



# Essays in Labor Economics

## Citation

Sarsons, Heather. 2018. Essays in Labor Economics. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

## Link

<http://nrs.harvard.edu/urn-3:HUL.InstRepos:41128383>

## Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

## Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

## Share Your Story

The Harvard community has made this article openly available.  
Please share how this access benefits you. [Submit a story](#)

# Essays in Labor Economics

A dissertation presented  
by

Heather Sarsons

to

The Department of Economics

in partial fulfillment of the requirements  
for the degree of  
Doctor of Philosophy  
in the subject of  
Economics

Harvard University  
Cambridge, Massachusetts  
May 2018

© 2018 Heather Sarsons  
All rights reserved.

*Dissertation Advisors:*  
**Professor Lawrence Katz**  
**Professor David Laibson**

*Author:*  
**Heather Sarsons**

## **Essays in Labor Economics**

### **Abstract**

This dissertation explores sources of gender inequality in labor markets. It first empirically documents how individuals form and update beliefs about themselves and others. It then asks how differences in beliefs and belief-updating about men and women lead to gender inequality in labor markets. The first chapter uses data from the medical field to show that a person's gender influences the way others interpret information about his or her ability. The second chapter provides evidence that we use gender when allocating credit for group work when individual contributions are unobservable. The third chapter asks whether women who reach the top of their careers need to be as confident or more confident than their male counterparts.

# Contents

Abstract . . . . .	iii
Acknowledgments . . . . .	viii
<b>Introduction</b>	<b>1</b>
<b>1 Interpreting Signals: Evidence from Medical Referrals</b>	<b>2</b>
1.1 Introduction . . . . .	2
1.2 Motivating Example . . . . .	6
1.3 Empirical Setting and Data . . . . .	6
1.3.1 Referral Decisions . . . . .	7
1.3.2 Data . . . . .	7
1.4 Empirical Strategy . . . . .	12
1.4.1 Matching Procedure . . . . .	14
1.4.2 Estimating Equations and Identification Assumptions . . . . .	15
1.5 Results . . . . .	18
1.5.1 Responses to Signals . . . . .	19
1.5.2 What Influences a Physician’s Reaction to a Signal? . . . . .	28
1.6 Alternative Interpretations . . . . .	33
1.6.1 Differences in Patient Risk . . . . .	34
1.6.2 Are Outcomes Differentially Predictive of Future Outcomes? . . . . .	35
1.6.3 Do Physicians Stop Referring for Certain Procedures? . . . . .	35
1.7 Welfare Analysis and Career Effects . . . . .	37
1.7.1 Surgeon Ability . . . . .	37
1.7.2 Surgeon Pay and Skill Accumulation . . . . .	37
1.8 Theoretical Framework . . . . .	40
1.8.1 Physician’s Decision Problem . . . . .	40
1.8.2 Bayesian Updating: Physician Cares about Mean Ability . . . . .	41
1.8.3 Bayesian Updating: Physician Cares about Variance . . . . .	43
1.8.4 Alternative Models . . . . .	45
1.9 Conclusion . . . . .	47

<b>2</b>	<b>Gender Differences in Recognition for Group Work</b>	<b>49</b>
2.1	Introduction . . . . .	49
2.2	Data . . . . .	51
2.2.1	Sample Selection and Data Overview . . . . .	51
2.2.2	Construction of Tenure . . . . .	52
2.2.3	Summary Statistics . . . . .	52
2.3	Empirical Strategy and Results . . . . .	54
2.3.1	Main Results . . . . .	54
2.3.2	Testing Against Other Coauthoring Conventions . . . . .	61
2.3.3	Robustness Checks . . . . .	62
2.4	Channels . . . . .	66
2.4.1	Ability-Based Sorting . . . . .	66
2.4.2	Preference-Based Sorting . . . . .	70
2.4.3	Women Not Claiming Credit for Papers . . . . .	73
2.4.4	Taste-Based Discrimination . . . . .	73
2.5	Further Discussion: Are Things Improving and Where? . . . . .	73
2.6	Conclusion . . . . .	76
<b>3</b>	<b>Confidence Men</b>	<b>79</b>
3.1	Introduction . . . . .	79
3.2	Data . . . . .	81
3.3	Results . . . . .	84
3.3.1	Main Results . . . . .	84
3.3.2	Mechanisms . . . . .	86
3.4	Conclusion . . . . .	92
	<b>References</b>	<b>95</b>
	<b>Appendix A Appendix to Chapter 1</b>	<b>100</b>
A.1	Data Appendix . . . . .	100
A.1.1	Example of Procedure Code Groupings . . . . .	100
A.1.2	Specialties included in matched sample . . . . .	101
A.2	Appendix to Theoretical Framework . . . . .	101
A.2.1	Calculation of Work Hours Adjustment . . . . .	101
A.2.2	Proofs . . . . .	102
A.3	Additional Tables . . . . .	104
A.4	Additional Figures . . . . .	105
	<b>Appendix B Appendix to Chapter 2</b>	<b>109</b>

