases	-0.019 (0.01)	-0.019 (0.01)	-0.020 (0.01)	-0.030** (0.01)	-0.030** (0.01)
POI Rejects	0.005 (0.01)	0.012 (0.01)	0.012 (0.01)	0.009 (0.01)	0.009 (0.01)
Observations	1517	1517	1517	1314	1314
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Model (4) and (5) excludes 203 knock-out round speeches without judge panel information. Robust clustered standard errors at debate level are in parentheses.

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

Fillers and hedges in speeches. To investigate where speakers use fillers and hedges in their speeches, I aggregate all words in the *fillers* and *hedge* categories into 40-word moving average from start to finish of their speeches of each gender, as shown in Figure 2.2 below.⁵⁶ There exists very little difference in fillers, and virtually no difference in the number of times men and women were asked questions during their speeches and the amount of BP-specific words and phrases. A closer look at the occurrence of hedges and fillers in this Figure 2.2, show that women have a higher proportion of fillers in the beginning and before the end of their speeches; whereas men have persistently lower and less varied patterns. This result partially confirms hypothesis **H1.1 (Speech Patterns)**: *Speeches given by women contain more hedges, BUT not more fillers, than those given by men.*

⁵⁶i.e. A point with (200,0.0035) for female coordinate means that, in the immediate 161th to 200th word, 0.0035 is the weighted average proportion of fillers against total word count/speech across all female speeches.

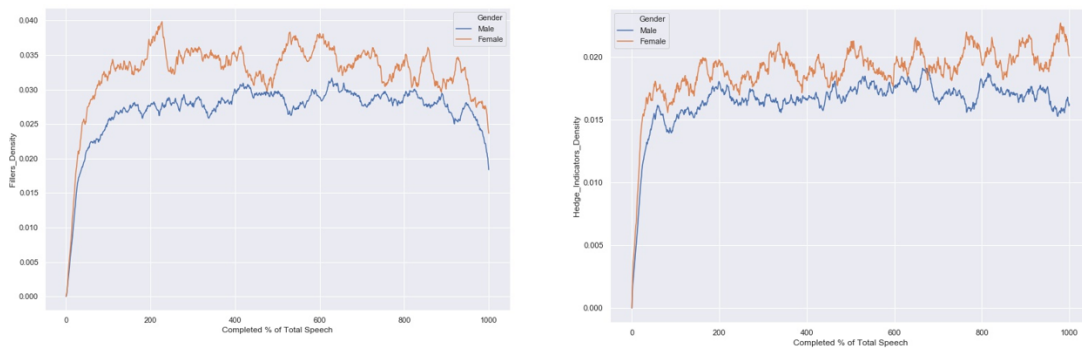


Figure 2.2: 40-word moving average time series of fillers and hedges in male vs. female speeches

2.5.1.2 Parts of Speech

Table 2.2 shows that speeches given by women have lower proportions of nouns and adjectives, both by 5.5 p.p, with and without control variables. Noteworthy, a one SD increase in personal pronouns increases the likelihood that the speech is given by a woman by 2.7 p.p. Such statistically significant differences are robust across all controls. These findings partially confirm hypothesis **H1.2 (Speech Patterns)**: *Speeches given by women do NOT have more adverbs, but they have more personal pronouns compared to those given by men.* The finding on women using more personal pronouns than men aligns with early small-scale studies into gender differences in public speaking [Mulac and Lundell, 1986; Mulac et al., 1986], computer conferences [Fahy, 2002], televised interviews [Brown, 1957] and sports [Kuo, 2003]. Nonetheless, the finding is at odds with larger-scale work on political speeches [Yu, 2014; Lenard, 2017], i.e. female politicians use fewer personal pronouns due to the formal political setup. Since debate tournaments blend features of formal parliamentary debate settings with extemporaneous speech requirements for speakers, this finding is more in line with the dialogue-based evidence, rather than the formal congressional speech forums.

Table 2.2: Average Marginal Effects from Logistic regression of parts of speech (POS)

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Noun	-0.070*** (0.01)	-0.046*** (0.01)	-0.046*** (0.01)	-0.055*** (0.01)	-0.055*** (0.01)
Verb	-0.021 (0.01)	-0.008 (0.01)	-0.008 (0.01)	-0.008 (0.01)	-0.008 (0.01)
Adjectives	-0.057*** (0.02)	-0.054*** (0.02)	-0.054*** (0.02)	-0.055*** (0.02)	-0.055*** (0.02)
Adverbs	0.019 (0.01)	0.013 (0.01)	0.013 (0.01)	0.010 (0.01)	0.010 (0.01)
Personal Pronouns	0.025** (0.01)	0.026** (0.01)	0.026** (0.01)	0.027** (0.01)	0.027** (0.01)
Observations	1517	1517	1517	1314	1314
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Model (4) and (5) excludes 203 knock-out round speeches without judge panel information. Robust clustered standard errors at debate level are in parentheses.

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

2.5.1.3 LIWC psychometric features

Table 2.3 reports the results regarding LIWC psychometric features. All columns highlight that, compared to speeches given by men, those given by women are more authentic (i.e. personal, disclosing discourse style) yet less analytical (i.e. more structured, hierarchical thinking) and contain slightly fewer uncertain words. There is no difference between the emotional tone (i.e. net sentiment) of male and female speeches. This is possibly due to the random allocation of pro- or against the debate topic. These findings conclude hypothesis **H1.2 (Speech Patterns):** *Speeches give by women do NOT have wider range of sentiment, yet they have more personal, disclosing styles compared to speeches given by men.*

Table 2.3: Average Marginal Effects from Logistic regression of LIWC psychometric features

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Authentic	0.036*** (0.01)	0.031*** (0.01)	0.031*** (0.01)	0.040*** (0.01)	0.040*** (0.01)
Analytic	-0.078*** (0.01)	-0.074*** (0.01)	-0.074*** (0.01)	-0.079*** (0.01)	-0.079*** (0.01)
Emotional Tone	0.005 (0.01)	0.007 (0.01)	0.007 (0.01)	0.006 (0.01)	0.006 (0.01)
Certain Words	-0.008 (0.01)	0.003 (0.01)	0.003 (0.01)	-0.001 (0.01)	-0.002 (0.01)
Uncertain Words	-0.019* (0.01)	-0.027** (0.01)	-0.027** (0.01)	-0.035*** (0.01)	-0.034*** (0.01)
Observations	1517	1517	1517	1314	1314
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Model (4) and (5) excludes 203 knock-out round speeches without judge panel information. Robust clustered standard errors at debate level are in parentheses.

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

2.5.2 Are men and women evaluated differently?

The following sections summarize the regression results on the evaluation scores of male and female speakers given their linguistic and debate room characteristics. Since individual speech scores are only given in preliminary rounds, and not in elimination rounds, the total observation points in this section are 984 speeches.

2.5.2.1 Raw difference in scores across control groups

Table 2.4 summarizes the regression results of standardized speech score against speaker's gender, taking into account demographic control variables. Column (1) gives the raw score difference between male and female debaters, Column (2) controls for room variables,⁵⁷ Column (3) and (4) adds gender of the chair judge and judge panel gender composition⁵⁸ respectively. Column (5) includes all room, chair and panel gender composition as control

⁵⁷i.e. (i) Language status, (ii) Speaking position, (iii) Competition, (iv) Room gender composition (i.e. whether the room has ≥ 4 female speakers), (v) Institution ranking, and (vi) Motion type.

⁵⁸i.e. whether the judge panel has *at least* equal, or higher number of female judges vs. male judges.

variables, and finally, Column (6) checks whether there is score differential across genders *within* a debate.

Column (1) shows that, there exists a significant unconditional difference of 16 p.p SD lower score ($p - value = 0.03$) for speeches given by women, compared to those given by men. This confirms hypothesis **H2.1 (Speech Evaluations)**. *On average, speeches given by female speakers receive lower scores than male speakers.* Nevertheless, this raw score difference disappears once controlling for other variables, as seen in Column (2) to (5). At the debate room level, Column (6) shows that there is a 4.2 p.p SD reduction in score for women, yet this result is not significant.⁵⁹

Regarding other control variables, rooms with more female speakers are associated with a consistently lower scores across Column (2) to (5). Female chair judges or female-dominated panels are not associated with lower evaluation scores. A separate extension in Section 2.6 investigates the role of female judges in evaluation scores. Relative to the last speaking position, speaking first in a debate is associated with a 21 p.p SD lower score. This is intuitive given the debate format, where the first speaker lays the argument framework that are subsequently challenged by the maximum number of participants. Non-native speakers are also associated with a significant 34.3 p.p SD lower scores, even when controlling for all other variables. Speakers from top-50-ranked institutions receive significantly higher scores, at 40.8 p.p SD higher scores, compared to the rest. With respect to competitions, speakers in EUDC 2015 have significantly lower scores, whereas those in WUDC 2015 and HWS 2017 have notably higher scores. Noteworthy, speeches in feminism and military motions⁶⁰ get significantly higher scores. However, the number of speeches debating these motions is too limited to observe whether male and female speakers receive significantly different scores.

⁵⁹Debate-fixed effects control for many unobservable factors. However, given the limited number of speeches per debate (minimum 6, maximum 8) and the power-matching in team allocation to debates, variation in scores within debates is relatively limited.

⁶⁰Further checks show that this is due to these motions being in HWS and higher-ranked debates of EUDC and WUDC.

Table 2.4: Regression of Control Variables on Speech Score (N = 984)

	Dependent Variable: Score					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.160** (0.07)	0.016 (0.05)	0.018 (0.05)	0.013 (0.05)	0.015 (0.05)	-0.042 (0.05)
Female-dominated Room		-0.228* (0.13)	-0.239* (0.13)	-0.236* (0.13)	-0.240* (0.13)	
Female Chair Judge			0.141 (0.10)		0.073 (0.12)	
Female-dominated Panels				0.170 (0.11)	0.132 (0.12)	
NonNative		-0.350*** (0.12)	-0.348*** (0.12)	-0.343*** (0.12)	-0.343*** (0.12)	
Top 50 Institutions		0.416*** (0.07)	0.408*** (0.07)	0.410*** (0.07)	0.408*** (0.07)	
1 st Speaking Position		-0.208** (0.09)	-0.209** (0.09)	-0.210** (0.09)	-0.210** (0.09)	
WUDC 2015		0.633** (0.26)	0.658** (0.26)	0.640** (0.25)	0.651** (0.26)	
EUDC 2015		-0.512** (0.22)	-0.547** (0.21)	-0.534** (0.21)	-0.547** (0.21)	
HWS 2016		0.327* (0.20)	0.306 (0.19)	0.232 (0.20)	0.242 (0.20)	
HWS 2017		0.602*** (0.22)	0.630*** (0.22)	0.600*** (0.22)	0.615*** (0.22)	
Feminism Motions		2.078** (0.94)	2.036** (0.92)	2.146** (0.94)	2.109** (0.94)	
Military Motions		1.023*** (0.27)	0.943*** (0.27)	0.874*** (0.27)	0.866*** (0.27)	
R^2	0.006	0.360	0.364	0.365	0.366	0.029
Observations	984	984	984	984	984	984
Room controls		✓	✓	✓	✓	
Chair Judge Gender			✓		✓	
Panel Gender				✓	✓	
Debate fixed effect						✓

Score is standardized. Only significant variables are reported for Competition Type & Year (31 dummies) and Motion Type (17 dummies) and Speaking position (8 dummies). Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (6) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.5.2.2 Score evaluation across genders

To check whether there exists gender-specific evaluation standards between genders, Table 2.5, 2.6 and 2.7 provide the regression results for three groups: (i) basic speech features, (ii) parts of speech (POS) and (iii) LIWC psychometric features. In every table, Column

(1) reports unconditional results of these variables; Column (2) adds room controls;⁶¹ Column (3) considers room controls and chair judge gender; Column (4) accounts for room controls and panel gender composition; Column (5) take into account room, chair and panel controls; and Column (6) uses debate-room fixed effect to account for any room-specific unobserved heterogeneities.

As can be seen in Table 2.5, across the board, longer speeches with more complex words and BP procedural words and phrases correlate positively with evaluation scores. On the other hand, longer sentences, more fillers, and hedges are associated with a negative reduction in scores.⁶² The positive correlation between longer speeches and higher scores aligns with the finding of [Tan et al., 2016] on argumentation success in the Change My View online forum, where they explained that longer replies can be more explicit and contain more information. In terms of evaluation patterns, I find no significant differences in how speeches given by men and women are evaluated, except for fillers and BP procedural words and phrases.

Table 2.6 summarizes the findings with respect to evaluations on composition of different parts of speech. Across the board, controlling for all room and judge characteristics, Column (5) shows that the proportion of nouns and adverbs positively correlate with higher scores. Compared to male speakers, a one standard deviation increase in the proportion of personal pronouns is associated with an 8.5 p.p SD higher score for female speakers. At the debate room level, results in Column (6) show that, in terms of parts of speech, there is no difference in terms of how judges evaluate speeches. In fact, apart from the significant positive correlation between the proportion of nouns in speeches to score, other elements do not matter for evaluation at the debate level.

Moving onto the LIWC psychometric features, Table 2.7 shows some interesting evaluation patterns. Without any demographic control variables, Column (1) shows that certainty indicators, uncertainty indicators, analytic and authentic styles positively correlate with higher scores. Compared to male speakers, an increase of one SD in certainty and

⁶¹i.e. (i) Language status, (ii) Speaking position, (iii) Competition, (iv) Room gender composition (i.e. whether the room has ≥ 4 female speakers) (v) Institution ranking, and (vi) Motion type.

⁶²Table 2.26 in Section 2.8.3.6 provides a clearer overview of how these speech elements matter for speech scores, independent of gender.

uncertainty indicators leads to a 9.99 p.p and 12.3 p.p SD increase in score. In Columns (2) to (5) where room and judge control variables are taken into account gradually, except for authentic speaking style, the effects of other variables become insignificant. Noteworthy, a one SD increase in analytical speaking style reduces women's scores by 12.8 p.p, compared to speeches given by men. Conversely, women are rewarded for having an authentic style and positive emotional tone. At the debate room level, Column (6) shows that such differences in evaluation standards become insignificant. Since the debate fixed-effect model substantially restricts the degrees of freedom (123 clusters of six to eight observations per debate), the disappearing significance could be due to insufficient variation in speeches within a debate, to estimate these heterogeneous effects by gender within debates. In sum, together with the observation of female speech behavior, these results suggest that the raw score gender difference is more attributable to more women using score-reducing features, rather than gender-specific evaluation standards.

These findings suggest that, if these results carry over to workplace settings in workplace negotiations and interviews, any gender differences in outcomes are due to differences in persuasion tactics, rather than how negotiation is evaluated. Given that the lexical features investigated in this high-stake, competitive, male-dominated context correlate with confidence and charisma,⁶³ these findings provide actionable verbal tactics to enhance diversity and inclusion training, especially in narrowing the gender gap in self-promotion [Exley and Kessler, 2019], leadership tendency [Born et al., 2020] and workplace authority [Wright et al., 1995]. The finding on how analytical speaking style could backfire for women also speaks to the potential backlash of "lean in" advice [Exley et al., 2020], thus highlighting the double-bind dilemma in persuasive communication styles for women [Jamieson et al., 1995].

⁶³e.g: fillers [Dinkar et al., 2020], hedges [Mihatsch, 2012; Holmes, 1990], speaking tone and lexical complexity [Yang et al., 2020; Hirschberg and Rosenberg, 2005]

Table 2.5: Linear & Fixed Effects Regression of Basic Features (interacting with Gender) (N = 984)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.006 (0.06)	0.079 (0.06)	0.079 (0.06)	0.077 (0.06)	0.077 (0.06)	0.032 (0.05)
Word Count	0.381*** (0.07)	0.259*** (0.05)	0.258*** (0.05)	0.258*** (0.05)	0.258*** (0.05)	0.195*** (0.03)
Female × Word Count	-0.045 (0.07)	-0.049 (0.05)	-0.049 (0.05)	-0.049 (0.05)	-0.049 (0.05)	0.021 (0.04)
Words per Sentence	-0.040 (0.04)	-0.080** (0.04)	-0.080** (0.04)	-0.086** (0.04)	-0.087** (0.04)	-0.099*** (0.03)
Female × Words per Sentence	0.007 (0.07)	-0.011 (0.06)	-0.011 (0.06)	-0.009 (0.06)	-0.009 (0.06)	-0.050 (0.04)
Complex Words	0.215*** (0.05)	0.171*** (0.05)	0.172*** (0.05)	0.174*** (0.05)	0.174*** (0.05)	0.146*** (0.04)
Female × Complex Words	0.015 (0.07)	-0.040 (0.06)	-0.043 (0.06)	-0.047 (0.06)	-0.046 (0.06)	-0.092 (0.06)
Argument Indicators	0.008 (0.04)	0.011 (0.04)	0.010 (0.04)	0.010 (0.03)	0.010 (0.03)	0.007 (0.03)
Female × Argument Indicators	-0.002 (0.07)	-0.003 (0.05)	-0.004 (0.05)	-0.009 (0.05)	-0.009 (0.05)	-0.025 (0.04)
Fillers	-0.204*** (0.07)	-0.117* (0.06)	-0.112* (0.06)	-0.109* (0.06)	-0.110* (0.06)	-0.037 (0.06)
Female × Fillers	-0.053 (0.08)	-0.086 (0.07)	-0.089 (0.07)	-0.093 (0.07)	-0.092 (0.07)	-0.098* (0.05)
Hedges	-0.116*** (0.03)	-0.076** (0.03)	-0.075** (0.03)	-0.071** (0.03)	-0.071** (0.03)	-0.053* (0.03)
Female × Hedges	0.009 (0.06)	0.010 (0.06)	0.010 (0.06)	0.007 (0.06)	0.007 (0.06)	-0.006 (0.05)
BP Words & Phrases	0.076** (0.03)	0.059** (0.03)	0.059** (0.03)	0.057** (0.03)	0.057** (0.03)	0.068*** (0.03)
Female × BP Words & Phrases	-0.149** (0.06)	-0.104* (0.05)	-0.104* (0.05)	-0.102* (0.05)	-0.102* (0.05)	-0.093** (0.04)
POI Reject	0.028 (0.04)	0.005 (0.04)	0.004 (0.04)	0.006 (0.04)	0.006 (0.04)	-0.005 (0.02)
Female × POI Reject	0.011 (0.06)	0.024 (0.06)	0.024 (0.06)	0.020 (0.06)	0.020 (0.06)	0.029 (0.04)
R^2	0.319	0.455	0.455	0.457	0.458	0.432
Observations	984	984	984	984	984	984
Room controls		✓	✓	✓	✓	
Chair Judge Gender			✓		✓	
Panel Gender				✓	✓	
Debate fixed effect						✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses. R^2 of model (6) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.6: Linear & Fixed Effects Regression of Parts of Speech (interacting with Gender) (N = 984)

	Dependent Variable: Score(standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.046 (0.07)	0.022 (0.06)	0.025 (0.06)	0.020 (0.06)	0.022 (0.06)	-0.032 (0.05)
Noun	0.291*** (0.08)	0.137** (0.05)	0.138** (0.05)	0.142*** (0.05)	0.141*** (0.05)	0.090** (0.04)
Female × Noun	0.104 (0.08)	0.022 (0.06)	0.018 (0.06)	0.017 (0.06)	0.016 (0.06)	0.016 (0.06)
Verb	0.134** (0.07)	0.046 (0.05)	0.046 (0.05)	0.045 (0.05)	0.046 (0.05)	0.025 (0.04)
Female × Verb	0.019 (0.10)	0.004 (0.07)	0.003 (0.07)	0.003 (0.07)	0.003 (0.07)	0.033 (0.06)
Adjective	-0.095 (0.07)	-0.046 (0.05)	-0.045 (0.05)	-0.050 (0.05)	-0.049 (0.05)	-0.015 (0.04)
Female × Adjective	0.038 (0.08)	-0.016 (0.07)	-0.023 (0.07)	-0.024 (0.07)	-0.025 (0.07)	-0.011 (0.06)
Adverb	0.026 (0.06)	0.094** (0.05)	0.091** (0.05)	0.090** (0.04)	0.090** (0.04)	0.042 (0.03)
Female × Adverb	-0.007 (0.08)	-0.064 (0.07)	-0.062 (0.07)	-0.064 (0.07)	-0.063 (0.07)	-0.054 (0.06)
Personal Pronouns	-0.137*** (0.05)	-0.051 (0.04)	-0.053 (0.04)	-0.055 (0.04)	-0.055 (0.04)	-0.004 (0.04)
Female × Personal Pronouns	0.142** (0.06)	0.090* (0.05)	0.086* (0.05)	0.086* (0.05)	0.085* (0.05)	0.022 (0.04)
R^2	0.149	0.387	0.390	0.393	0.394	0.323
Observations	984	984.0	984	984	984	984
Room controls		✓	✓	✓	✓	
Chair Judge Gender			✓		✓	
Panel Gender				✓	✓	
Debate fixed effect						✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses. R^2 of model (6) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.7: Linear & Fixed Effects Regression of LIWC psychometric features (interacting with Gender) (N = 984)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	-0.181** (0.08)	-0.021 (0.06)	-0.018 (0.06)	-0.022 (0.06)	-0.020 (0.06)	-0.043 (0.05)
Certainty Indicators	0.114*** (0.04)	0.044 (0.03)	0.037 (0.03)	0.037 (0.03)	0.035 (0.03)	0.030 (0.02)
Female × Certainty Indicators	0.099* (0.06)	-0.001 (0.05)	0.004 (0.05)	0.001 (0.05)	0.003 (0.05)	0.032 (0.04)
Uncertainty Indicators	0.011 (0.04)	0.028 (0.03)	0.027 (0.03)	0.027 (0.03)	0.027 (0.03)	0.020 (0.03)
Female × Uncertainty Indicators	0.123** (0.06)	0.025 (0.05)	0.027 (0.05)	0.029 (0.05)	0.030 (0.05)	0.061 (0.04)
Analytic	0.147*** (0.04)	0.021 (0.04)	0.022 (0.04)	0.022 (0.04)	0.022 (0.04)	-0.006 (0.03)
Female × Analytic	-0.161** (0.07)	-0.129** (0.06)	-0.128** (0.06)	-0.129** (0.06)	-0.128** (0.06)	-0.004 (0.04)
Authentic	0.175*** (0.05)	0.065* (0.03)	0.063* (0.03)	0.063* (0.03)	0.063* (0.03)	0.041 (0.04)
Female × Authentic	0.043 (0.07)	-0.019 (0.06)	-0.022 (0.06)	-0.023 (0.06)	-0.024 (0.06)	-0.017 (0.04)
Tone	-0.044 (0.05)	-0.049 (0.04)	-0.047 (0.04)	-0.047 (0.04)	-0.047 (0.04)	-0.057 (0.04)
Female × Tone	0.167** (0.07)	0.123* (0.06)	0.124* (0.06)	0.127* (0.06)	0.126* (0.06)	0.031 (0.05)
R^2	0.099	0.374	0.377	0.378	0.379	0.082
Observations	984	984	984	984	984	984
Room controls		✓	✓	✓	✓	
Chair Judge Gender			✓		✓	
Panel Gender				✓	✓	
Debate fixed effect						✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses. R^2 of model (6) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.6 Extensions

2.6.1 Do men and women receive different questions? Do they answer them differently?

Understanding the types of received strategic questions and associated answers by men and women provides a crucial insight into the types of questions posed to each sex and how they handle them. Given the availability of 1828 pairs of Q& As during these debate speeches, I extracted the linguistic features listed in Table 2.21 and 2.22 to understand

the extemporaneous reaction of speakers to the posed questions.⁶⁴ This provides a direct comparison to online forum language sentiment observed in [Bohren et al., 2018]. Furthermore, the large-scale competitive debate speech contexts significantly enrich the small-scale conversation analyses in socio-linguistic literature [Ghilzai and Baloch, 2015].

Table 2.27 gives the logistic regression of linguistic features of accepted questions against the speaker's gender.⁶⁵ Similar to the descriptive statistics in Section 2.3.3, questions posed for female debaters tend to have more hedges. A closer look at these hedges reveals that these are evasive hedging, i.e. hedges that serves to augment the uncertainty of a claim. There appears to be no difference in other measured dimensions, compared to questions posed to male debaters.

On answering behavior, there are notable differences in how male and female speakers answered these questions. Table 2.28 shows that the answers of female debaters during their speeches are longer, have slightly more hedges and second and third-person pronouns, but less negation, compared to those of male debaters. In both questions and answers, there are no LIWC psychometric differences between the groups, with respect to speaking style or the proportion of certainty or uncertainty indicators in their speeches.

The finding of more hedges and personal pronouns in answers of female speakers might indicate persuasiveness of such speeches, in light of the work of [Tan et al., 2016] on interaction dynamics on online forum discussions. They found that persuasive arguments tend to have a significantly larger amount of personal pronouns and hedges, and openness is strongly correlated with a higher level of personal pronouns. Nonetheless, in this back-and-forth, competitive face-to-face debate context, we need further evidence to conclusively determine whether these answers are indeed more persuasive. On top of this, given that there are a wide variety of hedging strategies (evasive hedges, politeness hedges, etc) that depend heavily on contexts, a more fine-grain hedge classification would improve our understanding of hedging behavior in dialogues.

⁶⁴i.e. (i) word count, (ii) hedges, (iii) fillers and pauses, (iv) personal (first and second) and impersonal pronouns, (v) negations, (iv) sentiment (emotional tone), (v) uncertainty words, (vi) certainty words, in addition to (v) analytical style and (vi) authentic style.

⁶⁵The logit estimation equation is the same as the one given in Section 2.4.1.1.

2.6.2 Do female judges evaluate speeches differently?

Insights into how female lead judges and female-dominated panels evaluate speeches, and in particular, women speeches, ties into the growing literature on gender's role in committee evaluation decisions [Régner et al., 2019; Bagues et al., 2017; De Paola and Scoppa, 2015]. This section highlights the evaluation patterns of female chair judges and female-dominated judge panels in a two-fold manner. To understand whether female judges value different speech elements differently compared to male judges, I run a similar regression analysis to the main analysis with speaker's gender, i.e. interact female judge dummy with speech elements, controlling for room characteristics and debate fixed effect, as summarized below.

$$S_{ik} = \alpha_{ik}\mathbb{I}_{FemJ} + \beta_{ik}\mathbb{I}_{FemJ} \sum_{j=1}^n \alpha_j \mathbf{X}_{jik} + \sum_{j=1}^n \gamma_j \mathbf{Y}_{ik} + (\eta_k) + \varepsilon_{ik}$$

To investigate whether female judges evaluate *female speakers* differently compared to male speakers, I interact female judge dummy with speaker's gender dummy variable, controlling for room characteristics, then speech elements and finally debate fixed effects:

$$S_{ik} = \alpha_{ik}\mathbb{I}_{FemS} + \eta_{ik}\mathbb{I}_{FemJ} + \beta_{ik}\mathbb{I}_{FemS}\mathbb{I}_{FemJ} + \sum_{j=1}^n \alpha_j \mathbf{X}_{jik} + \sum_{j=1}^n \gamma_j \mathbf{Y}_{ik} + (\eta_k) + \varepsilon_{ik}$$

In both cases, \mathbf{X}_{jik} represents the term document frequency matrix of frequency count of words and phrases in category j in speech i of debate k , \mathbf{Y}_{ik} are the control categorical variables and η_k is the debate fixed effects.

Since the chair judge in a debate is vested with significantly more decision power,⁶⁶ I run two separate regression sets for two dummy variables: chair judge gender and panel gender composition. Since the number of female judges for each panel varies unevenly across debates, I classify a judge panel as female-dominated if it contains at least equal or higher number of female judges compared to male judges, and vice versa.

⁶⁶Chair judges control the debate flows and judge deliberation discussions, along with holding veto power in case of a tie. They also plays a major role in determining the individual scores of speakers. They are also the judges responsible for giving the oral adjudication justification to debate teams after every debate.

2.6.2.1 Do female judges evaluate speeches differently?

For female chair judges, Table 2.29, 2.30 and 2.31 give the results from the linear and fixed effects regression on how chair judges value speech elements across three linguistic variable groups: (i) Basic Features, (ii) Parts of Speech and (iii) LIWC Psychometric Measures, corresponding to the first equation in 2.6.2. Each table for these groups reports four regression models: (1) no controls; (2) with room controls; (3) With room controls and whether the judge-panel is female-dominated and (4) debate fixed effect. Across all feature groups, female chair judges do not appear to judge these speech elements differently, compared male chair judges.

For female-dominated panels, Table 2.32, 2.33 and 2.34 show the findings from the linear and fixed effects regression on how female-dominated judge panels value speech elements across three linguistic variable groups: (i) Basic Features, (ii) Parts of Speech and (iii) LIWC Psychometric Measures. Each table for these groups reports four regression models: (1) no controls; (2) with room controls; (3) With room controls and whether the chair judge is a female; and (4) debate fixed effect. Controlling for room and chair judge gender characteristics, compared to male-dominated panels, female-dominated panels reward one STD increase in the proportion of BP Words Phrases with 13.7 p.p in score. Nonetheless, at the debate room level, they punish one SD increase in the use of personal pronouns by 12.2 p.p in score. There is no difference in how female-dominated judge panels evaluate LIWC linguistic features of speeches. Altogether, these results suggest that, linguistic-wise, there is virtually no substantial difference in speech evaluation standards between male and female judges.

2.6.2.2 Do female judges evaluate speeches from female speakers differently?

This section summarizes the estimation results of the second equation in 2.6.2, to understand how female chair judges and female-dominated judge panels evaluate speeches given by female speakers vs. those given by male speakers.⁶⁷ I first report the unconditional evaluation patterns, taking into account room controls and debate fixed effect, and then proceed onto controlling for speech elements across three linguistic variable groups: (i)

⁶⁷This corresponds to 128 female speeches judged by female chair judges, and 134 female speeches judged by female-dominated panels.

Basic Features, (ii) Parts of Speech, and (iii) LIWC Features.

Table 2.35 gives the unconditional evaluation patterns of female judges towards female speakers, as well as controlling for room variables and debate fixed effects. Column (C1) to (C3) reports these results for chair judges, whereas column (P1) to (P3) for female-dominated panels. Regarding chair judges, female speeches judged by a female-chaired committee received almost 30. p.p lower SDs in scores. Similar results are found upon considering the entire panel gender composition. Compared to the 179 female speakers who were judged by male-dominated panels, the 134 female speakers judged by female-dominated speakers receive 30.8 p.p lower SDs in scores.

Upon controlling for speech elements, for female chair judges, Table 2.36, 2.37 and 2.38 summarize the results on whether female chair judge evaluate speeches given by women differently. All models (1) to (4) in each table control for speech elements corresponding to specified groups. Table 2.36 controls for Basic Features, Table 2.37 controls for Parts of Speech and Table 2.38 controls for LIWC features. When taking into account speech features and relevant debate room characteristics, having a female chair judge does not appear to impact scores of speeches given by women in the debate. However, at the debate room level, compared to male chair judges, female chair judges punish speeches given to women at a magnitude of 27 p.p to 29 p.p SDs in scores.

For female-dominated judge panels, Tables 2.39, 2.40 and 2.41 summarize the results on evaluation patterns of female-dominated judge panels on speeches given by women. Similar as above, all models (1) to (4) in each table control for speech elements corresponding to specified groups. Table 2.39 controls for Basic Features, Table 2.40 controls for Parts of Speech and Table 2.41 controls for LIWC features. Compared to the results on female chair judges, the impact of having a female-dominated judge panel is notably more significant across all controlled linguistic groups. Even at the debate room level, Column (4) shows that, if there is a female-dominated judge panel, female speakers receive between 29 p.p and 31 p.p SD reduction in scores across all linguistic control groups.

There are three possible hypotheses to explain these results. First, regarding the average lower scores for women across debates, more female judges may be slotted to judge

lower-ranked rooms, which, as we know from Table 2.25 in Appendix 2.8.3.5, are more likely to have more female speakers. In the final three rounds, as a practiced norm, judges who receive consistently good feedback are often "pooled" together to judge either: (1) "battleground" rooms i.e. those with highest accumulated scores, or (2) "bubble" rooms i.e. those with at least one team who could make it to knock-out rounds. As different chief adjudicators (CAs) have different ideas in how to allocate comparably excellent judges, more female judges are likely slotted to "bubble" rooms, instead of "battleground" rooms.

Nonetheless, the robust finding on low scores given only for women at the debate level speaks to the "queen bee effect", i.e. more successful women, who often see younger women as competitors, can actively take steps to hinder the advancement of younger women [Derks et al., 2016; Arvate et al., 2018].⁶⁸ An alternative hypothesis for the "queen bee effect" in debate tournament context could be that successful female judges apply harsh standards that were applied to them by previous generation judges, who are much more predominantly male than in recent years.⁶⁹ Given increasingly applied gender quota policies in boards and committees [Adams and Funk, 2012; Green and Homroy, 2018], along with evidence on gendered behavior and stereotypes in groups [Coffman, 2014; Sarsons, 2017b; Coffman et al., 2019], if female judges indeed judge female speakers more harshly, having more women in evaluation committees does not necessarily translate into gender-neutral evaluation standards.

2.6.3 Do speakers in preliminary rounds speak differently than those in elimination rounds?

To check for internal validity of speech measures I split the data into preliminary round speeches ($N = 984$) and elimination round speeches ($N = 533$) and ran logistic regression of each feature groups against gender. Table 2.44 and Table 2.45 report the results for preliminary round and elimination round speeches, respectively. Overall, most linguistic differences detected between men and women in the previous section exhibit mostly between preliminary round speeches, as noted in Table 2.44. This is particularly the case for sentence length, hedges and proportion of adjectives in elimination round speeches, with

⁶⁸See this [news article](#) for the discussions on such behavior in the workplace

⁶⁹To investigate this possibility, we need data on the judge gender composition of previous years, along with the speeches, which unfortunately we do not have.

speeches given by women have slightly longer sentences, more hedges and fewer adjectives. In terms of LIWC psychometric features, there appears to be no difference between male and female speeches in elimination rounds. These results confirm the internal validity of these linguistic measures, which are in line with the existing literature on lexical correlates of persuasiveness in speeches [Petukhova et al., 2017b,a].

2.7 Discussion & Conclusion

Given the importance of oral persuasion skill [Buser and Yuan, 2020] and the gap on studies on spoken linguistic tactics and evaluations, this paper analyzes gender differences in speech behavior and evaluation patterns of 1517 speeches from top-tier debate tournaments. I find significant variation in speech patterns of male and female speakers. Female speakers use more personal and disclosing speaking style, with more hedging phrases and non-fluencies in their speeches. In their answers to questions from opponents during their speeches, they negate less while having notably longer and more vague answers. Evaluation-wise, within debates, except for non-fluencies, there is no robust evidence of gender-specific evaluation standards. Controlling for speech patterns, even at the debate room level, female-dominated judge panels are significantly harsher towards female speakers.

Overall, these results suggest that the raw score gender differences is more attributable to more women using score-reducing features, rather than gender-specific evaluation standards. Importantly, these results cannot exclude the possibility of differing "optimal" persuasion styles between men and women. Conditional on argumentation quality, the current evaluation standards might have evolved to reward more the speech features that men are, on average, better at. In future work, given the speech pattern findings in this paper, an experiment with variation in speaker's gender can provide causal interpretation on whether particular speech elements are more credited towards male or female speakers.

The combined dictionary-based method to study persuasion style in this paper, drawing on validated studies in persuasiveness and politeness, offers an easy-to-apply tool to extract lexical insights from text data with limited observations in experimental settings. Examples include recent work on gender dynamics in economic seminars [Dupas et al.,

2021], [Dupas et al., 2021], opinion aggregation [Mengel, 2020], [Coffman et al., 2019] or group deliberation exchanges [Stoddard et al., 2020]. Given the sparse debate data set at hand, it gives a meaningful overview on the correlations between linguistically meaningful variables and evaluation scores. However, the current approach is subject to two limitations. First, one caveat in interpreting significant results in this paper is the multiple hypothesis testing issue, arising from the large number of linguistic variables. Text as data is inevitably high-dimensional, and the hypothesis-driven approach in this paper arguably delivers the smallest set of interpretable linguistic variables between speeches of men and women.

A second limit is that it cannot disambiguate the contexts surrounding the tagged words and phrases, thus missing out on important information.⁷⁰ For future work, it is crucial to cross-validate various hedges [Ulinski et al., 2018] and fillers detection techniques. Furthermore, dictionary-based approach cannot account for: (i) words and phrases not included in these categories; and (ii) argument components and relations. Addressing the first issue is possible with lasso, ridge and elastic net regularization [Zou and Hastie, 2005] on the high-dimensional term frequency matrices of words and phrases⁷¹ to predict which words/phrases are the best predictors of: (i) speeches given by men or women and (ii) higher evaluation scores. These methods take into account the multi-collinearity across these term-frequency predictors i.e. lasso weeds out low-relevant variables; ridge shrinks coefficients of variables that do not significantly help explaining variations across speeches; while elastic net model blends both features. Nonetheless, they do not work well on the current data set.⁷²

A more promising approach is argument mining, given the recent success of BERT model [Chakrabarty et al., 2019] in detecting argument components and relations in smaller data sets. A major prerequisite though is an annotated corpus of these debates to serve as training data set, which requires an extensive amount of manual annotation of different coders [Lippi and Torroni, 2016]. Hence, a natural extension of this paper is adapting the annotation guidelines of [Stab and Gurevych, 2014; Stab and Habernal, 2016] to tag argument components and relations in debate speeches. Following the prevailing annotation

⁷⁰For instance, the word "like" can be an adjective, a verb and also serves as a hedge in various occasions.

⁷¹word and phrases of two or three words (i.e 2-gram, 3-gram respectively)

⁷²The top predictive features from this method are stop words and debate-procedural phrases. Removing these words drastically reduces its predictive power.

schemes [Palau and Moens, 2009], three aspects from the speeches are annotated: (1) Point of Information⁷³; (2) Argument components: Claims⁷⁴ and Premises;⁷⁵ and (3) Argument Relations.⁷⁶ In terms of the manual argument annotation process, based on the guideline, we use [ATLAS.ti](#) to code the above-mentioned elements. The 80% manually labelled tags of debate speeches serve as training data to feed into the BERT model [Chakrabarty et al., 2019]. This model is pre-trained on a large number of online forum exchanges Change My View, persuasive essays [Habernal and Gurevych, 2016] and speeches [Petukhova et al., 2017b]. I leave this as a natural extension for this chapter in future work.

⁷³Question and Answer parts taken/given during the speech.

⁷⁴A claim [Palau and Moens, 2009] is a debatable statement that can either be true or false.

⁷⁵Premises are reasons/justifications given to persuade listeners of the claim, including, but not limited to: P1 = examples, P2 = facts/statistics/news bits, P3 = analogies, P4 = anecdotes and P5 = others.

⁷⁶i.e. Support/Attack/Discuss relationship. see Appendix 2.8.2.11 for an example on ATLAS.ti.

2.8 Appendix

2.8.0.1 Debate format

British Parliamentary (BP) is the most widely adopted debate format in top-tier intercollegiate competitions and youth public speaking training programs worldwide.⁷⁷ BP debate topics relate to a broad range of current issues in politics, animal rights or social justice. With respect to motion types, the motion is either a policy which changes the status quo (e.g. This House Would Provide All Police Officers With Firearms) or a statement, the truth or falsehood of which is examined in the debate (e.g. This House Regrets the Decline of Marxism in Western Liberal Democracies).⁷⁸

In terms of the debate format, participants enter the competition in fixed teams of two, whereby they will be *randomly allocated* : (i) a TEAM speaking position (Opening Government (OG), Opening Opposition (OO), Closing Government (CG), Closing Opposition (CO)) and (ii) opponent teams, in every debate round. After given a topic, teams are given fifteen minutes to prepare; no online resources are allowed. During preparation time, speakers *within* a team can strategically decide who takes which roles given their assigned team position. Afterwards, everyone gives a 7-minute speech, sequentially, as shown in Figure 2.3 below:

⁷⁷For instance, the World Universities Debating Championship, Pan African Universities Debate Championship and European Universities Debating Championship and numerous regional tournaments in Europe, Canada, United States, Hong Kong, Shanghai, Philippines, Australia, New Zealand and Africa.

⁷⁸For a list of motions and topic pools in debate tournaments, see [Hello Motions](#) and [European Debating Blogspot](#)

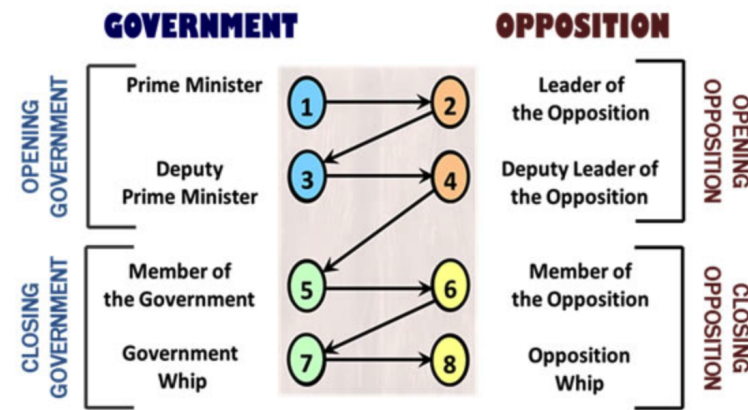


Figure 2.3: British Parliamentary debate speech order

Speech duration is capped at 7 minutes per person. The first and last minute of the speech is *protected*, i.e. no opposition teams could offer a point of information (POI). The POI is a formalized interjection of any speaker from the opposite side, which often lasts *no* longer than fifteen seconds. The speaking debater can choose to hear the POI or to dismiss it. It is generally considered good practice to accept at least one POI during a speech. After the debate, the adjudication panel, often consists of highly accomplished debaters, will discuss in 15-20 minutes and decide upon: team ranking (from 1st to 4th place), individual speech evaluation scores and justifications for the ranking decision. In BP debating, speech evaluation refers primarily to the comparative strength of the argument analysis, with respect to logical proofs and rebuttals to substantive materials of opponents.

2.8.0.2 Judge allocation mechanism

Every tournament has an appointed Chief Adjudicator (CA) team of four to six internationally accomplished debaters who are in charge of judge recruitment, quality screening, monitoring and overall panel allocation throughout the tournament. The CA team often recruits and ranks judges in two rounds. In the first round, prospective judges can opt in to do an advanced judge test to be selected as *Independent Adjudicators (IA)*. If they are chosen, they can potentially get offers of travel and/or accommodation funding from the CA team. In the second round, all judges, IAs or not, must finish a comprehensive judge test, which, in combination with survey data on past judging & debate experiences, enable the CAs to

slot judges as chairs, panelists or trainee⁷⁹ judges in Round 1s. A tabulation algorithm,⁸⁰ supervised by a Tabmaster, randomly assigns judges and debaters simultaneously to different rooms, taking into account possible conflicts of interests.⁸¹ From Round 2 onward, judge performance feedback for each judge, given by peer judges and debaters in previous rounds⁸² are incorporated into the allocation algorithm to demote or promote⁸³ judges in the next round. From Round 7 to Round 9, to ensure highest quality judging in strategically important debate rooms, as a conventional practice, Chief Adjudicators stack highest-performing judges to these rooms; while promoting promising high-performing new wing judges to chair in lower-ranked rooms. In other words, given the debate quality assigned in these rounds, judges could form educated guess of whether or not they have performed well in previous rounds. In general, cumulative judge performance in all preliminary rounds determines the highest 20 – 30% ranked judges to adjudicate the out-rounds, i.e. being acknowledged as the best judges of Europe or the world. Hence, judging in EUDC/WUDC out-rounds is considered a significant achievement, which often yield invitations to travel and chief adjudicate major competitions.

2.8.1 Data Collection & Text Pre-processing Procedure

2.8.1.1 LIWC psychometric variables in detail

The following equations provide the linear combination of the psychometric variables used in LIWC, while Figure 2.4 shows the qualitative meaning of these scales in spectrum:

$\text{Analytic} = 30^{84} + \text{Article} - \text{Preposition} - \text{Personal Pronoun} - \text{"I" Pronoun} - \text{Auxiliary Verb} - \text{Conjugation} - \text{Adverb} - \text{Negation}$

⁷⁹i.e judges without any judging experience and/or do not do the judge test. They have no voting right.

⁸⁰e.g. Tabbie2, Tabbycat

⁸¹e.g: judges and speakers from the same current or past institutions, ex-debate partners, romantic partners, close friends.

⁸²An adjudicator's overall score is as follows: $Score = (1 - w) * Testscore + w * AverageCumulativeFeedbackScore$ where w is the feedback weight for the round. Test score includes weighted average between judge test and previous judge & speaking achievements.

⁸³e.g. bad-performing chairs become wing judges, good-performing wings become chair judges, good-performing chair judges are promoted to higher rooms.

⁸⁴According to the authors [Pennebaker et al., 2015], the value 30 was added to the word percentages so that the resultant score becomes typically positive.