## List of Tables

1.1	Summary Statistics: Unmatched Sample . . . . .	12
1.2	Balance Table: Matched Sample for Bad Outcomes . . . . .	15
1.3	Balance Table: Matched Sample for Good Outcomes . . . . .	16
1.4	Impact of Event on Referrals to Performing Surgeon . . . . .	20
1.5	Impact of an Event on Referrals to Others . . . . .	27
1.6	Experience with Surgeons and Physician Response . . . . .	31
1.7	Variables Correlated with Surgeon Response . . . . .	32
1.8	Are Events Predictive of Future Events? . . . . .	36
2.1	Summary Statistics . . . . .	53
2.2	Relationship Between Papers & Tenure . . . . .	57
2.3	Coauthor Gender . . . . .	60
2.4	Sociology Summary Statistics . . . . .	62
2.5	Sociology: Papers and Tenure . . . . .	63
2.6	Robustness Checks . . . . .	64
2.7	Survey Results . . . . .	67
2.8	Accounting for Sorting on Ability . . . . .	69
2.9	Paper Split by Top 5 . . . . .	71
2.10	Coauthor Seniority . . . . .	72
2.11	Results Over Time . . . . .	74
2.12	Interaction w/ School Rank . . . . .	75
3.1	Comparison of background characteristics and confidence by gender . . . . .	83
3.2	Propensity to Provide Extreme Judgements and Gender . . . . .	85
3.3	Self-reported Confidence Level and Gender . . . . .	85
3.4	Differential Confidence in Answering Foreign Field Questions and Gender . . . . .	87
3.5	Confidence in Answering Foreign Field Questions . . . . .	89
3.6	Confidence and Disagreement by Gender . . . . .	90
3.7	Robustness Checks . . . . .	93
A1	Balance for Placebo Matched Samples . . . . .	104

## List of Figures

1.1 Gender Breakdown of Specialties . . . . .	11
1.2 Distribution of Surgeon Ability . . . . .	13
1.3 Quarterly Estimates of Physician’s Reaction to Death . . . . .	19
1.4 Riskiness of Future Procedures and Patients . . . . .	22
1.5 Comparison with Placebo Surgeons . . . . .	23
1.6 Quarterly Estimates for Unexpectedly Good Outcomes . . . . .	25
1.7 Spillovers to Other Surgeons after Bad Outcome . . . . .	26
1.8 Information Spillovers to Other Physicians . . . . .	29
1.9 Physician Response by Patient Risk . . . . .	34
1.10 Change in Surgeon Quality . . . . .	38
1.11 Medicare Payments . . . . .	39
2.1 Total Papers and Tenure . . . . .	55
2.2 Solo Authored Papers and Tenure . . . . .	56
2.3 Coauthored Papers and Tenure . . . . .	58
2.4 Relationship Between Paper Composition and Tenure . . . . .	59
2.5 Ability and Sorting . . . . .	77
2.6 Assortative Matching . . . . .	78
3.1 Distribution of Responses (Likert scale) . . . . .	82
3.2 Distribution of Confidence . . . . .	82
3.3 Confidence and disagreement . . . . .	91
A1 Spillovers to Female Surgeons, >10 Prior Refs . . . . .	105
A2 Spillovers to Other Surgeons after Good Outcome . . . . .	106
A3 Referrals for Procedure . . . . .	106
A4 Referrals for Other Procs. to Performing Surgeon . . . . .	107
A5 Deaths that Occur on Day of Surgery . . . . .	107
A6 Unexpectedly Good Outcomes: Top 5% Risk Level . . . . .	108

## Acknowledgments

The number of people I have to thank cannot fit on this page, so I'll try to do an overview.