Authentic = I + She/He + They + Differing Group - Negative Emotion - Motion Group

Emotional Tone = Positive Emotion - Negative Emotion

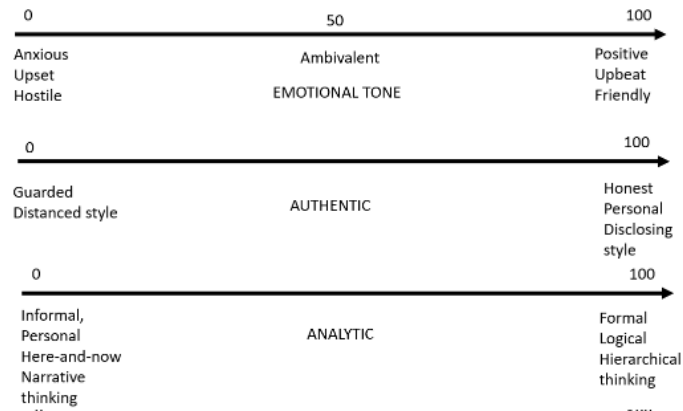


Figure 2.4: LIWC psychometric measures of speech [Pennebaker et al., 2015]

2.8.1.2 Speech transcripts

I tracked down 250 debates i.e. 1104 preliminary round speeches and 896 knock-out round speeches from YouTube, tournament websites and Facebook channels. Given the limited recording resources at tournaments, especially in earlier years, and the prevailing community demand for above-average debates, the majority of recorded debates are from 2014 afterwards in upper-average debate rooms.⁸⁵ All debates with ≤ 6 out of 8 speeches are omitted, to ensure sufficient power for within debate analyses. Furthermore, as we prioritize debate speeches with evaluation scores, I kept all preliminary round speeches; for EUDC and WUDC, I kept 596 speeches from knock-out rounds of native speaker category. This results in 1700 debate speeches, i.e. 225+ hours of usable debate speeches, which are converted into encoded audio files and sent to a professional human-based transcription company. Using three layers of independent editing and proofreading by four professional transcribers, the process ensures 98+ % accuracy standard. Upon receiving the transcripts, I watched the videos of associated transcripts to check: (i) transcript completeness and speech comprehensibility and (ii) speaker gender, team identity, speaking position, debate round and judges. In terms of transcript completeness, I omitted speeches that have ei-

⁸⁵HWS only starts recording final rounds from 2010 onwards, and in-rounds from 2014 onwards.

ther: (i) $\geq 4\%$ of inaudible segments; or (ii) significant portion of cutoff audios (i.e. 20% lower word count than average) due to audio recording issues. Two research assistants independently repeated these steps. Overall, we omitted 183 speeches and kept $N = 1517$ speeches for analysis. Finally, I removed inaudible/clapping/laughter remarks, punctuation and brackets, and lower-cased the entire corpus for analysis [Gentzkow et al., 2019a]. Table 2.8 and 2.9 report the selected set of speeches, breaking down by year, competition and gender. The number of speeches given by females are approximately 25 – 30% across all subcategories, which is in line with debate tournament participation rate across genders.

2.8.1.3 Questions and Answers (Q&A) during speeches

By British Parliamentary (BP) debate rule, between 2nd and 6th minute of a speech, opponents from the other side could ask permission to give a short interjection question. There is no limit on how many questions can be asked and answered; and the common rule of thumb is that the speaker accepts one to two offered requests during their speech. Given the strategic relevance of these questions in persuasive speech evaluations, we extracted 1828 pairs of clearly audible accepted questions and associated answers given during each speech. Out of 1517 speeches, 1333 questions are answered by male debaters, and 495 are by female debaters. In other words, 50 male debaters and 31 female debaters did not accept any offered questions during their speeches. Table 2.18 gives the breakdown of these accepted Q & As by speaker’s gender per competition.

2.8.1.4 Demographic data

Regarding speech evaluation score and relevant demographic characteristics, for EUDC and WUDC tournaments from 2015 to 2018, I scraped information about individual names, language skill statuses⁸⁶ and judge panels from [Tabbie2](#) and [Tabbycat](#), the open-source tabulation systems for parliamentary debate tournaments. For the period of 2008 to 2014, I scrape the data of EUDC and WUDC from: (i) WayBack Machine, a digital archive of

⁸⁶In large competitions i.e. WUDC and EUDC, people are classified into three categories (i) EPL: English as a proficient language; (ii) ESL: English as a second language; and (iii) EFL: English as a foreign language. These statuses are determined by the Language Committee team based on the submitted information from debaters on: (i) the age at which one is exposed to English; (ii) the content, structure and quality of English used for any relevant instruction or exchange; and (iii) the fluency and background of the people in frequent contact. Due to the small data set of EFL-classified speakers and the nebulous categorization between EFL and ESL speakers, in this paper, individuals are classified as *native* (i.e. EPL-status speakers) ($N_{Native} = 1212$) or *non-native* (i.e. ESL- and EFL-status speakers) ($N_{Non-native} = 305$).

World Wide Web; and (ii) organizers of these tournaments. For HWS Round Robin Champion League, I solicited the data directly from the tournament director, as it is organized annually by Hobart William Smith College in New York, United States.

Debaters. In terms of individual identity matching, given that team names reflect the institutions and countries that the respective debaters represent,⁸⁷ I sorted all variations of such team names and identified 176 unique institutions from 43 countries in the entire administrative data set. Since team name uniquely corresponds to two debaters in each competitions, I can match the individual names⁸⁸ from the above constructed datasets from Tabbie2, Tabbycat and Wayback Machine. This procedure resulted in $N_{names} = 748$ unique individuals in the data set, given which we can directly obtain individual language skill statuses from the yearly tabulated data sets. To better understand the different study majors of debaters, we collected data from social media and professional profiles of these individuals, by matching their names, institutions, debate activities and appearance from the videos. This procedure allows us to identify 507 out of 748 unique speakers, whereby Figure 2.6 summarizes this information. Overall, speakers come from a wide variety of background, with predominantly students of social sciences (e.g: law, political sciences, government studies) and economics, with roughly 20% come from natural sciences & medicine.

To determine individual gender and speaking order within a team,⁸⁹ I watched the videos. When teams have the same gender representation (i.e. both speakers are males/females), running their names on popular social media channels to identify their faces, education background, and ethnicity to determine their speaking position within the team in the debate. For five videos without any team information or faces, extensive deduction exercises are done based on the speaker's accent, the judge's gender, the team composition of the room, and scores on Tabbie/Tabbycat. I also reach out to tab masters, whenever possible, confirm the institution and speaking positions. Two research assistants independently verified this information to ensure correctness.

⁸⁷e.g.: Harvard A, Oxford B, Rotterdam C, etc.

⁸⁸Many individuals used varying versions of their names e.g: Elizabeth - Lizzy, Tom - Thomas, etc, in which case I matched their institutions and language statuses and occasionally re-watched the video to confirm any overlapping identities over the years.

⁸⁹Individuals debate in teams of two and can *endogenously* decide who speaks first or second within the team.

Gender composition of debaters in the room. Given the conditionally exogenous assignment⁹⁰ of debate opponents and the growing literature documenting women underperforming in mixed-sex/male-dominated environment [Apesteguia et al., 2012; Booth and Yamamura, 2018], another source of potential explanation for women having lower scores than men could be due to this reason. Figure 2.10 shows that the majority of rooms are male-dominated, with the highest concentration of rooms where there are 2 female and 6 male debaters. There are no rooms with exclusively female speakers, while there are 24 rooms with exclusively male speakers.

Judges. For EUDC and WUDC competitions from 2015 to 2018, web scraping from Tablicat and Tabbie 2 gave the information on judge's full name, their associated institution,⁹¹ judge panel size per debate and the debates they have judged for each competition. For HWS competitions, such information is solicited directly from the tournament director. Noteworthy, from 2016 to 2018, HWS has adopted a two-panel judge setup, whereby two panels adjudicate the same debate independently. Thus, for these debates, the dichotomous variable for chair judge gender and female-dominated judge panel takes the value of 1 if there is *at least* one panel satisfying the criteria. In terms of the female ratio on the panel, this would be the total number of women divided by the total number of judges of *both* panels.

Gender composition of judges on panel. Given the growing evidence of the impact of having women in the committee on evaluations [Bagues et al., 2017] and in corporate board decisions [Kim and Starks, 2016], I identify the gender of the chair judge and panel gender composition for each debate from 407 unique judges in these debates, to serve as control variables in the regression analyses. From the web scraped information on judge panels and score ballots from HWS debate director, the gender representation of chair judge is identified for all 1517 debate speeches. Nonetheless, since the camera did not capture the entire judge panel in some debates, we can only document the entire panel gender composition of 1314 out of 1517 speeches. Looking at Table 2.10 and Figure 2.11, we noted that one-third of speeches are adjudicated by a female chair judge, whereas

⁹⁰teams are power matched given the accumulative points they receive throughout the rounds.

⁹¹By rule, debate institutions must send $N - 1$ number of judges, where N = number of teams they send to the competition. E.g: Oxford sends 3 teams in EUDC 2018, hence they must send 2 judges. On top of institutional judges, the chief adjudicator team recruits a lot of experienced and established debaters as "Independent Adjudicators" in these large-scale tournaments.

the rest are by a male chair judge. Similar to judge panels, Table 2.11 and Figure 2.12 show that they are primarily male-dominated, with the most common setup including one-third of judges as female. The only exception is HWS debates, where almost 40% are adjudicated by female-dominated panels.⁹²

Institutions. Given that the academic institution that a person represents could potentially carry reputation/prestige that impacts evaluations, we collected unique institution data⁹³ from every tab file and obtained 176 distinct institutions for this data set. Since there exists no formal ranking of universities worldwide based on their debate achievements,⁹⁵ these institutions are categorized by their average academic ranking from QS World Universities Ranking in the past 11 years (2008-2018) into two groups: top-50 and the non-top-50 universities. These universities are represented across 170 countries, where the top 10 are documented in Figure 2.6. After collecting institution information from team names of debaters, I clustered 170 institutions by their reputation ranking into two groups: top 50 or non-top-50 institutions. Figure 2.14 shows the distribution of speakers given their gender across these two groups. Noteworthy, whereas male and female debaters in the top-50 universities have practically similar score distribution, female debaters not in the top-50 universities have significantly lower speech scores compared to their male counterparts.

Debate topics. Across the 198 debates of these 1517 speeches, there are 130 unique debate motions discussed across a wide range of topics. All topics provided a balanced, in-depth but polarized distribution of views, as empirically tested by chief adjudicators in earlier regional competitions. I manually classified these motions into 17 debate topics, based on the classification at [International Debate Education Association](#), the distribution of debate speeches across these topics is in Figure 3.4. Most debate topics fall into Society, International Relations, Law, Politics, and Economy, with virtually no topics on Science and only one debate on Sports motion. Given the imbalanced distribution across topics, in robustness checks analysis, to understand whether men and women speak are evaluated differently across topics, I exclude speeches discussing Science, Sports, Health, and Education (i.e. topics with ≤ 0.05 proportion of speeches per motion type).

⁹²HWS is a 16-team round-robin tournament format, hence it is a lot easier for the organizers to arrange more gender-balanced judge panels without compromising the judge pool quality.

⁹³In the case of multiple abbreviations by the same institutions, we checked the data manually⁹⁴

⁹⁵Apart from a top 5 and top 10 list of universities to master debate skills in the [US](#) and [UK](#)

2.8.2 Figures

2.8.2.1 Example of individual speech score scale (50 -100): Tallinn EUDC 2017



SPEAKER SCALE¹

The mark bands below are rough and general descriptions; speeches need not have every feature described to fit in a particular band. Throughout this scale, 'arguments' refers both to constructive material and responses. Please use the full range of the scale. Speaker marks determine many of the breaking teams, and tab finishes can be big achievements, so please give them the serious thought they require.

95-100	<ul style="list-style-type: none"> Plausibly one of the best debating speeches ever given; It is incredibly difficult to think up satisfactory responses to any of the arguments made; Flawless and compelling arguments.
92-94	<ul style="list-style-type: none"> An incredible speech, undoubtedly one of the best at the competition; Successfully engaging with the core issues of the debate, arguments exceptionally well made, and it would take a brilliant set of responses to defeat the arguments; There are no flaws of any significance.
89-91	<ul style="list-style-type: none"> Brilliant arguments successfully engage with the main issues in the round; Arguments are very well-explained and illustrated, and demand extremely sophisticated responses in order to be defeated; Only very minor problems, if any, but they do not affect the strength of the claims made.
86-88	<ul style="list-style-type: none"> Arguments engage with core issues of the debate, and are highly compelling; No logical gaps, and sophisticated responses required to defeat the arguments; • Only minor flaws in arguments.
83-85	<ul style="list-style-type: none"> Arguments address the core issues of the debate; Arguments have strong explanations, which demand a strong response from other speakers in order to defeat the arguments; May occasionally fail to fully respond to very well-made arguments; but flaws in the speech are limited.
79-82	<ul style="list-style-type: none"> Arguments are relevant, and address the core issues in the debate; Arguments well made without obvious logical gaps, and are all well explained; May be vulnerable to good responses.
76-78	<ul style="list-style-type: none"> Arguments are almost exclusively relevant, and address most of the core issues; Occasionally, but not often, arguments may slip into: i) deficits in explanation, ii) simplistic argumentation vulnerable to competent responses or iii) peripheral or irrelevant arguments; Clear to follow, and thus credit.
73-75	<ul style="list-style-type: none"> Arguments are almost exclusively relevant, although may fail to address one or more core issues sufficiently; Arguments are logical, but tend to be simplistic and vulnerable to competent responses; Clear enough to follow, and thus credit.
70-72	<ul style="list-style-type: none"> Arguments are frequently relevant; Arguments have some explanation, but there are regular significant logical gaps; Sometimes difficult to follow, and thus credit fully.
67-69	<ul style="list-style-type: none"> Arguments are generally relevant; Arguments almost all have explanations, but almost all have significant logical gaps; Sometimes clear, but generally difficult to follow and thus credit the speaker for their material.
64-66	<ul style="list-style-type: none"> Some arguments made that are relevant; Arguments generally have explanations, but have significant logical gaps; Often unclear, which makes it hard to give the speech much credit.
61-63	<ul style="list-style-type: none"> Some relevant claims, and most will be formulated as arguments; Arguments have occasional explanations, but these have significant logical gaps; Frequently unclear and confusing; which makes it hard to give the speech much credit.
58-60	<ul style="list-style-type: none"> Claims are occasionally relevant; Claims are not formulated as arguments, but there may be some suggestion towards an explanation; Hard to follow, which makes it hard to give the speech much credit.
55-57	<ul style="list-style-type: none"> One or two marginally relevant claims; Claims are not formulated as arguments, and are instead are just comments; Hard to follow almost in its entirety, which makes it hard to give the speech much credit.
50-55	<ul style="list-style-type: none"> Content is not relevant; Content does not go beyond claims, and is both confusing and confused; Very hard to follow in its entirety, which makes it hard to give the speech any credit.

¹ The scale is consistent with the one used at Warsaw EUDC

2.8.2.2 Distribution of speech scores: Male vs. female speakers

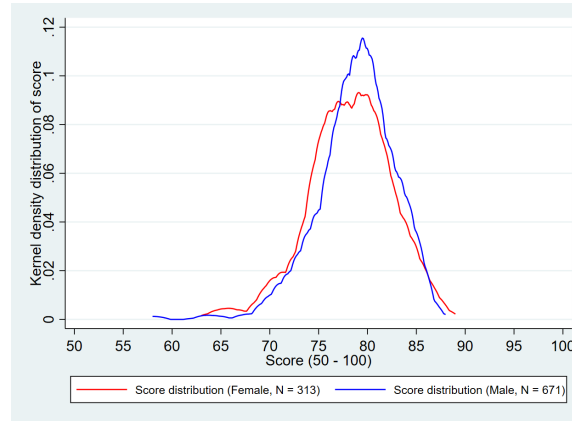


Figure 2.5: Kernel density distribution of speech scores ($N_{Female} = 313$, $N_{Male} = 671$)

2.8.2.3 Debaters' background: study majors & countries of institutions

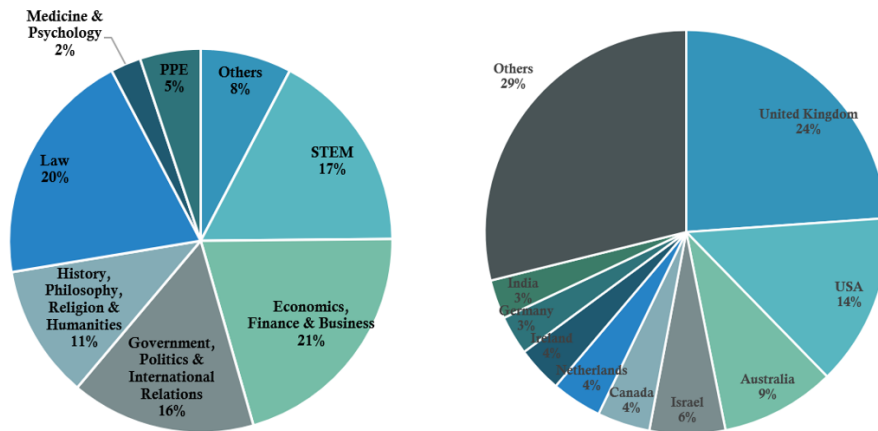


Figure 2.6: Study Major & Country of institutions (N = 507 persons)

2.8.2.4 Example: feature extraction of fillers

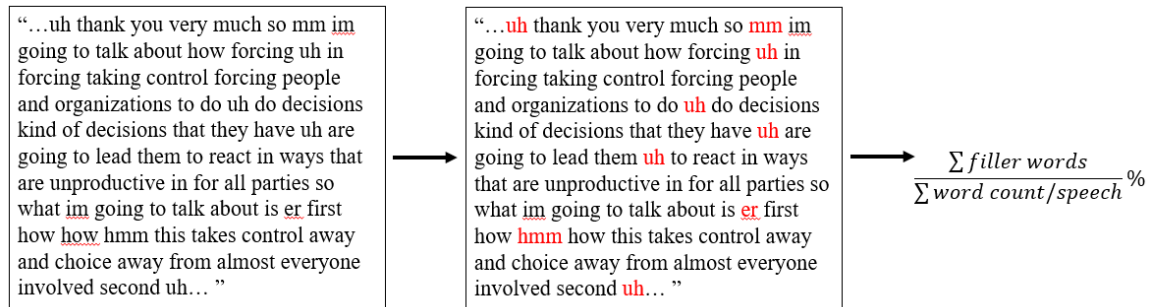


Figure 2.7: Example of constructing variable "Fillers", which includes words such as "uh, umm, hmm, er"

2.8.2.5 Box plots of score differentials across groups

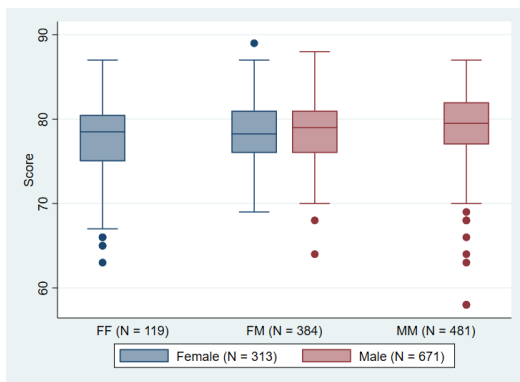


Figure 2.8: Score distribution by team gender composition

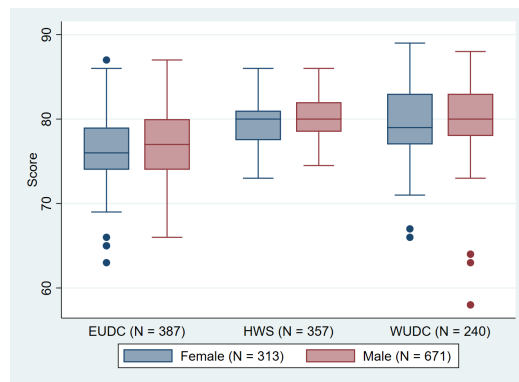


Figure 2.9: Score distribution across competitions

2.8.2.6 Distribution of speeches given number of female speakers in a debate

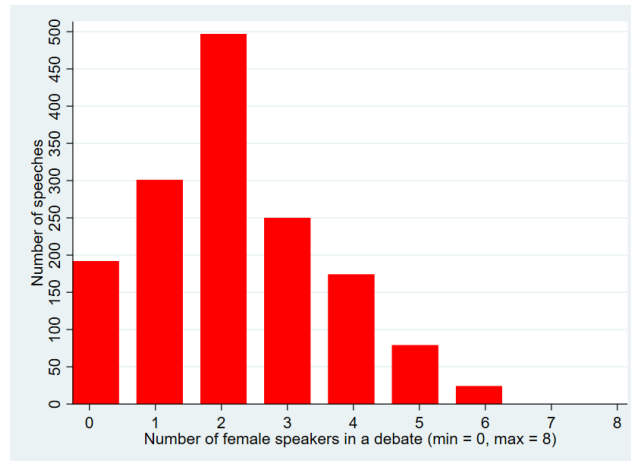


Figure 2.10: Distribution of speeches given number of female speakers in a debate

2.8.2.7 Proportion of speeches given gender of chair judge

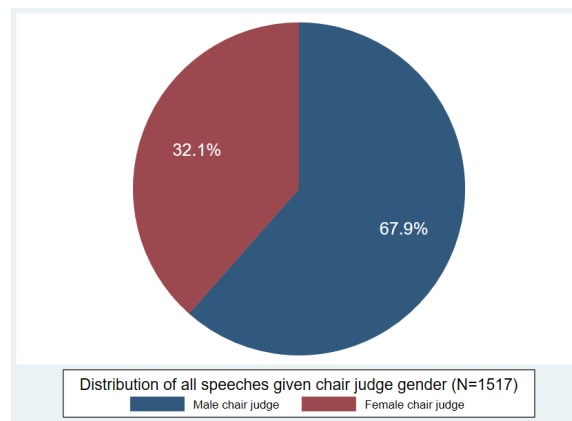


Figure 2.11: Proportion of speeches given given chair judge gender

2.8.2.8 Proportion of speeches given gender of judge panels

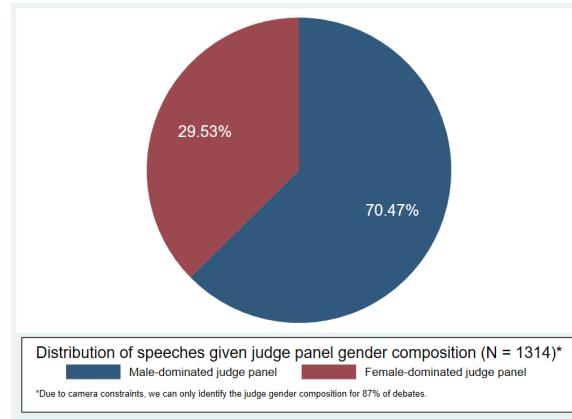


Figure 2.12: Proportion of speeches given judge panel gender composition

2.8.2.9 Distribution of debate motions

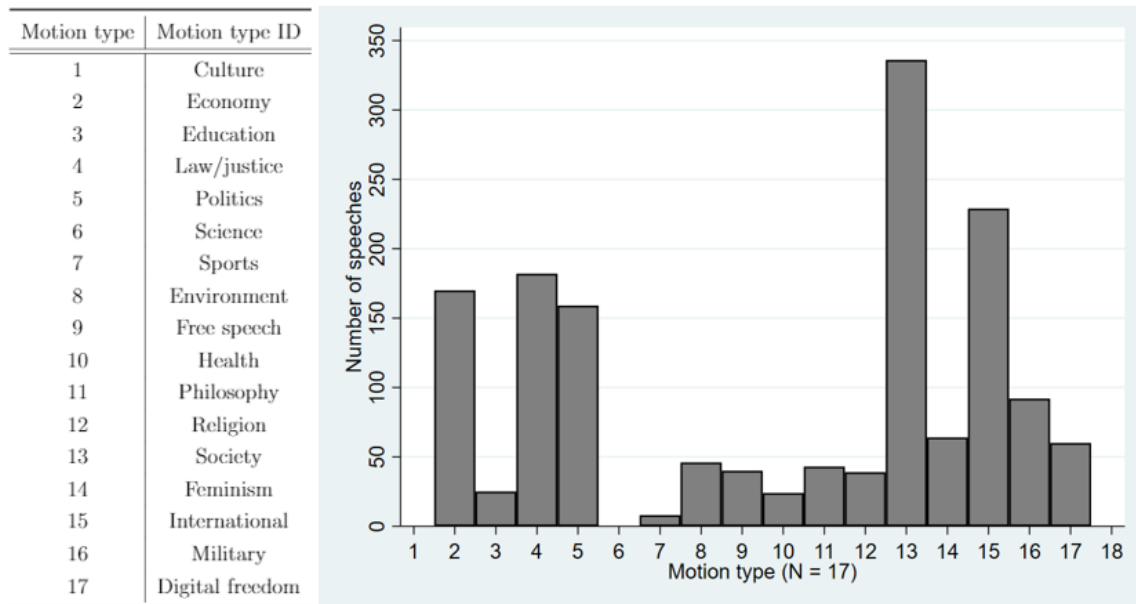


Figure 2.13: Distribution of 1517 speeches according to motion types (17 motion groups)

2.8.2.10 Distribution of speech scores given institution ranking

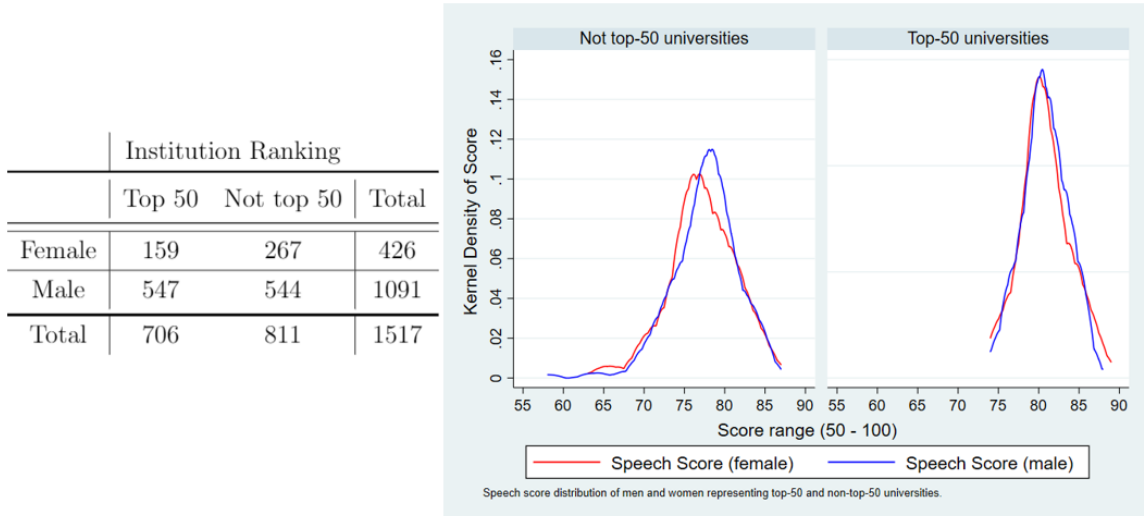
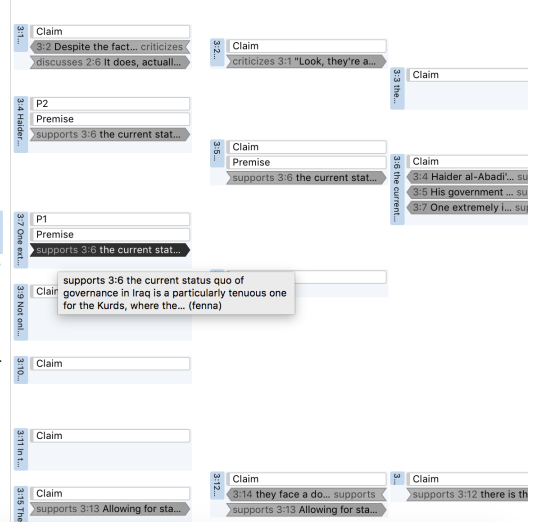


Figure 2.14: Distribution of speech scores of men and women given their institution ranking

2.8.2.11 Tagging overview example

Thank you, honorable Speaker. I want to talk about two things that [INAUDIBLE 00:16:04], that will directly deal with this point we get from opposition, that, "Look, they're autonomous anyway, what are the benefits of this policy?" Despite the fact that the Kurdish people in Iraq live in a state of semi-autonomous control over their region, there are a number of distinct benefits of this policy that [INAUDIBLE 00:16:19] already talked about, that I want to expand on. First, the idea that they are no longer dependent to the same degree on the whims, and the political fortunes, of the central government in Baghdad. Haider al-Abadi's whole governing coalition is 56% Shia. This is actually more than his predecessor, Amaliki [Phonetic], who was generally considered a Shia extremist, who was much more concerned with benefits to his own sect of Islam, as opposed to spreading these benefits to other minorities, and other people within Iraq, such as the Kurds. His government coalition is even more skeptical of decentralization to the region than his predecessor. What this means is, the current status quo of governance in Iraq is a particularly tenuous one for the Kurds, where the autonomy they talk about is more likely than ever to be rescinded...not in totality, obviously...but on a sliding scale that means that Kurd...Kurdistan...is likely to lose control over [INAUDIBLE 00:17:08] powers. One extremely important example of this is oil revenue. Currently, a major source of contention between the central government region is over exploration and development contracts for oil companies, and whether or not Kurdistan gets to allocate them on their own, or whether or not the central government has veto power over the kinds of companies who enter that region. The problem is that, in a situation of uncertainty, the continued ability of...for...the Kurds to control as basic an aspect of who is allowed to do business within their borders is not certain. Not only is this a problem for the economic prospects of the region, not only is it a problem for their self-control, it is a, like, self-fulfilling prophecy, [INAUDIBLE 00:17:41] bad for business, because of the uncertainty that exists over who controls that critical aspect of getting contracts.

That means that the control that exists right now in the status quo, while...no, thank you...it does cover several important powers, is an uncertain one, and that is unacceptable. Next, on regional alliances, Joe talks to you about why an independent Kurdish state could be an extremely useful regional ally for the West. He talks to you about the good will you build up when youth in the region, and leaders in the region, have seen the West, and [INAUDIBLE 00:18:06] motivations, pushing and advocating for them to have a state. But there's a second reason why I think it makes them better allies. In the status quo, governing leaders and politicians of the Kurdish regional government have to balance competing considerations when assessing how to cooperate both with Iraq, and international actors. Given that they face a domestic population clamoring for independence, there is that immediate check on the degree to which they can cooperate without having something to show for it in return. There is an immediate check to the extent to which they can help the West, and the extent that...to which...they can help Iraq, before the question comes about what these people they're helping are doing to assist



2.8.3 Tables

2.8.3.1 Data Overview

Table 2.8: Speeches by year & competition

Year	WUDC	EUDC	HWS	Total
2008	8	7	0	15
2009	24	15	0	39
2010	32	8	8	48
2011	16	16	8	40
2012	32	8	8	48
2013	56	8	8	72
2014	22	206	111	339
2015	48	189	24	261
2016	105	94	128	327
2017	123	24	110	257
2018	42	13	16	71
Total	508	588	421	1517

Table 2.9: Speeches by gender & competition

Speech	Male	Female	Total
In-rounds	671	313	984
Out-rounds	420	113	533
EUDC	407	181	588
WUDC	366	142	508
HWS	318	103	421
Total	1091	426	1517

Table 2.10: Distribution of speeches given chair gender across competitions

Competition	M-chair	F-chair	Total
In-rounds	606	378	984
Out-rounds	424	109	533
WUDC	330	178	508
EUDC	407	181	588
HWS	293	128	421
All	1030	487	1517

Table 2.11: Distribution of speeches given panel gender composition across competitions

Competition	M-dom	F-dom	Total
In-rounds	618	366	984
Out-rounds	308	22	330
WUDC	292	121	413
EUDC	419	125	544
HWS	215	142	357
All	926	388	1314

2.8.3.2 All tournament score descriptive statistics

Table 2.12: Tournament score descriptive statistics ($N = 107705$ speeches)

Competition code	Mean	Min	Max	Median	SD	No. of speeches	Note
WUDC08	74.02	50	93	75	6.16	6804	
WUDC09	73.78	50	91	74	5.36	5400	
WUDC10	74.94	50	92	75	5.14	6837	
WUDC11	74.95	53	90	75	4.65	5580	
WUDC12	74.33	51	90	75	4.73	6813	
WUDC13	74.63	50	91	75	4.31	6877	
WUDC14	75.15	56	90	75	4.32	6030	
WUDC15	75.94	58	90	76	4.20	6308	
WUDC16	75.72	52	88	76	3.95	6732	
WUDC17	76.47	58	88	77	3.79	6561	
WUDC18	76.44	58	88	77	4.10	5417	
EUDC08	72.42	50	90	73	6.10	2114	only 7 rounds
EUDC09	73.77	51	93	74	5.86	2296	only 7 rounds
EUDC10	74.25	51	91	75	5.26	3008	only 8 rounds
EUDC11	74.85	51	89	75	4.52	3240	
EUDC12	74.41	54	89	75	4.51	3771	
EUDC13	75.12	58	88	75	4.52	3762	
EUDC14	75.17	59	88	75	3.98	3735	
EUDC15	75.03	54	89	75	4.43	3734	
EUDC16	75.66	57	91	76	4.39	3969	
EUDC17	75.97	55	88	76	4.03	3762	
EUDC18	76.01	55	87	76	3.93	3195	
HWS08	78.20	69	89	78	4.59	160	
HWS09	78.11	67	87	78	3.81	160	
HWS10	78.82	69	90	78.5	3.74	160	
HWS11	79.29	55	87	79	3.64	160	
HWS12	79.86	72	90	80	3.78	160	
HWS13	79.76	72	88	80	3.04	160	
HWS14	79.97	73	86	80	2.66	160	
HWS15	79.82	74	88	80	2.57	160	
HWS16	79.26	73	85	79	2.52	160	two judging panels
HWS17	80.74	74	88	81	2.68	160	two judging panels
HWS18	81.0	76	89	81	2.45	160	two judging panels
WUDC	75.12	50	93	75	4.69	69359	
EUDC	74.79	50	93	75	4.69	36586	
HWS	79.38	55	90	79.50	3.30	1760	

2.8.3.3 Gender, Team Gender Composition with and without scores

Table 2.13: Team Gender Composition (overall)

Team Gender Composition (overall)				
Gender	FF	FM	MM	Total
Female	145	281	0	426
Male	0	279	812	1091
Total	145	560	812	1517

Table 2.14: Team Gender Composition (with Scores)

Team Gender Composition (with score)				
Gender	FF	FM	MM	Total
Female	119	194	0	313
Male	0	190	481	671
Total	119	384	481	984

Table 2.15: Competition Group (overall)

Competition Group (overall)				
Gender	EUDC	HWS	WUDC	Total
Female	181	103	142	426
Male	407	318	366	1091
Total	588	421	508	1517

Table 2.16: Competition Group (with Score)

Competition Group (with score)				
Gender	EUDC	HWS	WUDC	Total
Female	137	93	83	313
Male	250	264	157	671
Total	387	357	240	984

Table 2.17: Score breakdown by gender group, in total and per competition ($N = 984$ speeches)

Competition	Gender	Observations	Max	Min	Mean	Median	STD
WUDC	Male	157	91	58	79.80	80	4.59
	Female	83	89	66	79.24	79	4.35
EUDC	Male	250	87	66	77.17	77	4.17
	Female	137	90	63	76.92	76	5.05
HWS	Male	264	86	75	80.10	80	2.44
	Female	93	86	73	79.77	80	2.37
TOTAL	Male	671	91	58	78.86	79.00	4.00
	Female	313	63	89	78.20	78.50	4.28
	All	984	91	58	78.65	79.00	4.01

Table 2.18: Total number of accepted Q&As :Male vs. Female Speakers

Q & As	Male	Female	Total
In-rounds	876	374	1250
Out-rounds	457	121	578
EUDC	426	185	611
WUDC	377	141	518
HWS	530	169	699
All	1333	495	1828

2.8.3.4 Descriptive Statistics: Speeches, Q&As, Control Variables

Table 2.19: Summary statistics of linguistic and psychometric variables of speeches (N = 1517)

VARIABLE	Description	Mean	Median	SD	Min	Max
Word Count	Number of words/ speech	1,495	1,502	191.4	661	2,196
Character Count	Number of characters/ speech	6,715	6,789	889.4	2,977	10,760
Sentence Count	Number of sentences/ speech	73.55	72	18.60	25	149
Words per Sentence	Average number of words/sentence	21.49	20.510	5.841	9.369	63.28
≥6-letter Words	Fraction of words ≥6-letter/speech (% point)	18.30	18.310	2.871	8.930	28.11
Argument Indicators	Fraction of claim- premise- flag indicators/speech	3.284	3.150	1.322	0.290	8.950
Fillers	Fraction of fillers and non-fluencies/speech (% point)	4.231	2.840	4.395	0.0600	41.89
Hedges	Fraction of hedge words and phrases/speech (% point)	3.297	3.210	1.116	0.480	9.300
Verb	Fraction of verbs/speech (% point)	16.62	16.59	1.623	10.93	23.41
Adjective	Fraction of adjective/speech (% point)	9.867	9.63	1.945	5.236	28.20
Adverb	Fraction of adverbs/speech (% point)	7.864	7.79	1.463	3.761	14.5
Noun	Fraction of nouns/speech (% point)	21.28	21.41	2.038	12.35	27.57
Personal Pronouns	Fraction of personal pronouns/speech (% point)	7.512	7.36	1.644	3.160	14.93
Certainty Indicators	LIWC Certain Word and Phrase list	2.894	1.55	0.787	0.740	7.160
Uncertainty Indicators	LIWC Certain Word and Phrase list	2.892	2.71	0.901	0.541	7.120
Analytic	LIWC score (0-100)	50.07	50.29	14.52	5.160	92.07
Authentic	LIWC score (0-100)	26.92	25.55	12.02	1	71.88
Tone	LIWC score (0-100)	46.87	46.55	22.81	1	98.63
BP Phrases	Fraction of BP phrases/speech (% point)	0.77	0.98	0.545	0	6.230
POI Reject	Fraction of question rejects/speech (% point)	0.091	0	0.176	0	1.740

Table 2.20: Linguistic Features: Male vs. Female ($N = 1517$, $N_F = 426$, $N_M = 1091$)

Group	Marker	mean _M	mean _F	t-statistics	p-value
Basic features	Word Count	1498.92	1484.78	1.299	0.195
	Character Count	6748.84	6629.16	2.389	0.0172**
	Sentence Count	75.12	69.51	5.484	0.000***
	Words per Sentence	21.08	22.54	-4.362	0.000***
	Character per Word	4.50	4.47	3.509	0.000***
	Words > 6 letters	18.45	17.90	3.249	0.001***
Linguistic markers	Argument Indicators	3.31	3.23	1.053	0.2928
	Fillers & Hesitations	3.97	4.89	-3.212	0.001***
	Hedges	3.209	3.522	-4.901	0.000***
LIWC	Certainty Indicators	1.62	1.60	0.365	0.715
	Uncertainty Indicators	2.83	2.82	0.178	0.859
	Analytic	51.54	46.33	6.317	0.000***
	Authentic	26.38	28.28	-2.739	0.006***
	Emotional Tone	46.79	47.07	-0.215	0.830
Sentiment	Positive	12.09	12.25	-1.068	0.286
	Negative	8.34	7.85	2.570	0.0104**
	Neutral	79.62	79.82	-1.094	0.274
Parts of Speech	Noun	21.46	20.82	5.348	0.000***
	Verb	16.56	16.77	-2.210	0.027**
	Adjective	9.97	9.60	3.026	0.003***
	Adverb	7.77	8.10	-3.902	0.000***
	Personal Pronoun	7.39	7.83	-4.581	0.000***
BP-specific strategies	BP Words& Phrases	0.78	0.72	1.944	0.052*
	POI Rejects	0.09	0.09	0.3392	0.7346

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

Table 2.21: Summary statistics of linguistic and psychometric variables of accepted questions during speeches ($N = 1828$)

VARIABLE	Description	Mean	Median	SD	Min	Max
Word Count	Number of words/ speech	37.68	36.00	18.25	2	287
Fillers	Fillers and non-fluencies/speech (% point)	1.29	0.01	4.40	0.06	25.53
Hedges	Hedge words and phrases/speech (% point)	2.26	1.82	2.24	0.00	16.67
Negation	Contradiction words/speech (% point)	2.21	1.92	2.57	0.00	20.00
Certainty Indicators	LIWC Certain Word and Phrase list	1.57	1.55	0.00	0.00	18.75
Uncertainty Indicators	LIWC Certain Word and Phrase list	3.19	2.63	3.51	0.00	50.00
Analytic	LIWC score (0-100)	49.22	50.33	30.74	1.00	99.00
Authentic	LIWC score (0-100)	28.06	17.46	28.46	1.00	99.00
Tone	LIWC score (0-100)	46.87	46.55	22.81	1	98.63
Impersonal Pronouns	Non-person reference (% point)	7.89	4.35	5.09	0.00	10.71
First-person Pronouns	I, we, us, myself, ourselves, etc (% point)	1.72	0.00	2.96	0.00	23.53
Second and third-person Pronouns	You, heshe, they, yours, etc (% point)	5.51	5.00	4.19	0.00	33.33

Table 2.22: Summary statistics of linguistic and psychometric variables of answers during speeches (N = 1828)

VARIABLE	Description	Mean	Median	SD	Min	Max
Word Count	Number of words/ speech	93.47	80.00	61.10	1	489
Fillers	Fillers and non-fluencies/speech (% point)	2.55	2.840	4.59	0.00	50.00
Hedges	Hedge words and phrases/speech (% point)	2.35	3.210	1.116	0.480	9.300
Negation	Contradiction words/speech (% point)	2.60	3.210	1.116	0.480	9.300
Certainty Indicators	LIWC Certain Word and Phrase list	1.85	1.55	0.787	0.740	7.160
Uncertainty Indicators	LIWC Certain Word and Phrase list	3.00	2.71	0.901	0.541	7.120
Analytic	LIWC score (0-100)	40.97	50.29	14.52	5.160	92.07
Authentic	LIWC score (0-100)	34.93	25.55	12.02	1	71.88
Tone	LIWC score (0-100)	50.10	46.55	22.81	1	98.63
Impersonal Pronouns	Non-person reference (% point)	9.60	16.59	1.623	10.93	23.41
First-person Pronouns	I, we, us, myself, ourselves, etc (% point)	4.22	9.63	1.945	5.236	28.20
Second and third-person Pronouns	You, heshe, they, yours, etc (% point)	4.58	7.79	1.463	3.761	14.5

Table 2.23: Linguistic features for accepted questions during speeches: Male vs. Female (N = 1828, $N_M = 1333$; $N_F = 495$)

Group	Marker	mean _M	mean _F	t-statistics	p-value
Basic features	Word Count	37.461	38.251	-0.740	0.460
Linguistic	Hedges	2.195	2.429	-1.952	0.051*
	Fillers and Hesitations	1.308	1.234	0.495	0.621
	Negation	2.242	2.112	0.967	0.334
LIWC	Uncertainty Indicators	3.131	3.359	-1.226	0.221
	Certainty Indicators	1.535	1.653	-0.935	0.350
	Analytic	49.624	48.125	0.911	0.362
	Authentic	27.991	28.243	-0.164	0.869
	Emotional Tone	46.375	46.137	0.119	0.906
Sentiment	Positive	3.181	3.326	-0.816	0.414
	Negative	2.2624	2.298	-0.208	0.835
Parts of Speech	Impersonal Pronouns	7.830	8.068	-0.874	0.382
	First-person Pronouns	1.669	1.867	-1.223	0.222
	Second and Third-person Pronouns	5.535	5.462	0.329	0.742

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

Table 2.24: Linguistic features for answers during speeches: Male vs. Female (N = 1828, $N_M = 1333$; $N_F = 495$)

Group	Marker	mean _M	mean _F	t-statistics	p-value
Basic features	Word Count	90.445	101.634	-3.347	0.000***
Linguistic	Hedges	2.288	2.519	-1.853	0.064*
	Fillers and Hesitations	2.443	2.838	-1.451	0.147
	Negation	2.705	2.327	3.533	0.000***
LIWC	Uncertainty Indicators	3.047	2.855	1.535	0.125
	Certainty Indicators	1.856	1.849	0.048	0.962
	Analytic	41.332	39.987	1.002	0.317
	Authentic	34.417	36.316	-1.272	0.204
	Emotional Tone	49.442	51.859	-1.348	0.178
Sentiment	Positive	3.266	3.387	-0.730	0.466
	Negative	1.911	1.752	1.511	0.131
Parts of Speech	Impersonal Pronouns	9.643	9.486	0.792	0.428
	First-person Pronouns	4.214	4.232	-0.106	0.915
	Second and Third Person Pronouns	4.431	4.979	-3.098	0.002***

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

2.8.3.5 Single regression of room control variables on speech scores

Table 2.25: Single regression of room control variables on speech score

	Dependent Variable: Score						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Female	-0.160** (0.07)						
Non-native		-0.921*** (0.12)					
Top 50 Institutions			0.821*** (0.10)				
Female-dominated Room				-0.398** (0.17)			
EUDC					-0.748*** (0.20)		
HWS					0.046 (0.17)		
1 st Speaking Position						-0.242*** (0.09)	
Environment Motions							-0.581*** (0.22)
Philosophy Motions							0.823*** (0.27)
Feminism Motions							1.566*** (0.42)
Military Motions							0.867*** (0.22)
No. of observations	984	984	984	984	984	984	984

Score is standardized. For speaking position (8 dummies) and motion type (17 dummies), only statistically significant variables are reported. Robust clustered standard errors at debate level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

2.8.3.6 How do different speech elements matter for evaluations?

Table 2.26: Linear & Fixed Effects Regression of linguistic features (N = 984 speeches)

	Dependent Variable : Score					
	(1)	(2)	(3)	(4)	(5)	(6)
Basic Features						
Word Count	0.366*** (0.06)	0.243*** (0.04)	0.242*** (0.04)	0.242*** (0.04)	0.242*** (0.04)	0.204*** (0.03)
Words per Sentence	-0.029 (0.04)	-0.076** (0.04)	-0.076** (0.04)	-0.081** (0.04)	-0.082** (0.03)	-0.101*** (0.03)
Complex Words	0.217*** (0.05)	0.156*** (0.04)	0.156*** (0.04)	0.156*** (0.04)	0.156*** (0.04)	0.112*** (0.03)
Argument Indicators	0.010 (0.04)	0.010 (0.03)	0.010 (0.03)	0.008 (0.03)	0.008 (0.03)	-0.000 (0.02)
Fillers	-0.224*** (0.06)	-0.149*** (0.05)	-0.145** (0.06)	-0.144** (0.06)	-0.145** (0.06)	-0.071 (0.05)
Hedges	-0.111*** (0.03)	-0.073*** (0.03)	-0.072*** (0.03)	-0.069** (0.03)	-0.069** (0.03)	-0.056** (0.02)
BP Words & Phrases	0.028 (0.03)	0.027 (0.03)	0.027 (0.03)	0.026 (0.03)	0.026 (0.03)	0.040* (0.02)
POI Rejects	0.028 (0.03)	0.010 (0.03)	0.009 (0.03)	0.010 (0.03)	0.010 (0.03)	0.001 (0.02)
R^2	0.313	0.451	0.452	0.454	0.454	0.495
POS						
Noun	0.336*** (0.06)	0.147*** (0.05)	0.146*** (0.05)	0.150*** (0.05)	0.149*** (0.05)	0.100*** (0.03)
Verb	0.141*** (0.04)	0.048 (0.03)	0.048 (0.04)	0.048 (0.03)	0.048 (0.03)	0.039 (0.03)
Adjective	-0.079 (0.06)	-0.052 (0.04)	-0.053 (0.04)	-0.058 (0.04)	-0.058 (0.04)	-0.016 (0.04)
Adverb	0.020 (0.05)	0.072* (0.04)	0.070* (0.04)	0.068* (0.04)	0.068* (0.04)	0.027 (0.02)
Personal Pronouns	-0.084** (0.04)	-0.017 (0.04)	-0.020 (0.04)	-0.022 (0.04)	-0.022 (0.04)	0.004 (0.03)
R^2	0.144	0.384	0.387	0.390	0.391	0.326
LIWC Psychometric Features						
Certainty Indicators	0.154*** (0.04)	0.043 (0.03)	0.038 (0.03)	0.037 (0.03)	0.036 (0.03)	0.041* (0.02)
Uncertainty Indicators	0.052 (0.04)	0.035 (0.03)	0.035 (0.03)	0.036 (0.03)	0.035 (0.03)	0.039* (0.02)
Analytic	0.115*** (0.04)	-0.020 (0.04)	-0.018 (0.04)	-0.019 (0.04)	-0.018 (0.04)	-0.002 (0.03)
Authentic	0.183*** (0.04)	0.059* (0.03)	0.056* (0.03)	0.056* (0.03)	0.055* (0.03)	0.031 (0.03)
Tone	0.014 (0.05)	-0.011 (0.04)	-0.010 (0.04)	-0.009 (0.04)	-0.008 (0.04)	-0.046 (0.03)
R^2	0.076	0.367	0.370	0.372	0.372	0.047
Room controls		✓	✓	✓	✓	
Chair Judge Gender			✓		✓	
Panel Gender				✓	✓	
Debate fixed effect						✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. R^2 of model (6) is $R^2_{between}$. Robust clustered standard errors at debate level are in parentheses.

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

2.8.3.7 Extensions

Q& A during speeches: Male vs. Female (Question part)

Table 2.27: Average Marginal Effects from logistic regression of linguistic features in questions ($N_F = 595$, $N_M = 1333$)

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Basic Features					
Word Count	0.010 (0.01)	0.009 (0.01)	0.009 (0.01)	0.009 (0.01)	0.009 (0.01)
Hedges	0.020** (0.01)	0.019** (0.01)	0.019** (0.01)	0.019** (0.01)	0.020** (0.01)
Fillers	-0.004 (0.01)	-0.018 (0.01)	-0.019* (0.01)	-0.017 (0.01)	-0.018 (0.01)
Negation	-0.010 (0.01)	-0.012 (0.01)	-0.012 (0.01)	-0.012 (0.01)	-0.012 (0.01)
POS					
Impersonal Pronoun	0.009 (0.01)	0.014 (0.01)	0.014 (0.01)	0.014 (0.01)	0.015 (0.01)
First-person Pronouns	0.013 (0.01)	0.009 (0.01)	0.009 (0.01)	0.009 (0.01)	0.009 (0.01)
Second-person Pronouns	-0.000 (0.01)	-0.001 (0.01)	-0.001 (0.01)	-0.002 (0.01)	-0.003 (0.01)
LIWC Psychometric Features					
Uncertainty Indicators	0.012 (0.01)	0.014 (0.01)	0.014 (0.01)	0.014 (0.01)	0.014 (0.01)
Certainty Indicators	0.011 (0.01)	0.007 (0.01)	0.008 (0.01)	0.006 (0.01)	0.007 (0.01)
Analytic	-0.008 (0.01)	-0.001 (0.01)	-0.001 (0.01)	-0.001 (0.01)	-0.001 (0.01)
Authentic	0.002 (0.01)	0.004 (0.01)	0.005 (0.01)	0.005 (0.01)	0.005 (0.01)
Emotional Tone	-0.002 (0.01)	0.001 (0.01)	0.001 (0.01)	0.002 (0.01)	0.003 (0.01)
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓
Observations	1828	1828	1828	1828	1828

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Q& A during speeches: Male vs. Female (Answer part)Table 2.28: Average Marginal Effects from logistic regression of linguistic features in answers ($N_F = 595$, $N_M = 1333$)

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Basic Features					
Word Count	0.030*** (0.01)	0.026*** (0.01)	0.026*** (0.01)	0.026*** (0.01)	0.026*** (0.01)
Hedges	0.022** (0.01)	0.021** (0.01)	0.021** (0.01)	0.020** (0.01)	0.021** (0.01)
Fillers	0.016 (0.01)	-0.002 (0.01)	-0.003 (0.01)	-0.002 (0.01)	-0.003 (0.01)
Negation	-0.032*** (0.01)	-0.034*** (0.01)	-0.034*** (0.01)	-0.034*** (0.01)	-0.033*** (0.01)
POS					
Impersonal Pronoun	-0.003 (0.01)	0.008 (0.01)	0.008 (0.01)	0.008 (0.01)	0.008 (0.01)
First-person Pronouns	0.010 (0.01)	0.004 (0.01)	0.004 (0.01)	0.004 (0.01)	0.002 (0.01)
Second and Third-person Pronouns	0.033*** (0.01)	0.026** (0.01)	0.026** (0.01)	0.026** (0.01)	0.027** (0.01)
LIWC Psychometric Features					
Uncertainty Indicators	-0.017 (0.01)	-0.015 (0.01)	-0.015 (0.01)	-0.015 (0.01)	-0.017 (0.01)
Certainty Indicators	-0.002 (0.01)	0.000 (0.01)	0.000 (0.01)	-0.000 (0.01)	-0.003 (0.01)
Analytic	-0.012 (0.01)	-0.004 (0.01)	-0.003 (0.01)	-0.004 (0.01)	-0.002 (0.01)
Authentic	0.012 (0.01)	0.008 (0.01)	0.007 (0.01)	0.008 (0.01)	0.007 (0.01)
Emotional Tone	0.014 (0.01)	0.012 (0.01)	0.012 (0.01)	0.012 (0.01)	0.014 (0.01)
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓
Observations	1828	1828	1828	1828	1828

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Female Chair Judge and Speech Evaluation Patterns

Table 2.29: Linear & Fixed Effects Regression of Basic Features (interact with Chair Judge Gender) (N = 984, $N_{femalechair} = 378$, $N_{malechair} = 606$)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female Chair Judge	-0.026 (0.10)	0.043 (0.09)	-0.036 (0.11)	
Word Count	0.383*** (0.08)	0.259*** (0.06)	0.256*** (0.06)	0.175*** (0.04)
Female Chair Judge × Word Count	-0.044 (0.10)	-0.040 (0.08)	-0.033 (0.08)	0.065 (0.07)
Words per Sentence	-0.031 (0.06)	-0.088* (0.05)	-0.093* (0.05)	-0.117** (0.05)
Female Chair Judge × Words per Sentence	0.003 (0.08)	0.017 (0.06)	0.013 (0.06)	0.021 (0.06)
Complex Words	0.254*** (0.07)	0.202*** (0.06)	0.198*** (0.06)	0.090** (0.04)
Female Chair Judge × Complex Words	-0.082 (0.09)	-0.109 (0.09)	-0.096 (0.08)	0.035 (0.06)
Argument Indicators	0.037 (0.05)	0.025 (0.04)	0.024 (0.04)	-0.002 (0.03)
Female Chair Judge × Argument Indicators	-0.050 (0.07)	-0.028 (0.06)	-0.030 (0.06)	-0.001 (0.04)
Fillers	-0.156** (0.07)	-0.096 (0.06)	-0.093 (0.06)	-0.062 (0.06)
Female Chair Judge × Fillers	-0.168 (0.11)	-0.129 (0.09)	-0.137 (0.09)	-0.031 (0.11)
Hedges	-0.131*** (0.04)	-0.088*** (0.03)	-0.083*** (0.03)	-0.054* (0.03)
Female Chair Judge × Hedges	0.042 (0.06)	0.039 (0.05)	0.034 (0.05)	0.002 (0.05)
BP Words & Phrases	0.000 (0.05)	-0.013 (0.03)	-0.016 (0.04)	0.019 (0.03)
Female Chair Judge × BP Words & Phrases	0.074 (0.06)	0.102** (0.05)	0.109** (0.05)	0.052 (0.04)
POI Reject	0.049 (0.05)	0.040 (0.05)	0.042 (0.05)	-0.014 (0.03)
Female Chair Judge × POI Reject	-0.038 (0.06)	-0.061 (0.06)	-0.065 (0.06)	0.034 (0.04)
R^2	0.320	0.458	0.461	0.444
Room controls		✓	✓	
Panel Gender			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types; (vii) Speaker's gender. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.30: Linear & Fixed Effects Regression of Parts of Speech (interact with Chair Judge Gender) (N = 984, $N_{femalechair} = 378$, $N_{malechair} = 606$)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female Chair Judge	0.089 (0.12)	0.139 (0.10)	0.057 (0.11)	
Noun	0.329*** (0.09)	0.150** (0.07)	0.149** (0.07)	0.085** (0.04)
Female Chair Judge × Noun	0.027 (0.12)	-0.012 (0.10)	-0.000 (0.10)	0.045 (0.07)
Verb	0.144*** (0.05)	0.065 (0.04)	0.064 (0.04)	0.034 (0.03)
Female Chair Judge × Verb	-0.007 (0.09)	-0.048 (0.08)	-0.042 (0.08)	0.016 (0.07)
Adjective	-0.028 (0.09)	-0.012 (0.06)	-0.014 (0.06)	-0.005 (0.04)
Female Chair Judge × Adjective	-0.110 (0.12)	-0.099 (0.08)	-0.101 (0.08)	-0.030 (0.07)
Adverb	0.004 (0.08)	0.049 (0.05)	0.054 (0.05)	0.008 (0.03)
Female Chair Judge × Adverb	0.061 (0.10)	0.064 (0.07)	0.049 (0.07)	0.062 (0.05)
Personal Pronouns	-0.041 (0.06)	0.003 (0.05)	0.001 (0.05)	0.039 (0.04)
Female Chair Judge × Personal Pronouns	-0.110 (0.08)	-0.067 (0.07)	-0.064 (0.07)	-0.092 (0.06)
R^2	0.152	0.391	0.394	0.279
Observations	984	984	984	984
Room controls		✓	✓	
Panel Gender			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types; (vii) Speaker's gender. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.31: Linear & Fixed Effects Regression of LIWC psychometric features (interact with Chair Judge Gender) (N = 984, $N_{femalechair} = 378$, $N_{malechair} = 606$)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female Chair Judge	0.010 (0.13)	0.120 (0.10)	0.057 (0.12)	
Certainty Indicators	0.122*** (0.04)	0.041 (0.04)	0.035 (0.04)	0.043 (0.03)
Female Chair Judge × Certainty Indicators	0.066 (0.08)	-0.004 (0.06)	0.005 (0.06)	-0.005 (0.04)
Uncertainty Indicators	0.038 (0.05)	0.028 (0.03)	0.029 (0.03)	0.027 (0.03)
Female Chair Judge × Uncertainty Indicators	0.035 (0.07)	0.017 (0.05)	0.014 (0.05)	0.028 (0.05)
Analytic	0.064 (0.05)	-0.046 (0.05)	-0.045 (0.05)	-0.039 (0.04)
Female Chair Judge × Analytic	0.125* (0.07)	0.069 (0.06)	0.066 (0.06)	0.089 (0.06)
Authentic	0.231*** (0.05)	0.065 (0.04)	0.062 (0.04)	0.038 (0.04)
Female Chair Judge × Authentic	-0.124 (0.09)	-0.022 (0.06)	-0.016 (0.06)	-0.021 (0.06)
Emotional Tone	0.054 (0.06)	0.027 (0.05)	0.028 (0.05)	-0.022 (0.04)
Female Chair Judge × Emotional Tone	-0.113 (0.10)	-0.103 (0.07)	-0.102 (0.07)	-0.060 (0.07)
R^2	0.087	0.373	0.375	0.058
Observations	984	984	984	984
Room controls		✓	✓	
Panel Gender			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types; (vii) Speaker's gender. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Female-dominated Judge Panel and Speech Evaluation Patterns

Table 2.32: Linear & Fixed Effects Regression of Basic Features (interact with Female-dominated Judge Panel) (N = 984, $N_{female_dominated} = 366$, $N_{male_dominated} = 618$)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female-dominated Judge Panel	0.076 (0.10)	0.125 (0.10)	0.128 (0.11)	
Word Count	0.415*** (0.07)	0.288*** (0.06)	0.288*** (0.06)	0.205*** (0.05)
Female-dominated Judge Panel × Word Count	-0.141 (0.09)	-0.119 (0.08)	-0.120 (0.08)	-0.002 (0.06)
Words per Sentence	0.016 (0.06)	-0.037 (0.05)	-0.037 (0.05)	-0.090* (0.05)
Female-dominated Judge Panel × Words per Sentence	-0.083 (0.07)	-0.075 (0.07)	-0.075 (0.07)	-0.024 (0.06)
Complex Words	0.262*** (0.06)	0.197*** (0.06)	0.197*** (0.06)	0.116*** (0.04)
Female-dominated Judge Panel × Complex Words	-0.149 (0.09)	-0.137 (0.09)	-0.137 (0.09)	-0.021 (0.06)
Argument Indicators	0.013 (0.05)	0.023 (0.04)	0.023 (0.04)	0.003 (0.03)
Female-dominated Judge Panel × Argument Indicators	0.004 (0.06)	-0.026 (0.06)	-0.026 (0.06)	-0.003 (0.04)
Fillers	-0.176** (0.07)	-0.097 (0.07)	-0.097 (0.07)	-0.061 (0.07)
Female-dominated Judge Panel × Fillers	-0.127 (0.11)	-0.112 (0.11)	-0.112 (0.11)	-0.027 (0.10)
Hedges	-0.120*** (0.04)	-0.063* (0.03)	-0.063* (0.03)	-0.061** (0.03)
Female-dominated Judge Panel × Hedges	0.015 (0.06)	-0.019 (0.05)	-0.018 (0.05)	0.012 (0.04)
BP Words & Phrase	-0.003 (0.05)	-0.029 (0.04)	-0.029 (0.04)	0.019 (0.03)
Female-dominated Judge Panel × BP Words & Phrase	0.089 (0.05)	0.138*** (0.05)	0.137*** (0.05)	0.054 (0.04)
POI Rejects	0.034 (0.04)	0.016 (0.05)	0.016 (0.05)	-0.004 (0.03)
Female-dominated Judge Panel × POI Rejects	-0.005 (0.06)	-0.017 (0.06)	-0.017 (0.06)	0.017 (0.04)
R^2	0.324	0.464	0.464	0.462
Observations	984	984	984	984
Room controls		✓	✓	
Female Chair Judge			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types; (vii) Speaker's gender. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.33: Linear & Fixed Effects Regression of Parts of Speech (interact with Female-dominated Judge Panel) ($N = 984$, $N_{female_dominated} = 366$, $N_{male_dominated} = 618$)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female-dominated Judge Panel	0.188*	0.190*	0.162	
	(0.11)	(0.10)	(0.12)	
Noun	0.351***	0.174**	0.173**	0.094**
	(0.09)	(0.07)	(0.07)	(0.05)
Female-dominated Judge Panel \times Noun	-0.044	-0.077	-0.075	0.036
	(0.11)	(0.10)	(0.10)	(0.06)
Verb	0.156***	0.053	0.053	0.013
	(0.06)	(0.05)	(0.05)	(0.04)
Female-dominated Judge Panel \times Verb	-0.054	-0.026	-0.025	0.074
	(0.08)	(0.07)	(0.07)	(0.06)
Adjective	-0.052	-0.040	-0.039	0.001
	(0.08)	(0.06)	(0.06)	(0.04)
Female-dominated Judge Panel \times Adjective	-0.085	-0.056	-0.057	-0.045
	(0.12)	(0.09)	(0.09)	(0.07)
Adverb	0.024	0.070	0.071	0.021
	(0.07)	(0.05)	(0.05)	(0.03)
Female-dominated Judge Panel \times Adverb	-0.010	-0.006	-0.010	0.034
	(0.09)	(0.07)	(0.07)	(0.05)
Personal Pronouns	-0.077	-0.015	-0.015	0.051
	(0.06)	(0.05)	(0.05)	(0.04)
Female-dominated Judge Panel \times Personal Pronouns	-0.024	-0.017	-0.018	-0.122**
	(0.08)	(0.07)	(0.07)	(0.06)
R^2	0.153	0.391	0.392	0.182
Observations	984	984	984	984
Room controls		✓	✓	
Female Chair Judge			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types; (vii) Speaker's gender. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (4) is $R_{between}^2$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.34: Linear & Fixed Effects Regression of LIWC Psychometric Features (interact with Female-dominated Judge Panel) (N = 984, $N_{female_dominated} = 366$, $N_{male_dominated} = 618$)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female-dominated Judge Panel	0.057 (0.13)	0.158 (0.11)	0.128 (0.13)	
Certainty Indicators	0.101* (0.05)	0.022 (0.04)	0.020 (0.04)	0.022 (0.03)
Female-dominated Judge Panel × Certainty Indicators	0.107 (0.08)	0.035 (0.06)	0.037 (0.06)	0.045 (0.04)
Uncertainty Indicators	0.027 (0.05)	0.032 (0.03)	0.032 (0.03)	0.025 (0.03)
Female-dominated Judge Panel × Uncertainty Indicators	0.064 (0.07)	0.010 (0.05)	0.007 (0.05)	0.046 (0.05)
Analytic	0.099* (0.05)	-0.020 (0.05)	-0.019 (0.05)	-0.034 (0.04)
Female-dominated Judge Panel × Analytic	0.036 (0.07)	0.004 (0.06)	0.003 (0.06)	0.083 (0.06)
Authentic	0.246*** (0.05)	0.088** (0.04)	0.087** (0.04)	0.021 (0.04)
Female-dominated Judge Panel × Authentic	-0.169* (0.09)	-0.083 (0.07)	-0.081 (0.07)	0.028 (0.06)
Emotional Tone	0.062 (0.07)	-0.011 (0.05)	-0.011 (0.05)	-0.026 (0.04)
Female-dominated Judge Panel × Emotional Tone	-0.124 (0.09)	0.010 (0.08)	0.010 (0.08)	-0.062 (0.07)
R^2	0.091	0.373	0.374	0.034
Observations	984	984	984	984
Room controls		✓	✓	
Female Chair Judge			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types; (vii) Speaker's gender. Robust clustered standard errors at debate level are in parentheses.

R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Female Judges on Female Speakers: Chair & Panel

Table 2.35: Linear & Fixed Effects Regression of Score on Female Chair & Female-dominated Judge Panel

	Dependent Variable: Score					
	(C1)	(C2)	(C3)	(P1)	(P2)	(P3)
Female	-0.057 (0.09)	0.105 (0.07)	0.082 (0.07)	-0.059 (0.10)	0.127* (0.08)	0.087 (0.07)
Female Chair Judge	0.141 (0.14)	0.212** (0.10)				
Female × Female Chair Judge	-0.263* (0.14)	-0.215* (0.12)	-0.298*** (0.10)			
Female-dominated Judge Panel				0.244* (0.13)	0.265** (0.10)	
Female × Female-dominated Judge Panel				-0.283** (0.14)	-0.280** (0.12)	-0.308*** (0.10)
R^2	0.010	0.366	0.000	0.015	0.369	0.002
Observations	984	984	984	984	984	984
Room controls		✓			✓	
Debate fixed effect			✓			✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Robust clustered standard errors at debate level are in parentheses. R^2 of model (C3) and (P3) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Female Chair Judges on Female Speakers (control for speech elements)

Table 2.36: Linear & Fixed Effects Regression of Score on Female Speakers and Chair Judges (controlling for Basic Speech features)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female	0.037 (0.10)	0.143* (0.07)	0.139* (0.07)	0.127* (0.07)
Female Chair Judge	0.069 (0.12)	0.128 (0.10)	0.060 (0.11)	
Female × Female Chair Judge	-0.190 (0.13)	-0.181 (0.11)	-0.176 (0.11)	-0.270*** (0.10)
R^2	0.196	0.412	0.415	0.219
Observations	984	984	984	984
Basic Features controls	✓	✓	✓	✓
Room controls		✓	✓	
Female-dominated Panel			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Basic Features are: (i) Words per Sentence; (ii) Complex Words; (iii) Argument Indicators; (iv) Fillers; (v) Hedges; (vi) BP Words & Phrases; and (vii) POI Rejects. Robust clustered standard errors at debate level are in parentheses. R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.37: Linear & Fixed Effects Regression of Score on Female Speakers and Chair Judges (controlling for Parts of Speech)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female	0.048 (0.09)	0.112 (0.07)	0.107 (0.07)	0.086 (0.07)
Female Chair Judge	0.165 (0.12)	0.209** (0.10)	0.124 (0.11)	
Female × Female Chair Judge	-0.237* (0.13)	-0.203* (0.12)	-0.195 (0.12)	-0.287*** (0.10)
R^2	0.149	0.390	0.393	0.173
Observations	984	984	984	984
Parts of Speech controls	✓	✓	✓	✓
Room controls		✓	✓	
Female-dominated Panel			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Parts of Speech Controls are: (i) Noun; (ii) Verbs; (iii) Adjective; (iv) Adverbs; (v) Personal Pronouns. Robust clustered standard error at debate level are in parentheses. R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.38: Linear & Fixed Effects Regression of Score on Female Speakers and Chair Judges (controlling for LIWC features)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female	-0.065 (0.10)	0.080 (0.07)	0.076 (0.07)	0.074 (0.06)
Female Chair Judge	0.098 (0.13)	0.189* (0.10)	0.125 (0.12)	
Female × Female Chair Judge	-0.230 (0.14)	-0.199 (0.12)	-0.194 (0.12)	-0.289*** (0.10)
R^2	0.084	0.372	0.374	0.026
Observations	984	984	984	984
LIWC controls	✓	✓	✓	✓
Room controls		✓	✓	
Female-dominated Panel			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. LIWC Controls are: (i) Certainty Words; (ii) Uncertainty Words; (iii) Authentic; (iv) Analytic; (v) Emotional Tone. Robust clustered standard error at debate level are in parentheses. R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Female-dominated Panels on Female Speakers (control for speech elements)

Table 2.39: Linear & Fixed Effects Regression of Score on Female Speakers and Female-dominated Judge Panels (controlling for Basic Speech features)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female	0.175** (0.08)	0.175** (0.08)	0.174** (0.08)	0.142* (0.07)
Female-dominated Judge Panel	0.223** (0.10)	0.223** (0.10)	0.223* (0.12)	
Female × Female-dominated Judge Panel	-0.262** (0.12)	-0.262** (0.12)	-0.262** (0.12)	-0.304*** (0.10)
R^2	0.416	0.416	0.416	0.192
Observations	984	984	984	984
Basic Features controls	✓	✓	✓	✓
Room controls		✓	✓	
Female Chair Judge			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Basic Features are: (i) Words per Sentence; (ii) Complex Words; (iii) Argument Indicators; (iv) Fillers; (v) Hedges; (vi) BP Words & Phrases; and (vii) POI Rejects. Robust clustered standard errors at debate level are in parentheses. R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.40: Linear & Fixed Effects Regression of Score on Female Speakers and Female-dominated Panels (controlling for Parts of Speech)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female	0.050 (0.10)	0.135* (0.08)	0.136* (0.08)	0.098 (0.07)
Female-dominated Judge Panel	0.281** (0.11)	0.285*** (0.10)	0.255** (0.12)	
Female × Female-dominated Judge Panel	-0.269** (0.13)	-0.267** (0.12)	-0.266** (0.12)	-0.311*** (0.10)
R^2	0.157	0.394	0.394	0.133
Observations	984	984	984	984
Parts of Speech controls	✓	✓	✓	✓
Room controls		✓	✓	
Female Chair Judge			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Parts of Speech Controls are: (i) Noun; (ii) Verbs; (iii) Adjective; (iv) Adverbs; (v) Personal Pronouns. Robust clustered standard error at debate level are in parentheses. R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2.41: Linear & Fixed Effects Regression of Score on Female Speakers and Female-dominated Judge Panels (controlling for LIWC features)

	Dependent Variable: Score			
	(1)	(2)	(3)	(4)
Female	-0.039 (0.10)	0.109 (0.08)	0.110 (0.08)	0.085 (0.07)
Female-dominated Judge Panel	0.184 (0.13)	0.251** (0.10)	0.220* (0.12)	
Female × Female-dominated Judge Panel	-0.312** (0.14)	-0.281** (0.13)	-0.281** (0.13)	-0.316*** (0.10)
R^2	0.087	0.375	0.376	0.020
Observations	984	984	984	984
LIWC controls	✓	✓	✓	✓
Room controls		✓	✓	
Female Chair Judge			✓	
Debate fixed effect				✓

All variables are standardized. Room Controls are: (i) Language Skill Status; (ii) Speaking position; (iii) Competition Type & Year; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. LIWC Controls are: (i) Certainty Words; (ii) Uncertainty Words; (iii) Authentic; (iv) Analytic; (v) Emotional Tone. Robust clustered standard error at debate level are in parentheses. R^2 of model (4) is $R^2_{between}$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

2.8.3.8 Robustness Checks

Table 2.42: Average Marginal Effects from Linear Probability regression of linguistic features

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Basic Features					
Word Count	-0.026** (0.01)	-0.019 (0.01)	-0.018 (0.01)	-0.017 (0.01)	-0.016 (0.01)
Words per Sentence	0.056*** (0.02)	0.047*** (0.01)	0.047*** (0.01)	0.051*** (0.02)	0.051*** (0.02)
Complex Words	-0.028** (0.01)	-0.015 (0.01)	-0.015 (0.01)	-0.018 (0.01)	-0.018 (0.01)
Argument Words	-0.008 (0.01)	-0.003 (0.01)	-0.003 (0.01)	-0.002 (0.01)	-0.002 (0.01)
Fillers	0.028** (0.01)	0.014 (0.01)	0.014 (0.01)	0.013 (0.01)	0.012 (0.01)
Hedges	0.054*** (0.01)	0.040*** (0.01)	0.040*** (0.01)	0.034** (0.01)	0.034** (0.01)
BP Words and Phrases	-0.019 (0.01)	-0.019 (0.01)	-0.020 (0.01)	-0.029** (0.01)	-0.029** (0.01)
POI Rejects	0.005 (0.01)	0.012 (0.01)	0.012 (0.01)	0.009 (0.01)	0.009 (0.01)
POS					
Noun	-0.072*** (0.01)	-0.047*** (0.01)	-0.047*** (0.01)	-0.056*** (0.02)	-0.056*** (0.02)
Verb	-0.021 (0.01)	-0.008 (0.01)	-0.008 (0.01)	-0.008 (0.01)	-0.008 (0.01)
Adjectives	-0.055*** (0.02)	-0.053*** (0.01)	-0.053*** (0.01)	-0.054*** (0.02)	-0.054*** (0.02)
Adverbs	0.020 (0.01)	0.013 (0.01)	0.013 (0.01)	0.010 (0.01)	0.010 (0.01)
Personal Pronouns	0.026* (0.01)	0.027** (0.01)	0.028** (0.01)	0.028* (0.01)	0.028* (0.01)
LIWC Psychometric Features					
Authentic	0.036*** (0.01)	0.031*** (0.01)	0.031*** (0.01)	0.041*** (0.01)	0.042*** (0.01)
Analytic	-0.078*** (0.01)	-0.074*** (0.01)	-0.074*** (0.01)	-0.080*** (0.01)	-0.080*** (0.01)
Emotional Tone	0.005 (0.01)	0.008 (0.01)	0.008 (0.01)	0.007 (0.01)	0.007 (0.01)
Certain Words	-0.008 (0.01)	0.003 (0.01)	0.003 (0.01)	-0.001 (0.01)	-0.001 (0.01)
Uncertain Words	-0.019* (0.01)	-0.026** (0.01)	-0.026** (0.01)	-0.035*** (0.01)	-0.035*** (0.01)
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓
Observations	1517	1517	1517	1314	1314

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Model (4) and (5) excludes 203 knock-out round speeches without judge panel information. Robust clustered standard errors at debate level are in parentheses.

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

Table 2.43: Average Marginal Effects from Probit regression of linguistic features

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Basic Features					
Word Count	-0.025** (0.01)	-0.018 (0.01)	-0.018 (0.01)	-0.016 (0.01)	-0.015 (0.01)
Words per Sentence	0.053*** (0.01)	0.044*** (0.01)	0.044*** (0.01)	0.047*** (0.01)	0.048*** (0.01)
Complex Words	-0.027** (0.01)	-0.014 (0.01)	-0.014 (0.01)	-0.017 (0.01)	-0.017 (0.01)
Argument Words	-0.008 (0.01)	-0.003 (0.01)	-0.003 (0.01)	-0.002 (0.01)	-0.002 (0.01)
Fillers	0.025** (0.01)	0.013 (0.01)	0.013 (0.01)	0.011 (0.01)	0.011 (0.01)
Hedges	0.054*** (0.01)	0.039*** (0.01)	0.039*** (0.01)	0.033*** (0.01)	0.033*** (0.01)
BP Words and Phrases	-0.018 (0.01)	-0.019 (0.01)	-0.019 (0.01)	-0.030** (0.01)	-0.030** (0.01)
POI Rejects	0.006 (0.01)	0.011 (0.01)	0.012 (0.01)	0.008 (0.01)	0.008 (0.01)
POS					
Noun	-0.070*** (0.01)	-0.046*** (0.01)	-0.046*** (0.01)	-0.054*** (0.01)	-0.054*** (0.01)
Verb	-0.020 (0.01)	-0.007 (0.01)	-0.007 (0.01)	-0.007 (0.01)	-0.006 (0.01)
Adjectives	-0.053*** (0.02)	-0.051*** (0.01)	-0.051*** (0.01)	-0.053*** (0.02)	-0.053*** (0.02)
Adverbs	0.020 (0.01)	0.014 (0.01)	0.014 (0.01)	0.011 (0.01)	0.011 (0.01)
Personal Pronouns	0.026** (0.01)	0.027** (0.01)	0.027** (0.01)	0.028** (0.01)	0.028** (0.01)
LIWC Psychometric Features					
Authentic	0.036*** (0.01)	0.031*** (0.01)	0.031*** (0.01)	0.041*** (0.01)	0.041*** (0.01)
Analytic	-0.077*** (0.01)	-0.072*** (0.01)	-0.072*** (0.01)	-0.078*** (0.01)	-0.078*** (0.01)
Emotional Tone	0.004 (0.01)	0.007 (0.01)	0.007 (0.01)	0.005 (0.01)	0.006 (0.01)
Certain Words	-0.009 (0.01)	0.003 (0.01)	0.003 (0.01)	-0.002 (0.01)	-0.002 (0.01)
Uncertain Words	-0.019* (0.01)	-0.024** (0.01)	-0.024** (0.01)	-0.033*** (0.01)	-0.032*** (0.01)
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓
Observations	1517	1517	1517	1314	1314

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types.

Model (4) and (5) excludes 203 knock-out round speeches without judge panel information.

Robust clustered standard errors at debate level are in parentheses.

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

Table 2.44: (Preliminary Rounds) Average Marginal Effects from Logistic regression on linguistic features: Male vs. Female Speakers ($N_{prelim} = 984$)

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Basic Features					
Word Count	-0.029*	-0.016	-0.015	-0.016	-0.016
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Words per Sentence	0.052***	0.042**	0.042**	0.040**	0.041**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Complex Words	-0.050***	-0.032**	-0.032**	-0.032**	-0.032**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Argument Indicators	-0.011	-0.001	-0.001	-0.002	-0.003
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Fillers	0.016	0.008	0.008	0.008	0.008
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Hedges	0.055***	0.041***	0.040***	0.041***	0.041***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)
BP Words & Phrases	-0.017	-0.023	-0.023	-0.023	-0.024
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
POI Rejects	-0.000	0.004	0.004	0.004	0.004
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
POS					
Noun	-0.088***	-0.066***	-0.066***	-0.066***	-0.065***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Verb	-0.013	-0.006	-0.005	-0.006	-0.006
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Adjectives	-0.054**	-0.045***	-0.044***	-0.046***	-0.045***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Adverbs	0.024	0.020	0.020	0.019	0.019
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Personal Pronouns	0.041***	0.046***	0.047***	0.046***	0.046***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
LIWC Psychometric Features					
Authentic	0.049***	0.048***	0.048***	0.047***	0.047***
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Analytic	-0.103***	-0.101***	-0.101***	-0.101***	-0.101***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Tone	0.008	0.007	0.007	0.008	0.008
	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Certainty Indicators	0.006	0.011	0.011	0.010	0.010
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Uncertainty Indicators	-0.048***	-0.047***	-0.047***	-0.047***	-0.047***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Observations	984	976	976	976	976
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types. Only 8 speeches (all male) debated Environment motion, thus they are dropped. Robust clustered standard errors at debate level are in parentheses.

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

Table 2.45: (Elimination Rounds) Average Marginal Effects from Logistic regression of linguistic features: Male vs. Female Speakers ($N_{elim} = 533$)

	Dependent variable: Gender (Female = 1)				
	(1)	(2)	(3)	(4)	(5)
Basic Features					
Word Count	-0.011 (0.02)	-0.023 (0.02)	-0.024 (0.02)	-0.037* (0.02)	-0.037* (0.02)
Words per Sentence	0.041** (0.02)	0.034* (0.02)	0.034* (0.02)	0.065*** (0.02)	0.064*** (0.02)
Complex Words	0.020 (0.02)	0.010 (0.02)	0.010 (0.02)	0.005 (0.03)	0.004 (0.03)
Argument Indicators	0.003 (0.02)	-0.006 (0.02)	-0.006 (0.02)	0.003 (0.02)	0.003 (0.02)
Fillers	0.007 (0.02)	-0.009 (0.02)	-0.009 (0.02)	-0.010 (0.02)	-0.009 (0.02)
Hedges	0.041** (0.02)	0.041** (0.02)	0.041** (0.02)	0.022 (0.03)	0.022 (0.03)
BP Words & Phrases	-0.012 (0.02)	-0.013 (0.02)	-0.013 (0.02)	-0.044** (0.02)	-0.043** (0.02)
POI rejects	0.024 (0.02)	0.031* (0.02)	0.030* (0.02)	0.029 (0.02)	0.028 (0.02)
POS					
Noun	-0.012 (0.02)	0.002 (0.03)	0.002 (0.03)	-0.018 (0.03)	-0.018 (0.03)
Verb	-0.008 (0.02)	-0.005 (0.02)	-0.005 (0.02)	-0.009 (0.03)	-0.009 (0.03)
Adjective	-0.073*** (0.03)	-0.077*** (0.03)	-0.077*** (0.03)	-0.114*** (0.03)	-0.112*** (0.04)
Adverb	-0.013 (0.02)	0.007 (0.02)	0.007 (0.02)	-0.011 (0.03)	-0.009 (0.03)
Personal Pronouns	-0.015 (0.02)	0.003 (0.02)	0.003 (0.02)	-0.015 (0.02)	-0.014 (0.02)
LIWC Psychometric Features					
Authentic	-0.006 (0.02)	-0.009 (0.02)	-0.009 (0.02)	0.001 (0.03)	0.003 (0.03)
Analytic	-0.017 (0.02)	-0.034 (0.02)	-0.034 (0.02)	-0.026 (0.03)	-0.028 (0.03)
Tone	-0.008 (0.02)	0.012 (0.02)	0.012 (0.02)	-0.000 (0.02)	-0.001 (0.02)
Certainty Indicators	-0.019 (0.02)	-0.008 (0.02)	-0.009 (0.02)	-0.034 (0.02)	-0.034 (0.02)
Uncertainty Indicators	0.021 (0.01)	0.004 (0.02)	0.004 (0.02)	0.000 (0.02)	-0.000 (0.02)
Observations	533	533	533	330	330
Room controls		✓	✓	✓	✓
Chair Judge Gender			✓		✓
Panel Gender				✓	✓

Variables are standardized. Room Controls are: (i) Language Status; (ii) Speaking position; (iii) Competition; (iv) Room gender composition; (v) Institution Ranking; (vi) Motion types.

Out-round speeches without panel gender identity are dropped in Model (4) and (5).

Robust clustered standard errors at debate level are in parentheses.

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

2.8.4 Word List

2.8.4.1 Argument Indicators

I/we think	as	in light of	moreover
I/we believe	as indicated by	in that	furthermore
consequently	as shown	in view of	therefore
accordingly	besides	in view of the fact	hence
as a consequence	because	that	though
as a result	deduced	in particular	thus
it follows that	derived from	indicated by	just because
overall	due to	is supported by	primarily because
to sum up	first/second/third(ly)	may be inferred	especially because
in conclusion	follows from	(from)	presumably because
in short	for example	nevertheless	simply because
in addition	for instance	moreover	particularly because
in particular	for one thing	owing to	merely because
specifically	for the reason that	this can be seen from	only because
nevertheless	for this/that reason	since	not because
so that	furthermore	since the evidence is	
after all	given that	what's more	
although	inasmuch as	whereas	
assuming that	in addition	while	

2.8.4.2 Hedges

I/we think	sth around	principally	practically
I/we mean	approximate(ly)	pretty much	relatively
I suspect	more or less	in a sense	roughly
I suppose	so to say	in a way	essentially
I guess	exceptionally	virtually	typically
We say	regularly	rather	literally
a little bit	often	almost	particularly
on a tall side	technically	strictly	especially

mostly	let's say	nothing but	anyway
largely	kind of	maybe	
basically	sort of	perhaps	
actually	like	somewhat	
so again	anything but	almost	

2.8.4.3 Fillers & disfluencies

ah	oh	umm	Oh well
ahh	ohh	well	You know
er	sigh	anyway	woah
hm	sighed	blah	
huh	ugh	dunno	
mm	uh	I don't know	
mmm	um	I mean	

2.8.4.4 British Parliamentary debate procedural phrases

madam/mr. Chairman/Chair/Speaker	ladies and gentlemen
(honorable/esteemed) panel	thank you/thanks for the floor
opening/first half	I am/we are (very/extremely) proud to propose/oppose
closing/second/back half	never been prouder to propose/oppose
side Government/Gov	beg you to propose/oppose
side Opposition/Opp	on the comparative
Prime Minister	the point at which
Leader of Opposition	the stakeholders are
Deputy Prime Minister	since the evidence is
Deputy Leader of Opposition	what we are proposing
Member of Government	this can be seen from
Member of Opposition	what I am/we are proposing
Government Whip	we told you
Opposition Whip	
Closing/Opening	
Our/Their side	

" When I'm sometimes asked when will there be enough women on the Supreme Court and I say, 'When there are nine,' people are shocked. But there'd been nine men, and nobody's ever raised a question about that."

Ruth Bader Ginsburg

3 Gender Composition of Committees and Performance Evaluation: Evidence from Debate Tournaments

3.1 Introduction

Despite significant progress in improving women's access to the labor market over the last decades, gender gaps in the representation and performance at high-ranked, influential positions remain a stylized fact [Blau and Kahn, 2007; Adams and Kirchmaier, 2016; Goldin et al., 2017; Blau and Kahn, 2017; Eckel et al., 2020]. To mitigate these gaps, an increasing number of countries¹ have adopted gender quotas [Schwindt-Bayer, 2009] to increase the number of women in high-level committees, ranging from board rooms [Adams and Funk, 2012; Matsa and Miller, 2013], local elections [Baltrunaite et al., 2014] to judiciary in the courtrooms [Bagues and Esteve-Volart, 2010] and professorship committees in universities [Bagues et al., 2017; Deschamps et al., 2018]. Nevertheless, evidence on the effectiveness of female leadership and female-majority committees in racking up support for female candidates are, at best, mixed [Bertrand et al., 2019; Azmat et al., 2020].

While various studies on corporate board committees observe that increasing women on boards improves firm performance [Adams and Ferreira, 2009; Adams and Funk, 2012; Matsa and Miller, 2013; Régner et al., 2019], spillover benefits to women in lower ranks [Kunze and Miller, 2017] or closes the gender wage gap [Bertrand et al., 2019], its

¹Norway mandates 40% representation of each gender on the board of public limited liability companies since 2003 [Bertrand et al., 2019; Hoel, 2008], where non-compliance entails liquidation or de-listing. In 2012, Italy passes the Golfo-Mosca law to impose gender quotas on boards of directors of publicly listed companies. In 2016, Germany requires the top 100 largest public traded firms to fulfill a 30% quota for their supervisory boards [Huang et al., 2020]. In politics, 21 countries have adopted gender quota laws that require between 20% and 50% of all legislative candidates to be women [Baldez, 2004].

trickle-down impact to promote qualified women remains contested.² In settings where women are exogenously assigned to evaluation committees, [Bagues et al., 2017] found no impact on women's promotion chances, while [Bagues and Esteve-Volart, 2010] noted an adverse impact of higher female representation in judiciary contexts.

This chapter exploits the random gender composition of evaluation committees in World and European inter-varsity debate tournaments to understand the causal impact of women in committees on success chances of female speakers. The institutional arrangements of these competitions along with its participants offer several uniquely attractive features. The data set consists of 39 168 individual speech evaluations by 4986 adjudication panels from high-stake, multi-round annual debate competitions from 2015 to 2018, where 400 to 800+ university-level students compete to be crowned debate champion of Europe or the world. This diverse and ambitious subject pool of next generation elites, who hone and spar their persuasion skills against one another regularly at debate competitions, makes this setup externally relevant to real-life competitive contexts in corporations and the political world. In fact, many famous politicians, lawyers and judges trained their persuasion skills in competitive debating during their high school and university studies.³ Regarding the setup of debate competitions, in every preliminary round, fixed teams of two speakers are exogenously assigned a speaking position⁴ and opponents to argue for or against a series of controversial, policy-relevant topics.⁵ Using transparent evaluation criteria and scale,⁶ a panel of three to five judges assess the comparative argumentation strength of these speeches. Conditional on adjudication experience and judge test performance, judges are

²See, for instance, the failure of hiring quotas [Woolston, 2019] and hiring committees [Deschamps et al., 2018]; no increased share of top-earning women in Italian companies after 2011 board reforms [Maida and Weber, 2019]; the U-shape relationship among US bank holding companies by [Owen and Temesvary, 2018]; the inconsistent shifts in public policies of Norwegian executive gender quota [Geys and Sørensen, 2019] or its indiscernible effect on other highly qualified women [Bertrand et al., 2019; Bagues and Campa, 2021]; no to limited effect of German gender quota rule on advisory board [Burrow et al., 2018; Huang et al., 2020]; and the conditional deliberation rule given the number of women on the committees [Karpowitz et al., 2012; Mendelberg et al., 2014].

³Examples include the current UK Prime Minister Boris Johnson, US House Speaker Nancy Pelosi and Senator Elizabeth Warren, economist John-Maynard Keynes, former Pakistani prime minister Benazi Bhutto, US Circuit Judge Stephanos Bibas, 29th Australian Prime Minister Malcom Turnbull.

⁴Four positions per debate: two proposition teams vs. two opposition teams. See Appendix 2.8.0.1 for details on the British Parliamentary Debate Format.

⁵See Figure 3.4 for the motion types across debates.

⁶The evaluation guideline and scoring scale are available to all participants prior and during the debate. For the detailed guideline, please refer to page 4 - 10 of [Novi Sad EUDC 2018 Judge Briefing](#). For the score scale with description, please see Appendix 2.8.2.1.

randomly allocated to debate rooms, with exogenously assigned decision power. The *chair* judge leads the deliberation discussion and has the ultimate decision over individual speech scores; whereas the *wing* judges only hold the basic adjudication duty.⁷ This randomized setup of female leadership in male-majority versus female-majority panels overcomes the endogeneity problem of power and gender balance in real-life committee formation processes. Moreover, the competitive *tournament-for-judges* system based on cumulative subjective performance feedback mirrors real-life committees, where career concerns, authority play and social pressures matter. In debate tournaments, round-by-round assessments by (i) judges who adjudicated with and (ii) debaters who were judged by them is instrumental to room allocation mechanism⁸ of subsequent rounds. As this cumulative feedback determines the best judges of Europe or the world, they arguably have strong incentives to exert their best adjudication effort.

I find that committees with a female chair judge give lower scores to both male and female speakers. This result holds regardless of whether they are paired with male- or female-majority wing judges. No significant score gap is detected between male- versus female-majority committees. Further investigation into the evaluation patterns of committees chaired by accomplished judges⁹ and debate room quality reveal differential standards among accomplished judges, especially in higher-ranked debates. Compared to committees with a novice chair, those headed by an accomplished judge give relatively high scores to both male and female speakers. This is driven by accomplished male chair judges, who do so to a larger extent for male speakers than for female speakers. In contrast, panels with an accomplished female chair do not give higher scores to either male or female speakers. Although panels with female-majority wing judges are positively preferential towards female speakers, the magnitude is insufficient to improve their success chances. Overall, only part of the unconditional gender score gap is explained by differential evaluation standards, i.e. committees chaired by accomplished male judges in high-ranked rooms give male speakers relatively higher scores than female speakers. Hence, having more female judges on a committee or a female chair does not necessarily improve the success chances

⁷Their roles are to determine team ranking, speaker scores and justifications during the deliberation.

⁸Wing judges with consistently excellent feedback are promoted to chair judges, while excellent chair judges are promoted to chair in better rooms, and vice versa. A detailed explanation of judge allocation mechanism is in Section 4.2.