I would not have gone to graduate school without the encouragement, patience, and support of several professors at UBC. Thank you for being fantastic mentors, and for being willing to write reference letters multiple years in a row when I got rejected from grad schools multiple years in a row.

I would not have made it through the first year of graduate school without the friendship and round-the-clock tutoring of Team Maxwell. Thank you for teaching me what a derivative is.

I would not have finished graduate school without the encouragement and support of my incredible advisors: Claudia Goldin, Larry Katz, David Laibson, and Amanda Pallais. Thank you for your patience, advice, and insistence that I learn to spell things the American way. I cannot overstate what a wonderful group of advisors you are, or how much I learned from you.

I would not have finished graduate school with my sanity somewhat intact were it not for the friendship of Mitra Akhtari, Rahul Bhui, Kyrah Daniels, Ellora Derenoncourt, Laura Derksen, Raissa Fabregas, Siddharth George, Jessica Laird, Jonathan Libgober, Rachael Meager, Tabatha Robinson, and Chenzi Xu, as well as countless others in and outside of the economics department (maybe not countless – I don't have that many friends).

I would not have gone through graduate school being relatively unstressed were it not for the good people of SK/AB. Thanks to my friends and family there for reminding me that "things aren't that serious" and "why don't you have a job yet".

I would not have made it through anything without my parents, partly because I needed them to be born, but also because they have always stood by and behind me. I finally have a job, but I still think you should live with one of your sons when you're older.

Which brings me to my brothers. Jim, your kindness is unmatched. Paul, you have nice shoes. Thank you, everyone.

To Ms. Belsey and Mrs. Bracken, for encouraging curiosity.

# Introduction

This dissertation explores sources of gender inequality in labor markets. The first chapter provides evidence that a person's gender influences the way others interpret information about his or her ability. Using data on physicians' referrals to surgeons, I find that physicians view patient outcomes differently depending on the performing surgeon's gender. Specifically, physicians become more pessimistic about a female surgeon's ability than a male's after a patient death, indicated by a sharper drop in the number of referrals she receives. Physicians become more optimistic about a male surgeon's ability after a good patient outcome, indicated by a larger increase in the number of referrals that he receives. Physicians also change their behavior toward other female surgeons after a bad experience with one female surgeon, becoming less likely to form new referral connections with other women in the same specialty. I show that the empirical results are reconcilable with Bayesian updating only if they do not have rational expectations about the true ability distribution of surgeons.

The second chapter asks how credit for group work is allocated when individual contributions are not perfectly observed. Do demographic traits like gender influence the allocation of credit? Using data from academic economists' CVs, I test whether coauthored and solo-authored publications matter differently for tenure for men and women. Because coauthors are listed alphabetically in economics, coauthored papers do not provide specific information about each contributor's skills or ability. Solo-authored papers, on the other hand, provide a relatively clear signal of ability. I find that men are tenured at roughly the same rate regardless of whether they coauthor or solo-author. Women, however, become less likely to receive tenure the more they coauthor. The result is most pronounced for women coauthoring with men and less pronounced among women who coauthor with other women.

The third chapter explores whether a confidence gap exists between men and women who make it to the top of their careers. Using data from a select group of economists working in top U.S. universities, we find that women are still less confident than men along two margins. First, when asked about their level of agreement on survey questions about the economy, women are less likely to give "extreme" answers in which they strongly agree or disagree. Second, women are less confident in the accuracy of their answer. We provide suggestive evidence that the confidence gap is driven by women being less confident when asked questions that are outside their field of expertise.

# Chapter 1

## Interpreting Signals: Evidence from Medical Referrals

### 1.1 Introduction

Does a person's gender influence the way we interpret information about his or her ability? The answer to this question has important implications for gender inequality in labor markets, particularly for how women move up the career ladder relative to men. Research shows that in many industries, women are promoted at lower rates than their male counterparts.<sup>1</sup> Gender gaps in wages persist in upper-level positions (Blau and Kahn, 2016). Employers use information that is often subjective or imprecise to evaluate individuals. Their evaluations influences hiring, wage, and promotion decisions. Such decisions become distorted if there are systematic differences in how information about a man and a woman is processed. If an employee's gender influences the way an employer views his or her performance, for example, the employer might end up with different beliefs about a man and a woman's ability even if their objective performance is the same.