⁹i.e. chair judges whose cumulative performance in nine preliminary rounds have qualified to the elimination rounds *at least* once, as a speaker or a judge, in previous EUDC/WUDC tournaments.

of female debaters.

Contribution-wise, this research provides causal findings on the ambiguous impact of having more women on committees, in a repeated *tournament-for-judges* setting with deliberation discussions among future elites of the world. In comparison to the existing causal evidence on the randomly assigned professor and judicial promotion committees [Bagues et al., 2017; Bagues and Esteve-Volart, 2010], in gender quota on candidate lists in Spain [Bagues and Campa, 2021], in academic hiring in French academia [Deschamps et al., 2018], or in sports [Sandberg, 2015], debate evaluation committees have comparatively more exogenously assigned female members and leaders, thereby augmenting the null finding of using gender quota on boards to fix the "leaky pipeline" [Clark Blickenstaff, 2005]. The deliberation element among the evaluation committees in this paper speaks to the literature advocating the holistic approach to close the gender gap in authority, rather than just a gender quota requirement. In the lab, [Goeree and Yariv, 2011] found that free-form communication increases decision efficiency, whereas recent work by [Mengel, 2020] and [Coffman et al., 2019] documented strong evidence on gender biases in opinion aggregation under open communication. Committees who are unaware of their implicit biases are found to promote fewer women [Régner et al., 2019]. In the field, [Green and Homroy, 2018] found a positive correlation between firm performance and increasing female representation on board committees in large European firms, while [Kunze and Miller, 2017] found that female bosses have positive spillover benefits to women in lower ranks. In politics, [Baltrunaite et al., 2014] found that gender quotas lead to the higher election of higher-educated women and fewer low-educated elected men. However, higher performance is only achievable with at least 30% of women on the board of firms, as noted by [Joecks et al., 2013]. Similarly in politics, [Schwindt-Bayer, 2009] found that candidate quota law only increases the election of women conditional on quota designs; whereas [Karpowitz et al., 2012], [Haire et al., 2013] and [Mendelberg et al., 2014] show that gender gap in voice and authority during deliberation only vanishes given appropriate decision rules and group gender composition.¹⁰ As norms around what constitutes persuasiveness might be male-skewed i.e. reward more the speech features that men are better at, accomplished evaluators, male or female, inevitably adopt such norms. Therefore, my findings raise doubts about the effectiveness of the gender quota law on

¹⁰Specifically, the gap vanishes under unanimous rule in groups with few women, or majority rule in groups with more women

breaking the glass ceiling for women in the highest offices.

Second, the subjective evaluation of speech persuasiveness in these male-dominated debate tournaments relates to the literature on the highly gendered evaluation of competence. Particularly in male-dominated settings with stereo-typically male tasks, an extensive body of theoretical and lab studies have documented persistent statistical and preference-based discrimination against women [Reuben et al., 2014; Bordalo et al., 2016; Karpowitz and Mendelberg, 2014]. While such gendered perceptions might not impede actual team performance [Heursen et al., 2020], male members still underrated female leadership effectiveness, regardless of superior team performance of a female leader [De Paola et al., 2018]. In the field, be it online forums [Bohren et al., 2019], physicians referrals [Sarsons, 2017a], orchestra sex-based hiring [Goldin and Rouse, 2000] or entrepreneurial pitches to investors [Brooks et al., 2014; Balachandra et al., 2019], women are systematically discriminated against, regardless of their performance and abilities. My finding that accomplished male chair judges are more preferential towards male speakers resonates with in-group favoritism and gate-keeping evidence [Brooks et al., 2014]; while the harsher evaluation of female judges sides with the null impact of increasing female representation in committees [Bagues et al., 2017; Deschamps et al., 2018].

Third, as speech evaluation scores result from deliberative committee decisions,¹¹ this research speaks to the literature on committee decision-making dynamics. While groups have been shown to make more rational decisions in strategic situations, social pressure [Fershtman et al., 2020] and reputation concerns [Prendergast, 1993; Levy, 2007] undeniably play a role, especially in a *tournament-for-judges* based on the feedback of peer judges and speakers considered in this paper. Specifically, along with the desire to be liked by other group members, [Isenberg, 1986] and [Visser and Swank, 2007] show that committee members manipulate information to align with the member who has the decisive vote. Since chair judges have moderation and tie-breaking vote power, these mechanisms could explain why only the chair judge's gender, and not the gender composition of other committee members, matters for speakers' evaluations.

This chapter proceeds as follows. Section 3.2 summarizes the debate competition setup.

¹¹For a step-by-step description of the deliberation procedure, please see Section 3.2.2.

Section 3.3 provides data set overview, followed by summary statistics in Section 3.4. Empirical strategies and hypotheses are given in Section 3.5. Section 3.6 highlights the main results, with extension findings on the accomplishment of chair judges and debate room quality in Section 3.7. Section 3.8 concludes with discussions on future research avenues.

3.2 Institutional Setup

3.2.1 Tournament format and allocation mechanism

Tournament Format. Every year, around 200+/- teams across Europe attend the European Universities Debate Championship (EUDC) and 450+/- teams across the world participate in the World Universities Debate Championship. These two-person teams represent their institutions to compete over nine preliminary rounds (i.e. *in-rounds*) over three days with exogenously assigned controversial topics, speaking positions, judges, and opponents in every round. All debates are conducted in British Parliamentary (BP) Debate style.¹² After each round, individuals receive two results: (1) team ranking¹³ and (2) individual speaker scores.¹⁴ Importantly, individual speaker scores must reflect the ordinal team ranking i.e. the cumulative score of two speakers whose team ranked first must be higher than that of the team ranked second. In Round 1, teams are randomly matched. From Round 2 onward, teams are power-matched i.e. teams debate teams with similar cumulative team *and* speech evaluation points from previous rounds.¹⁵ Therefore, the transparent and universal individual speech score scale is meant to ensure consistent speech quality evaluation across rooms; i.e. winning from a lower-ranked room does not necessarily mean higher individual speaker scores than, for instance, taking a 2nd or 3rd in a higher ranked room. The accumulated team points and speaker points¹⁶ across 9 preliminary rounds determine the top 10 – 15% performing teams to enter elimination rounds (i.e. *out-rounds*). In these rounds, teams that are ranked 1st and 2nd advanced into further rounds, whereas those on 3rd and 4th place are eliminated. In the final debate, the best team becomes the champion.

¹²For more details on BP debate style and format, please check Appendix 2.8.0.1.

¹³i.e. team that ranks 1st gets 3 points, 2nd gets 2 points, 3rd gets 1 points and 4th gets no point.

¹⁴50-to-100 score scale, with 50 as the lowest. See Appendix 2.8.2.1 for a speaker score scale example of European Universities Debate Championship 2017.

¹⁵For more information about power-pairing, see this discussion thread on [Monash Debate Review](#).

¹⁶Speaker points are used for: (i) award best performing speakers in the form of top 10 speaker awards; and (ii) determine teams that advance into elimination rounds in case of ties.

The best speaker of the tournament is an individual with the highest cumulative individual speech scores across all preliminary rounds.

Judge Recruitment Mechanism. An appointed Chief Adjudicator (CA) team of four to six internationally accomplished debaters in every tournament is in charge of judge recruitment, quality screening, monitoring, and overall panel allocation throughout the tournament. The CA team often recruits and ranks judges in two rounds. In the first round, prospective judges can opt in to do an advanced judge test to be selected as *Independent Adjudicators (IA)*. For those who do well, the CA team can offer partial or full travel and/or accommodation funding. In the second round, all judges, IAs or not, must finish a comprehensive judge test. These results along with survey data on past judging and debate experiences enable the CAs to slot judges as chairs, panelists, or trainee¹⁷ judges in Round 1s.

Judge Allocation Mechanism. Based on the judge test results and survey data on debate experience, an algorithm,¹⁸ supervised by a tabulation team, randomly assigns judges and debaters simultaneously to different rooms, taking into account possible conflicts of interests.¹⁹ From Round 2 onward, individual judge performance feedback, given by peer judges and debaters in previous rounds²⁰ is automatically incorporated into the allocation algorithm to demote or promote²¹ judges in the next round. This procedure is overseen by an independent tabulation team in Round 2 to Round 6.²² From Round 7 to Round 9, in strategically important debate rooms, to ensure the highest quality judging, the CA team manually stacks the highest-performing judges to these rooms; while promoting promising high-performing new wing judges to chair in lower-ranked rooms. Another notable feature of these rounds is that judges do not announce and justify team ranking results to debaters after their deliberation, unlike from Round 1 to Round 6. Overall,

¹⁷i.e judges without any judging experience and/or do not do the judge test. They have no voting right.

¹⁸e.g. Tabbie2, Tabbycat

¹⁹e.g: judges and speakers from the same current or past institutions, ex-debate partners, romantic partners, close friends.

²⁰An adjudicator's overall score is as follows: $Score = (1 - w) * Testscore + w * AverageCumulativeFeedbackScore$ where w is the feedback weight for the round. Test score includes weighted average between judge test and previous judge & speaking achievements.

²¹e.g. bad-performing chairs become wing judges, good-performing wings become chair judges, good-performing chair judges are promoted to higher rooms.

²²Due to a large number of debate rooms and the intensive time schedule, unless there are very serious clashes reported, the CA team barely ever intervenes with judge/team allocation during these rounds.

cumulative judge performance in all preliminary rounds determines the highest 20 – 30% ranked judges to adjudicate the out-rounds, i.e. being acknowledged as the best judges of Europe or the world. Hence, judging in EUDC/WUDC out-rounds is considered a significant achievement, which often yields invitations to chief adjudicate major competitions.

Allocation Fairness Mechanism: Teams & Judges. Three mechanisms are set in place to ensure fairness in judgment across rounds. First, no judges who come from the same past or present institutions as any debaters in the room can be allocated to judge that debate. Second, prior to the competition, judges and debaters are required to notify any potential conflicts with other participants.²³ To disincentivize strategic clashing, an independent committee conducts confidential interviews with the requested persons to verify their reasons. Third, the intensive nature of a 3-day, 9-round competition makes it difficult for any strategic collusion to be formed between judges, the CA team, and speakers from different institutions.

3.2.2 Adjudication structure and deliberation rules

A Judge's Role.²⁴ Judges need to act as an *informed global citizen*, who evaluate the argumentative cases *holistically*, given their relevance and plausibility. Given *impartial* reading of the entire debate, judging is done comparatively, i.e. decide which team, amongst the four teams, gives the most persuasive case. The knowledge standard is restricted to the front pages of major articles in the national or international newspapers.²⁵ A qualified judge must accurately weigh what was *actually said* by teams in the debate, without inserting one's preconceptions or expert knowledge into their decisions.

Adjudication Panel Structure. A debate room is adjudicated by a *chair (C)* judge (one chair/room) and several *wing (W)* judges. Typically two to four wing judges are allocated in preliminary rounds, whereas four to eight in elimination rounds. All judges are responsible for keeping track of the key arguments and determine the team ranking, speaker scores,

²³Reasons include, but are not limited to, close friendship/partnership, past romantic encounters, or negative past experience.

²⁴For a detailed description of judge's role, please refer to page 4 - 10 of [Novi Sad EUDC 2018 Judge Briefing](#).

²⁵For instance, discussing the reparations for WWII, the Iraq conflict, or AI ethics would be a fair game, not on the technical or esoteric knowledge about these issues.

and justifications thereof. The chair (C) judge has the *ultimate power* and responsibility to assign the definitive ranking,²⁶ and speaker scores, as well as delivering verbal ranking explanations to debaters after the panel discussion.

Deliberation Procedure. During the debate, all judges carefully take notes of the speeches and determine their ranking of teams without interacting with one another. Upon making up their mind, judges reveal their decisions to one another. The chair judge then moderates the discussion, with the unanimous voting rule if time permits. If the panel exceeds the given time, majority voting i.e. 'split' rule kicks in. In case of a tie,²⁷ the chair judge has the tie-breaking vote to determine the final call. Afterward, chair judges determine the individual speech scores, which must reflect the ordinal team ranking. In Round 1 to Round 6, the chair judge then orally justifies the team ranking decision to speakers. No such information is released in Round 7 to Round 9. All individual speech scores and round ranking in Round 7 to Round 9 are announced to speakers after the tournament ends.

3.3 Data

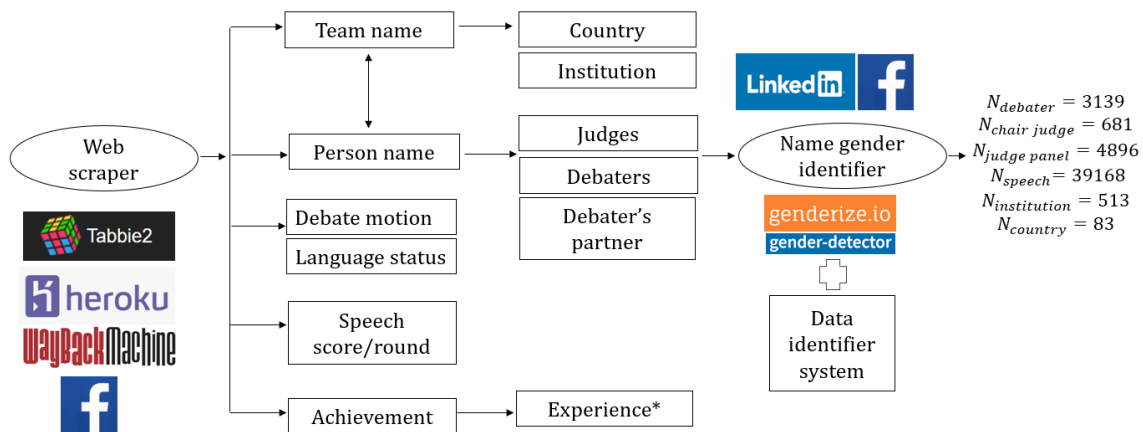


Figure 3.1: Overview: Data collection and construction procedure

²⁶In case of ranking conflict, the vote of the chair judge will be the tie-breaker vote.

²⁷e.g. one wing gives a second to Team X, and the other gives a fourth to Team X.

This section describes the data set construction procedure of evaluation scores, speakers, and judges, as illustrated in Figure 4.1. Overall, names of individuals, judges and institutions, the roles of judges²⁸ and opponents,²⁹ speech evaluation scores for every debate, language skill status of speakers, and debate motions are available from tabulated archival sources.³⁰ Detailed data collection procedures on other control variables such as debate topics, language skills, and institutions are provided in Appendix 4.8.1. Section 4.3.2 then gives descriptive statistics on score differentials across judges and speakers given their demographics.

3.3.1 Outcome Variable: Speech Scores

The main outcome variable in this chapter is the individual speech evaluation scores, which range from 50 to 100, with 50 being the lowest. Across 5081 debates from eight competitions, 185 debates are omitted due to the missing identity of speakers or speech scores.³¹ Table 3.5 reports the score descriptive statistics across tournaments. We noted a slight increase in the average scores from EUDC 2014 to EUDC 2018 tournament, yet the average of scores in WUDC hover largely around 76. Overall, WUDC tournaments are associated with slightly higher scores than EUDC tournaments.

The correlation heat map in Figure 3.5 shows three interesting associations across characteristics of judges and speakers. First, there is no significant dependence between cumulative speaker score (i.e. measure of the respective debate room quality) and gender of the chair judge or wing judges. This is further confirmed with the Spearman correlation test [de Winter et al., 2016] with Bonferroni's adjustment, whereby the correlation coefficient between gender of speaker and chair judge is $\rho = 0.0003$ ($p_value = 0.9461$), and that between gender of speakers and wing gender composition is $\rho = 0.0043$ ($p_value = 0.3903$). No correlation between the gender of the chair judge and the number of female wing judges on the panel, nor the number of female speakers in the debate is detected either. Second, the more EUDC/WUDC speaking achievements a chair judge has accumulated, the more likely

²⁸i.e. chair judge, wing judge, and trainee judge.

²⁹i.e. Opening Government, Opening Opposition, Closing Government, Closing Opposition

³⁰Such data is released given the consent of speakers and judges unless otherwise redacted, in which case they are omitted from the sample.

³¹To ensure full understanding of judge's dynamics, any debate with either of the following reasons are omitted: (i) swing speakers;³² (ii) speakers who redacted their identity after the tournament; or (iii) one speaker spoke for both roles. An overview is summarized in Appendix 3.4.

he/she is allocated to debate rooms with more female speakers. Third, rooms with higher debate quality are positively correlated with chair judges with higher EUDC/WUDC judging achievements. Overall, past EUDC/WUDC judging or speaking experience matters in judge allocation, whereas this does not hold for the gender of speakers or judges.

3.3.2 Judges & Evaluation Panels

Identity & gender representation. From the tabulation tournament data archive in [Tabbie2](#) and [Tabbycat](#), we obtained the full names of chair (C), wing (W) and trainee (T) judges for every debate. To determine their identity and represented gender, we sorted the names of all judges per tournament by their functions. For chair (C) judges, I can determine unique identity and gender for everyone, since their identity is easily tracked given their high-profile statuses and profiles in debate channels. This procedure results in the identification of $N_{chairjudge} = 681$, with $N_{malechair} = 440$ and $N_{femalechair} = 241$.

For wing (W) and trainee (T) judges, I combined results from gender inference algorithms on their first names, affiliated institutions, countries, and region with final confirmation with respective tournament tab directors. I used two commonly used gender inference packages: [gender guesser](#) and [genderize.io](#).³³ Both algorithms return the most likely gender of a first name, given its manually coded labels, and a frequency count of such names in their database as male or female.³⁴ With this method, I identified the gender of 92% of (W) and (T) judges. For the remaining 8%, which either are: (i) African, South East Asian, Indian, and Israeli and gender-neutral names or (ii) conflicting gender assignments, I manually checked them using social media connections. Finally, the completed gender list for each judge is confirmed with respective tab directors.

Overall, 20 – 35% of speeches across tournaments are adjudicated by female chair judges, as noted in [Figure 3.14](#). Most adjudication panels have 3 to 4 judges, with roughly one-third of the judges as female, shown in [3.17](#). Around 40% of the committees have at least as many female as male members, as noted in [Figure 3.15](#).

³³This API contains 216286 distinct names across 79 countries and 89 languages.

³⁴For a comparison of features and performance of different gender inference algorithms, please refer to the report of [[Menéndez et al., 2020](#)] and [[Santamaría and Mihaljević, 2018](#)].

Adjudication/Speaking Experience. To proxy for the reputation of chair judges on the international circuits, I collected from the archived tabulation system the information on the number of times they have, in *at least* the immediate past two years:³⁵ (1) qualified for the out-rounds as a judge and/or as a speaker in the EUDC and/or WUDC tournaments;³⁶ and (2) been part of the Chief Adjudicator (CA). This information is summarized in Table 3.6. 40 % of chair judges without prior EUDC/WUDC achievements indicate both the dynamic weighing of judging performance feedback throughout the tournament, as well as the role of the judge test performance and past local/regional achievements.³⁷

The cumulative distribution plots of speeches given speaking and judging achievements in Figure 3.18 show that, in both categories, male chair judges have slightly more achievements than their female counterparts. Noteworthily, the correlation heat map of Figure 3.5 shows the importance of CA experience and WUDC/WUDC achievements in room allocation. Chairs with previous *judging* EUDC/WUDC achievements or Chief Adjudicator (CA) experience are positively correlated with cumulative average speech score (i.e. proxy for debate room quality), at a magnitude of 0.34 and 0.29 respectively. Furthermore, rooms with more female speakers correlate positively with previous *speaking* achievements of chair judges.

3.3.3 Speakers

Upon scraping, cleaning full names of speakers³⁸ and matching their identities across the years,³⁹ we obtained $N = 3180$ unique persons. To identify the gender of speakers, similar to the judges, I ran gender inference algorithms over their first names. This procedure results

³⁵e.g: Experience level for judges in 2015 is their total speaking, judging and CA achievements from 2014 and 2013. For 2016, 2017, and 2018, cumulative experience from 2015, 2016, and 2017 is considered, respectively.

³⁶The timeline of WUDC tournaments (December/January) and EUDC (August) is accounted for.

³⁷Judge test performance & individual feedback is strictly confidential data. As for past local/regional achievements, the potentially inconsistent mechanism of how different CA teams weighed these experiences over time makes it an extremely noisy control variable. On the other hand, the size and prestige of EUDC/WUDC tournaments makes achievements at these tournaments predominantly important in the judge ranking metric of Chief Adjudicators.

³⁸i.e. odd characters from non-English names, reversed first and last names, abbreviated names are properly restored across tournaments by matching with their institutions and, where applicable, their social media profiles.

³⁹To minimize discretionary judgment, we used a conservative method: a name is considered a duplicate only if the name, institution, EUDC language status, and WUDC language status are exactly the same.

in 89.23% of names assigned gender with certainty. The remaining 10.77% names, which consists of mostly African, South Asian, Israeli, and Eastern European names, were manually checked using social media. Altogether, after omitting 27 unisex names without any social media sources and possible confirmation from tab masters, we have $N = 3153$ unique speakers for analysis: whereby $N_{MaleSpeaker} = 1949$ and $N_{FemaleSpeaker} = 1190$. Figure 4.2 shows the proportion of speakers by gender for each competition. Across all competitions, female speakers account for 35% to 41% of all participants. Furthermore, the world map distribution of speakers given their gender in Figure 3.3 shows that most countries sent disproportionately more male speakers than female speakers, except for China/Hong Kong. The US, UK, and Australia sent the highest number of speakers, understandably so, given their established debate training culture and civic participation.

3.4 Descriptive Statistics

Male vs. Female Chair Judges. Table 3.6 summarizes the count and percentage of speeches adjudicated by male versus female chairs given judge, speaker, and debate room characteristics. With respect to judge accomplishment, there are slightly more female chairs with no previous EUDC/WUDC achievement, at 43.62 % overall, compared to 38.18 % of male chairs. Male chair judges are matched with slightly more female-majority wings, at 58.57 % overall in comparison to 53.59 % for female chairs. There are no other notable differences between observable characteristics of speakers, debate room, or judges between male vs. female chair judges. Regarding evaluation patterns, the t-test statistics in Table 3.9 and the kernel density distribution of Figure 3.7 show that female chairs give significantly lower scores compared to their male counterparts. When breaking down the speeches by male vs. female speakers, Figure 3.8 shows that male chair judges give slightly higher scores to female speakers, regardless of whether they are in female-only or mixed-gender teams, as noted in Figure 3.9. Yet, for female chairs, they particularly gave lower scores to female-only teams, while scoring patterns for mixed gender and male-only teams have the same distribution.

Male - vs. Female-majority Panels. Table 3.7 summarizes the count and percentages of speeches adjudicated by male- vs. female-majority judge panels given chair judge accomplishment, speaker, and debate room characteristics. In 2018, there seems to be a

different proportion of wing judges given their gender. Specifically, in EUDC 2018, there is double the number of speeches adjudicated by female-majority panels (10.84 % vs. 5.91 %), while the opposite occurs for WUDC 2018: 16.17 % for male-majority panels vs. only 9.52 % for female-majority panels. For the rest, there are no notable differences between speeches adjudicated by male vs. female-majority panels. Regarding evaluation patterns, the t-test statistics in Table 3.9 and the kernel density distribution in Figure 3.10 show similar pattern as with the chair judge gender. Upon looking further between male vs. female speakers, Figure 3.11 does not show any differing patterns in scoring, but Figure 3.12 showed that male-only teams and mixed-gender teams have slightly higher scores than female-only teams from male-majority panels. Overall, the gap between speech evaluation scores is much larger between male vs. female chair judges, and not so much between male vs. female-majority panels over the rounds.

Male vs. Female Speakers. Table 3.8 provides a comprehensive breakdown of the proportion of speeches by male and female speakers given relevant speaker and room control characteristics. Given the male-dominated environment of debate tournaments, there are disproportionately more male speakers in male-majority rooms than female speakers, and vice versa. Other than that, there does not appear to be any differences in terms of the proportion of speeches by male vs. female speakers across these characteristics. Regarding scores, both Table 3.10 and the kernel density of speech scores in Figure 4.8 show that male speakers scored significantly higher than those given by females. This pattern holds regardless of the language skill statuses or whether the speaker belongs to the top 50-ranked institutions.

3.5 Methodology & Hypotheses

3.5.1 Empirical Strategies

To analyze whether adjudication panels evaluate speeches differently given their own gender composition and speaker's gender, I run linear and fixed effects regression on standardized speech score, interacting speaker's gender indicator variable with chair or panel gender composition, as shown in Equation 3.5.1 below:

$$S_{ik} = \alpha_{ik}\mathbb{I}_{FemS} + \theta_k\mathbb{I}_{C/P} + \beta_{ik}\mathbb{I}_{FemS}\mathbb{I}_{C/P} + \sum_{j=1}^n \gamma_j \mathbf{Y}_{ik} + \eta_k + \eta_S + \varepsilon_{ik}$$

The dependent variable S_{ik} is the standardized evaluation score of speech i in debate k . $FemS$ refers to the speaker's gender, whereas $\mathbb{I}_{C/P}$ is a dummy variable for gender of chair judge and the whole panel. Specifically, $\mathbb{I}_C = 1$ indicates the chair judge is a female and $\mathbb{I}_P = 1$ indicates that the panel (P) is female-majority; i.e it contains the same or higher number of female judges compared to male judges.

The coefficient of interest β_{ik} measures any significant relationship between being a female speaker and the gender of judges. Debate fixed effect η_k and individual speaker fixed effect η_S are added to the analysis to take care of any unobserved heterogeneities. ε_{ik} is the error term of speech i in debate k . Throughout all analyses, standard errors are clustered at the debate room level. The rest of speaker and room control variables \mathbf{Y}_{ik} are as follows:

1. $\mathbb{I}_L = 1$ if a debater is a non-native English speaker.
2. $\mathbb{I}_R = 1$ if a debater represents a top 50 academically ranked institutions worldwide.
3. $\mathbb{I}_{ChairExp} = 1$ indicates that the chair judge has achieved ≥ 1 EUDC/WUDC speaking or judging achievement.
4. \mathbb{I}_C is the group competition type, with $\mathbb{I}_C = 0$ as WUDC and $\mathbb{I}_C = 1$ as EUDC.
5. Speaking position (1^{st} to 8^{th}) in any given debate.
6. Motion topic type (17 topics) in any given debate.⁴⁰
7. Debate round (1 to 9) for any given debate.
8. $\mathbb{I}_D = 1$ if a room has ≥ 4 female speakers i.e. the room has equal or higher number of female speakers compared to male speakers.
9. $\mathbb{I}_W = 1$ if at least half of the evaluation panel, excluding the chair judge, are female. This control variable only applies for analysis on chair judges, and not on panel gender composition.

⁴⁰See Figure 3.4 for the list of motions.

Given the random pairing of chair and wing judges⁴¹ and the sufficiently large number of panels, earlier work on male-majority environment on female leadership [Born et al., 2020; Hoogendoorn et al., 2013] suggests taking a look into smaller sub-samples to properly understand the dynamics between chair and wing judges given their gender. Specifically, I ran similar regression analysis as above, interacting the gender of chair judges with the gender composition of the wing judges.

Finally, the dynamic power matching mechanism for both teams and judges from Round 2 onward makes it necessary to establish whether there exists any relationship between gender of judges (chair and wings) and debate room quality. For this purpose, I use the average cumulative speaker score up to the respective round as a control variable. This serves as a proxy for current round standing and thus determining team match-up composition from Round 2 onward. Hence, I can compare the evaluation patterns in conditionally random rooms vs. Round 1s, where allocation of debaters across the rooms is unconditionally random.

3.5.2 Hypotheses

Similar to real-life corporate, politics and academic settings, panels with female chair judges are under-represented in debate tournaments.⁴² More women on corporate boards is positively correlated with more desirable investment decisions [Adams and Funk, 2012; Matsa and Miller, 2013; Régner et al., 2019], better evaluations [Woolley et al., 2010] and narrower gender wage gaps [Bertrand et al., 2019]. Furthermore, women on committees are believed to enhance the success chances of lower-ranked [Kunze and Miller, 2017] and qualified women [Adams and Kirchmaier, 2016]. Nonetheless, recent evidence on scientific [Bagues et al., 2017], judiciary and gender quota on candidate lists [Bagues and Campa, 2021] found no evidence that more women on boards break the glass ceiling for highly qualified women to succeed. A similar null result is documented in the introduction of gender quota in French academic hiring committees [Deschamps et al., 2018]. Since chair judges have the highest authority in a panel, in line with the positive spillover evidence of female leadership, I expect that panels with female chair judges give relatively

⁴¹The correlation matrix 3.5 shows virtually no relationship between gender of chair judge and the number of female wing judges on a panel.

⁴²See details on the proportion of chair and female-majority judge panels in Summary Table 3.6, Figure 3.14, 3.15 and 3.17.

higher scores to female speakers.

H1 (Chair: Male vs. Female Judges): *Compared to committees chaired by male judges, those chaired by female judges give relatively higher scores to women than to men.*

In debate tournament contexts, given the overwhelming evidence on out-group discrimination in evaluation of competence in male-dominated settings [De Paola et al., 2018; Brooks et al., 2014; Balachandra et al., 2019], female judges could mitigate the possible implicit biases against female speakers by male judges. In addition to taste-based discrimination mechanism [Goldin, 2014], another possible mechanism is homophily in communication, whereby members of one's group i.e. gender in this case, find it easier to understand members of their own groups [Kets and Sandroni, 2019]. Given indicative evidence on the differing patterns of communication between men and women [Lakoff, 1973; Holmes, 1990; Leaper and Robnett, 2011], I expect that an increasing number of female judges on the panel, on average, understand speeches given by women better, thereby increasing the chance of higher scores for female speakers.

H2 (Panel: Male- vs. Female-majority Panels): *Compared to male-majority panels, female-majority panels give relatively higher scores to female speakers than to male speakers.*

Female leadership is found effective in giving more voice to women in teams [Heursen et al., 2020] and political debates [Latu et al., 2013; Blumenau, 2019], yet, they could have a harder time in managing the discussions [Vial et al., 2016] due to stereotype threats [Hoyt and Murphy, 2016], especially in male-majority committees. Given that assigned chair judges are either more accomplished or have performed notably better in the judge tests and previous rounds, along with their vested decisive vote power, I expect that on average, female chair judges still have the commanding voice, even when paired with male-majority wings. In committees with female-majority wings, given the homophily mechanism evidence and more salient in-group biases among women [Rudman and Goodwin, 2004; Carlsson and Eriksson, 2019], I expect that even when paired with a male chair judge, on average, the total vote counts and evaluation scores for female speakers from female-majority committees are higher than in male-majority committees.

H3 (Chair Gender ft. Wing Gender Composition): *Given the chair judge’s gender, committees with a majority of female (male) wing judges give relatively higher (lower) scores to female speakers than male speakers.*

3.6 Results

3.6.1 Gender of Chair Judge and Speech Evaluation

Table 3.1 reports the regression results of chair judge’s gender on speaker gender against standardized evaluation score S_{ik} . Columns (1) and (2) show the single regression result between S_{ik} and speaker and chair judge gender, respectively. Columns (3) to (8) summarize the results of five regression models that include the interaction between the gender of the speaker and chair judges, namely: (3) simple OLS regression (i.e. unconditional effect); (4) regression controlling for chair judge experience,⁴³ speaker and debate room characteristics Y_{ik} ;⁴⁴ (5) speaker η_S fixed effect regression; (6) speaker η_S fixed effect regression along with chair judge experience and debate room characteristics; (7) debate room η_k fixed effect regression; and (8) debate room η_k and speaker η_S fixed effect regression.

Column (1) of Table 3.1 shows that, unconditionally, female speakers on average receive 5.6 percentage point (p.p) standard deviation (SD) lower scores than male speakers. Upon interacting speaker gender with chair judge gender, Columns (3) and (4) show that the gap is similar among committees chaired by a male or female judge. Columns (5) and (6) also show no significant interaction effect, as is the case for Round 1s versus Round 2s to 9s in Table 3.11. Within debates, the gap vanishes, as noted in Column (7). Hence, the unconditional score difference is mostly due to more women sorting into relatively weaker debate rooms as the tournaments progressed. This finding from female chair judges rejects hypothesis **H1**. *Female chair judges do NOT give higher scores to female speakers, compared to male judges.* It is in line with the ambiguous or low impact of

⁴³i.e. whether or not the judge have advanced to elimination rounds, as a speaker or a judge, in previous EUDC/WUDC tournaments.

⁴⁴i.e. the number of female speakers in the debate, language skill status, number of female judges on the panel, speaking position, debate round, motion type, institution ranking, competition and year.

increasing female evaluators on committees in [Bagues et al., 2017; Bagues and Campa, 2021; Deschamps et al., 2018].

One important thing to note, while there is no significant evaluation difference between male and female chair judges across speakers, such difference appears upon controlling for individual speaker fixed effects. Specifically, for male speakers, Column (5) of Table 3.1 shows that, controlling for speaker fixed-effects, committees chaired by a female judge give on average significantly lower evaluation scores compared to committees chaired by a male judge. For male speakers, this effect is 5.1 p.p. SD; whereas for female speakers, this is 4.7 p.p. SD. Controlling for chair judge experience and debate room characteristics in column (6) does not meaningfully affect these estimates. This significant result also holds upon breaking down the analyses into Round 1s vs. Round 2 to Round 9 (controlling for debate room quality) in Table 3.11.

Three mechanisms could explain this result: (1) female chairs are allocated to weaker debate rooms, and speakers that are sorted into these rooms performed worse; (2) female chair judges are harsher than their male counterparts; (3) speakers perform worse if their chair judge is female. Channel (1) is unlikely since no correlation between debate room quality and chair judge gender is found in Figure 3.5. Disentangling between the channel (2) and (3) requires controlling for the speech content of contestants, which is not possible with the current data set.

Table 3.1: Regression Analysis of Chair Judge Gender against Speech Score (N = 39 168)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Speaker	-0.056*** (0.01)		-0.054*** (0.01)	-0.062*** (0.01)				-0.011 (0.01)
Female Chair		-0.035 (0.02)	-0.033 (0.02)	-0.029 (0.02)	-0.051*** (0.02)	-0.048*** (0.01)		
Female Speaker \times Female Chair			-0.007 (0.02)	0.000 (0.02)	-0.004 (0.02)	0.008 (0.02)	-0.012 (0.02)	0.001 (0.01)
Speaker Controls				✓				
Chair Experience				✓		✓		
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.001	0.000	0.001	0.183	0.568	0.588	0.587	0.768
Observations	39168	39168	39168	39168	39157	39157	39168	39157

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) round, (vi) motion type, (viii) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$. Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.6.2 Gender Composition of Committees and Speech Evaluation

Table 3.2 provides the results of the female-majority judge panel on speaker gender against standardized evaluation scores, with similar regression models across the columns as in the previous section. In contrast to findings on chair judges, there are no indications of differential standards by male- or female-majority judge panels, either for male speakers or for female speakers. In line with the findings on chair judges, I find no significant interaction effect between the gender composition of the committee and the speaker's gender. Breaking down the analysis between Round 1s and Round 2s to Round 9s (controlling for debate room quality) in Table 3.12 show qualitatively similar results. This finding rejects hypothesis **H2**. *Female-majority panels do NOT give higher speech scores to female speakers.*

3.6.3 Chair ft. Wing Gender Composition and Speech Evaluation

Table 3.3 studies the interaction between the gender of the chair judge and the gender composition of the wing judges. Most importantly, I find no systematic pattern in how committees of different makeup score men relative to women. The interaction effects of committee compositions with the female speaker dummy are all small and statistically

significant.⁴⁵ These results indicate that wing gender composition matters more in case the panel is chaired by a male judge, rather than a female judge, thereby collectively rejecting hypothesis **H3**. *Compared to panels of male chairs paired with male-majority wings, panels of female chairs and male majority wings do NOT give lower scores to female speakers. Panels with female-majority wings led by male chairs do NOT give higher scores to female speakers. Panels with female chairs give comparatively lower scores to both male and female speakers, regardless of the gender composition of wing judges.*

On the general scoring patterns across different gender compositions of wings and chair judges, I find that, unconditionally, compared to panels with a male chair paired with male-majority wing judges, male chairs paired with female-majority wing judges gave 7.5 p.p SD lower scores. At speaker and debate fixed effect levels, there is no significant difference in evaluation patterns between panels of male-chairs pairing with male- or female-majority wing judges. Overall, compared to panels with a male chair paired with male-majority wing judges, committees with female-majority wings give significantly lower scores to all speakers.

Regarding female chairs paired with female-majority wing judges, unconditionally, they gave 10.4 p.p SD. lower scores compared to panels with a male chair pairing with male-majority wing judges. Upon interacting with the speaker's gender, Column (3) shows that having a female-majority wing headed by female chair panels gives male speakers 8.8 p.p lower scores compared to speakers judged by male-majority, male chair panels. After adding relevant chair experience and room control variables onto speaker fixed effects, Column (4) to (6) shows that this gap remains significant.

⁴⁵Further breakdown of the analysis into Round 1s vs Round 2s to 9s (controlling for debate room quality) in Table 3.13 found similar patterns in Round 2s to Round 9s.

Table 3.2: Regression Analysis of Panel Gender Composition against Speech Score (N = 39168)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Speaker	-0.056*** (0.01)		-0.054*** (0.01)	-0.055*** (0.01)				-0.016 (0.01)
Female-majority Panel		0.004 (0.02)	0.007 (0.02)	0.034 (0.02)	-0.023 (0.01)	-0.007 (0.02)		
Female Speaker × Female-majority Panel			-0.005 (0.02)	-0.022 (0.02)	-0.000 (0.02)	0.001 (0.02)	0.003 (0.02)	-0.007 (0.01)
Chair Gender & Experience				✓		✓		
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.001	0.000	0.001	0.183	0.568	0.588	0.587	0.768
Observations	39168	39168	39168	39168	39157	39157	39168	39157

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) round, (vi) motion type, (viii) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$. Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.3: Regression Analysis of Chair ft. Wing Gender against Speech Score (N = 39168)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Speaker	-0.056*** (0.01)		-0.046** (0.02)	-0.049*** (0.02)				-0.016 (0.02)
M Chair ft. F-majority Wings		-0.075*** (0.03)	-0.070** (0.03)	-0.053** (0.02)	-0.013 (0.02)	-0.011 (0.02)		
F Chair ft. M-majority Wings		-0.051 (0.04)	-0.056 (0.04)	-0.055* (0.03)	-0.063*** (0.02)	-0.062*** (0.02)		
F Chair ft. F-majority Wings		-0.104*** (0.03)	-0.088** (0.04)	-0.059** (0.03)	-0.055*** (0.02)	-0.046** (0.02)		
Female Speaker × M Chair ft. F-majority Wings			-0.013 (0.03)	-0.022 (0.02)	0.007 (0.02)	0.008 (0.02)	0.008 (0.02)	0.010 (0.02)
Female Speaker × F Chair ft. M-majority Wings			0.015 (0.04)	0.014 (0.03)	0.011 (0.03)	0.013 (0.02)	0.008 (0.03)	0.028 (0.02)
Female Speaker × F Chair ft. F-majority Wings			-0.041 (0.03)	-0.043 (0.03)	0.012 (0.02)	0.013 (0.02)	-0.021 (0.03)	-0.012 (0.02)
Chair Experience				✓		✓		
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.001	0.001	0.002	0.183	0.568	0.588	0.587	0.768
Observations	39168	39168	39168	39168	39157	39157	39168	39157

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) speaking position, (iv) round, (v) motion type, (vi) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$. Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.7 Extensions

3.7.1 Do accomplished chair judges evaluate speeches differently?

3.7.1.1 Novice vs. Accomplished Chair Judges

Advancing to the elimination rounds as a speaker or a judge at EUDC/WUDC tournaments significantly boosts one's reputation in the debate community, which could affect persuasion dynamics in deliberation discussions [Manzoor et al., 2020] across debate rooms. Therefore, I investigate whether accomplished chair judges evaluate speeches differently given the speaker's gender, where $\mathbb{I}_{ExpChair}$ indicates whether the chair judge has succeeded in previous EUDC/WUDC tournaments.

$$S_{ik} = \alpha_{ik}\mathbb{I}_{FemSpeaker} + \theta_k\mathbb{I}_{ExpChair} + \beta_{ik}\mathbb{I}_{FemSpeaker}\mathbb{I}_{ExpChair} + \sum_{j=1}^n \gamma_j \mathbf{Y}_{ik} + (\eta_k) + (\eta_S) + \varepsilon_{ik}$$

Table 3.14 shows a large unconditional score gap between committees chaired by novice vs. accomplished judges. In particular, Column (2) shows that panels with accomplished chair judges gave 25.6 p.p SD higher scores compared to those chaired by novice judges. This is mainly due to the sorting of accomplished judges to higher-ranked debate rooms, as their past achievements are crucial for debate room allocation.

Columns (3) to (7) show no significant difference in terms of how accomplished chair judges evaluate speeches by male and female speakers. Nevertheless, within debates and for certain speakers, Column (8) found a small differential standard applied to female speakers. Compared to judge panels with novice chairs, on average, speeches given by women received 2.4 p.p. SD lower scores in debates adjudicated by panels with accomplished chairs.⁴⁶ These findings suggest that committees headed by accomplished chair judges are more generous towards all speakers, compared to those whose chair judges have no previous EUDC/WUDC achievements.

⁴⁶Further split of the analysis into Round 1 and Round 2 to Round 9 (controlling for debate room quality), as seen in Table 3.15 illustrate that, regardless of debate room quality, the score gap between committees headed by novice and accomplished chair judges remains significant.

3.7.1.2 Novice vs. Accomplished Chair Judges: Male vs. Female

Given the significant evaluation difference between novice and accomplished chairs, Table 3.16 investigates whether male and female accomplished chairs evaluate female speeches differently than novice ones. Column (2) shows that, on average, in comparison to committees with novice male chairs, accomplished chairs, both male and female, give higher scores to all speakers. This is largely due to the allocation of more accomplished judges into higher-quality rooms.

Noteworthy, after controlling for speaker fixed effects in Columns (5) and (6), I find that panels chaired by accomplished male chairs give higher scores to male speakers, but not to female speakers. No such patterns are found for panels with accomplished female chairs. In particular, controlling for speaker fixed effect and debate room characteristics, on average, committees headed by accomplished male chairs give 3.9 p.p SD lower scores to female speakers, compared to those headed by novice male chairs. No such evaluation gap is detected for committees chaired by novice or accomplished female judges. These results might indicate old boy's club gate-keeping practice by accomplished male chairs.

3.7.2 Do female judges evaluate speeches in higher vs. lower-ranked debates differently?

As the *tournament-for-judges* mechanism gives noisy performance feedback signals to judges in subsequent rounds,⁴⁷ it is crucial to understand evaluation patterns between male and female judges in higher- vs. lower-ranked debates. For every N^{th} round, I split the sample based on the *median* cumulative $(N - 1)$ round speech scores of two speakers in a team. Since the power-matching mechanism in EUDC and WUDC tournaments match teams with not only similar team records but also speak performance records, this split gives a good overview of team standings in any respective N^{th} round. Any debates with teams having higher than or equal to the *median threshold*⁴⁸ are classified as higher-ranked debates. The following subsections report regression results for chair judges and chair ft. wing gender composition in higher- and lower-ranked debates.

⁴⁷Judges do not know the exact room ranking. They can only noisily infer the assigned room level from the team names and speech quality during the debate.

⁴⁸Given the speech score scale of 50 to 100 in use, for any debate rounds, all debates with speeches scoring ≥ 76 (the median score) are classified as higher-ranked debates, and vice versa.

3.7.2.1 Chair Judges

Table 3.17 summarizes the results on how judge's experience and their gender matter in higher- vs. lower-ranked debates. Overall, the evaluation patterns found in the previous subsection hold for higher-ranked rooms, not lower-ranked debates. Specifically, in these rooms, committees chaired by novice female judges gave significantly lower scores to both male and female speakers, compared to those chaired by novice male judges. On the other hand, panels headed by an accomplished male judge are significantly more generous compared to novice male judges. Yet, they are significantly less so towards female speakers, giving 6.1 p.p SD lower score to female speakers compared to male speakers. In lower-ranked rooms, no significant gender gap in scores is detected.

For committees chaired by accomplished female judges, similar to accomplished male judges, these committees gave 7.4 p.p SD higher scores than those chaired by novice male judges. Yet, at the speaker fixed effect level, Columns (6) and (7) show no significant difference in scores between panels headed by accomplished female and novice male judges. In terms of evaluation standards between male and female speakers, Column (5) shows less generous, yet statistically insignificant score differences for speeches given by women from accomplished female chaired panels compared to those chaired by novice male chairs. This statistically insignificant gap shrinks further at the speaker fixed effect level in Columns (6) and (7), thus suggesting no gendered evaluation standards.

3.7.2.2 Chair ft. Wing Gender Composition

From Table 3.3, we already learn that more female judges on a committee are associated with lower evaluation scores. Table 3.18 shows that this pattern holds only in higher-ranked rooms. Specifically, compared to panels with male-majority wings and male chairs, panels with female-majority wings gave significantly lower scores, with a significantly larger gap for female-majority wings led by female judges, standing at 8.3 p.p SD. This harsher evaluation pattern stays significant across all specifications. For panels with female chairs paired with male-majority wings, at speaker fixed effect level, Columns (6) and (7) show that they gave significantly lower scores compared to panels with male chairs and male-majority wings.

With respect to evaluation standards between speeches given by men and women, I found that, at speaker fixed effect level, panels with female-majority wings are more lenient towards female speakers. Specifically, compared to debates adjudicated by male-majority wings and male chairs, female speeches judged by female-majority wings and male chairs get 0.68 p.p. SD higher scores than speeches of male speakers. This gap stands at 0.70 if the female-majority panel is led by a female judge. Overall, these findings confirm that, in higher-ranked debates, panels chaired by female judges give lower scores to all speakers. Even though panels with female-majority wing judges are positively preferential towards female speakers, the magnitude is insufficient to improve the success chances of female speakers. The unconditional gap between male and female speakers is only partially explained by committees chaired by accomplished male judges in high-ranked rooms, who give male speakers relatively higher scores than female speakers.

3.8 Conclusion

This chapter exploits the random matching of judges with varying decision power to adjudication committees of large-scale, high-stake debate tournaments to investigate the role of female judges in increasing the success chance of female speakers. My core finding is that the gender composition of committees has virtually no impact on the score gap between male and female speakers. This finding is in line with previous work on earlier debate tournaments by [Lundgren, 2017], i.e. there is no robust indication that gender biases account for the gender performance gap in these tournaments. Noteworthy, committees with female chair judges are associated with lower scores to all speakers, especially in higher-ranked debates, regardless of whether they are paired with male- or female-majority wing judges. Furthermore, the finding of slight differential treatment of accomplished male chair judges towards female speakers complies with the common gate-keeping mechanisms by the old boys' club at the top [Van den Brink and Benschop, 2014]. All in all, these findings raise doubts about the effectiveness of the gender quota law on breaking the glass ceiling for women in the highest offices. A gender quota law without accounting for potential biases and decision structures of committees might not deliver the intended outcome of reducing the "leaky pipeline" [Clark Blickenstaff, 2005], nor could it facilitate aspiring female voices.

Persuasion is a crucial skill to master, especially to climb the ladders of success in high-ranked, influential careers [Buser and Yuan, 2020; Deming, 2017; Fallows and Steven, 2000]. It is important to note that findings in this paper cannot exclude the possibility that the current evaluation standards reward more speech features that men are, on average, better at. In my related work investigating speech patterns between male and female debaters, I find significant differences across genders, with speeches by women exhibiting non-argumentative speech features that often correlates with lower confidence and credibility. One possibility is that norms around what constitutes persuasiveness are more male-skewed. Hence, even in the meticulously designed debate tournament structure to ensure an equitable and inclusive playground, such norms can be ingrained in evaluation standards, independent of the gender of the evaluator. Given that high-profile debate tournaments are a fruitful playground to learn the dos and don'ts of argumentation among future elites, these findings augment the null or negative impact of blanket gender quota implementation, without thorough consideration of systematic underlying issues around hiring and promotion e.g. the old boy's clubs [Cullen and Perez-Truglia, 2019], credit attribution in group work [Sarsons et al., 2021], and subjective assessment of potential abilities [Benson et al., 2021].

The current findings are qualified by two limitations that offer fruitful future research avenues. First, ex-ante and ex-post individual ranking decisions of committee members would allow us to understand how individual opinions translate into a final aggregate evaluation after deliberation. Text analysis of deliberation discussions of committees would further shed light on their receptiveness to being persuaded and the power dynamics across committee members. Second, the current work cannot disentangle whether the mechanism where speakers respond to judges' gender and one where female judges evaluate speeches differently for male and female speakers is responsible for the lower score patterns associated with female judges. Future work could create an experiment where speakers do not know the identity of judges to identify the responsible channels.

3.9 Appendix

3.9.1 Debater's Background, Institution Quota & Team Selection

Participants in these tournaments are undergraduate or graduate students who are active and dedicated in their respective debate societies. To prepare themselves for these prestigious tournaments, debaters participate in weekly meetings and travel to various local and international tournaments to sharpen their debate skills. The maximum number of eligible teams that an institution can send depends on three factors: (i) the hosting capacity of the organizing institution; (ii) whether they have teams making it to elimination round in the immediate past EUDC/WUDC;⁴⁹ and (iii) the number of first-time institutions requesting team slots at EUDC/WUDC.⁵⁰ In established debate societies at top-ranked institutions, a formal team trial competition often takes place to choose and fund the most competent speakers, who will commit to intensive training together until the respective EUDC/WUDC competition. In areas with less resourceful debate funds, participants in EUDC/WUDC tournaments are often debate enthusiasts and fund themselves in these competitions.

3.9.2 Data collection: Debate topics, Language skill, Institution & ranking

Language skill status. In EUDC/WUDC debate tournaments, individuals are classified into different language categories by an appointed independent language committees. This classification is meant to provide an inclusive playground to speakers with limited exposure to English language, which enables participants to break into open and/or non-native (ESL) speaker's league in knock-out rounds. The evaluation criteria are based on individual survey applications regarding: (i) the age at which they were exposed to English; and (ii) the content, structure and quality of English used for any relevant instruction or exchange.⁵¹ From the archival tab data, I documented 46.65 % of speeches given by non-native English

⁴⁹The more teams from their institutions have made it in the previous EUDC/WUDC tournament, the higher the number of teams they can send, up to a limit.

⁵⁰EUDC/WUDC tournaments have an inclusion policy, whereby they ensure that every institution who request team slots can be offered at least one team slot.

⁵¹For more detailed criteria to be qualified as ESL for EUDC and ESL & EFL for WUDC, see the Language Status section of [WUDC constitution](#) and [EUDC constitution](#).

speakers.

Debate topics. Across the 4896 debates, 72 unique debate motions discussed across a wide range of topics. All topics provided a balanced, in-depth but polarized distribution of views, as empirically tested by chief adjudicators in earlier regional competitions. I manually classified these motions into 17 debate topics, based on the classification at [International Debate Education Association](#), which are summarized by the distribution of debate speeches in Figure 3.4. Topics on society, international relations and military policies are the three most popular debate motions at these tournaments, followed closely by debates on the economy, law and justice systems, as well as topics on health, feminism and digital freedom.

Institution & Ranking. Since the academic institution that a speaker represents carries reputation/prestige that could impact evaluations, we collected institution information embedded in team name, in addition to registry data from tab masters, where possible. By pairing up speaker's identity with their team names, along with public social media and confirmation with the tab directors, we obtained 513 distinct institutions across 83 countries in this data set. Since there exists no university ranking given their debate achievements,⁵² these institutions are categorized by their average academic ranking from QS World Universities Ranking from 2013 to 2017 into two groups: top-50-ranked and the non-top-50-ranked universities. The breakdown of institutions in Appendix 3.2 shows that participants affiliated with top-50-ranked institutions account for roughly 10 - 20% of all participants, with the slight exception of WUDC 2017 and WUDC 2018, where this proportion is above 20%. Among these variables, more male speakers tend to represent top-50-ranked institutions and be native English speakers, as noted in the correlation heat map in Figure 3.5.

⁵²Apart from a top 5 and top 10 list of UK & UK universities to master debate skills in the [US](#) and [UK](#)

3.9.3 Figures

3.9.3.1 Proportion of top-ranked institutions over competitions

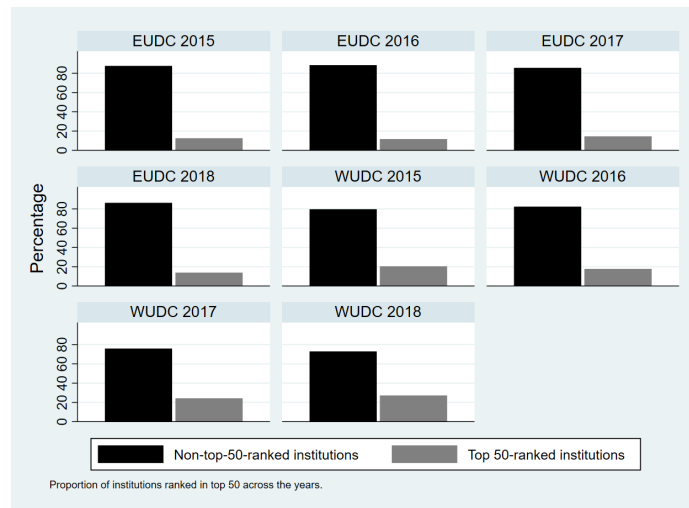


Figure 3.2: Proportion of top-50-ranked vs. non-top-50-ranked institutions (2015 - 2018)

3.9.3.2 Proportion of male vs. female speakers across institutions



Figure 3.3: Proportion of male vs. female speakers across participating institutions worldwide. The larger the circle, the more participants and institutions from that country represented in the tournaments. Blue refers to male speakers, whereas red refers to female speakers.

3.9.3.3 Distribution of speeches across debate topics

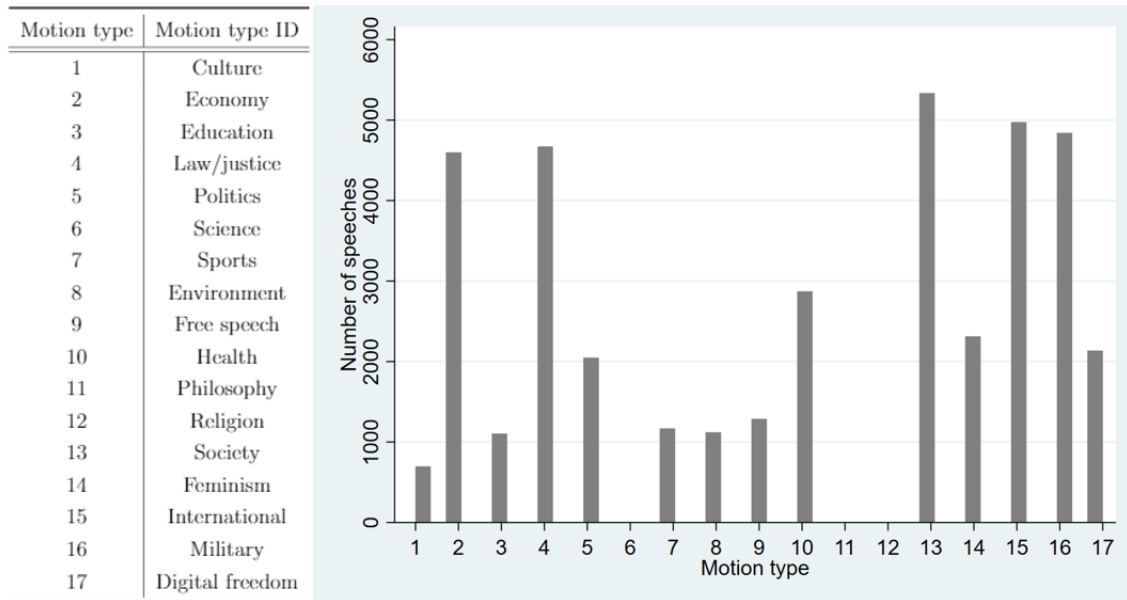


Figure 3.4: Distribution of speeches given motion types

3.9.3.4 Correlation matrix between judges and speakers' characteristics

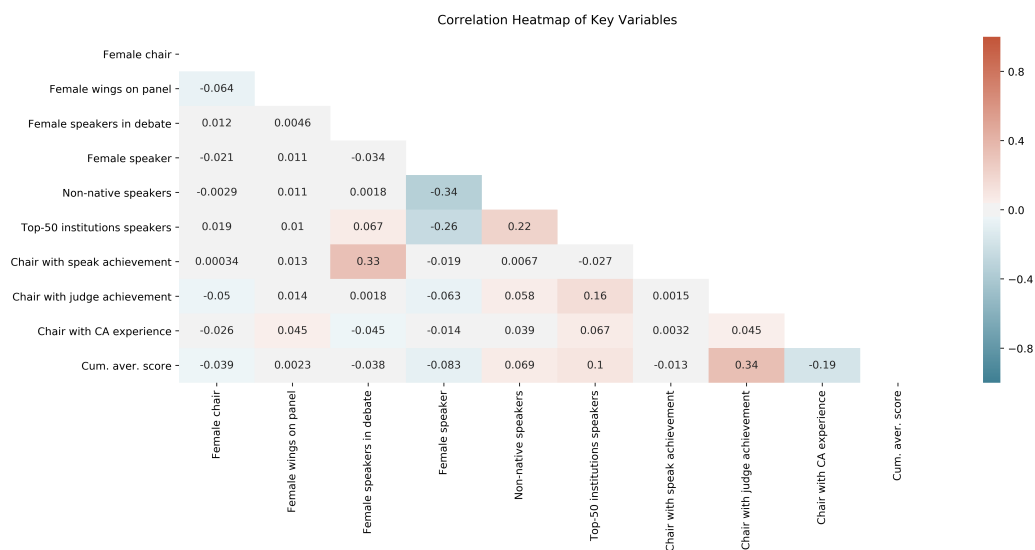


Figure 3.5: Correlation heat map across characteristics of judges and speakers

3.9.3.5 Speech Score Distribution: Male vs. female speakers

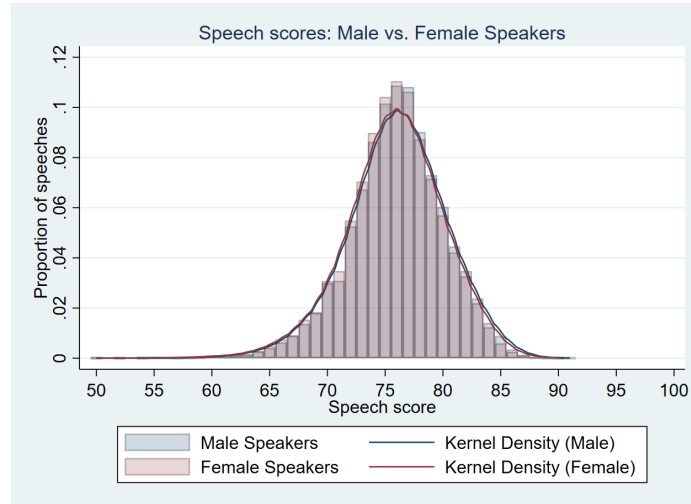


Figure 3.6: Speech score distribution by speaker's gender

3.9.3.6 Speech Score Distribution: Male vs. female chair judges

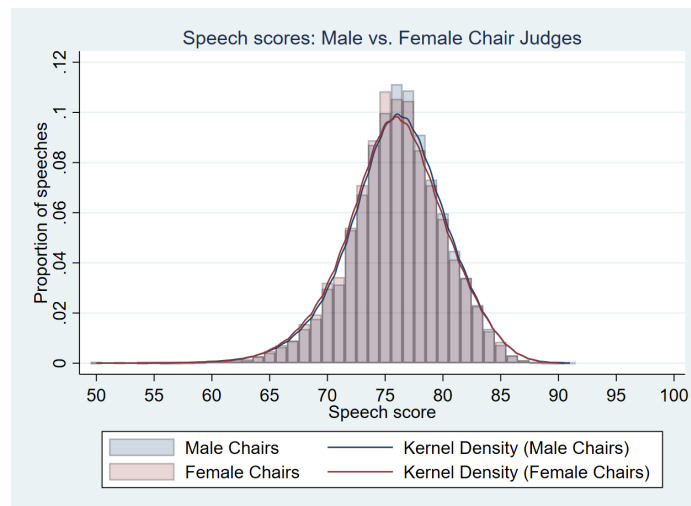


Figure 3.7: Distribution of speech scores, by chair judge's gender

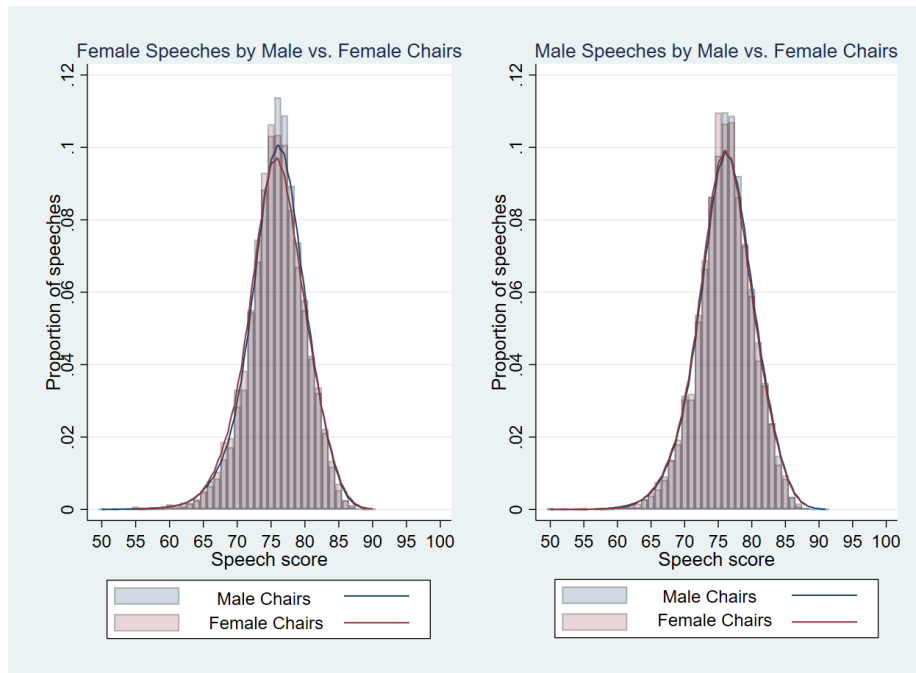


Figure 3.8: Distribution of speech scores for men vs. women, given chair judge's gender

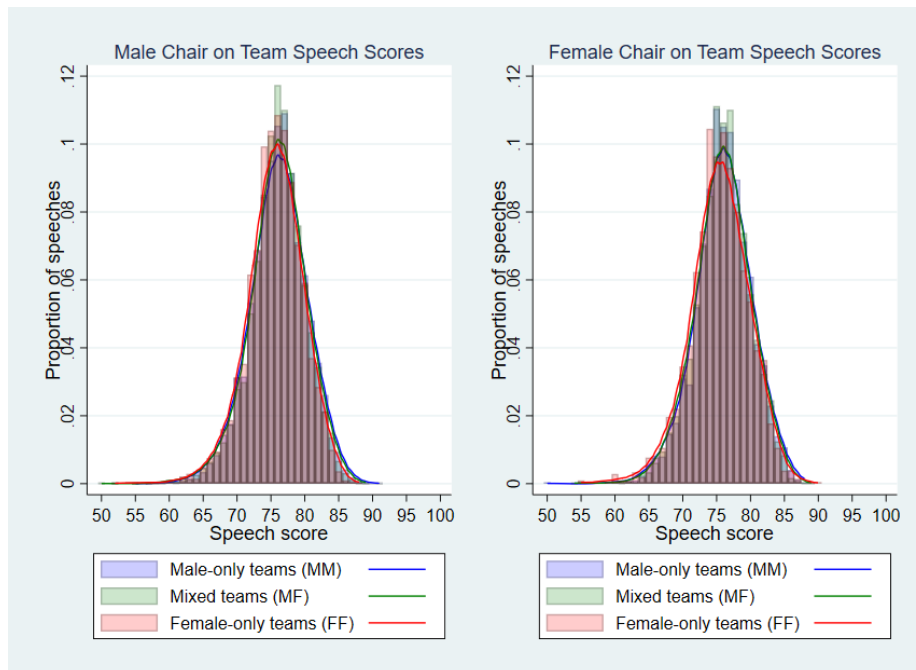


Figure 3.9: Distribution of speech scores given gender composition of teams

3.9.3.7 Speech Score Distribution: Male- vs. female-majority panels

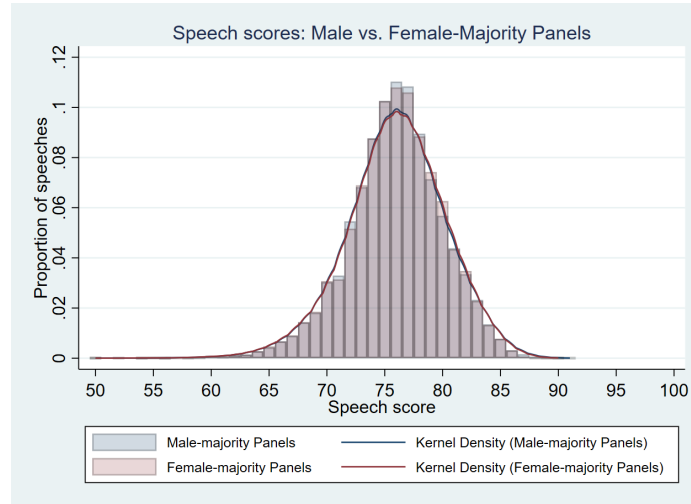


Figure 3.10: Distribution of speech scores by panel gender composition

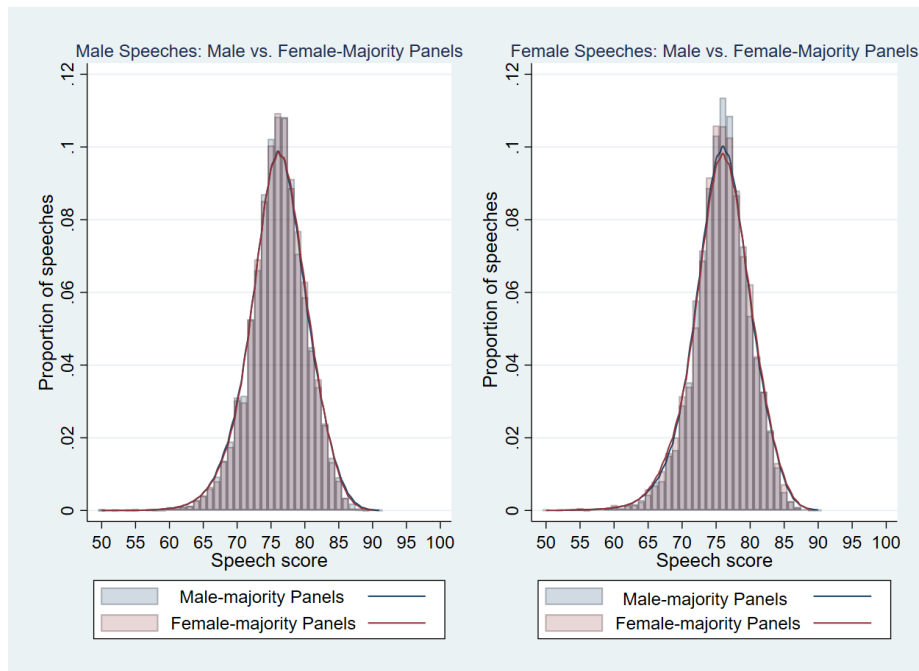


Figure 3.11: Distribution of speech scores for men vs. women, given chair panel gender composition

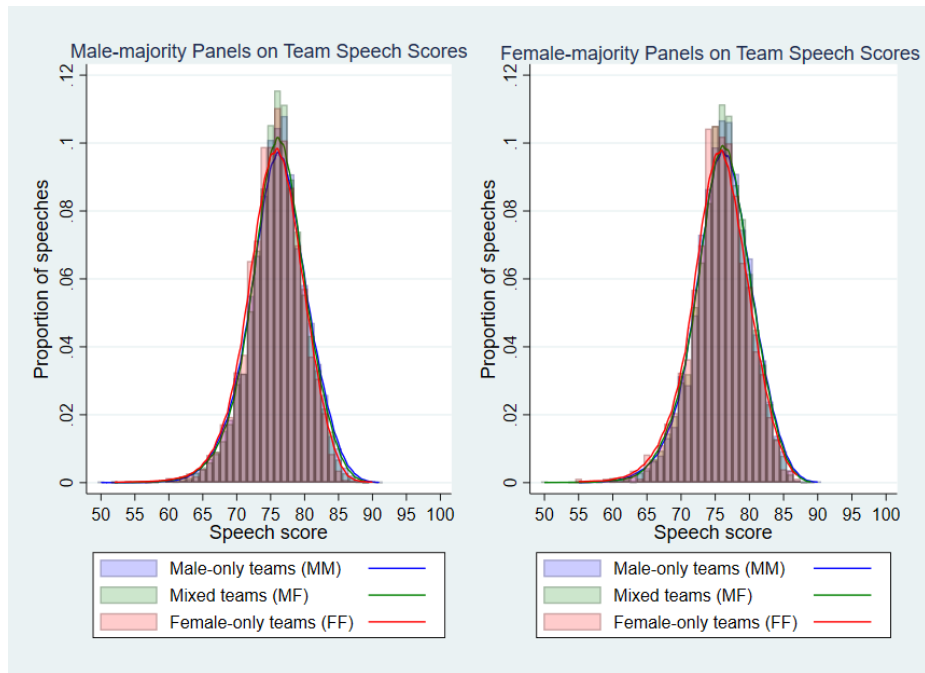


Figure 3.12: Distribution of speech scores given gender composition of teams

3.9.3.8 Speech Score Distribution: Novice vs. Experienced chair judges

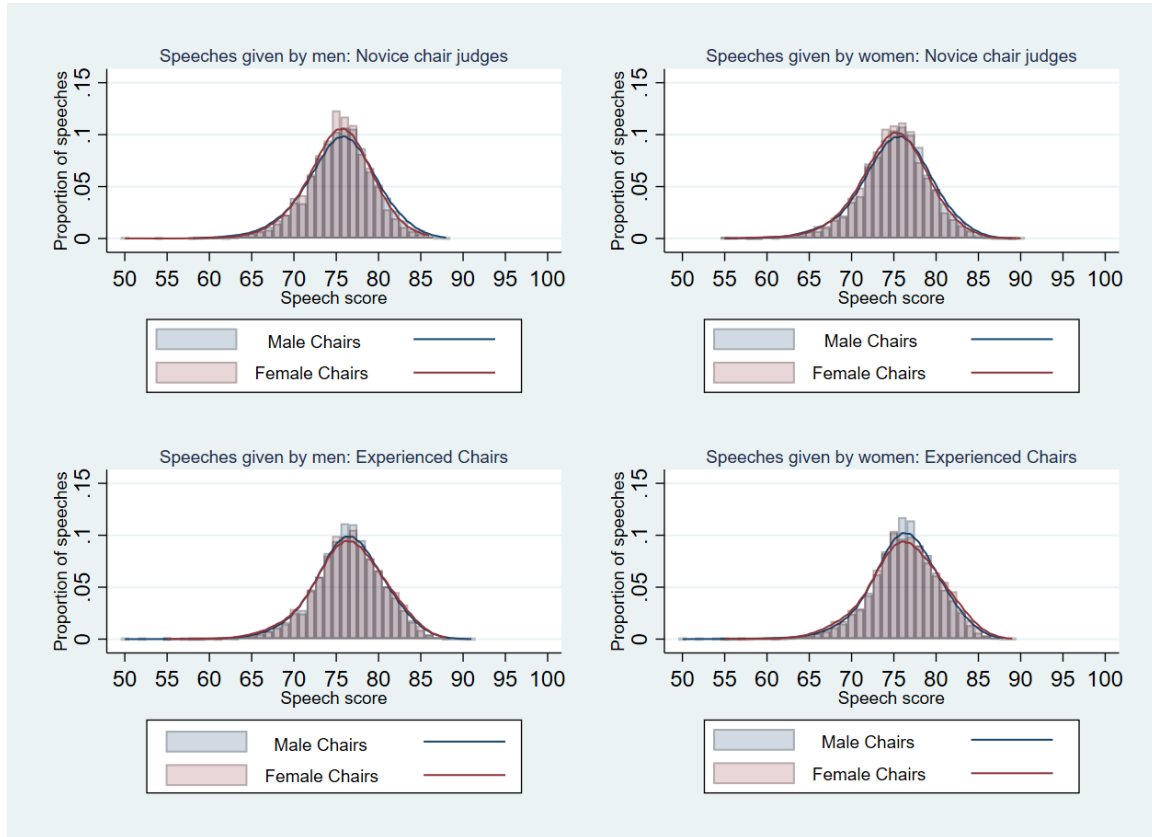


Figure 3.13: Histogram distribution of speech scores given past EUDC/WUDC achievements of chair judges. Chair judges are classified as experienced if they have ≥ 1 EUDC/WUDC speaking or judging achievements.

3.9.3.9 Proportion of speeches adjudicated by male vs. female chair judges

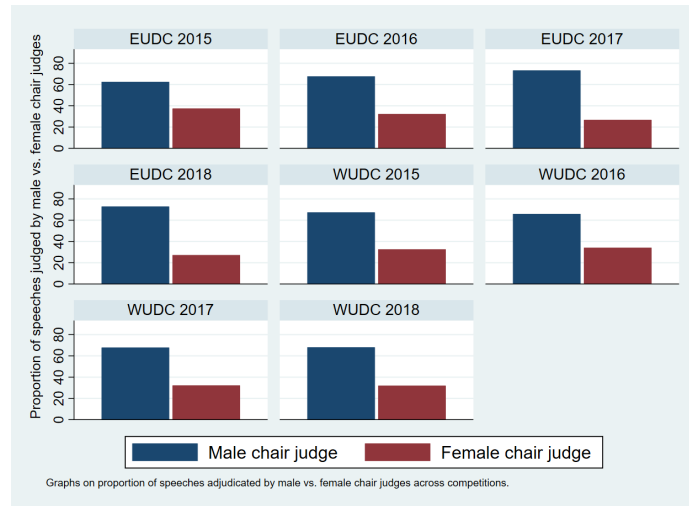


Figure 3.14: Proportion of speeches by chair judge’s gender across competitions

3.9.3.10 Proportion of speeches adjudicated by male vs. female-majority judge panels

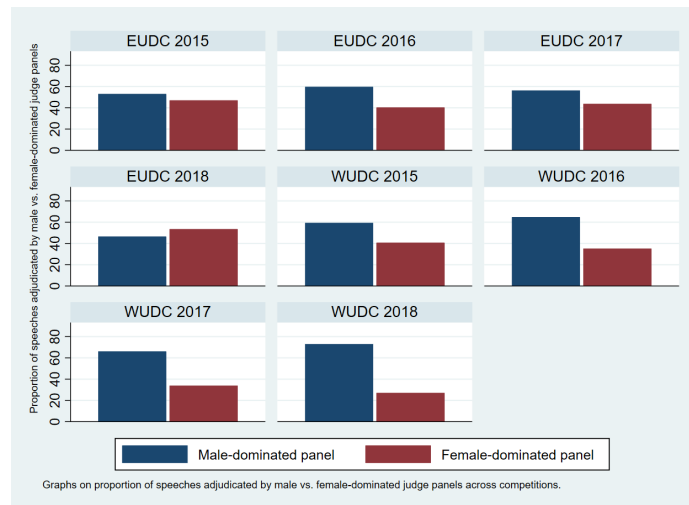


Figure 3.15: Proportion of speeches by judge panel gender composition across competitions

3.9.3.11 Proportion of male vs. female speakers across competitions

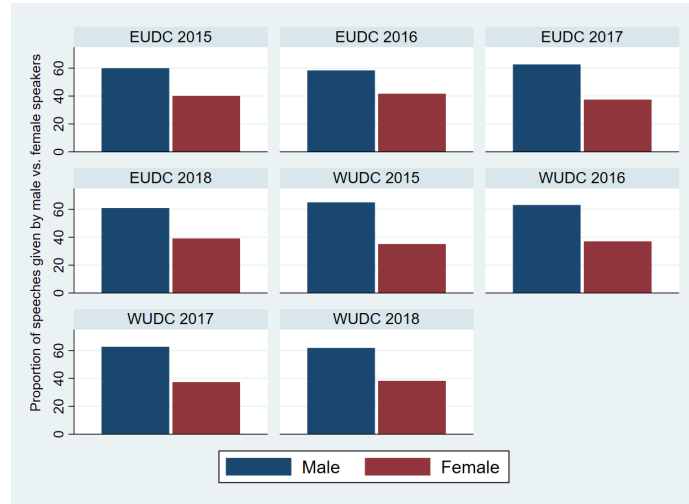


Figure 3.16: Number of male vs. female speakers per competitions ($N_{male} = 24334$, $N_{female} = 14834$)

3.9.3.12 Judge panel size & panel gender composition

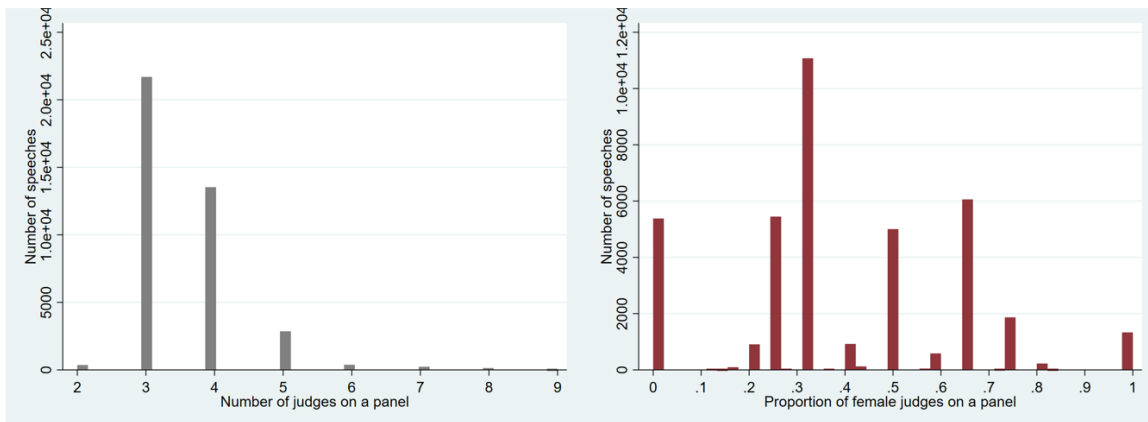


Figure 3.17: Total number of judges (left) and proportion of female judges (right) on an adjudication panel

3.9.3.13 Cumulative distribution of speeches given previous EUDC/WUDC achievements of chair judges

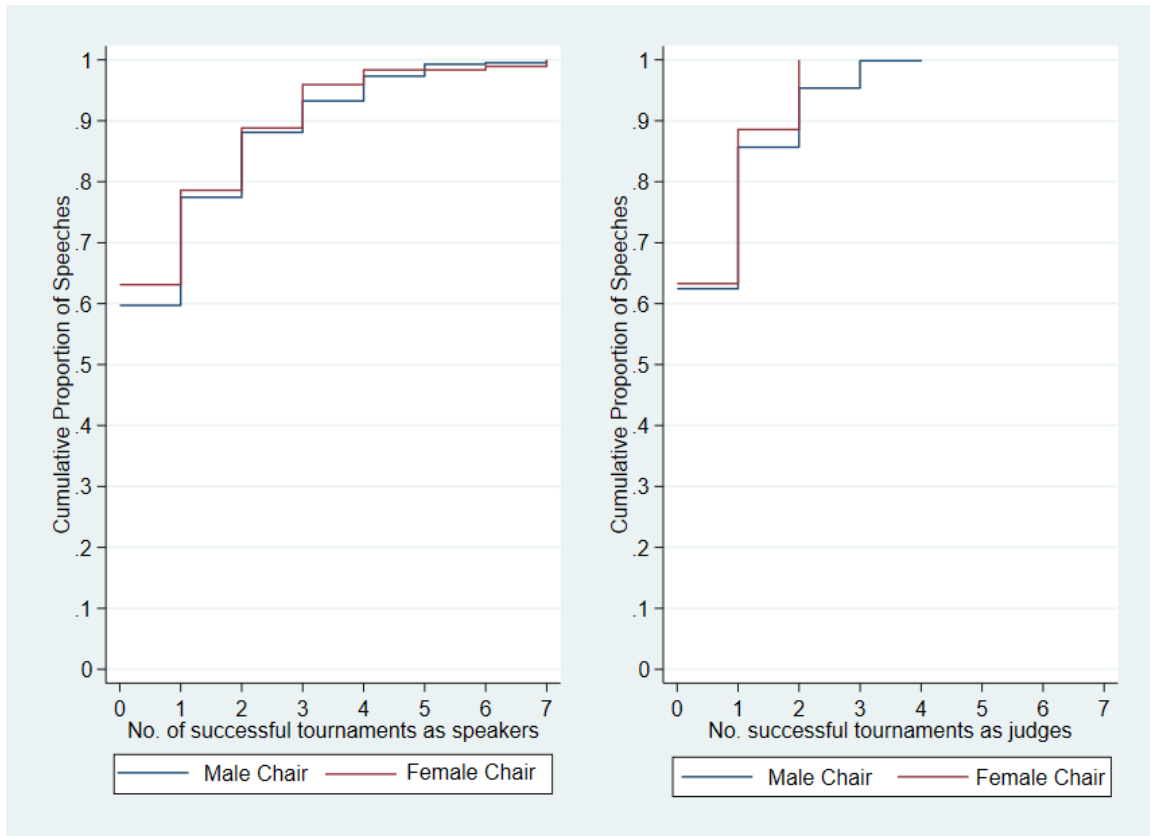


Figure 3.18: Cumulative distribution of speeches given past achievement of chair judges. The left graph refers to the number of tournaments that a chair judge has advanced to elimination round as a speaker, whereas the right graph is for judging.

3.9.4 Tables

3.9.4.1 Number of debates per tournament and omitted debates

Table 3.4: Speeches & debates by competition and missing debates

Tournament	Number of speeches	Number of debates	Omitted debates
EUDC 2015	3760	470	6
EUDC 2016	3968	496	9
EUDC 2017	3736	467	12
EUDC 2018	3064	383	35
WUDC 2015	6096	762	39
WUDC 2016	6776	847	16
WUDC 2017	6440	805	32
WUDC 2018	5328	666	36
Total	39168	4896	185

3.9.4.2 Descriptive statistics

Table 3.5: Tournament score descriptive statistics ($N = 39168$ speeches)

Competition code	Mean	Min	Max	Median	SD	Total speeches
WUDC15	76.04	58	90	76	4.09	6096
WUDC16	75.67	52	88	76	4.01	6776
WUDC17	76.47	58	88	77	3.83	6440
WUDC18	76.48	59	88	77	3.95	5328
EUDC15	74.96	54	89	75	4.46	3760
EUDC16	75.63	52	91	76	4.45	3968
EUDC17	75.98	55	88	76	4.03	3736
EUDC18	76.07	50	87	76	3.99	3064
WUDC total	76.14	52	90	76	3.99	24640
EUDC total	75.64	50	91	76	4.28	14528

Table 3.6: Descriptive statistics of speeches by control variables given the gender of chair judges ($N_{speech} = 39168$, $N_{MaleChair} = 440$, $N_{FemaleChair} = 241$)

Control variables	Speeches adjudicated by panels with...				Total	%
	male chair judges		female chair judges			
	Count	%	Count	%		
Chair Judge Accomplishment						
No EUDC/WUDC achievement	10144	38.18	5496	43.62	15640	39.93
One EUDC/WUDC achievement	5864	22.07	2408	19.11	8272	21.12
Multi-time EUDC/WUDC achievement	10560	39.75	4696	37.27	15256	38.95
Speaker Characteristics						
Male	16509	62.14	7825	62.10	24334	62.13
Female	10059	37.86	4775	37.90	14834	37.87
Non-native	12589	47.38	5683	45.10	18272	46.65
Native	13979	52.62	6917	54.90	20896	53.35
Top-50-ranked institutions	5000	18.82	2341	18.58	7341	18.74
Non-top-50-ranked institutions	21568	81.18	10259	81.43	31827	81.26
Room & Wing Gender Composition						
Female-majority Room	8704	32.76	4256	33.78	12960	33.09
Male-majority Room	17864	67.24	8344	66.22	26208	66.91
Female-majority Wing Judges	15560	58.57	6752	53.59	22312	56.96
Male-majority Wing Judges	11008	41.43	5848	46.41	16856	43.04
Tournament & year						
EUDC 2015	2352	8.85	1408	11.17	3760	9.60
EUDC 2016	2688	10.12	1280	10.16	3968	10.13
EUDC 2017	2736	10.30	1000	7.94	3736	9.54
EUDC 2018	2232	8.40	832	6.60	3064	7.82
WUDC 2015	4104	15.45	1992	15.81	6096	15.56
WUDC 2016	4464	16.80	2312	18.35	6776	17.30
WUDC 2017	4368	16.44	2072	16.44	6440	16.44
WUDC 2018	3624	13.64	1704	13.52	5328	13.60
Motion topic type						
Culture	512	1.93	184	1.46	696	1.78
Economy	3176	11.95	1424	11.30	4600	11.74
Education	800	3.01	304	2.41	1104	2.82
Law/justice	3000	11.29	1672	13.27	4672	11.93
Politics	1440	5.42	608	4.83	2048	5.23
Sports	792	2.98	376	2.98	1168	2.98
Environment	736	2.77	384	3.05	1120	2.86
Free Speech	872	3.28	416	3.30	1288	3.29
Health	1992	7.50	880	6.98	2872	7.33
Society	3568	13.43	1768	14.03	5336	13.62
Feminism	1624	6.11	688	5.46	2312	5.90
International Relations	3344	12.59	1632	12.95	4976	12.70
Military	3288	12.38	1552	12.32	4840	12.36
Digital Freedom	1424	5.36	712	5.65	2136	5.45
TOTAL SPEECHES	26568	67.83	12600	32.17	39168	100

Table 3.7: Descriptive statistics of speeches by control variables given panel gender composition ($N_{speech} = 39168$)

Control variables	Speeches adjudicated by...				Total	%
	male-majority panels		female-majority panels			
	Count	%	Count	%		
Chair Judge Accomplishment						
No EUDC/WUDC achievement	9528	39.63	6112	40.40	15640	39.93
One EUDC/WUDC achievement	5296	22.03	2976	19.67	8272	21.12
Multi-time EUDC/WUDC achievement	9216	38.34	6040	39.93	15256	38.95
Speaker Characteristics						
Male	15023	62.49	9311	61.55	24334	62.13
Female	9017	37.51	5817	38.45	14834	37.87
Native	12825	53.35	8071	53.35	20896	53.35
Non-native	11215	46.65	7057	46.65	18272	46.65
Top-50-ranked institutions	4443	18.48	2898	19.16	7341	18.74
Non-top-50-ranked institutions	19597	81.52	12230	80.84	31827	81.26
Room & Chair Gender Composition						
Female-majority Room	7760	32.28	5200	34.37	12960	33.09
Male-majority Room	16280	67.72	9928	65.63	26208	66.91
Female Chair Judges	3944	58.57	8656	53.59	22312	56.96
Male Chair Judges	20096	41.43	6472	46.41	16856	43.04
Tournament & year						
EUDC 2015	1992	8.29	1768	11.69	3760	9.60
EUDC 2016	2368	9.85	1600	10.58	3968	10.13
EUDC 2017	2104	8.75	1632	10.79	3736	9.54
EUDC 2018	1420	5.91	1640	10.84	3064	7.82
WUDC 2015	3616	15.04	2480	16.39	6096	15.56
WUDC 2016	4392	18.27	2384	15.76	6776	17.30
WUDC 2017	4256	17.70	2184	14.44	6440	16.44
WUDC 2018	3888	16.17	1440	9.52	5328	13.60
Motion topic type						
Culture	320	1.33	376	2.49	696	1.78
Economy	2992	12.45	1608	10.63	4600	11.74
Education	664	2.76	440	2.91	1104	2.82
Law/justice	3000	12.48	1672	11.05	4672	11.93
Politics	1288	5.36	760	5.02	2048	5.23
Sports	696	2.90	472	3.12	1168	2.98
Environment	616	2.56	504	3.33	1120	2.86
Free Speech	744	3.09	544	3.60	1288	3.29
Health	1688	7.02	1184	7.83	2872	7.33
Society	3376	14.04	1960	12.96	5336	13.62
Feminism	1480	6.16	832	5.49	2312	5.90
International Relations	2904	12.08	2072	13.70	4976	12.70
Military	2960	12.31	1880	12.43	4840	12.36
Digital Freedom	1312	5.46	824	5.45	2136	5.45
TOTAL SPEECHES	24040	61.38	15128	38.62	39168	100

Table 3.8: Descriptive statistics of control variables by speaker's gender ($N_{speech} = 39168$, $N_{MaleSpeaker} = 1949$, $N_{FemaleSpeaker} = 1190$)

Control variables	Speeches given by...				Total %	
	male speakers		female speakers			
	Count	%	Count	%		
Speaker's characteristics						
Non-native speakers	11530	47.38	6742	45.45	18272	46.65
Native speakers	12804	52.62	8092	54.55	20896	53.35
Top-50-ranked institutions	4511	18.54	2830	19.08	7341	18.74
Non-50-ranked institutions	19823	81.46	12004	80.92	31827	81.26
Female debate partner	8640	35.51	6194	41.76	14834	37.87
Male debate partner	15694	64.49	8640	58.24	24334	62.13
Room & Judge Gender Composition						
Female-majority Room	5533	22.74	7427	50.07	12960	33.09
Male-majority Room	18801	77.26	7407	49.93	26208	66.91
Female chair Judge	7825	32.16	4775	32.19	12600	32.17
Male chair Judge	16509	67.84	10059	67.81	26568	67.83
Female-majority Panel	9311	38.26	5817	39.21	15128	38.62
Male-majority Panel	15023	61.74	9017	60.79	24040	61.38
Tournament & year						
EUDC 2015	2254	9.26	1506	10.15	3760	9.60
EUDC 2016	2317	9.52	1651	11.13	3968	10.13
EUDC 2017	2338	9.61	1398	9.42	3736	9.54
EUDC 2018	1866	7.67	1198	8.08	3064	7.82
WUDC 2015	3960	16.27	2136	14.40	6096	15.56
WUDC 2016	4272	17.56	2504	16.88	6776	17.30
WUDC 2017	4035	16.58	2405	16.21	6440	16.44
WUDC 2018	3292	13.53	2036	13.73	5328	13.60
Motion topic type						
Culture	424	1.74	272	1.83	696	1.78
Economy	2915	11.98	1685	11.36	4600	11.74
Education	704	2.89	400	2.70	1104	2.82
Law/justice	2876	11.82	1796	12.11	4672	11.93
Politics	1259	5.17	789	5.32	2048	5.23
Sports	710	2.92	458	3.09	1168	2.98
Environment	705	2.90	415	2.80	1120	2.86
Free Speech	766	3.15	522	3.52	1288	3.29
Health	1780	7.31	1092	7.36	2872	7.33
Society	3285	13.49	2051	13.83	5336	13.62
Feminism	1450	5.96	862	5.81	2312	5.90
International Relations	3082	12.67	1894	12.76	4976	12.70
Military	3039	12.49	1801	12.14	4840	12.36
Digital Freedom	1339	5.50	797	5.35	2136	5.45
TOTAL SPEECHES	24334	62.13	14834	37.87	39168	100

Table 3.9: (JUDGES) Two sample t-test with unequal variances on speech scores given speaker's genders across gender composition of judge panels ($N_{MaleChair} = 440$, $N_{FemaleChair} = 241$, $N_{Speech} = 39618$)

Group Variable	Mean _M	Mean _F	SD _M	SD _F	t-test	p-value
Female Chair Judge	75.95	75.70	4.08	4.19	3.30	0.00***
Male Chair Judge	76.09	75.87	4.10	4.08	4.29	0.00***
Female-majority Panels	76.06	75.81	4.07	4.17	3.54	0.00***
Male-majority Panels	76.03	75.81	4.11	4.08	4.10	0.00***

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

Table 3.10: (SPEAKERS) Two sample t-test with unequal variances on speech scores across demographics ($N_{Speech} = 39168$, $N_{MaleSpeaker} = 1949$, $N_{FemaleSpeaker} = 1190$)

Group Variable	Mean _M	Mean _F	SD _M	SD _F	t-test	p-value
Speaker Gender	76.04	75.81	4.09	4.12	5.41	0.00***
Non-native Speakers	74.63	74.21	3.86	3.96	7.01	0.00***
Native Speakers	77.32	77.15	3.87	3.75	3.11	0.00***
Top-50-ranked Institutions	78.71	78.24	3.48	3.52	5.56	0.00***
Non-top-50-ranked Institutions	75.44	75.24	3.98	4.04	4.25	0.00***

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

3.9.4.3 Results: Round 1s vs. Round 2s to 9s

Table 3.11: Regression Analysis of Chair Judge Gender against Speech Score, Round 1s vs. Round 2s to 9s

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round 1s								
Female Speaker	-0.083*** (0.03)		-0.094*** (0.03)	-0.087*** (0.03)			-0.057* (0.03)	
Female Chair		-0.038 (0.04)	-0.050 (0.05)	-0.059 (0.04)	-0.090* (0.05)	-0.073 (0.05)		
Female Speaker × Female Chair			0.031 (0.07)	0.009 (0.07)	0.110 (0.08)	0.094 (0.08)	0.035 (0.07)	0.068 (0.08)
R^2	0.002	0.000	0.003	0.188	0.668	0.713	0.266	0.870
Observations	4376	4376	4376	4376	2172	2172	4376	2120
Round 2s to 9s								
Female Speaker	-0.030*** (0.01)		-0.024* (0.01)	-0.035*** (0.01)			-0.005 (0.01)	
Female Chair		-0.052** (0.02)	-0.046* (0.02)	-0.034 (0.02)	-0.050*** (0.02)	-0.048*** (0.02)		
Female Speaker × Female Chair			-0.018 (0.02)	-0.011 (0.02)	0.004 (0.02)	0.006 (0.02)	-0.018 (0.02)	-0.003 (0.02)
R^2	0.171	0.171	0.171	0.289	0.573	0.592	0.617	0.773
Observations	34792	34792	34792	34792	34786	34786	34792	34786
Speaker Controls				✓				
Chair Experience				✓		✓		
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓

All analyses in Round 2s to Round 9s control for average cumulative team standings.