This paper empirically tests whether gender influences the way information about others is interpreted. Using Medicare data on referrals from physicians to surgical specialists, I find that the referring physicians view their patients' surgical outcomes differently depending on the performing surgeon's gender.<sup>2</sup> Physicians increase their referrals more to a male surgeon than to a female surgeon after a good patient outcome but lower their referrals more to a female surgeon than a male surgeon after a bad outcome. Furthermore, a physician's experience with one female surgeon influences his or her referrals to other female surgeons in the same specialty.

---

<sup>1</sup>For example, women represent half of all managers in Fortune 500 companies but only 14% of all executive officers and 4% of all CEOs (Blau and Kahn, 2016). Women are less likely to make partner within law firms (Azmat and Ferrer; 2017; Noonan, Corcoran, and Courant; 2005). More than 40% of all medical students are female but only 24% of hospital division chiefs and 15% of medical department chairs are women (Lautenberger et al., 2014). Within academia, women are less likely to move to associate and full professorship positions than men (Blau and Kahn, 2016).

<sup>2</sup>Throughout the paper, I use the term physician for the referring physician and surgeon for the surgeon who performs the surgery even though surgeons are also physicians.

An experience with a male surgeon has no impact on a physician's behavior toward other male surgeons. These asymmetric responses imply that even if women are hired at the same rate as men, they receive fewer chances to show that they can be successful which could lead to lower promotion rates and wages.

Three features of the medical field and Medicare data allow me to address whether gender influences belief updating. First, medical research suggests that a primary factor influencing physicians' referral choices is their beliefs about a surgeon's ability (Barnett et al., 2007; Forrest et al., 2006; and Kinchen et al., 2004). The volume of a physician's referrals to a surgeon thus provides a proxy for the physician's belief about that surgeon's ability. Second, the high frequency of referral decisions makes it possible to document how physicians react to individual signals, something that is difficult to do in many work contexts.<sup>3</sup> Finally, using detailed Medicare data allows me to control for variables that could influence a physician's reaction, such as the patient and procedure risk, or the surgeon's experience. I can thereby isolate the portion of a physician's reaction that is attributable to the surgeon's gender.

To study the influence that gender has on a physician's reaction, I identify and match observably similar male and female surgeons who perform the same procedure on similar patients. Performing an event study on this matched sample, I document how the referring physician's belief about the performing surgeon, as well as other surgeons, changes after good and bad outcomes. The analysis reveals two asymmetries.

First, physicians update asymmetrically about individual male and female surgeons. Following a bad patient outcome (a patient death), physicians lower their beliefs about a female surgeon's ability more than they do for male surgeons. Referrals from the physician drop by 34% after a bad outcome when the surgeon is female compared with only a slight stagnation in referrals when the surgeon is male. Following a good outcome (an unanticipated survival), however, physicians become more optimistic about a male surgeon's ability than a female surgeon's, indicated by a doubling of referrals to the man versus a 70% increase in referrals to the woman

The second asymmetry exists in how physicians treat groups. Physicians appear to use information about individual female surgeons to update their beliefs about other female surgeons in the same specialty. Physicians become less likely to form new referral connections with women after a bad experience with one female surgeon. In contrast, a bad experience with one male surgeon does not affect physicians' behavior toward other men. I find weak evidence of positive spillovers to other women after a physician has a good experience with a female surgeon.

Two pieces of evidence suggest that the drop in referrals to women following a bad outcome is largely driven by the physician's behavior rather than female surgeons turning down referrals. First, the fact that physicians become less likely to refer to other women in the same specialty following a death suggests that physicians change their beliefs about female surgeons. Second, female surgeons do not receive fewer referrals from other physicians following a patient death,

---

<sup>3</sup>Promotion decisions and pay raises, for example, are viewed only after employers have received a series of signals about an employee, making it difficult to discern how they interpret each signal.

suggesting that they do not generally become more risk averse and refuse to perform surgeries in the future.