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) round, (vi) motion type, (viii) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$.

Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.12: Regression Analysis of Panel Gender Composition against Speech Score, Round 1s vs. Round 2s to 9s

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round 1s								
Female Speaker	-0.083*** (0.03)		-0.107*** (0.04)	-0.082** (0.03)				-0.064* (0.04)
Female-majority Panel		-0.022 (0.04)	-0.045 (0.05)	0.006 (0.05)	-0.009 (0.05)	0.021 (0.06)		
Female Speaker × Female-majority Panel			0.062 (0.06)	-0.004 (0.06)	0.029 (0.07)	0.041 (0.07)	0.046 (0.06)	0.024 (0.07)
R^2	0.002	0.000	0.002	0.188	0.667	0.711	0.266	0.870
Observations	4376	4376	4376	4376	2172	2172	4376	2120
Round 2s to 9s								
Female Speaker	-0.030*** (0.01)		-0.023* (0.01)	-0.033*** (0.01)				-0.010 (0.01)
Female-majority Panel		0.008 (0.02)	0.015 (0.02)	0.022 (0.02)	-0.020 (0.02)	-0.006 (0.02)		
Female Speaker × Female-majority Panel			-0.019 (0.02)	-0.017 (0.02)	-0.004 (0.02)	0.000 (0.02)	-0.002 (0.02)	-0.014 (0.01)
R^2	0.171	0.171	0.171	0.289	0.572	0.592	0.617	0.773
Observations	34792	34792	34792	34792	34786	34786	34792	34786
Chair Gender & Experience				✓		✓		
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓

All analyses in Round 2s to Round 9s control for average cumulative team standings.

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) round, (vi) motion type, (viii) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$. Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.13: Regression Analysis of Chair ft. Wing Gender Composition against Speech Score, Round 1s vs. Round 2s to 9s

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round 1s								
Female Speaker	-0.083*** (0.03)		-0.139** (0.06)	-0.103** (0.05)			-0.099* (0.05)	
M Chair ft. F-majority Wings		0.029 (0.05)	0.003 (0.05)	0.023 (0.05)	0.136** (0.06)	0.113* (0.06)		
F Chair ft. M-majority Wings		-0.056 (0.06)	-0.085 (0.07)	-0.097 (0.07)	-0.087 (0.08)	-0.080 (0.07)		
F Chair ft. F-majority Wings		0.007 (0.06)	-0.021 (0.06)	-0.009 (0.06)	0.028 (0.07)	0.043 (0.07)		
Female Speaker × M Chair ft. F-majority Wings			0.075 (0.07)	0.028 (0.06)	-0.029 (0.08)	-0.022 (0.08)	0.069 (0.07)	0.108 (0.08)
Female Speaker × F Chair ft. M-majority Wings			0.083 (0.10)	0.080 (0.10)	0.219* (0.12)	0.203* (0.12)	0.084 (0.10)	0.290** (0.12)
Female Speaker × F Chair ft. F-majority Wings			0.075 (0.11)	-0.017 (0.09)	-0.009 (0.11)	-0.008 (0.11)	0.072 (0.10)	0.027 (0.10)
R^2	0.002	0.001	0.003	0.189	0.670	0.714	0.266	0.871
Observations	4376	4376	4376	4376	2172	2172	4376	2120
Round 2s to 9s								
Female Speaker	-0.030*** (0.01)		-0.018 (0.02)	-0.034* (0.02)			-0.006 (0.02)	
M Chair ft. F-majority Wings		-0.065** (0.03)	-0.061** (0.03)	-0.057** (0.02)	-0.021 (0.02)	-0.019 (0.02)		
F Chair ft. M-majority Wings		-0.081** (0.03)	-0.085** (0.04)	-0.075** (0.03)	-0.062** (0.03)	-0.061** (0.02)		
F Chair ft. F-majority Wings		-0.099*** (0.03)	-0.078** (0.03)	-0.056* (0.03)	-0.063*** (0.02)	-0.057** (0.02)		
Female Speaker × M Chair ft. F-majority Wings			-0.011 (0.03)	-0.002 (0.02)	0.008 (0.02)	0.010 (0.02)	0.001 (0.02)	0.004 (0.02)
Female Speaker × F Chair ft. M-majority Wings			0.011 (0.03)	0.013 (0.03)	-0.001 (0.03)	0.000 (0.03)	-0.001 (0.03)	0.011 (0.02)
Female Speaker × F Chair ft. F-majority Wings			-0.056* (0.03)	-0.035 (0.03)	0.017 (0.03)	0.022 (0.03)	-0.032 (0.03)	-0.011 (0.02)
R^2	0.171	0.172	0.172	0.289	0.573	0.592	0.617	0.773
Observations	34792	34792	34792	34792	34786	34786	34792	34786
Chair Experience				✓		✓		
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓

All analyses in Round 2s to Round 9s control for average cumulative team standings.

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) round, (vi) motion type,

(viii) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$.

Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.9.4.4 Extension Results

Table 3.14: (OVERALL) Regression Analysis of Chair Judge Experience against Speech Score given Speaker's Gender (N = 39 168)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Speaker	-0.056*** (0.01)		-0.072*** (0.02)	-0.067*** (0.02)			-0.007 (0.01)	
Accomplished Chair Judge		0.256*** (0.02)	0.246*** (0.02)	0.220*** (0.02)	0.047*** (0.02)	0.031** (0.02)		
Female Speaker × Accomplished Chair Judge			0.027 (0.02)	0.007 (0.02)	-0.022 (0.02)	-0.022 (0.02)	-0.013 (0.02)	-0.024* (0.01)
Chair Gender	✓	✓	✓	✓	✓	✓		
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.001	0.016	0.017	0.183	0.568	0.588	0.587	0.768
Observations	39168	39168	39168	39168	39157	39157	39168	39157

Accomplished Chair Judge is a dummy variable indicating whether the judge has advanced to elimination rounds as a Speaker or a Judge at past EUDC/WUDC tournaments. Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include:

(i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) motion type, (vi) round, (vii) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$.

Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.15: (ROUND 1s vs. ROUND 2s to 9s) Regression Analysis of Chair Judge Experience against Speech Score given Speaker's Gender (controlling for team standing)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Round 1s								
Female Speaker	-0.083*** (0.03)		-0.073 (0.05)	-0.066 (0.04)				-0.018 (0.05)
Accomplished Chair Judge		-0.006 (0.04)	0.001 (0.04)	0.013 (0.05)	0.080 (0.06)	-0.022 (0.06)		
Female Speaker × Accomplished Chair Judge			-0.017 (0.06)	-0.032 (0.06)	0.021 (0.08)	0.049 (0.07)	-0.048 (0.06)	0.003 (0.07)
R^2	0.002	0.000	0.002	0.188	0.667	0.713	0.266	0.870
Observations	4376	4376	4376	4376	2172	2172	4376	2120
Round 2s to 9s								
Female Speaker	-0.030*** (0.01)		-0.037** (0.02)	-0.036** (0.02)				-0.006 (0.01)
Accomplished Chair Judge		0.171*** (0.02)	0.167*** (0.02)	0.152*** (0.02)	0.046*** (0.02)	0.032** (0.02)		
Female Speaker × Accomplished Chair Judge			0.011 (0.02)	-0.004 (0.02)	-0.028 (0.02)	-0.028 (0.02)	-0.009 (0.02)	-0.029* (0.02)
R^2	0.171	0.177	0.177	0.289	0.573	0.592	0.617	0.774
Observations	34792	34792	34792	34792	34786	34786	34792	34786
Chair Gender	✓	✓	✓	✓	✓	✓		
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓

All models in Round 2s to 9s control for average cumulative team standings.

Accomplished Chair Judge is a dummy variable indicating whether the judge has advanced to elimination rounds as a Speaker or a Judge at past EUDC/WUDC tournaments. Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include:

(i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) motion type, (vi) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$.

Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.16: Regression Analysis of Chair Judge Gender and Experience against Speech Score given Speaker's Gender (N = 39 168)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Female Speaker	-0.056*** (0.01)		-0.061*** (0.02)	-0.046** (0.02)				
Novice Female Chair		-0.055 (0.03)	-0.043 (0.04)	-0.038 (0.03)	-0.022 (0.02)	-0.019 (0.02)	-0.019 (0.02)	
Accomplished Male Chair		0.236*** (0.03)	0.233*** (0.03)	0.217*** (0.03)	0.060*** (0.02)	0.047** (0.02)	0.047** (0.02)	
Accomplished Female Chair		0.239*** (0.03)	0.226*** (0.04)	0.194*** (0.03)	-0.008 (0.02)	-0.020 (0.02)	-0.020 (0.02)	
Female Speaker × Novice Female Chair			-0.028 (0.04)	-0.011 (0.03)	-0.024 (0.03)	-0.025 (0.03)	-0.025 (0.03)	-0.023 (0.02)
Female Speaker × Accomplished Male Chair			0.008 (0.03)	-0.001 (0.02)	-0.041** (0.02)	-0.039** (0.02)	-0.039** (0.02)	-0.037** (0.02)
Female Speaker × Accomplished Female Chair			0.034 (0.03)	0.004 (0.03)	-0.012 (0.02)	-0.010 (0.02)	-0.010 (0.02)	-0.023 (0.02)
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.001	0.016	0.017	0.182	0.568	0.588	0.588	0.771
Observations	39168	39168	39168	39168	39157	39157	39157	39157

Accomplished Chair Judge is a dummy variable indicating whether the judge has advanced to elimination rounds as a Speaker or a Judge at past EUDC/WUDC tournaments. Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (iii) wing judge gender composition, (iv) speaking position, (v) motion type, (vi) round, (vi) competition & year. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is $R^2_{between}$. Singleton observations are dropped in model (5), (6) and (8). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.17: Regression Analysis of Chair Judge Gender & Experience against Speech Score: Higher vs. Lower-ranked Debates (N = 34 788)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Higher-ranked Rooms								
Female Speaker	-0.034*** (0.01)		0.014 (0.03)	-0.004 (0.03)	0.025 (0.02)			
Novice Female Chair		-0.119*** (0.04)	-0.099** (0.04)	-0.062 (0.04)		-0.068* (0.03)	-0.055 (0.03)	
Accomplished Male Chair		0.069** (0.03)	0.096*** (0.03)	0.119*** (0.03)		0.097*** (0.03)	0.079*** (0.02)	
Accomplished Female Chair		0.074** (0.04)	0.090** (0.04)	0.115*** (0.04)		0.040 (0.03)	0.024 (0.03)	
Female Speaker × Novice Female Chair			-0.055 (0.05)	-0.046 (0.05)	-0.043 (0.04)	-0.022 (0.04)	-0.023 (0.04)	-0.014 (0.04)
Female Speaker × Accomplished Male Chair			-0.071** (0.03)	-0.070** (0.03)	-0.052* (0.03)	-0.060** (0.03)	-0.061** (0.03)	-0.039 (0.03)
Female Speaker × Accomplished Female Chair			-0.043 (0.04)	-0.045 (0.04)	-0.041 (0.03)	-0.037 (0.03)	-0.035 (0.03)	-0.031 (0.03)
R^2	0.000	0.006	0.006	0.108	0.548	0.426	0.455	0.719
Observations	18270	18270	18270	18270	18270	18046	18046	18043
Lower-ranked Rooms								
Female Speaker	-0.019 (0.02)		-0.035 (0.03)	-0.040 (0.03)	0.004 (0.02)			
Novice Female Chair		-0.004 (0.04)	0.009 (0.04)	-0.001 (0.04)		0.008 (0.04)	0.005 (0.04)	
Accomplished Male Chair		0.090*** (0.03)	0.075** (0.03)	0.087** (0.03)		0.001 (0.03)	0.001 (0.03)	
Accomplished Female Chair		0.033 (0.04)	0.015 (0.05)	0.026 (0.04)		-0.074* (0.04)	-0.070* (0.04)	
Female Speaker × Novice Female Chair			-0.029 (0.05)	-0.011 (0.05)	-0.039 (0.04)	-0.030 (0.04)	-0.032 (0.04)	-0.025 (0.03)
Female Speaker × Accomplished Male Chair			0.039 (0.04)	0.053 (0.04)	-0.015 (0.03)	-0.036 (0.03)	-0.038 (0.03)	-0.059** (0.03)
Female Speaker × Accomplished Female Chair			0.046 (0.05)	0.037 (0.05)	-0.004 (0.04)	0.003 (0.04)	-0.004 (0.04)	-0.024 (0.03)
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.000	0.002	0.002	0.063	0.532	0.399	0.426	0.734
Observations	16518	16518	16518	16518	16518	16331	16331	16324

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (ii) wing judge gender composition, (iii) speaking position, (iv) round, (v) motion type, (vi) competition & year. $R^2_{between}$.

Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is

Singleton observations are dropped in model (6) and (7). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3.18: Regression Analysis of Chair ft. Wing Gender Composition against Speech Score: Higher vs. Lower-ranked Debates (N = 34 788)

	Dependent Variable: Score (standardized)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Higher-ranked Rooms								
Female Speaker	-0.034*** (0.01)		-0.058** (0.02)	-0.071*** (0.02)	-0.011 (0.02)			
M Chair ft. F-majority Wings		-0.046* (0.03)	-0.061** (0.03)	-0.056** (0.03)		-0.066*** (0.02)	-0.061*** (0.02)	
F Chair ft. M-majority Wings		-0.045 (0.04)	-0.052 (0.04)	-0.037 (0.04)		-0.091*** (0.03)	-0.081*** (0.03)	
F Chair ft. F-majority Wings		-0.083** (0.03)	-0.095** (0.04)	-0.067** (0.03)		-0.108*** (0.03)	-0.096*** (0.03)	
Female Speaker × M Chair ft. F-majority Wings			0.042 (0.03)	0.029 (0.03)	-0.001 (0.03)	0.068*** (0.03)	0.068*** (0.02)	0.020 (0.02)
Female Speaker × F Chair ft. M-majority Wings			0.019 (0.04)	0.022 (0.04)	-0.002 (0.04)	0.025 (0.03)	0.030 (0.03)	0.027 (0.03)
Female Speaker × F Chair ft. F-majority Wings			0.031 (0.04)	0.021 (0.04)	-0.009 (0.03)	0.070** (0.03)	0.071** (0.03)	0.001 (0.03)
R^2	0.000	0.001	0.002	0.108	0.548	0.425	0.455	0.719
Observations	18270	18270	18270	18270	18270	18046	18046	18043
Lower-ranked Rooms								
Female Speaker	-0.019 (0.02)		0.019 (0.03)	0.018 (0.03)	-0.009 (0.03)			
M Chair ft. F-majority Wings		-0.009 (0.03)	0.012 (0.03)	-0.005 (0.03)		0.034 (0.03)	0.026 (0.03)	
F Chair ft. M-majority Wings		-0.028 (0.04)	-0.021 (0.04)	-0.051 (0.04)		-0.014 (0.04)	-0.025 (0.04)	
F Chair ft. F-majority Wings		-0.051 (0.04)	-0.022 (0.04)	-0.016 (0.04)		-0.009 (0.04)	-0.012 (0.04)	
Female Speaker × M Chair ft. F-majority Wings			-0.052 (0.04)	-0.049 (0.04)	0.007 (0.03)	-0.055 (0.03)	-0.051 (0.03)	-0.017 (0.03)
Female Speaker × F Chair ft. M-majority Wings			-0.018 (0.05)	-0.023 (0.05)	0.022 (0.04)	-0.026 (0.04)	-0.030 (0.04)	0.010 (0.04)
Female Speaker × F Chair ft. F-majority Wings			-0.074 (0.05)	-0.066 (0.05)	-0.037 (0.04)	-0.028 (0.04)	-0.029 (0.04)	-0.015 (0.03)
Speaker Controls				✓				
Room Controls				✓		✓		
Speaker FE					✓	✓		✓
Debate FE							✓	✓
R^2	0.000	0.000	0.001	0.063	0.532	0.398	0.425	0.734
Observations	16518	16518	16518	16518	16518	16331	16331	16324

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) number of female speakers in room (including team partner), (ii) wing judge gender composition, (iii) speaking position, (iv) round, (v) motion type, (vi) competition & year. $R^2_{between}$. Robust clustered standard errors at debate level are in parentheses. R^2 of model (5) to (8) is Singleton observations are dropped in model (6) and (7). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

"Know thy self, know thy enemy. A thousand battles, a thousand victories."

Sun Tzu

4 Choking upon Facing (Fe)male Opponents? Evidence from Debate Tournaments

4.1 Introduction

It is a well-established fact that the higher up the rungs of career ladders, the fewer women are present, especially in competitive occupations [Goldin et al., 2017; Blau and Kahn, 2017; Eckel et al., 2020]. A significant body of research have linked women's distaste for competition [Niederle and Vesterlund, 2007; Buser et al., 2014; Niederle and Vesterlund, 2011; Villeval, 2012; Datta Gupta et al., 2013] and under-performance in competitive settings [Niederle, 2017; Gneezy et al., 2003; Gneezy and Rustichini, 2004; Antonovics et al., 2009; Shurchkov, 2012] to this persistent gap. One prominent hypothesis revolves around the idea that the gender of opponents affects competitive performance [Shurchkov and Eckel, 2018]. Yet, evidence on this hypothesis is mixed, and especially scarce when it comes to real-life, dynamic competitions on multi-dimensional and complex tasks.

On the one hand, women have been shown to perform worse, especially when facing men, in the seminal lab evidence of [Gneezy et al., 2003] and [Niederle and Vesterlund, 2007], and across high-stake field studies [van Dolder et al., 2020; Säve-Söderbergh and Sjögren Lindquist, 2017]. On the other hand, numerous lab and field studies demonstrate that women compete similarly or better against men, from one-on-one lab experiments [Moely et al., 1979; Conti et al., 2001; Mago and Razzolini, 2019], real-effort team competition [Ivanova-Stenzel and Kübler, 2011] to high-stake field settings [Antonovics et al., 2009; Jetter and Walker, 2018; Iriberry and Rey-Biel, 2019]. Noteworthily, most of the relevant studies assess either one-on-one competitive settings or the overall gender composition in a static competition environment. Since competitive success often requires

repeated interaction in larger groups, over multiple rounds of applications and assessments in labor markets, it raises the question: To what extent can the observed gendered performance patterns given the gender composition of opponents be generalized to other real-life contests?

This chapter exploits the random assignment of 3153 contestants to multiple rounds of international university debate tournaments¹ to causally investigate how the gender composition of opponents affects speech performance. Competitive debating is a complex, multi-dimensional skill² that is externally relevant to careers in law, politics, business, and academia, where oral persuasion skill is instrumental to success [Buser and Yuan, 2020]. Three institutional features make these competitions an attractive setting to investigate whether the gender composition of opponents affects one's competitive performance. First, with 3153 contestants giving 39 168 speeches across nine debate rounds in each tournament, I can include individual fixed effects to control for unobserved factors, such as the innate ability of contestants. Second, in fixed teams of two, for every round, participants are exogenously matched to compete against three other teams (i.e. six other competitors).³ This system creates a randomly allocated set of opponents in terms of gender across debates. Finally, except for Round 1 where team matching is completely randomized, in each N^{th} round, every debate consists of contestants with similar $(N - 1)$ speech performance records. This *power-matching mechanism* mirrors the labor market contests, where repeated relative performance evaluations are used to assign jobs and promote employees. It also enables me to study whether the impact on speech performance given the gender composition of opponents is consistent across debate room levels, especially with heightened competitive pressure in high-ranked debates⁴ as the preliminary rounds progress.

I find that, on average, the performance of *neither* male nor female debaters is affected by the gender composition of opponents. Overall, while an additional female opponent is

¹Four annual World Universities Debating Championships (WUDC), and four annual European Universities Debating Championships (EUDC), from 2015 to 2018.

²Debating is considered one of the most effective activities to train four major language skills [Green III and Klug, 1990; Li et al., 2019] and leadership skill [Chikeleze et al., 2018].

³In general, opponents, judges, topics, and speaking position to argue for or against a policy-relevant topic are exogenously assigned. See Appendix 2.8.0.1 for details on the Debate Format.

⁴i.e. For N^{th} round, higher-ranked debates are those where teams with speech performance record equal to or higher than the *median* cumulative performance record in $(N - 1)$ rounds.

associated with a reduction of 2.0 percentage point standard deviation in speech score this unconditional score gap vanishes upon controlling for speaker fixed effect. This result is confirmed in the non-parametric specification of the number of female opponents. In other words, within individuals, the speech performance of neither men nor women responds to the number of female opponents they face in a debate.

In higher-ranked debates, women perform comparatively worse when facing more *female* opponents, whereas the performance of male speakers is unaffected by the gender composition of opponents. Controlling for speaker fixed effects, I estimate that for female speakers, an additional female opponent yields a 2.1 percentage point standard deviation reduction in score. No significant finding regarding the gender of opponents is detected in lower-ranked debates. An alternative analysis with a non-parametric specification of opponents' gender among these rooms shows negative and significant results for female speakers given *any* number of female speakers they face. For male speakers in higher-ranked debates, their speech performances are only comparatively worse in rooms with four female opponents or more. Furthermore, this gender score gap concerning female opponents is observed in women in female-only teams, and not those in mixed-gender teams.

This chapter contributes a causal finding on the interplay between one's gender and the gender composition of opponents in high-profile debate tournaments. This task and tournament setup particularly complements the current literature, which usually involves tasks without oral persuasion elements. A copious body of literature in one-on-one settings has shown that women perform comparatively worse when they face men, for instance, in the seminal work of [Gneezy et al., 2003], [Niederle and Vesterlund, 2007]. In [Delfgaauw et al., 2013], they show that sales competition among employees increases sales growth, but only in stores where the majority have the same gender. Meanwhile, [Datta Gupta et al., 2013] found that men choose to compete for less against other men than against women. In the field, [van Dolder et al., 2020] use data from the Dutch *Jeopardy!* shows to demonstrate female contestants perform worse when facing men, especially when taking into account the competitiveness of others. Conversely, men become more competitive in anticipation of decreasing competitiveness of their female contestants. In [Säve-Söderbergh and Sjögren Lindquist, 2017], female juniors employ inferior wagering strategies when randomly assigned to male opponents.

Nevertheless, a series of evidence, ranging from low-stake lab studies to high-stake field experiments suggest otherwise. [Moely et al., 1979; Conti et al., 2001] documented that girls perform better when competing against boys than girls. Most recently, the best-of-five repeated contest by [Mago and Razzolini, 2019] found that women exert significantly higher effort only when competing against other women, while women are just as competitive as men in mixed-gender sessions. In the field, across five sequential elementary math contests, [Cotton et al., 2013] found that the male advantage is at best short-lived, while females even outperform males in later periods. In TV shows, in contrast to findings of [van Dolder et al., 2020], [Jetter and Walker, 2018] and [Antonovics et al., 2009] found that women are more competitive when facing men in the US *Jeopardy!* version and the high-stake rounds of the *Weakest Link* show, respectively. The closest work to my paper is [De Paola et al., 2015] on midterm exam performance of Italian students competing in pairs of equal predicted ability but different gender composition. Similar to their work and [Mago and Razzolini, 2019], I find that on average, the performance of *neither* men or women is affected by the gender composition of opponents.

Secondly, this research expands the empirical evidence on real-world contest literature with a piece of novel evidence in high-stake debate tournaments. To the best of my knowledge, other than school exams or TV shows, empirical studies on the gender differences in competitive performance are mostly restricted to one-on-one settings e.g. expert chess or tennis tournaments. Specifically, in chess tournaments, [Backus et al., 2016], [Dilmaghani, 2020] and [Gerdes and Gränsmark, 2010] consistently confirm that conditional on ELO ratings, the gender composition effect is driven by women performing worse against men, rather than by men playing better against women. Furthermore, the largest gender performance gap is among elite players. Comparatively, in debate tournaments, I find supporting evidence for a larger gender gap in higher-ranked debates. Yet, in contrast to chess tournaments, female debaters fare comparatively worse when facing more female opponents in higher-ranked debates. In same-sex only tennis tournaments where [Wozniak, 2012] studied the tournament entry decision given relative past performance feedback, he found that such information feedback has gender-specific effects. Since recruitment or promotion decisions in firms are often drawn on a pool of similarly able candidates across multiple rounds, insights from these mixed-sex, multi-round debate competitions, where participants compete head-to-head based on previous rankings, are more relatable to real-life competitions.

Finally, the literature on team gender composition and performance provides possible mechanisms to explain the descriptive result of the concentrated gender score gap among female speakers in female-only teams, and not those in mixed-gender teams. Since participants compete in their chosen teams of two, my descriptive finding that women-only teams perform worse than mixed-gender or male-only teams is in line with the observational study of [Apesteguia et al., 2012] in high-stake, online business game contests. [Dargnies, 2012] offers a likely explanation for this overall gender score gap, based on differential self-selection: low-performing women are more likely to enter tournaments with similar others in two-person tournaments. The descriptive finding that male-dominated teams perform similarly to mixed-gender teams is also in line with the causal result in the larger 12-person business team field experiment by [Hoogendoorn et al., 2013].

This chapter proceeds as follows. Section 4.2 summarizes the debate competition setup. Section 4.3 provides data set overview and summary statistics, followed by empirical strategies in Section 4.4. Section 4.5 highlights the main results, with extension findings on higher- vs. lower-ranked debates and competition performance given teammate's gender choice in Section 4.6. Section 4.7 concludes with discussions on future research avenues.

4.2 Institutional Setup

Tournament Format. Participants in these tournaments are undergraduate or graduate students who are active and dedicated in their respective debate societies. Debaters participate in weekly meetings and travel to various local and international tournaments to sharpen their debate skills. Every year, around 200+/- two-person-teams across Europe attend the European Universities Debate Championship (EUDC); 450+/- teams across the world participate in the World Universities Debate Championship. They represent their institutions to compete across nine preliminary rounds (i.e. *in-rounds*) with exogenously assigned controversial topics, speaking positions, judges, and opponents in every round. All debates are conducted in British Parliamentary (BP) Debate style.⁵ After each round, a panel of judges submit two results of each individual to the score tabulation organizer:

⁵For more details on BP debate style and format, please check Appendix 2.8.0.1.

(1) team ranking ⁶ and (2) individual speaker scores.⁷ Within a debate, individual speaker scores must reflect the ordinal team ranking i.e. the cumulative score of two speakers whose team ranked first must be higher than that of the team ranked second. The total team points and speaker points⁸ across all preliminary rounds determine the top 10 – 15% performing teams to enter elimination rounds (i.e. *out-rounds*)⁹. Since evaluation scores are only given in preliminary rounds, this research focuses exclusively on these rounds, and not the out-rounds.

Team Matching & Performance Feedback Mechanism. Every debate consists of four teams. In Round 1, team matching is unconditionally randomized. From Round 2 onward, teams are power-matched i.e. in N^{th} round, teams debate teams with similar cumulative team *and* speech evaluation points from $(N - 1)$ rounds.¹⁰ In other words, *within* each team score bracket, the teams with the highest speaker points in $(N - 1)$ rounds will meet one another in N^{th} round. Hence, the universal individual speech score scale aims to ensure consistent evaluation across rooms.¹¹ Regarding performance feedback to speakers, from Round 1 to Round 6 (*open rounds*), teams receive only their team ranking results and relative performance feedback after the debate and judge deliberation discussions. From Round 7 to Round 9 (*closed rounds*), no results are communicated to speakers right after the debate. Once all elimination rounds are completed, speakers receive team ranking results and feedback from judges. Finally, speakers will receive the public results of their evaluation scores across rounds when the tournament ends.

Judge Allocation Mechanism & Fairness. Every tournament has an appointed Chief Adjudicator (CA) team of four to six internationally accomplished debaters who are in charge

⁶i.e. team that ranks 1st gets 3 points, 2nd gets 2 points, 3rd gets 1 point and 4th gets no point.

⁷50-to-100 score scale, with 50 as the lowest. See Appendix 2.8.2.1 for a speaker score scale example of European Universities Debate Championship 2017.

⁸Speaker points are used for: (i) award best performing speakers in the form of top 10 speaker awards; and (ii) determine teams advancing to elimination rounds in case of ties.

⁹In these rounds, teams that are ranked 1st and 2nd advanced into further rounds, whereas those on 3rd and 4th place are eliminated. In the final debate, the best team becomes the champion. The best speaker of the tournament is an individual with the highest cumulative individual speech scores across all preliminary rounds.

¹⁰Note that *within* a debate, speech evaluation points must reflect team rankings i.e. the cumulative speech scores of two speakers whose team ranked first must be higher than that of the team ranked second. For more information about power-pairing, see this discussion thread on [Monash Debate Review](#).

¹¹i.e. winning in a lower-ranked room does not necessarily mean higher individual speaker scores than, for instance, taking a 2nd or 3rd in a higher ranked room

of judge recruitment, quality screening, monitoring, and overall panel allocation throughout the tournament. Three mechanisms are set in place to ensure fairness in judgment across rounds. First, no judges who come from the same institutions, in the past or present, as any debaters in the room can be allocated to judge that debate. Second, before the competition, judges and debaters are required to disclose any potential conflicts with other participants.¹² Third, the intensive nature of a 3-day, 9-round competition makes it difficult for any strategic collusion to be formed between judges, the CA team, and speakers from different institutions. Appendix 3.2.2 provides more details on the judge’s tasks, check-and-balance feedback mechanism, and adjudication procedure throughout these tournaments.

4.3 Data & Descriptive Statistics

4.3.1 Data

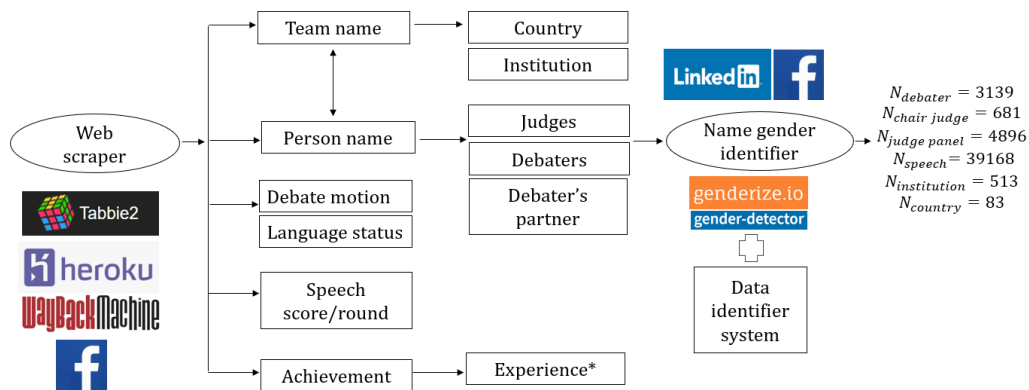


Figure 4.1: Overview: Data collection and construction procedure

This section describes the data set construction procedure and key descriptive statistics of speakers and judges. Figure 4.1 illustrates the entire data collection process. In general, names of individuals, judges and institutions, the roles of judges¹³ and opponents;¹⁴ individual evaluation scores for every debate, language skill status of speakers and debate

¹²Legitimate clashing reasons include, among others, close friendship/partnership, past romantic encounters, or negative experiences. To disincentivize strategic clashing, an independent committee conducts confidential interviews with the requested persons to verify their reasons.

¹³i.e. chair judge, wing judge and trainee judge.

¹⁴i.e. Opening Government, Opening Opposition, Closing Government, Closing Opposition

motions are available from tabulated archival sources.¹⁵ Detailed data collection procedure on other control variables such as judge panels, debate topics, language skills and institutions is provided in Appendix 4.8.1. Section 4.3.2 then gives descriptive statistics on score differentials across speakers given their characteristics.

4.3.1.1 Outcome Variable: Speech Scores

The main outcome variable is the speech evaluation scores given by the adjudication panel for every debate speech. Across the total of 5081 debates from eight competitions, 185 debates are omitted due to missing identity of speakers or speech scores. These are because of either of the following reasons: (i) swing speakers (i.e. last-minute fill-in volunteers in case speakers cannot speak); (ii) speakers who redacted their identity after the tournament; (iii) one speaker spoke for both roles, since the other speaker excused him/herself from speaking in that respective round.¹⁶ Since full information about gender composition of speakers and judges is crucial for the analysis, all rooms with at least one of such issues are omitted. This procedure results in 39168 speeches across 4896 debates, which is documented per competition in Appendix 3.4, along with omitted debates per competition.

4.3.1.2 Speakers

Full names of speakers¹⁷ and matching their identities across the years is done given the tabulation tournament data archive in [Tabbie2](#) and [Tabbycat](#). To avoid discretionary personal judgment as much as possible, we used a conservative method: a person is considered a duplicate only if their name, institution, EUDC language status, and WUDC language status are the same. Next, to identify gender of speakers, I ran gender inference algorithms: [gender guesser](#) and [genderize.io](#)¹⁸ on their first names. Both algorithms return the most likely gender, given its hand-coded data label, and a frequency count of such names in their database as male or female.¹⁹ This procedure results in 89.23% of names

¹⁵Such data is released given the consent of speakers and judges, unless otherwise redacted, in which case they are omitted from the sample.

¹⁶In this case, the missing speaker receives 0 point, whereas his/her partner who gave both speeches receive the higher score of the speeches he/she gave.

¹⁷We first clean out: strange characters from non-English names, reversed first and last names, abbreviated names are properly restored across tournaments by matching with their institutions and social media profile (where applicable).

¹⁸This API contains 216286 distinct names across 79 countries and 89 languages

¹⁹For a comparison of features and performance of different gender inference algorithms, please refer to the report of [Menéndez et al., 2020] and [Santamaría and Mihaljević, 2018].

assigned gender with certainty. The remaining 10.77% names, which consist of mostly African, South Asian, Israeli, and Eastern European names, were manually checked using social media.

Altogether, after omitting 27 unisex names without any social media sources and possible confirmation from tab masters, we have $N = 3153$ unique speakers for analysis: whereby $N_{MaleSpeaker} = 1949$ and $N_{FemaleSpeaker} = 1190$. Figure 4.2 shows the proportion of speakers by gender for each competition. Across all competitions, female speakers account for 35% to 41% of all participants. Furthermore, the world map distribution of speakers given their gender in Figure 3.3 shows that most countries sent disproportionately more male speakers than female speakers, except for China/Hong Kong. The US, UK, and Australia sent the highest number of speakers, understandably so, given their established debate training culture and civic participation.

4.3.2 Descriptive Statistics

Relationship between Gender of Speakers, Opponents, and Room Characteristics.

Table 3.8 provides a comprehensive breakdown of the proportion of speeches by male and female speakers, while Figure 4.4 gives the Spearman correlation coefficient heatmap across various characteristics of speakers, debate room, and judges. Table 3.8 shows that there does not appear to be any differences in terms of the proportion of speeches by male vs. female speakers across these characteristics. Most importantly, Figure 4.4 shows no correlation between the number of female opponents and any observable characteristic, including the speaker's gender.²⁰ Apart from a very mild positive correlation between the speaker's gender and the gender of their chosen debate partner, there is virtually no relationship between the speaker's gender and other characteristics. Regarding the distribution of female opponents in a debate (excluding partner's gender), Figure 4.2 notes that speakers face only one to three female opponents, thus reflecting the male-dominated nature of competitive debate tournaments.

Speech Scores: Male vs. Female Speakers. Table 3.5 reports the descriptive statistics of scores across all tournaments. The t-test statistics in Table 3.10 and the kernel density

²⁰This is confirmed in the Spearman's correlation coefficient test between speaker's gender and their opponents in a room (excluding debate partner), with $\rho = -0.0015$ and $p = 0.768$.

of speech scores in Figure 4.8 show that male speakers scored slightly higher than female speakers. This pattern holds regardless of whether it is male- or female-dominated debates,²¹ the language skill statuses or whether the speaker belongs to top 50-ranked institutions. Across rounds, Figure 4.5 plotting the mean standardized evaluation scores of men vs. women shows persistently lower scores from women than men, except for Round 3s.

Speech Score: Higher- vs. Lower-ranked Debates. At any given N^{th} round (except for Round 1s), I split the sample based on the *median* average cumulative $(N - 1)$ round speech scores of *two speakers in a team*. Specifically, higher-ranked debates are those where the score is higher than or equal to the *median* speech score and vice versa. Comparing speeches of male and female speakers in higher- vs. lower-ranked debates in Figure 4.10 shows notably lower scores of female speakers only in higher-ranked debates. A further breakdown given partner's gender in the histogram and kernel distribution in Figure 4.13 found slightly higher scores for women in mixed-gender teams than those in women-only teams, yet the pattern is more pronounced in lower-ranked debates.

Speech Score: Team Gender & Round Dynamics. Since speakers choose their respective partners to enter the tournament together, this subsection gives some descriptive graphs on score differentials across rounds given the team gender composition. Figure 4.7 gives a descriptive overview of the average speaker's score across rounds across male-only, mixed-gender, and female-only teams. We note that the gender score gap found in Figure 4.5 is predominantly driven by female speakers in female-only teams. For women in mixed-gender teams, compared to their male partner, except for Round 5s and 7s where they scored on average lower than their male partner, in the rest of the rounds, they either scored similarly or slightly higher than their partner.

4.4 Empirical Strategies

To understand whether the gender composition of opponents affects speech performance of male and female speakers, I run linear and fixed effects regression on standardized speech

²¹see the histogram in Figure 4.9.

score, interacting the indicator variable of speaker's gender with the number of female opponents in the debate, as shown below:

$$S_{sk} = \alpha \mathbb{I}_{FemS} + \theta \mathbb{I}_{FemO} + \beta \mathbb{I}_{FemS} \mathbb{I}_{FemO} + \sum_{i=1}^n \gamma \mathbf{Y}_{sk} + \eta_s + \varepsilon_{sk}$$

The dependent variable S_{sk} is the standardized evaluation score of the speech of speaker s in debate k . The coefficients of interest are θ and β , where θ measures any significant relationship between speech performance of male speakers and the number of female speakers in debate k , and β checks for any significant differences between male and female speakers therein.

\mathbb{I}_{FemS} is the gender of the speaker, whereas \mathbb{I}_{FemO} refers to the number of female opponents for speaker s in debate k . Given the male-skewed distribution of speakers in a room shown in Figure 4.3, I use both a linear specification of \mathbb{I}_{FemO} , where \mathbb{I}_{FemO} is the number of female opponents, and a non-parametric specification, where I add dummy variables for each possible number of female opponents i.e. $\mathbb{I}_{FemO} \in \{0, 6\}$. ε_{sk} is the error term of the speech given by speaker s in debate k . Throughout all analyses, standard errors are clustered at debate level.

Speaker fixed effect η_s is included to take care of any unobserved heterogeneity on the speaker's characteristics. Other control variables \mathbf{Y}_{sk} are as follows:

1. η_J is the chair judge fixed effect.²²
2. language skill level (non-native or native English speaker).
3. institution ranking group (i.e. whether if the speaker represents a top-50-ranked institution).
4. gender of speaker's debate partner.
5. group competition type (EUDC or WUDC).
6. Speaking position (1st to 8th) in any given debate.

²²Since chair judges have decisive power in determining the team and speaker outcomes in a debate, this fixed effect captures unobserved heterogeneity on chair judge's characteristics.

7. Motion topic type (17 topics) in any given debate.²³
8. Debate round (1 to 9) for any given debate.
9. whether the majority of wing judges are women.

In debate tournaments, the power-matching mechanism makes teams debate teams with a comparable cumulative performance from previous rounds, starting from Round 2 onward. Therefore, as an attempt to control for the average team standing from previous rounds i.e. selection effect on the interested variable, I include in some regression analyses the average cumulative speech scores over $(N - 1)$ rounds of *two speakers in a team* in the analysis of N^{th} round, for Round 2s to Round 9s.

4.5 Results

4.5.1 Overall

Column (1) of Table 4.1 shows that, unconditionally, female speakers get 5.6 percentage point (p.p) standard deviation (SD) lower scores compared to male speakers. On average, an additional female opponent is associated with a reduction of 2.0 p.p SD in speech scores, as noted in Column (2). Columns (3) and (4) show that there is no difference between male and female speakers in the relation between the number of female opponents and speech scores.

²³See Figure 3.4 for the list of motions.

Table 4.1: Regression Analysis: Gender of Speakers and Opponents (N = 39 168)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Speaker	-0.056*** (0.01)		-0.053** (0.02)	-0.051*** (0.02)		
Number of Female Opponents		-0.020** (0.01)	-0.020** (0.01)	-0.013* (0.01)	0.002 (0.01)	0.004 (0.01)
Female Speaker × Number of Female Opponents			-0.001 (0.01)	-0.001 (0.01)	-0.003 (0.01)	-0.002 (0.01)
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓
R^2	0.001	0.001	0.001	0.318	0.568	0.642
Observations	39168	39168	39168	39168	39157	39157

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year. Robust clustered standard errors at debate level in parentheses.

R^2 of model (5) and (6) is $R^2_{between}$. Singleton observations are dropped in model (5) and (6).

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

Noteworthy, the score gap in rooms with more or fewer female opponents vanishes upon controlling for speaker fixed effect, as seen in Columns (5) and (6). The difference in the estimated effect of the number of female opponents between column (2) to (4) and Column (5) to (6) suggests that the relationship in Columns (2) to (4) is driven by a selection effect. In other words, within individuals, the speech performance of neither men nor women responds to the number of female opponents in the room. Therefore, the effect in columns (2) to (4) is across individuals, and potentially due to the fact that female speakers perform slightly worse. Over time, because of the power-matching mechanism, gender segregation occurs, i.e. more women cluster to lower-ranked debates in later rounds compared to earlier rounds.

Next, Table 4.2 reports the non-parametric regression results of speaker's gender on the number of female opponents in the debate against speech scores. Column (2) shows the unconditional score difference across debates given the number of female opponents that a speaker faces. Compared to debates where speakers face no female opponents, speakers in debates with only one female opponent received 5.8 p.p. SD higher scores. As the number of female opponents increases, we noted a negative, yet insignificant speaker score gap between such debates and debates with no female opponents. Yet, given the limited number of rooms with 5 or 6 female opponents (see Figure 4.3), it is difficult to draw conclusions from these numbers. Importantly, at the speaker's fixed effect level, Column (5) shows

that there is no difference in the association between the number of female opponents and speech performance of both male and female speakers. This is consistent with the analysis using a continuous specification of the number of female opponents above.

Table 4.2: Regression Analysis : Gender of Speaker and Opponents (N = 39 168)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Speaker	-0.056*** (0.01)		-0.052 (0.05)	-0.014 (0.03)		
1 Female Opponent		0.058* (0.03)	0.059 (0.04)	0.051 (0.03)	0.000 (0.03)	-0.000 (0.02)
2 Female Opponents		0.054 (0.03)	0.053 (0.04)	0.040 (0.03)	-0.005 (0.03)	0.008 (0.02)
3 Female Opponents		0.022 (0.04)	0.029 (0.04)	0.027 (0.04)	-0.001 (0.03)	0.013 (0.03)
4 Female Opponents		-0.052 (0.04)	-0.052 (0.05)	-0.028 (0.04)	0.013 (0.03)	0.015 (0.03)
5 Female Opponents		-0.072 (0.06)	-0.074 (0.06)	-0.030 (0.05)	-0.013 (0.04)	-0.020 (0.04)
6 Female Opponents		-0.120 (0.10)	-0.136 (0.13)	-0.059 (0.11)	0.119 (0.08)	0.121* (0.07)
Female Speaker × 1 Female Opponent			-0.001 (0.05)	-0.040 (0.04)	-0.019 (0.04)	-0.013 (0.03)
Female Speaker × 2 Female Opponents			0.001 (0.05)	-0.050 (0.04)	-0.021 (0.04)	-0.028 (0.03)
Female Speaker × 3 Female Opponents			-0.019 (0.05)	-0.053 (0.04)	-0.015 (0.03)	-0.021 (0.03)
Female Speaker × 4 Female Opponents			0.003 (0.06)	-0.012 (0.05)	-0.033 (0.04)	-0.018 (0.03)
Female Speaker × 5 Female Opponents			0.004 (0.08)	-0.052 (0.06)	-0.007 (0.06)	0.010 (0.05)
Female Speaker × 6 Female Opponents			0.038 (0.16)	0.147 (0.14)	-0.015 (0.15)	-0.051 (0.10)
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓
R^2	0.001	0.002	0.002	0.318	0.568	0.642
Observations	39168	39168	39168	39168	39157	39157

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year. Robust clustered standard errors at debate level in parentheses.

R^2 of model (5) and (6) is $R^2_{between}$. Singleton observations are dropped in model (5) and (6).

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

4.5.2 Round 1s vs. Round 2s to 9s

Table 4.3 provides the regression analysis using a continuous specification of the number of female opponents, among Round 1s and those of Round 2s to 9s. The split is because the power-matching mechanism of teams is applied from Round 2 onward, whereas in Round 1s, the allocation is unconditionally random. In an attempt to capture the selection effect of the power-matching mechanism, the final rows in Table 4.3 control for the average team standing of $(N - 1)$ round from Round 2s to Round 9s, in the analysis of N^{th} round at a particular tournament. For Round 1s, no speaker fixed-effect model applies since there is only one observation per speaker for every tournament, and within-speaker comparisons across Round 1s of different tournaments are not comparable to other round analyses.

Overall, no significant difference in speech performance of men and women given the number of female opponents, which is consistent with the findings in Table 4.1. For Round 1s where room allocation is unconditionally random, Column (2) shows a similar, albeit insignificant, unconditional score gap of 1.9 p.p SD for speakers who face more female opponents, to the overall finding in Table 4.1. It is important to note that upon controlling for speaker and room characteristics and interacting the gender of speakers with the number of female opponents, I find a significant and negative relationship i.e. speakers who face more female opponents get 4.2 p.p SD lower scores.

Comparing the gender speech score gap between Round 1s and Round 2s to 9s, Column (1) shows that this gap from 8.3 p.p SD in Round 1s to only 5.3 p.p. SD across Round 2s to 9s. Upon controlling for team standing, this gap remains significant but shrinks to 3.0 p.p SD. This pattern illustrates the functioning power-matching mechanism, whereby teams of comparable ability compete against one another. Regarding the number of female opponents, Column (2) shows an unconditional score gap of 2.0 p.p SD for speakers who face more female opponents in Round 2s to Round 9s. Upon interacting speaker's gender with the number of female opponents, Column (3) finds that this relation is similar between male and female speakers, yet it vanishes upon controlling for speaker and debate room characteristics and speaker fixed effects. Noteworthy, once team standing is taken into account, speakers who face more female opponents get 4.0 p.p SD lower scores. A qualitatively similar gap of 3.5 p.p SD remains upon interacting with speaker's gender shows up, only to disappear upon further controls in Columns (3) to (5).

Table 4.3: Regression Analysis: Gender of Speaker and Opponents, Round 1s vs Round 2s - 9s (N = 39 168)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Round 1s						
Female Speaker	-0.083*** (0.03)		-0.061 (0.05)	-0.085 (0.05)		
Number of Female Opponents		-0.019 (0.01)	-0.014 (0.02)	-0.042* (0.02)		
Female Speaker × Number of Female Opponents			-0.009 (0.02)	-0.001 (0.02)		
R^2	0.002	0.001	0.003	0.361		
Observations	4376	4376	4376	4376		
Round 2s to 9s						
Female Speaker	-0.053*** (0.01)		-0.054** (0.03)	-0.045** (0.02)		
Number of Female Opponents		-0.020** (0.01)	-0.021** (0.01)	-0.009 (0.01)	0.005 (0.01)	0.006 (0.01)
Female Speaker × Number of Female Opponents			0.000 (0.01)	-0.001 (0.01)	-0.003 (0.01)	-0.003 (0.01)
R^2	0.001	0.001	0.001	0.337	0.572	0.648
Observations	34792	34792	34792	34792	34786	34786
Round 2s to 9s (controlled debate room quality)						
Female Speaker	-0.030*** (0.01)		-0.005 (0.02)	-0.025 (0.02)		
Number of Female Opponents		-0.040*** (0.01)	-0.035*** (0.01)	-0.009 (0.01)	0.005 (0.01)	0.006 (0.01)
Female Speaker × Number of Female Opponents			-0.012 (0.01)	-0.006 (0.01)	-0.003 (0.01)	-0.003 (0.01)
R^2	0.171	0.173	0.173	0.387	0.572	0.648
Observations	34792	34792	34792	34792	34786	34786
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year. Robust clustered standard errors at debate level in parentheses. R^2 of model (5) and (6) is $R^2_{between}$. Singleton observations are dropped in model (5) and (6). * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.4 investigates speech performance given the number of female opponents across Round 2s to Round 9s separately, controlling for team standing, speaker, and judge characteristics. We note that in Round 3s and 4s, female speakers facing more female opponents get comparatively lower scores compared to male speakers, at 3.5 and 5.0 p.p SD, respectively. As the rounds and the power-matching mechanism progress, no significant difference in speech performance of male and female speakers given the number of female opponents

is detected. Overall, the changes in magnitude could be due to the varying distribution of the number of female opponents across rounds. A non-parametric estimation of the effect of the number of female opponents in Round 1s of Table 4.5, and Round 2 to 9s in Table 4.6 and 4.7 (controlling for average team standing) in Appendix 4.8.3.1 show consistent findings with those in Table 4.3.

Table 4.4: Round-by-round regression Gender of Speaker and Opponents, controlling for room quality, speaker and judge characteristics (N = 34 792)

	Dependent variable: Score (standardized)							
	R2s	R3s	R4s	R5s	R6s	R7s	R8s	R9s
Female Speaker	-0.094 (0.06)	0.093* (0.05)	0.119** (0.05)	-0.058 (0.05)	0.010 (0.05)	-0.093* (0.05)	0.065 (0.06)	0.030 (0.05)
Number of Female Opponents	0.021 (0.03)	0.028 (0.03)	-0.016 (0.02)	0.002 (0.03)	-0.053** (0.02)	0.007 (0.02)	0.015 (0.03)	0.041 (0.03)
Female Speaker × Number of Female Opponents	-0.000 (0.02)	-0.035* (0.02)	-0.050** (0.02)	-0.019 (0.02)	-0.011 (0.02)	0.032 (0.02)	-0.020 (0.02)	0.005 (0.02)
R^2	0.421	0.531	0.519	0.561	0.585	0.592	0.609	0.679
Observations per round	4408	4376	4304	4320	4352	4336	4344	4352

All models control for team standing (i.e. average cumulative speech scores of two speakers in a team up to the respective round.). Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year.

Robust clustered standard errors at debate level in parentheses.

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

4.6 Extensions

4.6.1 Higher vs. lower-ranked debates

Given that the stakes are significantly higher in higher-ranked debates, for every N^{th} debate after Round 1, I split the sample based on the *median* average cumulative ($N - 1$) round speech scores of *two speakers in a team*. Since the power-matching mechanism in these tournaments matches teams with not only similar team records but also speech performance records, this split provides a good approximation of team standings in any respective N^{th} round. I then run the regression analysis in Table 4.8. For lower-ranked debates, controlling for average debate standing, I find no significant gender score gaps. An additional female opponent is associated with a reduction in speaker score of 2.2 pp SD. However, this score gap vanishes upon interacting with the speaker's gender and controlling for relevant speaker and room characteristics.

In contrast, among higher-ranked debates, female speakers get 3.1 p.p SD lower scores than their male counterparts. With respect to the number of female opponents, Column (2) shows that speakers who face more female opponents receive 2 p.p SD per additional female opponent. Upon interacting with the speaker's gender and controlling for speaker, judge, and room characteristics, I find that male speakers perform similarly when there are more or fewer female speakers. However, female speakers get significantly lower scores when faced with more female opponents. At the speaker fixed effect level controlling for relevant characteristics, one more female opponent leads to 2.2 p.p SD lower scores for female speakers, compared to their male counterparts.

To better understand the relationship between the speaker's gender and the specific number of female opponents, Table 4.9 gives the non-parametric regression analysis in higher-ranked debates. Column (2) shows that, compared to speakers who face no female opponents, speakers facing 4 and 5 female opponents get 11.1 p.p and 10.8 p.p SD lower scores respectively. Upon interacting speaker's gender with the number of opponents, Column (3) to (6) show that the number of female opponents only affect female speakers, but not male speakers. Specifically, compared to male speakers facing 1 to 4 female opponents, female speakers facing the same number of opponents receive robustly lower scores. Given the limited number of female speakers facing 5 or 6 female opponents in higher-ranked debates ($N_{5female} = 181$, $N_{6female} = 11$), there is insufficient power to draw firm conclusions from these results.

4.6.2 Speaker gender ft. partner's gender

Given the difference in speech performance between male-only (MM), mixed-gender (MF), and female-only teams (MF) shown in Figure 4.7, to better understand the relationship between speaker's performance given their partner's gender choice and the number of female opponents, Table 4.10 reports the regression analysis with a continuous variable of the number of female opponents. In this case, since teams are fixed within a tournament, models with speaker fixed effects estimate *across* tournaments. Therefore, the results in these models in Column (5) and (6) of Table 4.10 and 4.11 should be interpreted as cross-tournament estimation. Since many speakers compete in multiple tournaments with varying partner's choices, such analysis gives an insight into the overall scoring patterns across tournaments on the basis of the partner's gender choice.

Across all rounds, an important overall finding is that the gender score gap is mainly observed among female speakers in female-only teams; but the gender composition of opponents in a debate largely plays no role. A split analysis of Round 1s and Round 2s to 9s in Table 4.11, whereby the latter rounds controls for average cumulative team standings, find similar results across all rounds. An important note here is that these findings only serve as *descriptive evidence* on speaker's performance given their team gender composition because debate partner's choice is endogenous.

4.7 Conclusion

This chapter contributes a causal finding on the impact of the gender composition of opponents on the competitive performance of 3153 debaters in highest-profile debate tournaments. The multi-dimensional, complex debate task together with the multi-round, power-matching mechanism in these tournaments adds a piece of useful empirical evidence to the contest and gender competitive performance dynamics. The key finding is that the performance of *neither* male nor female debaters is affected by the gender composition of opponents. In higher-ranked debates, women perform comparatively worse when facing more *female* opponents. Descriptively, the raw gender score gap is mainly found among women in female-only teams, not those in mixed-gender teams. Overall, if these findings carry over to other real-life settings, they indicate that having more women competing for high-profile careers and positions does not necessarily reduce the thickness of the glass ceiling.

Three limitations potentially restrict the generalizability of these results. First, since the power-matching mechanism is a known feature among participants in debate tournaments, they have some certainty over the previous performance records of their opponents. In other real-world contexts, beliefs about the performance or ability of opponents are possibly biased or dependent on the gender of opponents, which in turn may affect one's own performance. Second, participants who select themselves into these debate competitions are young, talented university students with significantly public speaking training and international exposure. Since these competitions are held Europe-wide or worldwide, it is not directly applicable to local competition contexts. Third, while this paper can study the

causal impact of the gender composition of opponents on individual speech performance, endogenous team formation prevents any causal interpretation on how within-team gender composition interacts with the gender composition of opponents. As teams train and prepare speeches together, I cannot disentangle whether women who select into mixed-gender teams have a higher innate ability, or simply that their team dynamics differ from that of female-only teams. Finally, to enrich the analyses and descriptive evidence on partner's choices, future research can take into account previous debate experiences of participants and the progression of different team gender compositions in elimination rounds.

4.8 Appendix

4.8.1 Data collection: Judge panels, Debate topics, Language skill, Institution & ranking

4.8.1.1 Judges & Evaluation Panels

Upon scraping the archival tabulation data, we obtained the full names of chair (C), wing (W) and trainee (T) judges for each debate. To determine their identity uniqueness and represented gender, we first sorted the names of all judges per tournament by their function: C, W or T. We temporarily stored names of all judges in a different file to codify gender and deleted their names afterwards. For chair (C) judges, we managed to determine unique identity and gender for everyone, since: (1) they hold the most power in speech evaluation; and (2) their identity is easily tracked given their high-profile statuses and social media presence in debating channels. For wing (W) and trainee (T) judges, I combined results from gender inference algorithms on their first names and information on their affiliated institutions, countries and region. Using the gender inference algorithms similar to with speaker's names, we identified gender of 92% of (W) and (T) judges. For the remaining 8%, which either are: (i) African, South East Asian, Indian and Israeli and gender-neutral names or (ii) conflicting gender assignment, I manually checked them using social media connections. Finally, the completed gender list for each judge is confirmed with respective tab directors.

4.8.1.2 Language skill status

In EUDC/WUDC debate tournaments, individuals are classified into different language categories by an appointed independent language committees. This classification is meant to provide an inclusive playground to speakers with limited exposure to English language, which enables participants to break into open and/or non-native (ESL) speaker's league in knock-out rounds. The evaluation criteria are based on individual survey applications regarding: (i) the age at which they were exposed to English; and (ii) the content, structure and quality of English used for any relevant instruction or exchange.²⁴ From the archival

²⁴For more detailed criteria to be qualified as ESL for EUDC and ESL & EFL for WUDC, see the Language Status section of [WUDC constitution](#) and [EUDC constitution](#).

tab data, I documented 46.65 % of speeches given by non-native English speakers.

4.8.1.3 Debate topics

Across the 4896 debates, 72 unique debate motions discussed across a wide range of topics. All topics provided a balanced, in-depth but polarized distribution of views, as empirically tested by chief adjudicators in earlier regional competitions. I manually classified these motions into 17 debate topics, based on the classification at [International Debate Education Association](#), which are summarized by the distribution of debate speeches in Figure 3.4. Topics on society, international relations and military policies are the three most popular debate motions at these tournaments, followed closely by debates on the economy, law and justice systems, as well as topics on health, feminism and digital freedom.

4.8.1.4 Institution & Ranking

Since the academic institution that a speaker represents carries reputation/prestige that could impact evaluations, we collected institution information embedded in team name, in addition to registry data from tab masters, where possible. By pairing up speaker's identity with their team names, along with public social media and confirmation with the tab directors, we obtained 513 distinct institutions across 83 countries in this data set. Since there exists no university ranking given their debate achievements,²⁵ these institutions are categorized by their average academic ranking from QS World Universities Ranking from 2013 to 2017 into two groups: top-50-ranked and the non-top-50-ranked universities. Descriptive statistics table in Appendix 3.8 shows that participants affiliated with top-50-ranked institutions account for roughly 10 - 20% of all participants, with the slight exception of WUDC 2017 and WUDC 2018, where this proportion is above 20%. More male speakers tend to represent top-50-ranked institutions and be native English speakers.

²⁵Apart from a top 5 and top 10 list of UK & UK universities to master debate skills in the [US](#) and [UK](#)

4.8.2 Figures

4.8.2.1 Proportion of male vs. female speakers across competitions

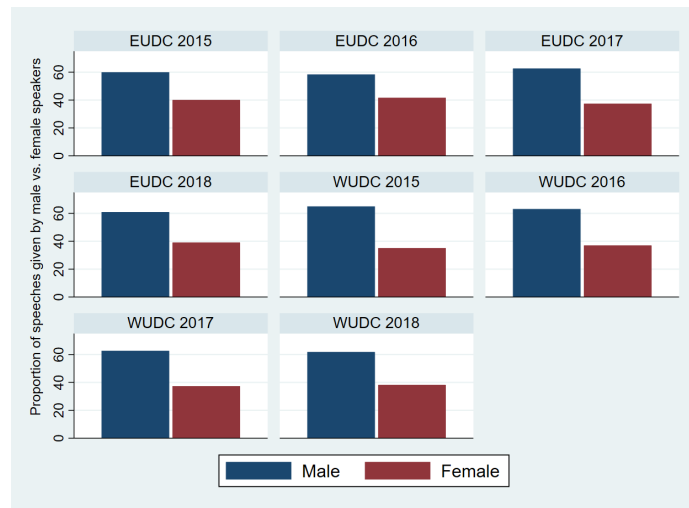


Figure 4.2: Number of male vs. female speakers by competition ($N_{male} = 24334$, $N_{female} = 14834$)

4.8.2.2 Distribution of female opponents in a debate

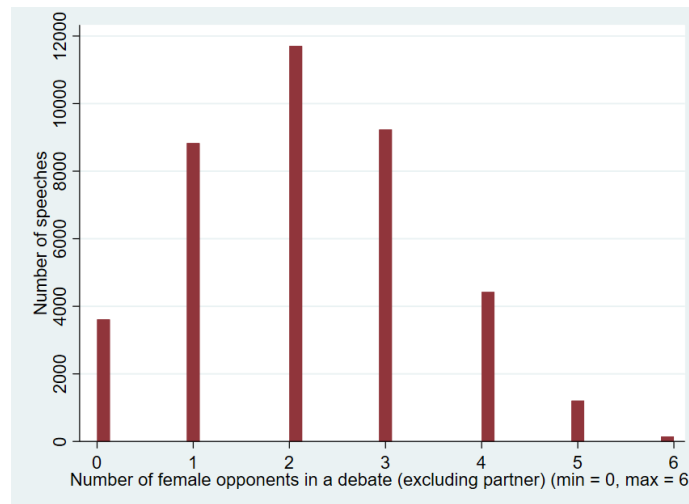


Figure 4.3: Number of female speakers in a debate room

4.8.2.3 Spearman's correlation coefficient heat map across characteristics of speakers, debate room and judges

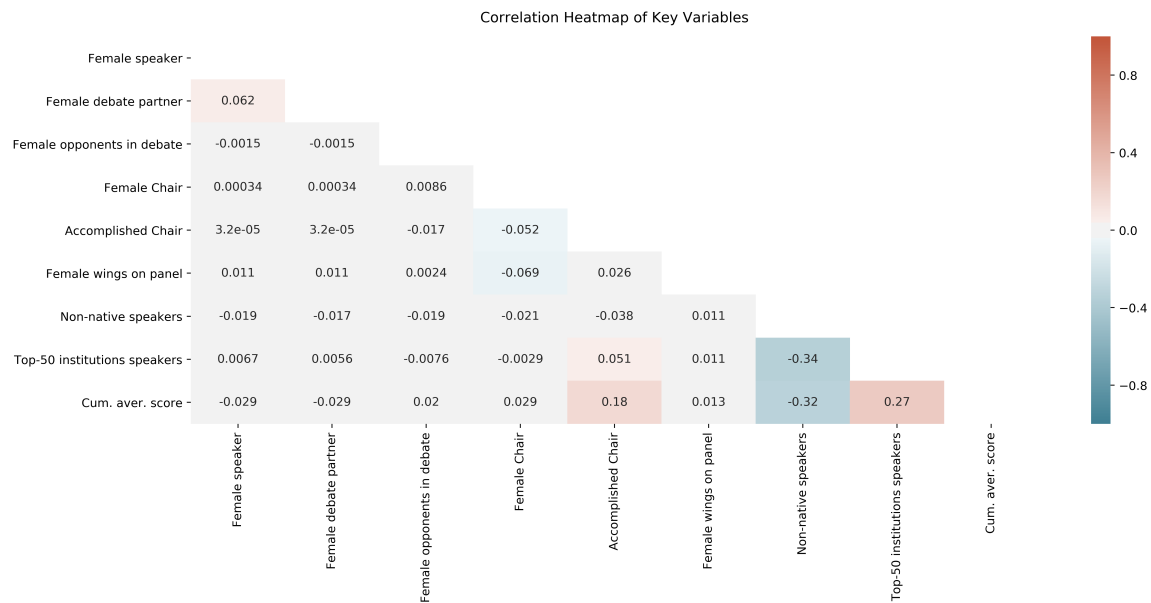


Figure 4.4: Spearman's correlation coefficient heat map across characteristics of speakers, debate room and judges. Accomplished chairs are judges who have advanced to at least one previous EUDC/WUDC tournaments.

4.8.2.4 Average scores across rounds

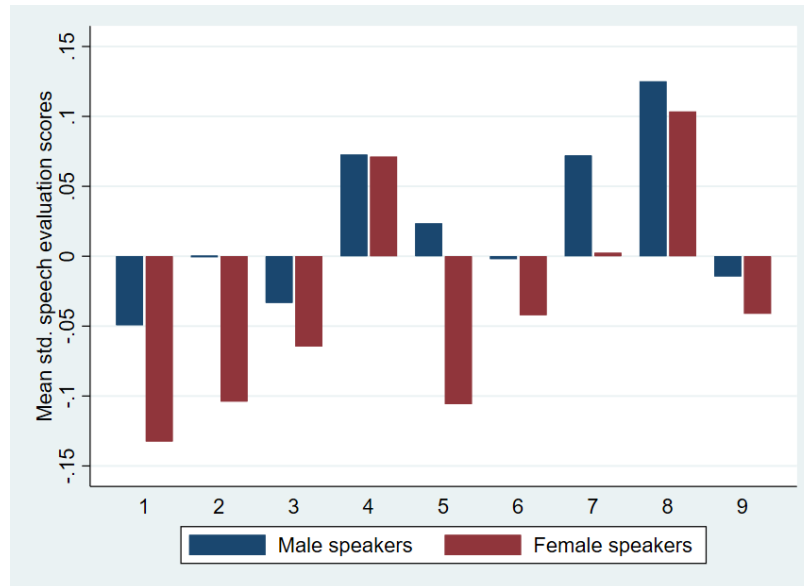


Figure 4.5: Average standardized scores of male vs. female speakers (R1 - R9)

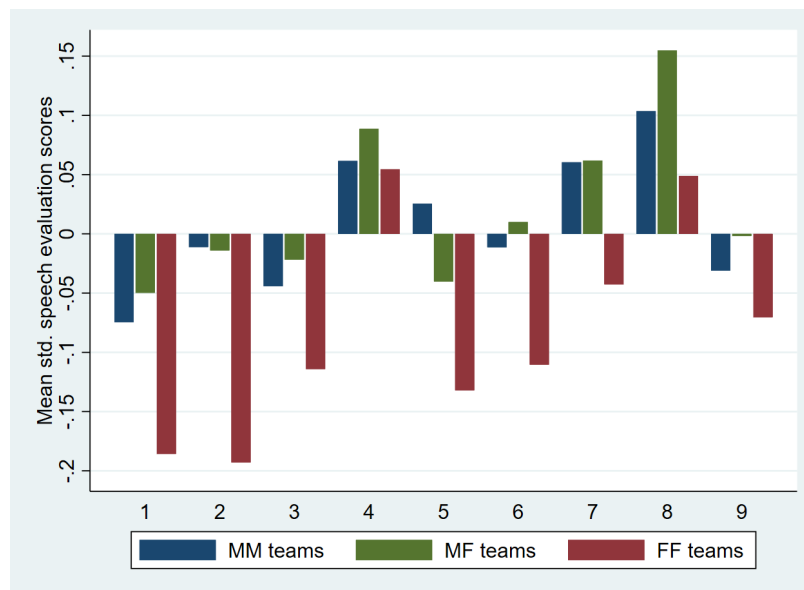


Figure 4.6: Average standardized scores of male-only vs. mixed vs. female-only teams (R1 - R9)

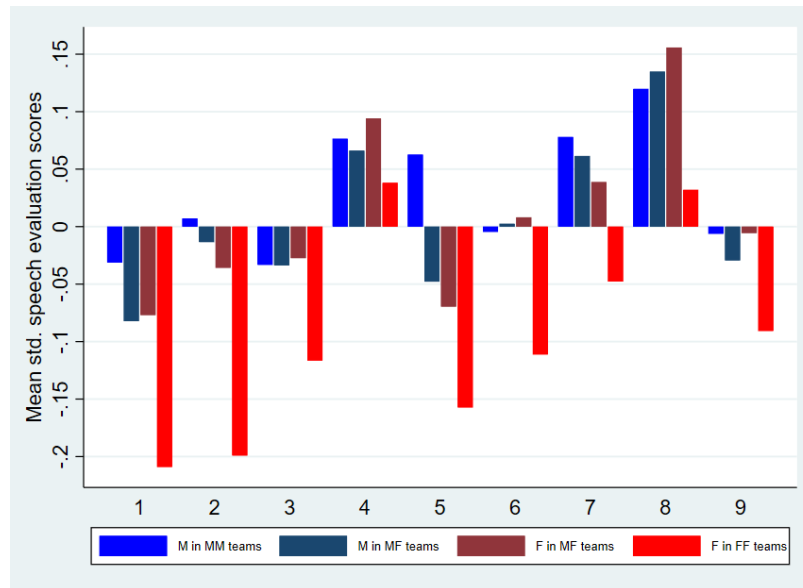


Figure 4.7: Average standardized scores speakers given teammate’s gender (R1 - R9)

4.8.2.5 Distribution of Speech Scores

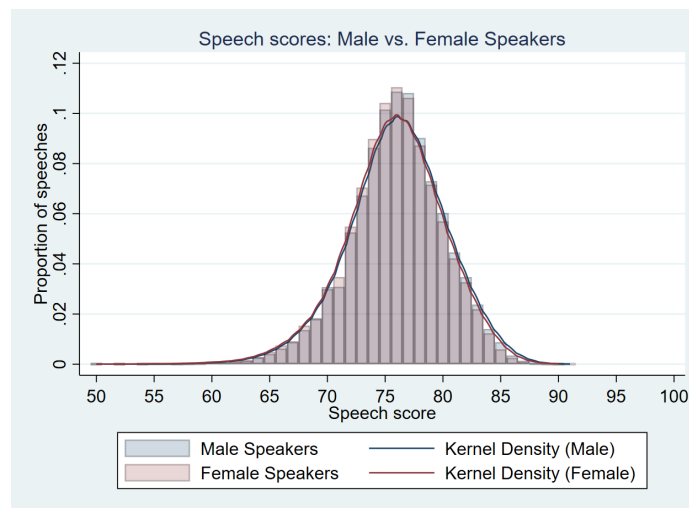


Figure 4.8: Speech score distribution by speaker’s gender

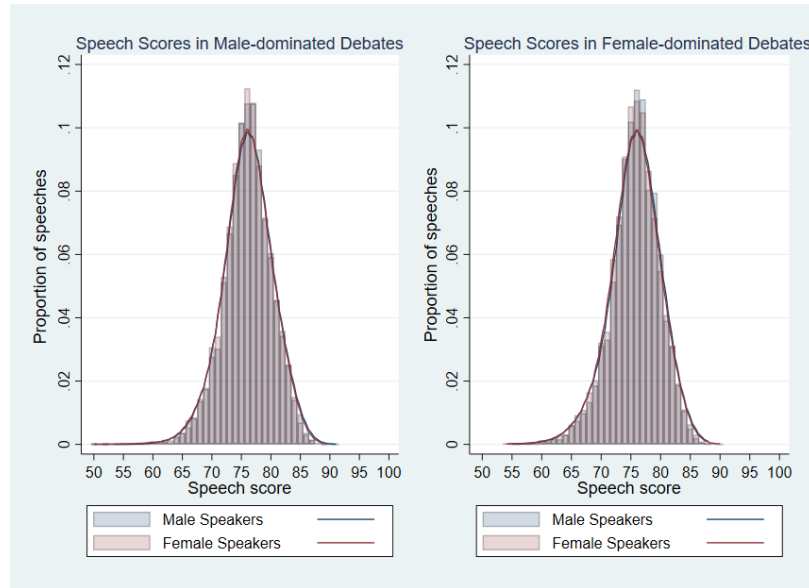


Figure 4.9: Speech score distribution by room gender composition

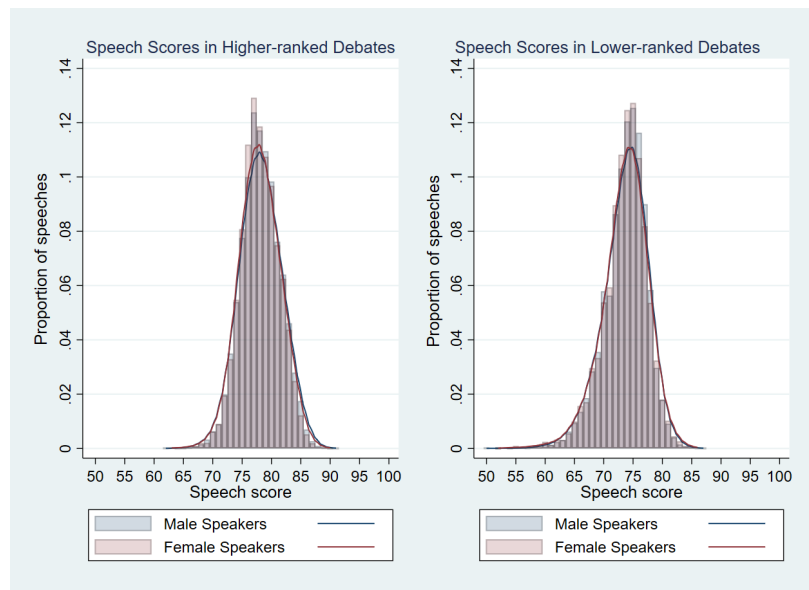


Figure 4.10: Speech score distribution by debate quality ranking

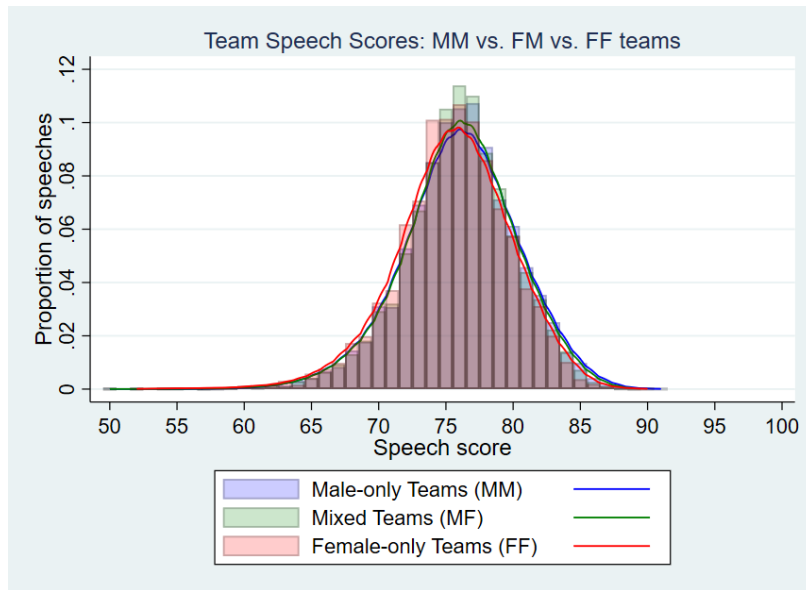


Figure 4.11: Distribution of team speech scores

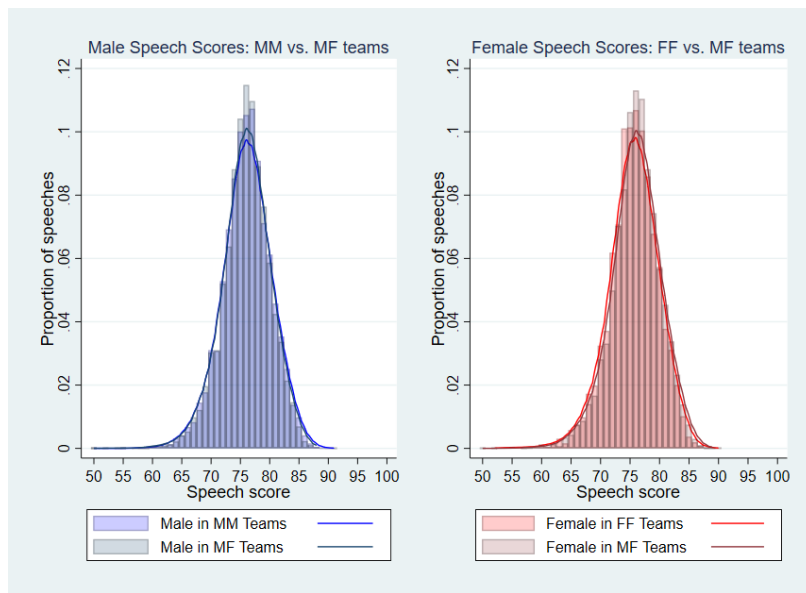


Figure 4.12: Distribution of individual speaker scores given their team gender composition

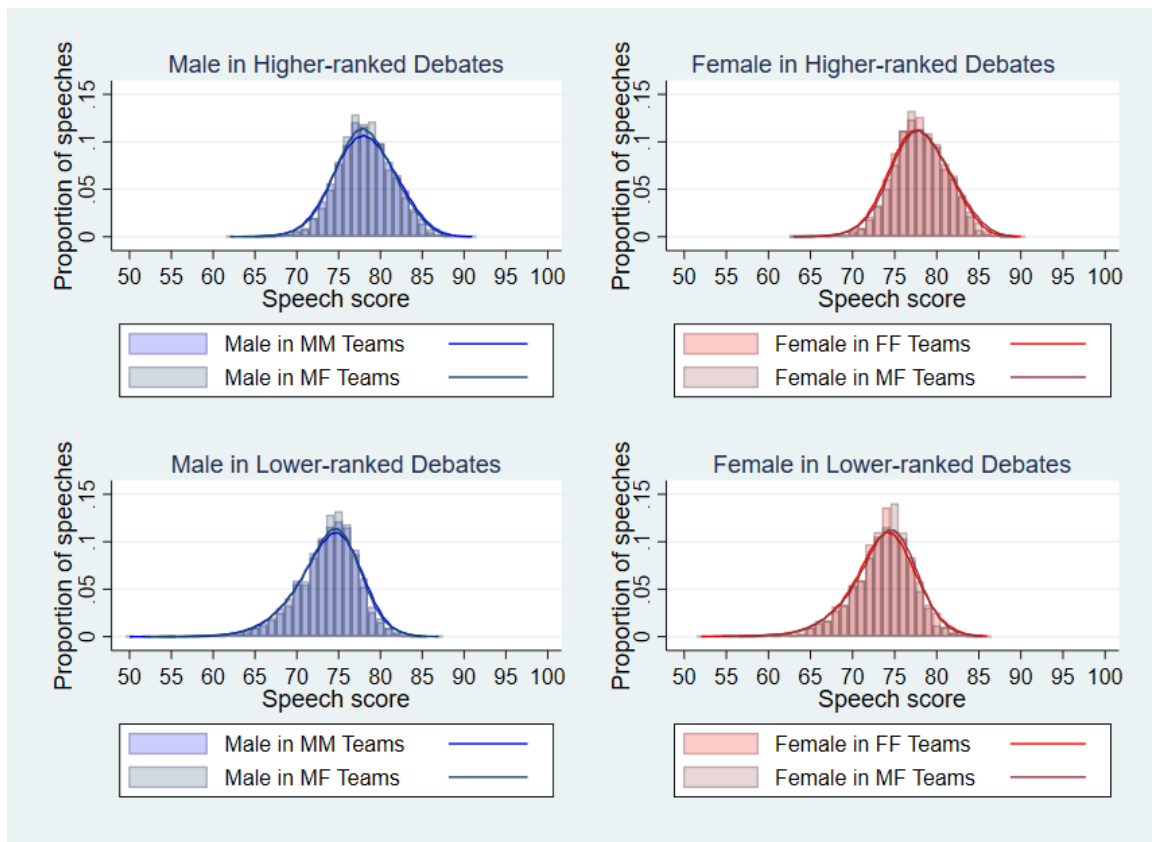


Figure 4.13: Distribution of speaker scores in higher vs. lower-ranked debates given team gender composition

4.8.3 Tables

4.8.3.1 Results: Gender of Speaker and Opponents, Round 1 vs. Round 2s to 9s (Indicator Variable)

Table 4.5: Regression Analysis: Gender of Speaker and Opponents, Round 1s only (N = 4376)

	Dependent Variable: Score (standardized)			
	(1)	(2)	(3)	(4)
Female Speaker	-0.083*** (0.03)		-0.075 (0.09)	-0.001 (0.08)
1 Female Opponent		0.030 (0.07)	0.013 (0.08)	0.035 (0.09)
2 Female Opponents		0.026 (0.07)	0.033 (0.08)	-0.130 (0.09)
3 Female Opponents		0.028 (0.07)	0.052 (0.08)	-0.083 (0.10)
4 Female Opponents		-0.051 (0.09)	-0.068 (0.10)	-0.225** (0.11)
5 Female Opponents		-0.176* (0.11)	-0.155 (0.12)	-0.135 (0.15)
6 Female Opponents		0.056 (0.18)	-0.034 (0.22)	-0.034 (0.34)
Female Speaker × 1 Female Opponent			0.045 (0.11)	-0.167* (0.10)
Female Speaker × 2 Female Opponents			-0.028 (0.11)	-0.078 (0.09)
Female Speaker × 3 Female Opponents			-0.056 (0.11)	-0.116 (0.10)
Female Speaker × 4 Female Opponents			0.049 (0.14)	0.046 (0.12)
Female Speaker × 5 Female Opponents			-0.029 (0.15)	-0.210* (0.12)
Female Speaker × 6 Female Opponents			0.303 (0.25)	0.186 (0.32)
Speaker Controls				✓
Room Controls				✓
R^2	0.002	0.003	0.005	0.363
Observations	4376	4376	4376	4376

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year.

Robust clustered standard errors at debate level in parentheses.

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

Table 4.6: Regression Analysis: Gender of Speaker and Opponents, Round 2s to 9s (N = 34 792)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Speaker	-0.053*** (0.01)		-0.049 (0.05)	-0.010 (0.04)		
1 Female Opponent		0.061* (0.04)	0.064 (0.05)	0.053 (0.04)	0.003 (0.03)	0.006 (0.03)
2 Female Opponents		0.056 (0.04)	0.054 (0.05)	0.049 (0.04)	0.002 (0.03)	0.016 (0.03)
3 Female Opponents		0.021 (0.04)	0.025 (0.05)	0.030 (0.04)	0.006 (0.03)	0.025 (0.03)
4 Female Opponents		-0.052 (0.04)	-0.051 (0.05)	-0.011 (0.04)	0.028 (0.03)	0.034 (0.03)
5 Female Opponents		-0.054 (0.06)	-0.063 (0.07)	-0.011 (0.05)	-0.005 (0.05)	-0.019 (0.04)
6 Female Opponents		-0.148 (0.11)	-0.152 (0.14)	-0.027 (0.12)	0.125 (0.09)	0.142* (0.08)
Female Speaker × 1 Female Opponent			-0.007 (0.06)	-0.038 (0.04)	-0.023 (0.04)	-0.016 (0.03)
Female Speaker × 2 Female Opponents			0.003 (0.06)	-0.047 (0.04)	-0.023 (0.04)	-0.026 (0.03)
Female Speaker × 3 Female Opponents			-0.014 (0.06)	-0.051 (0.04)	-0.014 (0.04)	-0.023 (0.03)
Female Speaker × 4 Female Opponents			-0.002 (0.07)	-0.012 (0.05)	-0.048 (0.04)	-0.033 (0.04)
Female Speaker × 5 Female Opponents			0.023 (0.09)	-0.030 (0.07)	0.004 (0.07)	0.020 (0.06)
Female Speaker × 6 Female Opponents			-0.002 (0.17)	0.088 (0.15)	0.015 (0.18)	-0.047 (0.11)
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓
R^2	0.001	0.002	0.002	0.338	0.572	0.648
Observations	34792	34792	34792	34792	34786	34786

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year. Robust clustered standard errors at debate level in parentheses.

R^2 of model (5) and (6) is $R^2_{between}$. Singleton observations are dropped in model (5) and (6).

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

Table 4.7: Regression Analysis: Gender of Speaker and Opponents, Round 2s to 9s, controlling for team standing (N = 34 792)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Speaker	-0.030*** (0.01)		0.041 (0.04)	0.021 (0.03)		
1 Female Opponent		-0.094*** (0.03)	-0.064 (0.04)	0.023 (0.03)	0.004 (0.03)	0.007 (0.03)
2 Female Opponents		-0.133*** (0.03)	-0.104** (0.04)	0.019 (0.03)	0.004 (0.03)	0.017 (0.03)
3 Female Opponents		-0.153*** (0.03)	-0.125*** (0.04)	0.005 (0.04)	0.008 (0.03)	0.026 (0.03)
4 Female Opponents		-0.210*** (0.04)	-0.177*** (0.05)	-0.027 (0.04)	0.029 (0.03)	0.035 (0.03)
5 Female Opponents		-0.191*** (0.05)	-0.156*** (0.06)	-0.025 (0.05)	-0.005 (0.05)	-0.018 (0.04)
6 Female Opponents		-0.271*** (0.10)	-0.227* (0.12)	-0.043 (0.11)	0.127 (0.09)	0.143* (0.08)
Female Speaker × 1 Female Opponent			-0.077 (0.05)	-0.063 (0.04)	-0.022 (0.04)	-0.015 (0.03)
Female Speaker × 2 Female Opponents			-0.076 (0.05)	-0.073* (0.04)	-0.022 (0.04)	-0.026 (0.03)
Female Speaker × 3 Female Opponents			-0.075 (0.05)	-0.068* (0.04)	-0.013 (0.04)	-0.023 (0.03)
Female Speaker × 4 Female Opponents			-0.084 (0.06)	-0.051 (0.05)	-0.047 (0.04)	-0.032 (0.04)
Female Speaker × 5 Female Opponents			-0.093 (0.08)	-0.070 (0.07)	0.007 (0.07)	0.022 (0.06)
Female Speaker × 6 Female Opponents			-0.136 (0.16)	0.017 (0.14)	0.017 (0.18)	-0.045 (0.11)
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓
R^2	0.171	0.173	0.174	0.387	0.572	0.648
Observations	34792	34792	34792	34792	34786	34786

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year. Robust clustered standard errors at debate level in parentheses.

R^2 of model (5) and (6) is $R^2_{between}$. Singleton observations are dropped in model (5) and (6).

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

4.8.3.2 Extensions: Higher vs. Lower-ranked Debates

Table 4.8: Regression Analysis: Gender of Speakers and Opponents, Higher vs. Lower-ranked Debates (controlling for team standing)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Higher-ranked Debates						
Female Speaker	-0.031** (0.01)		0.028 (0.03)	0.005 (0.02)		
Number of Female Opponents		-0.023*** (0.01)	-0.013 (0.01)	-0.008 (0.01)	-0.003 (0.01)	0.001 (0.01)
Female Speaker × Number of Female Opponents			-0.028** (0.01)	-0.026*** (0.01)	-0.023** (0.01)	-0.022** (0.01)
R^2	0.040	0.041	0.042	0.243	0.423	0.527
Observations	18270	18270	18270	18270	18046	18046
Lower-ranked Debates						
Female Speaker	-0.008 (0.02)		-0.022 (0.03)	-0.038 (0.03)		
Number of Female Opponents		-0.022** (0.01)	-0.024** (0.01)	-0.007 (0.01)	0.016 (0.01)	0.006 (0.01)
Female Speaker × Number of Female Opponents			0.006 (0.01)	0.012 (0.01)	0.016 (0.01)	0.015 (0.01)
R^2	0.051	0.052	0.052	0.276	0.399	0.551
Observations	16518	16518	16518	16518	16331	16331
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓

All models control for team standing (i.e. average cumulative speech scores of two speakers in a team up until the respective round.).

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type, (vii) competition & year. Robust clustered standard errors at debate level in parentheses. R^2 of model (5) and (6)

is $R^2_{between}$. Singleton observations are dropped in model (5) and (6).

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

Table 4.9: Regression Analysis: Gender of Speakers and Number of Opponents, Higher-ranked Debates only (controlling for debate room quality)

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Female Speaker	-0.031** (0.01)		0.074 (0.05)	0.060 (0.04)		
1 Female Opponent		-0.027 (0.04)	0.014 (0.05)	0.025 (0.04)	0.035 (0.04)	0.028 (0.03)
2 Female Opponents		-0.041 (0.04)	-0.007 (0.05)	0.016 (0.04)	0.025 (0.04)	0.027 (0.03)
3 Female Opponents		-0.050 (0.04)	-0.013 (0.05)	0.010 (0.04)	0.020 (0.04)	0.029 (0.03)
4 Female Opponents		-0.111** (0.04)	-0.044 (0.05)	-0.018 (0.05)	0.002 (0.04)	0.015 (0.04)
5 Female Opponents		-0.108* (0.06)	-0.049 (0.07)	-0.021 (0.06)	0.019 (0.06)	0.010 (0.05)
6 Female Opponents		-0.287 (0.20)	-0.064 (0.18)	-0.090 (0.18)	0.010 (0.15)	-0.000 (0.14)
Female Speaker × 1 Female Opponent			-0.113* (0.06)	-0.111** (0.05)	-0.127*** (0.05)	-0.113*** (0.04)
Female Speaker × 2 Female Opponents			-0.095 (0.06)	-0.112** (0.05)	-0.092** (0.05)	-0.095** (0.04)
Female Speaker × 3 Female Opponents			-0.102* (0.06)	-0.113** (0.05)	-0.113** (0.05)	-0.103** (0.04)
Female Speaker × 4 Female Opponents			-0.178*** (0.07)	-0.165*** (0.06)	-0.169*** (0.05)	-0.166*** (0.05)
Female Speaker × 5 Female Opponents			-0.169* (0.10)	-0.168* (0.09)	-0.163* (0.09)	-0.140* (0.08)
Female Speaker × 6 Female Opponents			-0.860** (0.38)	-0.716** (0.31)	-0.423 (0.27)	-0.410* (0.24)
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓
R^2	0.040	0.041	0.042	0.244	0.424	0.528
Observations	18270	18270	18270	18270	18046	18046

All models control for debate room quality (i.e. average cumulative speaker's score up until the respective round.).