Looking at which physicians update asymmetrically, I find that the effects are concentrated among physicians who just started referring to a particular surgeon. Physicians both react less strongly to signals and treat men and women more equally the longer is their referral history with a particular surgeon. Physicians thus appear to learn about surgeon ability over time and exhibit asymmetric learning only when they first start referring to a surgeon. Physicians who have more experience working with female surgeons are also more likely to treat men and women equally in response to an outcome. The physician's gender does not seem to play a role in how he or she reacts.

Examining the career implications of asymmetric updating, I find that in addition to receiving fewer referrals, women also receive less difficult procedures and less risky patients after a patient death. This change in the types of referrals female surgeons receive affects both skill accumulation and surgeon pay. Since surgeons learn by performing surgeries, women will have fewer chances to develop their surgical skills as they receive fewer referrals for difficult surgeries. Procedure risk is also correlated with pay. Overall, I find that women lose 60% of their Medicare billings from the referring physician per quarter when they experience a bad patient outcome, whereas men lose 30%.

Whether changes in referrals truly reflect differences in how physicians process signals depends on whether alternative interpretations can be ruled out. I consider three main alternative explanations and show that none is supported in the data. I first test whether male surgeons systematically receive riskier patients than women. If they do, physicians would naturally react less to a death under a male surgeon than a female surgeon. Although I match on observed patient risk, unobservable risk differences could exist. I bound what the difference in unobservable risk between men and women's patients would have to be to justify the physician's reaction. I find that men would have to receive patients who are 70 percentage points riskier on unobservables for risk to explain the gender difference in a physician's reaction.

A second alternative is that patient deaths are more predictive of future deaths for women while unanticipated survivals are more predictive of future survivals for men. In this case, physicians will become more certain that a woman is low ability after a death and more certain that a man is high ability after a survival, generating the asymmetry observed in the data. I check for differential predictability of future events conditional on a surgeon having one such event and conditional on the characteristics of past and future patients. I find that although a death or survival is predictive of a future death or survival, there is no difference by surgeon gender. In fact, women are slightly less likely to have another bad outcome after one patient dies.

Finally, physicians could become risk averse after a patient death and stop referring for a particular procedure. If women specialize in those procedures whereas men perform a wider variety of procedures, it will look like referrals to women fall after a death when in fact the physician stops referring to all surgeons for a particular procedure. I look at how the types of

procedures that physicians refer for change after a patient death and do not find evidence of any behavior change.

Turning to what the empirical results tell us about how physicians update, I draw on Zeltzer (2017) to develop a model of referrals decisions in which physicians try to maximize patient outcomes subject to idiosyncratic factors like patient preferences and wait times. I use the model to understand whether the empirical results are in line with Bayesian updating, considering cases in which physicians do and do not have risk preferences.

I present two propositions that place restrictions on what a physician's priors about men and women's abilities would have to look like for the results to be consistent with Bayesian updating. I show that a physician would either have to believe that women are on average higher ability than men, or that the difference in the variance of women and men's abilities increases as a physician receives more signals. I discuss whether these assumptions are plausible given the data and argue that although physicians could be Bayesian, their beliefs about ability would not be in line with men and women's measured ability distributions, ruling out rational expectations.

I then consider alternative models, first discussing how the empirical results are at odds with existing discrimination models, and then describing a model of attribution bias. In this model, physicians have an expected outcome for each surgery. When the actual outcome matches their expected outcome, physicians update as Bayesians. However, when the actual outcome is far from what they expected, physicians rationalize the event by attributing it either to the physician's ability or to noise, relying on existing stereotypes about men and women to do so. Therefore, if there is a broad stereotype that women are worse surgeons than men, physicians will attribute unexpectedly bad events to a woman's ability and unexpectedly good events to noise.

The model also explains the asymmetry in spillovers to other male and female surgeons. Physicians update iteratively about the group upon seeing outcomes from individuals from that group. Because women are underrepresented as surgeons, physicians receive fewer signals from them relative to men. As in a Bayesian model, each signal has a larger impact on how a physician updates about the female ability distribution than it does for men. In contrast to a Bayesian model, though, physicians update too little about the group in the positive direction as they attribute some of women's good signals to noise. The asymmetry in the attribution of signals from individual surgeons generates the empirical observation that bad outcomes spill over to other female surgeons while good outcomes have only a weak effect on other female surgeons.