Speaker controls include: (i) language skill status and (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) debate partner's gender, (iii) wing gender composition, (iv) speaking position, (v) round, (vi) motion type,

(vii) competition & year. Robust clustered standard errors at debate level in parentheses. R^2 of model (5) and (6)

is $R^2_{between}$. Singleton observations are dropped in model (5) and (6).

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

4.8.3.3 Extension: Speaker ft. Partner's Gender

Table 4.10: Regression Analysis: Speaker ft. Partner's Gender, All rounds

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Male Speaker in MF Team	-0.023 (0.02)		-0.033 (0.03)	-0.055** (0.03)	-0.029 (0.03)	-0.034 (0.03)
Female Speaker in MF Team	-0.021 (0.02)		-0.021 (0.03)	-0.041 (0.03)	0.087** (0.04)	0.070** (0.03)
Female Speaker in FF Team	-0.126*** (0.02)		-0.127*** (0.05)	-0.114*** (0.04)		
Number of Female Opponents		-0.020** (0.01)	-0.021* (0.01)	-0.015 (0.01)	0.007 (0.01)	0.007 (0.01)
Male Speaker in MF Team × Number of Female Opponents			0.005 (0.01)	0.005 (0.01)	-0.012 (0.01)	-0.008 (0.01)
Female Speaker in MF Team × Number of Female Opponents			0.000 (0.01)	-0.001 (0.01)	-0.013 (0.01)	-0.009 (0.01)
Female Speaker in FF Team × Number of Female Opponents			0.001 (0.02)	-0.001 (0.02)	0.000 (0.01)	-0.003 (0.01)
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓
\bar{R}^2	0.002	0.001	0.002	0.314	0.568	0.640
F	12.2	6.6	5.4	84.7	2.7	19.2
Observations	39168	39168	39168	39168	39157	39157

Speaker controls include: (i) language skill status, (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) wing gender composition, (iii) speaking position, (iv) round, (v) motion type, (vi) competition & year. Robust clustered standard errors at debate level in parentheses. R^2 of (5) and (6) is $R^2_{between}$. Singleton observations are dropped.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4.11: Regression Analysis: Speaker ft. Partner's Gender, Round by Round

	Dependent Variable: Score (standardized)					
	(1)	(2)	(3)	(4)	(5)	(6)
Round 1s						
Male Speaker in MF Team	-0.051 (0.04)		-0.148* (0.08)	-0.149* (0.09)		
Female Speaker in MF Team	-0.046 (0.04)		-0.108 (0.08)	-0.109 (0.08)		
Female Speaker in FF Team	-0.178*** (0.06)		-0.133 (0.11)	-0.226* (0.12)		
Number of Female Opponents		-0.019 (0.01)	-0.031 (0.02)	-0.060** (0.03)		
Male Speaker in MF Team × Number of Female Opponents			0.045 (0.03)	0.039 (0.03)		
Female Speaker in MF Team × Number of Female Opponents			0.030 (0.03)	0.022 (0.03)		
Female Speaker in FF Team × Number of Female Opponents			-0.017 (0.04)	0.015 (0.04)		
R^2	0.005	0.001	0.006	0.361		
Observations	4376	4376	4376	4376		
Round 2s to 9s						
Male Speaker in MF Team	-0.020 (0.02)		-0.019 (0.04)	-0.039 (0.03)	-0.012 (0.03)	-0.020 (0.03)
Female Speaker in MF Team	-0.018 (0.02)		-0.011 (0.04)	-0.029 (0.03)	0.089** (0.04)	0.070* (0.04)
Female Speaker in FF Team	-0.119*** (0.02)		-0.131** (0.05)	-0.109*** (0.04)		
Number of Female Opponents		-0.020** (0.01)	-0.020* (0.01)	-0.009 (0.01)	0.012 (0.01)	0.011 (0.01)
Male Speaker in MF Team × Number of Female Opponents			-0.001 (0.01)	-0.001 (0.01)	-0.019* (0.01)	-0.014 (0.01)
Female Speaker in MF Team × Number of Female Opponents			-0.003 (0.01)	-0.006 (0.01)	-0.016 (0.01)	-0.011 (0.01)
Female Speaker in FF Team × Number of Female Opponents			0.005 (0.02)	0.005 (0.02)	-0.001 (0.01)	-0.004 (0.01)
R^2	0.002	0.001	0.002	0.337	0.572	0.648
Observations	34792	34792	34792	34792	34786	34786
Round 2s to 9s (controlling for debate room quality)						
Male Speaker in MF Team	-0.012 (0.02)		0.018 (0.03)	-0.026 (0.03)	-0.013 (0.03)	-0.020 (0.03)
Female Speaker in MF Team	-0.009 (0.02)		0.026 (0.03)	-0.016 (0.03)	0.091** (0.04)	0.072* (0.04)
Female Speaker in FF Team	-0.069*** (0.02)		-0.032 (0.05)	-0.065* (0.04)		
Number of Female Opponents		-0.040*** (0.01)	-0.031*** (0.01)	-0.007 (0.01)	0.012 (0.01)	0.011 (0.01)
Male Speaker in MF Team × Number of Female Opponents			-0.014 (0.01)	-0.005 (0.01)	-0.019* (0.01)	-0.013 (0.01)
Female Speaker in MF Team × Number of Female Opponents			-0.017 (0.01)	-0.009 (0.01)	-0.016 (0.01)	-0.010 (0.01)
Female Speaker in FF Team × Number of Female Opponents			-0.017 (0.02)	-0.007 (0.02)	-0.001 (0.01)	-0.004 (0.01)
R^2	0.171	0.173	0.174	0.387	0.573	0.648
Observations	34792	34792	34792	34792	34786	34786
Speaker Controls				✓		
Room Controls				✓		✓
Speaker FE					✓	✓

Debate room quality is the average cumulative speaker's score up until the respective round. Speaker controls include: (i) language skill status, (ii) institution ranking. Room controls include: (i) chair judge fixed effect, (ii) wing gender composition, (iii) speaking position, (iv) round, (v) motion type, (vi) competition & year. Robust clustered standard errors at debate level in parentheses. R^2 of (5) and (6) is $R^2_{between}$. Singleton observations are dropped. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Summary

In this dissertation, I investigate gender disparities in speech patterns and how they matter in performance evaluation across genders, as well as how the gender composition of committees and opponents causally impact speech performance in real-life tournaments. Chapter 2 links the persuasion-relevant linguistic elements of debate speeches to speech evaluation scores, taking into account the interplay across genders of speakers and judges. Here, I find significant differences in persuasive speech patterns between men and women. Specifically, female speakers use more personal and disclosing speaking style, with more hedging phrases and non-fluencies in their speeches. In their answers to questions from opponents, women negate less, while having longer and more vague answers. On average, women receive lower evaluation scores than men. Across debates, having a less analytical speaking style and more positive sentiment is associated with higher scores for speeches by women, but not by men. Within debates, except for non-fluencies, there is no robust evidence of gender-specific evaluation standards. These findings suggest that the difference in average speech score between men and women arises because speeches of female speakers contain more score-reducing and fewer score-enhancing features, rather than discrimination.

In Chapter 3, I study how the gender composition and power hierarchy of judge committees causally impact performance evaluation patterns across male and female contestants. Committees with a female chair judge give lower scores to both male and female speakers, particularly in higher-ranked debates. Importantly, there is no difference between male and female speakers in how their scores are affected if the judge committee contains more women or is chaired by a woman. These results suggest that gender quotas on evaluation committees does not necessarily eliminate the glass ceiling for women.

Finally, Chapter 4 examines whether the gender composition of opponents affects the competitive performance of men and women in multi-round, high-stake contests. On average, neither male nor female contestants are affected by the gender composition of opponents. Nevertheless, in higher-ranked debates, female contestants perform comparatively worse in rooms with more female opponents. Therefore, these findings indicate that larger inflow of women into same competitions for high-profile positions does not necessarily reduce the thickness of the glass ceiling.

Nederlandse Samenvatting (Summary in Dutch)

In dit proefschrift onderzoek ik verschillen tussen mannen en vrouwen in spraakpatronen en hoe deze van belang zijn voor de evaluatie van prestaties van zowel mannen als vrouwen. Ook onderzoek ik hoe de samenstelling van commissies en tegenstanders in termen van geslacht een oorzakelijk effect heeft op prestaties in debat-toernooien. Hoofdstuk 2 koppelt de linguïstische elementen van toespraken die relevant zijn voor overredingskracht aan de score toegekend aan de toespraak, waarbij rekening wordt gehouden met de interactie in termen van geslacht tussen sprekers en juryleden. Hier vind ik significante verschillen in spraakpatronen tussen mannen en vrouwen. Over het algemeen geldt dat een minder analytische spreekstijl en een positiever sentiment geassocieerd is met hogere scores voor toespraken door vrouwen, maar niet voor toespraken door mannen. Binnen een debat is er echter geen robuust bewijs van genderspecifieke evaluatienormen, met uitzondering van vloeiend taalgebruik. Deze bevindingen suggereren dat het verschil in gemiddelde scores tussen mannen en vrouwen ontstaat doordat toespraken van vrouwelijke sprekers meer score-verlagende en minder score-verhogende kenmerken bevatten, in plaats van discriminatie.

In hoofdstuk 3 bestudeer ik hoe de gendersamenstelling en hiërarchie van jurycommissies van invloed zijn op hun evaluatie van mannelijke en vrouwelijke deelnemers. Commissies met een vrouwelijke juryvoorzitter geven lagere scores aan zowel mannelijke als vrouwelijke sprekers, vooral in hoger gerangschikte debatten. Een belangrijk resultaat is dat er geen verschil is tussen mannelijke en vrouwelijke sprekers in hoe hun score wordt beïnvloedt door het aantal vrouwen in de jurycommissie of het geslacht van de juryvoorzitter. Deze resultaten suggereren dat genderquota in evaluatiecommissies niet altijd een afdoende

maatregel is om het glazen plafond voor vrouwen te doorbreken.

Ten slotte wordt in Hoofdstuk 4 onderzocht of het geslacht van tegenstanders de competitieve prestatie van mannen en vrouwen beïnvloedt in toernooien met meerdere rondes en hoge belangen. Ik vind dat gemiddeld mannelijke noch vrouwelijke deelnemers beïnvloed worden door het geslacht van hun tegenstanders. Niettemin presteren vrouwelijke deelnemers in hoger gerangschikte debatten relatief slechter in kamers met meer vrouwelijke tegenstanders. Daarom geven deze bevindingen aan dat een grotere instroom van vrouwen in competitieve trajecten voor belangrijke posities niet vanzelf leidt tot een vermindering van de dikte van het glazen plafond.

Bibliography

- Abe, J. A. A. (2011). Changes in Alan Greenspan's language use across the economic cycle: A text analysis of his testimonies and speeches. *Journal of Language and Social Psychology*, 30(2):212–223.
- Adams, R. B. and Ferreira, D. (2009). Women in the boardroom and their impact on governance and performance. *Journal of Financial Economics*, 94(2):291–309.
- Adams, R. B. and Funk, P. (2012). Beyond the glass ceiling: Does gender matter? *Management Science*, 58(2):219–235.
- Adams, R. B. and Kirchmaier, T. (2016). Women on boards in finance and stem industries. *American Economic Review*, 106(5):277–81.
- Alan, S., Ertac, S., Kubilay, E., and Loranth, G. (2020). Understanding gender differences in leadership. *Economic Journal*, 130(626):263–289.
- Antonovics, K., Arcidiacono, P., and Walsh, R. (2009). The effects of gender interactions in the lab and in the field. *The Review of Economics and Statistics*, 91(1):152–162.
- Apesteguia, J., Azmat, G., and Iriberry, N. (2012). The impact of gender composition on team performance and decision making: Evidence from the field. *Management Science*, 58(1):78–93.
- Arvate, P. R., Galilea, G. W., and Todescat, I. (2018). The queen bee: A myth? the effect of top-level female leadership on subordinate females. *Leadership Quarterly*, 29(5):533–548.
- Ash, E., Chen, D. L., and Lu, W. (2017). Polarization of us circuit court judges: A machine learning approach.

- Ash, E., Chen, D. L., and Ornaghi, A. (2021). Gender attitudes in the judiciary: Evidence from us circuit courts. *Center for Law & Economics Working Paper Series*, 2019(02).
- Ashraf, N., Bau, N., Low, C., and McGinn, K. (2020). Negotiating a better future: How interpersonal skills facilitate inter-generational investment.
- Azmat, G., Boring, A., et al. (2020). Gender diversity in firms. *Oxford Review of Economic Policy*, 36(4).
- Azmat, G. and Ferrer, R. (2017). Gender gaps in performance: Evidence from young lawyers. *Journal of Political Economy*, 125(5):1306–1355.
- Backus, P., Cubel, M., Guid, M., Sanchez-Pages, S., and Mañas, E. (2016). Gender, competition and performance: Evidence from real tournaments.
- Bagues, M. (2017). Can gender quotas in candidate lists empower women? evidence from a regression discontinuity design.
- Bagues, M. and Campa, P. (2021). Can gender quotas in candidate lists empower women? evidence from a regression discontinuity design. *Journal of Public Economics*, 194:104315.
- Bagues, M., Sylos-Labini, M., and Zinovyeva, N. (2017). Does the gender composition of scientific committees matter? *American Economic Review*, 107(4):1207–38.
- Bagues, M. F. and Esteve-Volart, B. (2010). Can gender parity break the glass ceiling? evidence from a repeated randomized experiment. *Review of Economic Studies*, 77(4):1301–1328.
- Balachandra, L., Briggs, A. R., Eddleston, K., and Brush, C. (2013). Pitch like a man: Gender stereotypes and entrepreneur pitch success. *Frontiers of Entrepreneurship Research*, 33(8):2.
- Balachandra, L., Briggs, T., Eddleston, K., and Brush, C. (2019). Don't pitch like a girl! how gender stereotypes influence investor decisions. *Entrepreneurship Theory and Practice*, 43(1):116–137.
- Baldez, L. (2004). Elected bodies: The gender quota law for legislative candidates in Mexico. *Legislative Studies Quarterly*, 29(2):231–258.

- Baltrunaite, A., Bello, P., Casarico, A., and Profeta, P. (2014). Gender quotas and the quality of politicians. *Journal of Public Economics*, 118:62–74.
- Beg, S., Fitzpatrick, A., and Lucas, A. M. (2021). Gender bias in assessments of teacher performance. In *American Economic Review Papers and Proceedings*.
- Benson, A., Li, D., and Shue, K. (2021). “potential” and the gender promotion gap. Technical report, Working Paper.
- Bertrand, M., Black, S. E., Jensen, S., and Lleras-Muney, A. (2019). Breaking the glass ceiling? the effect of board quotas on female labour market outcomes in norway. *The Review of Economic Studies*, 86(1):191–239.
- Bertrand, M. and Hallock, K. F. (2001). The gender gap in top corporate jobs. *ILR Review*, 55(1):3–21.
- Biber, D., Douglas, B., Conrad, S., and Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Blau, F. D. and Kahn, L. M. (2007). The gender pay gap: Have women gone as far as they can? *Academy of Management Perspectives*, 21(1):7–23.
- Blau, F. D. and Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865.
- Blumenau, J. (2019). The effects of female leadership on women’s voice in political debate. *British Journal of Political Science*, pages 1–22.
- Bohnet, I. and Bowles, H. R. (2008). Gender in negotiation. *Negotiation Journal*, 24(4):389.
- Bohren, A., Imas, A., and Rosenberg, M. (2018). The language of discrimination: Using experimental versus observational data. In *American Economic Review Papers and Proceedings*, volume 108, pages 169–74.
- Bohren, J. A., Imas, A., and Rosenberg, M. (2019). The dynamics of discrimination: Theory and evidence. *American Economic Review*, 109(10):3395–3436.

- Booth, A. and Yamamura, E. (2018). Performance in mixed-sex and single-sex competitions: What we can learn from speedboat races in Japan. *Review of Economics and Statistics*, 100(4):581–593.
- Bordalo, P., Coffman, K., Gennaioli, N., and Shleifer, A. (2016). Stereotypes. *Quarterly Journal of Economics*, 131(4):1753–1794.
- Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145:27–41.
- Born, A., Raney, E., and Sandberg, A. (2020). Gender and willingness to lead: Does the gender composition of teams matter? *Review of Economics and Statistics*, pages 1–46.
- Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147.
- Bradac, J. J., Hemphill, M. R., and Tardy, C. H. (1981). Language style on trial: Effects of “powerful” and “powerless” speech upon judgments of victims and villains. *Western Journal of Speech Communication*, 45(4):327–341.
- Bradac, J. J. and Mulac, A. (1984). Attributional consequences of powerful and powerless speech styles in a crisis-intervention context. *Journal of Language and Social Psychology*, 3(1):1–19.
- Bradac, J. J., Mulac, A., and Thompson, S. A. (1995). Men’s and women’s use of intensifiers and hedges in problem-solving interaction: Molar and molecular analyses. *Research on Language and Social Interaction*, 28(2):93–116.
- Brooks, A. W., Huang, L., Kearney, S. W., and Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, 111(12):4427–4431.
- Brown, R. W. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology*, 55(1):1.
- Bruschke, J. and Johnson, A. (1994). An analysis of differences in success rates of male and female debaters. *Argumentation and Advocacy*, 30(3):162–174.

- Burrow, N., Fedorets, A., and Gibert, A. (2018). The effects of a gender quota on the board of german largest corporations. *Berlin: German Institute for Economic Research*.
- Buser, T., Niederle, M., and Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *Quarterly Journal of Economics*, 129(3):1409–1447.
- Buser, T. and Yuan, H. (2020). Public speaking aversion. *Available at SSRN 3724341. Tinbergen Institute Discussion Paper TI 2020 -074/I*.
- Carlsson, M. and Eriksson, S. (2019). In-group gender bias in hiring: Real-world evidence. *Economics Letters*, 185:108686.
- Chakrabarty, T., Hidey, C., Muresan, S., McKeown, K., and Hwang, A. (2019). Amper-sand: Argument mining for persuasive online discussions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP Empirical Methods in Natural Language Processing-IJCNLP)*, pages 2926–2936.
- Chambers, S. (2004). Behind closed doors: Publicity, secrecy, and the quality of deliberation. *Journal of Political Philosophy*, 12(4):389–410.
- Charteris-Black, J. (2011). *Politicians and rhetoric: The persuasive power of metaphor*. Springer.
- Charteris-Black, J. (2018). *Analysing political speeches*. Macmillan International Higher Education.
- Chikeleze, M., Johnson, I., and Gibson, T. (2018). Let’s argue: Using debate to teach critical thinking and communication skills to future leaders. *Journal of Leadership Education*, 17(2).
- Christine Banwart, M. and McKinney, M. S. (2005). A gendered influence in campaign debates? analysis of mixed-gender united states senate and gubernatorial debates. *Communication studies*, 56(4):353–373.
- Clark Blickenstaff, J. (2005). Women and science careers: leaky pipeline or gender filter? *Gender and Education*, 17(4):369–386.
- Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.

- Coffman, K. B. (2014). Evidence on self-stereotyping and the contribution of ideas. *Quarterly Journal of Economics*, 129(4):1625–1660.
- Coffman, K. B., Flikkema, C. B., and Shurchkov, O. (2019). *Gender Stereotypes in Deliberation and team decisions*. Harvard Business School.
- Cohn, M. A., Mehl, M. R., and Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding september 11, 2001. *Psychological Science*, 15(10):687–693.
- Conti, R., Collins, M. A., and Picariello, M. L. (2001). The impact of competition on intrinsic motivation and creativity: considering gender, gender segregation and gender role orientation. *Personality and individual differences*, 31(8):1273–1289.
- Cotton, C., McIntyre, F., and Price, J. (2013). Gender differences in repeated competition: Evidence from school math contests. *Journal of Economic Behavior & Organization*, 86:52–66.
- Crosby, F. and Nyquist, L. (1977). The female register: An empirical study of lakoff's hypotheses. *Language in Society*, 6(3):313–322.
- Cullen, Z. B. and Perez-Truglia, R. (2019). The old boys' club: Schmoozing and the gender gap. Technical report, National Bureau of Economic Research.
- Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., and Potts, C. (2013). A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.
- Dargnies, M.-P. (2012). Men too sometimes shy away from competition: The case of team competition. *Management Science*, 58(11):1982–2000.
- Datta Gupta, N., Poulsen, A., and Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, 51(1):816–835.
- De Paola, M., Gioia, F., and Scoppa, V. (2015). Are females scared of competing with males? results from a field experiment. *Economics of Education Review*, 48:117–128.
- De Paola, M., Gioia, F., and Scoppa, V. (2018). Teamwork, leadership and gender. Technical report, IZA Discussion Papers.

- De Paola, M. and Scoppa, V. (2015). Gender discrimination and evaluators' gender: evidence from Italian academia. *Economica*, 82(325):162–188.
- de Winter, J. C., Gosling, S. D., and Potter, J. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273.
- Delfgaauw, J., Dur, R., Sol, J., and Verbeke, W. (2013). Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326.
- Deming, D. J. (2017). The growing importance of social skills in the labor market. *Quarterly Journal of Economics*, 132(4):1593–1640.
- Derks, B., Van Laar, C., and Ellemers, N. (2016). The queen bee phenomenon: Why women leaders distance themselves from junior women. *Leadership Quarterly*, 27(3):456–469.
- Deschamps, P. et al. (2018). Gender quotas in hiring committees: a boon or a bane for women? *Sciences Po → LIEPP Working Paper*, 82.
- Dietrich, B. J., Hayes, M., and O'BRIEN, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review*, 113(4):941–962.
- Dillard, J. P. and Peck, E. (2000). Affect and persuasion: Emotional responses to public service announcements. *Communication Research*, 27(4):461–495.
- Dilmaghani, M. (2020). Gender differences in performance under time constraint: Evidence from chess tournaments. *Journal of Behavioral and Experimental Economics*, 89:101505.
- Dinkar, T., Vasilescu, I., Pelachaud, C., and Clavel, C. (2020). How confident are you? Exploring the role of fillers in the automatic prediction of a speaker's confidence. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8104–8108. IEEE.
- Dubiel, M., Halvey, M., Gallegos, P. O., and King, S. (2020). Persuasive synthetic speech: voice perception and user behaviour. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–9.

- Dubois, B. L. and Crouch, I. (1975). The question of tag questions in women's speech: They don't really use more of them, do they? *Language in Society*, 4(3):289–294.
- Dupas, P., Modestino, A. S., Niederle, M., Wolfers, J., et al. (2021). Gender and the dynamics of economics seminars. Technical report, National Bureau of Economic Research.
- Eckel, C. C., Gangadharan, L., Grossman, P. J., and Xue, N. (2020). *The Gender Leadership Gap: Insights from Experiments*. Monash University, Monash Business School, Department of Economics.
- Exley, C. L. and Kessler, J. B. (2019). The gender gap in self-promotion. Technical report, National Bureau of Economic Research.
- Exley, C. L., Niederle, M., and Vesterlund, L. (2020). Knowing when to ask: The cost of leaning in. *Journal of Political Economy*, 128(3):816–854.
- Fahy, P. J. (2002). Use of linguistic qualifiers and intensifiers in a computer conference. *The American Journal of Distance Education*, 16(1):5–22.
- Fallows, S. and Steven, C. (2000). Building employability skills into the higher education curriculum: a university-wide initiative. *Education+ training*.
- Fershtman, C., Segal, U., et al. (2020). *Social Influence in Committee Deliberation*. Boston College.
- Flinn, C., Todd, P., Zhang, W., et al. (2019). Personality traits, job search and the gender wage gap.
- Gatti, L., Guerini, M., Stock, O., and Strapparava, C. (2014). Sentiment variations in text for persuasion technology. In *International Conference on Persuasive Technology*, pages 106–117. Springer.
- Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, 57(3):535–74.
- Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

- Gerdes, C. and Gränsmark, P. (2010). Strategic behavior across gender: A comparison of female and male expert chess players. *Labour Economics*, 17(5):766–775.
- Geys, B. and Sørensen, R. J. (2019). The impact of women above the political glass ceiling: Evidence from a norwegian executive gender quota reform. *Electoral Studies*, 60:102050.
- Ghilzai, S. and Baloch, M. (2015). Conversational analysis of turn taking behavior and gender differences in multimodal conversation. *European Academic Research*, 3(9):10100–10116.
- Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *Quarterly Journal of Economics*, 118(3):1049–1074.
- Gneezy, U. and Rustichini, A. (2004). Gender and competition at a young age. *American Economic Review*, 94(2):377–381.
- Goeree, J. K. and Yariv, L. (2011). An experimental study of collective deliberation. *Econometrica*, 79(3):893–921.
- Goldin, C. (2014). A pollution theory of discrimination: male and female differences in occupations and earnings. In *Human Capital in History: The American Record*, pages 313–348. University of Chicago Press.
- Goldin, C., Kerr, S. P., Olivetti, C., and Barth, E. (2017). The expanding gender earnings gap: Evidence from the lehd-2000 census. *American Economic Review*, 107(5):110–14.
- Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741.
- Green, C. P. and Homroy, S. (2018). Female directors, board committees and firm performance. *European Economic Review*, 102:19–38.
- Green III, C. S. and Klug, H. G. (1990). Teaching critical thinking and writing through debates: An experimental evaluation. *Teaching Sociology*, pages 462–471.
- Grey, S. et al. (2002). Does size matter? critical mass and new zealand's women MP. *Oxford University Press*.

- Haas, A. (1979). Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, 86(3):616.
- Habernal, I. and Gurevych, I. (2016). What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1214–1223.
- Haire, S. B., Moyer, L. P., and Treier, S. (2013). Diversity, deliberation, and judicial opinion writing. *Journal of Law and Courts*, 1(2):303–330.
- Hanafiyeh, M. and Afghari, A. (2014). Gender differences in the use of hedges, tag questions, intensifiers, empty adjectives, and adverbs: A comparative study in the speech of men and women. *Indian Journal of Fundamental and Applied Life Sciences*, 4(4):1168–1177.
- Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *Quarterly Journal of Economics*, 133(2):801–870.
- Hargrave, L. and Langengen, T. (2020). The gendered debate: Do men and women communicate differently in the house of commons? *Politics & Gender*, pages 1–27.
- Hebert, C. (2018). Mind the gap: Gender stereotypes and entrepreneur financing. *Available at SSRN 3318245*.
- Hengel, E. (2020). Publishing while female (summary). *CEPR Press*.
- Heursen, L., Ranehill, E., and Weber, R. (2020). Are women less effective leaders than men? evidence from experiments using coordination games.
- Hirschberg, J. B. and Rosenberg, A. (2005). Acoustic/prosodic and lexical correlates of charismatic speech.
- Hoel, M. (2008). The quota story: Five years of change in norway. *Women on Corporate Boards of Directors: International Research and Practice*, pages 79–87.
- Holmes, J. (1984). Women’s language: A functional approach. *General Linguistics*, 24(3):149.

- Holmes, J. (1986). Functions of you know in women's and men's speech. *Language in Society*, 15(1):1–21.
- Holmes, J. (1990). Hedges and boosters in women's and men's speech. *Language & Communication*, 10(3):185–205.
- Holtgraves, T. (1997). Styles of language use: Individual and cultural variability in conversational indirectness. *Journal of Personality and Social Psychology*, 73(3):624.
- Hoogendoorn, S., Oosterbeek, H., and Van Praag, M. (2013). The impact of gender diversity on the performance of business teams: Evidence from a field experiment. *Management Science*, 59(7):1514–1528.
- Hosman, L. A. and Wright, J. W. (1987). The effects of hedges and hesitations on impression formation in a simulated courtroom context. *Western Journal of Communication (Includes Communication Reports)*, 51(2):173–188.
- Hoyt, C. L. and Murphy, S. E. (2016). Managing to clear the air: Stereotype threat, women, and leadership. *The Leadership Quarterly*, 27(3):387–399.
- Huang, J., Diehl, M.-R., and Paterlini, S. (2020). The influence of corporate elites on women on supervisory boards: female directors' inclusion in germany. *Journal of Business Ethics*, 165(2):347–364.
- Hutto, C. J. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.
- Iriberry, N. and Rey-Biel, P. (2019). Competitive pressure widens the gender gap in performance: Evidence from a two-stage competition in mathematics. *The Economic Journal*, 129(620):1863–1893.
- Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, 50(6):1141.
- Ivanova-Stenzel, R. and Kübler, D. (2011). Gender differences in team work and team competition. *Journal of Economic Psychology*, 32(5):797–808.
- Jamieson, K. H., Hall, K., et al. (1995). *Beyond the double bind: Women and leadership*. Oxford University Press on Demand.

- Jetter, M. and Walker, J. K. (2018). The gender of opponents: Explaining gender differences in performance and risk-taking? *European Economic Review*, 109:238–256.
- Joecks, J., Pull, K., and Vetter, K. (2013). Gender diversity in the boardroom and firm performance: What exactly constitutes a “critical mass?”. *Journal of Business Ethics*, 118(1):61–72.
- Karpowitz, C. F. and Mendelberg, T. (2014). *The Silent Sex: Gender, Deliberation, and Institutions*. Princeton University Press.
- Karpowitz, C. F., Mendelberg, T., and Shaker, L. (2012). Gender inequality in deliberative participation. *American Political Science Review*, pages 533–547.
- Kets, W. and Sandroni, A. (2019). A belief-based theory of homophily. *Games and Economic Behavior*, 115:410–435.
- Kim, D. and Starks, L. T. (2016). Gender diversity on corporate boards: Do women contribute unique skills? *American Economic Review*, 106(5):267–71.
- Kolev, J., Fuentes-Medel, Y., and Murray, F. (2020). Gender differences in scientific communication and their impact on grant funding decisions. In *American Economic Review Papers and Proceedings*, volume 110, pages 245–49.
- Kunze, A. and Miller, A. R. (2017). Women helping women? evidence from private sector data on workplace hierarchies. *Review of Economics and Statistics*, 99(5):769–775.
- Kuo, S.-H. (2003). Involvement vs detachment: gender differences in the use of personal pronouns in televised sports in Taiwan. *Discourse Studies*, 5(4):479–494.
- Lakoff, R. (1973). Language and woman’s place. *Language in Society*, 2(1):45–79.
- Latu, I. M., Mast, M. S., Lammers, J., and Bombari, D. (2013). Successful female leaders empower women’s behavior in leadership tasks. *Journal of Experimental Social Psychology*, 49(3):444–448.
- Leeper, C. and Robnett, R. D. (2011). Women are more likely than men to use tentative language, aren’t they? a meta-analysis testing for gender differences and moderators. *Psychology of Women Quarterly*, 35(1):129–142.

- Leibbrandt, A. and List, J. A. (2015). Do women avoid salary negotiations? evidence from a large-scale natural field experiment. *Management Science*, 61(9):2016–2024.
- Lenard, D. B. (2017). Gender differences in the personal pronouns usage on the corpus of congressional speeches. *Journal of Research Design and Statistics in Linguistics and Communication Studies*, 3(2):161.
- Levitan, S. I., Maredia, A., and Hirschberg, J. (2018). Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950.
- Levy, G. (2007). Decision making in committees: Transparency, reputation, and voting rules. *American Economic Review*, 97(1):150–168.
- Li, X., Han, Z., Fu, J., Mei, Y., and Liu, J. (2019). Debate: A new approach for improving the dialectical thinking of university students. *Innovations in Education and Teaching International*, pages 1–12.
- Lin, N. and Osnabrügge, M. (2018). Making comprehensible speeches when your constituents need it. *Research & Politics*, 5(3):2053168018795598.
- Lippi, M. and Torroni, P. (2016). Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.
- Lundgren, G. (2017). *Essays on job market screening, in-group bias and school competition*. Stockholm School of Economics.
- MacNeill, L., Driscoll, A., and Hunt, A. N. (2015). What’s in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4):291–303.
- Mago, S. D. and Razzolini, L. (2019). Best-of-five contest: An experiment on gender differences. *Journal of Economic Behavior & Organization*, 162:164–187.
- Maida, A. and Weber, A. (2019). Female leadership and gender gap within firms: Evidence from an italian board reform. *ILR Review*, page 0019793920961995.
- Manzoor, E., Chen, G. H., Lee, D., and Smith, M. D. (2020). Influence via ethos: On the persuasive power of reputation in deliberation online. *arXiv preprint arXiv:2006.00707*.

- Marcus, G. E., Neuman, W. R., and MacKuen, M. (2000). *Affective Intelligence and Political Judgment*. University of Chicago Press.
- Matsa, D. A. and Miller, A. R. (2011). Chipping away at the glass ceiling: Gender spillovers in corporate leadership. *American Economic Review*, 101(3):635–39.
- Matsa, D. A. and Miller, A. R. (2013). A female style in corporate leadership? evidence from quotas. *American Economic Journal: Applied Economics*, 5(3):136–69.
- Matz, S. and Bruschke, J. (2006). Gender inequity in debate, legal and business professions. *Contemporary Argumentation and Debate*, 27:29–47.
- McCroskey, J. C. and Mehrley, R. S. (1969). The effects of disorganization and nonfluency on attitude change and source credibility. *Communications Monographs*, 36(1):13–21.
- Mendelberg, T., Karpowitz, C. F., and Goedert, N. (2014). Does descriptive representation facilitate women’s distinctive voice? how gender composition and decision rules affect deliberation. *American Journal of Political Science*, 58(2):291–306.
- Menéndez, D. A., González-Barahona, J. M., and Robles, G. (2020). Damegender: Writing and comparing gender detection tools. In *SATToSE*.
- Mengel, F. (2020). Gender bias in opinion aggregation. *Available at SSRN 3572594*.
- Mengel, F., Sauermann, J., and Zölitz, U. (2018). Gender bias in teaching evaluations. *Journal of the European Economic Association*, 17(2):535–566.
- Mihatsch, W. (2012). Hedges. *The Encyclopedia of Applied Linguistics*.
- Miller, G. R. and Hewgill, M. A. (1964). The effect of variations in nonfluency on audience ratings of source credibility. *Quarterly Journal of Speech*, 50(1):36–44.
- Miller, N., Maruyama, G., Beaber, R. J., and Valone, K. (1976). Speed of speech and persuasion. *Journal of Personality and Social Psychology*, 34(4):615.
- Moely, B. E., Skarin, K., and Weil, S. (1979). Sex differences in competition—cooperation behavior of children at two age levels. *Sex Roles*, 5(3):329–342.
- Mulac, A. and Lundell, T. L. (1986). Linguistic contributors to the gender-linked language effect. *Journal of Language and Social Psychology*, 5(2):81–101.

- Mulac, A., Lundell, T. L., and Bradac, J. J. (1986). Male/female language differences and attributional consequences in a public speaking situation: Toward an explanation of the gender-linked language effect. *Communications Monographs*, 53(2):115–129.
- Newman, M. L., Groom, C. J., Handelman, L. D., and Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., and Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675.
- Niederle, M. (2017). A gender agenda: a progress report on competitiveness. *American Economic Review*, 107(5):115–19.
- Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *Quarterly Journal of Economics*, 122(3):1067–1101.
- Niederle, M. and Vesterlund, L. (2011). Gender and competition. *Annu. Rev. Econ.*, 3(1):601–630.
- Owen, A. L. and Temesvary, J. (2018). The performance effects of gender diversity on bank boards. *Journal of Banking & Finance*, 90:50–63.
- Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107. ACM.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. Technical report.
- Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., and Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.
- Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., and Burzo, M. (2015). Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 59–66.

- Petukhova, V., M. T., Malchanau, A., and Bunt, H. (2017a). Virtual debate coach design: assessing multimodal argumentation performance. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 41–50. ACM.
- Petukhova, V., Raju, M., and Bunt, H. (2017b). Multimodal markers of persuasive speech: designing a virtual debate coach. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH). International Speech Communication Association (ISCA), Baixas, France, Stockholm, Sweden*, pages 142–146.
- Phillimore, E. (2017). "i think that there is- um there might be a link here"- a linguistic analysis of powerless markers in intervasity debating: A way of accounting for the gender success gap. *Unpublished thesis (University of Western Australia)*.
- Pierson, E. (2013). Men outspoke women: Analysing the gender gap in competitive debate. *Monash Debating Review, 2013*.
- Prendergast, C. (1993). A theory of "yes men". *American Economic Review*, pages 757–770.
- Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., and Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*, 3(11):1171–1179.
- Reuben, E., Sapienza, P., and Zingales, L. (2014). How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences*, 111(12):4403–4408.
- Rudman, L. A. and Goodwin, S. A. (2004). Gender differences in automatic in-group bias: Why do women like women more than men like men? *Journal of Personality and Social Psychology*, 87(4):494.
- Rundquist, S. (1992). Indirectness: A gender study of flouting grice's maxims. *Journal of Pragmatics*, 18(5):431–449.
- Samar, R. G. and Alibakhshi, G. (2007). The gender linked differences in the use of linguistic strategies in face-to-face communication. *Linguistics Journal*, 2(3).
- Sandberg, S. (2015). Lean in-women, work and the will to lead.

- Santamaría, L. and Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, 4:e156.
- Sarsons, H. (2017a). Interpreting signals in the labor market: evidence from medical referrals. *Job Market Paper*.
- Sarsons, H. (2017b). Recognition for group work: Gender differences in academia. *American Economic Review*, 107(5):141–45.
- Sarsons, H., Gërkhani, K., Reuben, E., and Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political Economy*, 129(1):000–000.
- Säve-Söderbergh, J. and Sjögren Lindquist, G. (2017). Children do not behave like adults: Gender gaps in performance and risk taking in a random social context in the high-stakes game shows jeopardy and junior jeopardy. *The Economic Journal*, 127(603):1665–1692.
- Schacter, D. L., Gilbert, D. T., and Wegner, D. M. (2011). The accuracy motive: right is better than wrong-persuasion. *Psychology*, pages 105–119.
- Schwindt-Bayer, L. A. (2009). Making quotas work: The effect of gender quota laws on the election of women. *Legislative Studies Quarterly*, 34(1):5–28.
- Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10(5):1189–1213.
- Shurchkov, O. and Eckel, C. C. (2018). *Gender differences in behavioral traits and labor market outcomes*. Oxford, UK: Oxford University Press.
- Smith, S. M. and Shaffer, D. R. (1995). Speed of speech and persuasion: Evidence for multiple effects. *Personality and Social Psychology Bulletin*, 21(10):1051–1060.
- Stab, C. and Gurevych, I. (2014). Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.
- Stab, C. and Habernal, I. (2016). Detecting argument components and structures. In *Report of Dagstuhl Seminar on Debating Technologies (15512)*, volume 5, pages 32–32.

- Stepp, P. L. and Gardner, B. (2001). Ten years of demographics: Who debates in america. *Argumentation and Advocacy*, 38(2):69–82.
- Stoddard, O., Karpowitz, C., and Preece, J. (2020). Strength in numbers: A field experiment in gender, influence, and group dynamics. *IZA Discussion Paper*.
- Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., and Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, 118(2):364.
- Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, pages 613–624.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Ulinski, M., Benjamin, S., and Hirschberg, J. (2018). Using hedge detection to improve committed belief tagging. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–5.
- Van den Brink, M. and Benschop, Y. (2014). Gender in academic networking: The role of gatekeepers in professorial recruitment. *Journal of Management Studies*, 51(3):460–492.
- van Dolder, D., van den Assem, M. J., and Buser, T. (2020). Gender and willingness to compete for high stakes. *Available at SSRN 3537678*.
- Vial, A. C., Napier, J. L., and Brescoll, V. L. (2016). A bed of thorns: Female leaders and the self-reinforcing cycle of illegitimacy. *The Leadership Quarterly*, 27(3):400–414.
- Villeval, M. C. (2012). Ready, steady, compete. *Science*, 335(6068):544–545.
- Visser, B. and Swank, O. H. (2007). On committees of experts. *Quarterly Journal of Economics*, 122(1):337–372.
- Weisshaar, K. (2017). Publish and perish? an assessment of gender gaps in promotion to tenure in academia. *Social Forces*, 96(2):529–560.

- Wessel, J. L., Hagiwara, N., Ryan, A. M., and Kermond, C. M. (2015). Should women applicants “man up” for traditionally masculine fields? effectiveness of two verbal identity management strategies. *Psychology of Women Quarterly*, 39(2):243–255.
- Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., and Malone, T. W. (2010). Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688.
- Woolston, C. (2019). How a hiring quota failed. *Nature*, 566(7742):287–288.
- Wozniak, D. (2012). Gender differences in a market with relative performance feedback: Professional tennis players. *Journal of Economic Behavior & Organization*, 83(1):158–171.
- Wright, E. O., Baxter, J., and Birkelund, G. E. (1995). The gender gap in workplace authority: A cross-national study. *American Sociological Review*, pages 407–435.
- Wright, K. A. and Holland, J. (2014). Leadership and the media: Gendered framings of julia gillard’s ‘sexism and misogyny’ speech. *Australian Journal of Political Science*, 49(3):455–468.
- Yang, Z., Huynh, J., Tabata, R., Cestero, N., Aharoni, T., and Hirschberg, J. (2020). What makes a speaker charismatic? producing and perceiving charismatic speech. In *Proc. 10th International Conference on Speech Prosody 2020*, pages 685–689.
- Yeomans, M., Kantor, A., and Tingley, D. (2018). The politeness package: Detecting politeness in natural language. *R Journal*, 10(2).
- Yeomans, M., Minson, J., Collins, H., Chen, F., and Gino, F. (2020). Conversational receptiveness: Improving engagement with opposing views. *Organizational Behavior and Human Decision Processes*, 160:131–148.
- Yu, B. (2014). Language and gender in congressional speech. *Literary and Linguistic Computing*, 29(1):118–132.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: series B (Statistical Methodology)*, 67(2):301–320.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

739. R.J. DÖTTLING, *Essays in Financial Economics*
740. E.S. ZWIERS, *About Family and Fate: Childhood Circumstances and Human Capital Formation*
741. Y.M. KUTLUAY, *The Value of (Avoiding) Malaria*
742. A. BOROWSKA, *Methods for Accurate and Efficient Bayesian Analysis of Time Series*
743. B. HU, *The Amazon Business Model, the Platform Economy and Executive Compensation: Three Essays in Search Theory*
744. R.C. SPERNA WEILAND, *Essays on Macro-Financial Risks*
745. P.M. GOLEC, *Essays in Financial Economics*
746. M.N. SOUVERIJN, *Incentives at work*
747. M.H. COVENEY, *Modern Imperatives: Essays on Education and Health Policy*
748. P. VAN BRUGGEN, *On Measuring Preferences*
749. M.H.C. NIENTKER, *On the Stability of Stochastic Dynamic Systems and their use in Econometrics*
750. S. GARCIA MANDICÓ, *Social Insurance, Labor Supply and Intra-Household Spillovers*
751. Y. SUN, *Consumer Search and Quality*
752. I. KERKEMEZOS, *On the Dynamics of (Anti) Competitive Behaviour in the Airline Industry*

-
753. G.W. GOY, *Modern Challenges to Monetary Policy*
754. A.C. VAN VLODROP, *Essays on Modeling Time-Varying Parameters*
755. J. SUN, *Tell Me How To Vote, Understanding the Role of Media in Modern Elections*
756. J.H. THIEL, *Competition, Dynamic Pricing and Advice in Frictional Markets: Theory and Evidence from the Dutch Market for Mortgages*
757. A. NEGRIU, *On the Economics of Institutions and Technology: a Computational Approach*
758. F. GRESNIGT, *Identifying and Predicting Financial Earth Quakes using Hawkes Processes*
759. A. EMIRMAHMUTOGLU, *Misperceptions of Uncertainty and Their Applications to Prevention*
760. A. RUSU, *Essays in Public Economics*
761. M.A. COTOFAN, *Essays in Applied Microeconomics: Non-Monetary Incentives, Skill Formation, and Work Preferences*
762. B.P.J. ANDRÉE, *Theory and Application of Dynamic Spatial Time Series Models*
763. P. PELZL, *Macro Questions, Micro Data: The Effects of External Shocks on Firms*
764. D.M. KUNST *Essays on Technological Change, Skill Premia and Development*
765. A.J. HUMMEL, *Tax Policy in Imperfect Labor Markets*
766. T. KLEIN, *Essays in Competition Economics*
767. M. VIGH, *Climbing the Socioeconomic Ladder: Essays on Sanitation and Schooling*
768. YAN XU, *Eliciting Preferences and Private Information: Tell Me What You Like and What You Think*
769. S. RELLSTAB, *Balancing Paid Work and Unpaid Care over the Life-Cycle*
770. Z. DENG, *Empirical Studies in Health and Development Economics*

771. L. KONG, *Identification Robust Testing in Linear Factor Models*
772. I. NEAMȚU, *Unintended Consequences of Post-Crisis Banking Reforms*
773. B. KLEIN TEESELINK, *From Mice to Men: Field Studies in Behavioral Economics*
774. B. TEREICK, *Making Crowds Wiser: The Role of Incentives, Individual Biases, and Improved Aggregation*
775. A. CASTELEIN, *Models for Individual Responses*
776. D. KOLESNYK, *Consumer Disclosures on Social Media Platforms: A Global Investigation*
777. M.A. ROLA-JANICKA, *Essays on Financial Instability and Political Economy of Regulation*
778. J.J. KLINGEN, *Natural Experiments in Environmental and Transport Economics*
779. E.M. ARTMANN, *Educational Choices and Family Outcomes*
780. F.J. OSTERMEIJER, *Economic Analyses of Cars in the City*
781. T. ÖZDEN, *Adaptive Learning and Monetary Policy in DSGE Models*
782. D. WANG, *Empirical Studies in Financial Stability and Natural Capital*
783. L.S. STEPHAN, *Estimating Diffusion and Adoption Parameters in Networks New Estimation Approaches for the Latent-Diffusion-Observed-Adoption Model*
784. S.R. MAYER, *Essays in Financial Economics*
785. A.R.S. WOERNER, *Behavioral and Financial Change – Essays in Market Design*
786. M. WIEGAND, *Essays in Development Economics*
787. L.M. TREUREN, *Essays in Industrial Economics - Labor market imperfections, cartel stability, and public interest cartels*
788. D.K. BRANDS, *Economic Policies and Mobility Behaviour*