This paper relates to several literatures. First, a large body of work has sought to explain the gender gap in men and women's labor market outcomes. Factors such as family commitments, work preferences, personality traits, and discrimination have all been shown to contribute to the gap.<sup>4</sup> Yet a large portion remains unexplained. A growing literature also documents a pay gap between male and female physicians. Jena et al. (2016), for example, find that female

---

<sup>4</sup>For literature on each topic, see Antecol et al. (2016), Babcock and Laschever (2004), Blau and Kahn (2016), Bursztyn et al. (2017), Card et al. (2016), Ceci et al. (2014), Ginther and Kahn (2004), Goldin (2014a), Goldin (2014b), Goldin and Rouse (2000), Kleven et al. (2017), and Niederle and Vesterlund (2007)

physicians make \$19,878 less per year than male physicians after adjusting for specialty and a variety of physician characteristics. Among physicians accepting Medicare, Zeltzer (2017) documents significant gender homophily in referral decisions, leading to 5% lower demand for female physicians than for male physicians. This paper provides another mechanism that might explain part of the gender gap in wages and promotions, particularly among surgeons.

It also relates to the large literature on discrimination. Most papers in this area involve employers discriminating at the beginning of someone's career and then learning about a worker's true ability over time.<sup>5</sup> In these papers, employers have rational expectations about worker effort or ability. This paper offers a new angle to this discussion. Even if two workers are treated similarly early in their careers, employers might treat them unequally over time if they do not have rational expectations or if they discriminate in how they interpret signals. I cannot distinguish between the two, but neither fit with the current discrimination literature.

There is also empirical evidence that people from different demographic groups are punished differently for the same behavior or outcome. Female financial advisors, for example, are more likely to be fired for misconduct than are male advisors (Egan et al., 2017). Black students are more likely than white students to be suspended for misbehaving in class (Skiba et al., 2002; Gregory et al., 2010). Less studied, though, is whether there is differential treatment for good outcomes and whether these differences are driven by differences in individuals' behavior or differences in how others view their behavior. My results suggest that at part of the differential treatment is driven by the individual interpreting the event.

Finally, the paper relates to the behavioral and social psychology literature on biased updating. Asymmetric updating has been documented in the lab when individuals receive signals about themselves. Women are especially prone to updating too little when they receive good signals about their ability and updating too much when they receive bad signals. Eil and Rao (2011), for example, find that individuals avoid new information about themselves if it is negative but update using Bayes' Rule when they receive positive feedback. Mobius et al. (2014) also provide experimental evidence that individuals over-weight good signals and under-weight bad signals about themselves but that women are more conservative when doing so. This paper shows a similar result in a non-lab context and when updating about others rather than oneself.

## 1.2 Motivating Example

## 1.3 Empirical Setting and Data

I attempt to overcome the challenge of empirically measuring and tracking beliefs using Medicare data on referrals from physicians to surgeons. The approach relies on the assumption that a physician's referral choices are a valid measure of her beliefs about a surgeon's ability. In this

---

<sup>5</sup>See, for example, Becker (1957), Phelps (1972), Arrow (1973), Coate and Loury (1993), Fryer (2007), Altonji and Pierret (2001), and Bertrand and Mullainathan (2004).

section, I discuss the literature documenting how physicians make their referral decisions before describing the Medicare data.

### 1.3.1 Referral Decisions

Several medical studies have attempted to identify the factors influencing doctors' referral decisions. I focus specifically on the literature discussing how decisions about surgical referrals are made as I focus on surgeries in my analysis. While many factors influence a physician's referral choice, these studies, combined with informal interviews, suggest that surgeon quality is a primary consideration in a physician's choice. In fact, several studies ask physicians to rank the reasons for referring a patient to a specific specialist, *aside* from clinical expertise, suggesting that this is an obvious factor<sup>6</sup>. In a survey of 1200 physicians across the U.S., Kinchen et al. (2004) find that 88% of respondents considered the surgeon's skill (measured by medical skill and board certification) to be one of the most important factors in deciding whether to make a referral. Choudhry et al. (2014) also find that the perceived expertise of a specialist is a primary driver of referrals.

However, there are several other factors that influence referrals, many of which are not observed in my data and will add noise to my estimates. Surgeon availability and the urgency of the surgery, for example, are important determinants of surgeon choice (Forrest et al., 2006). Patients in worse health are significantly more likely to be referred to the first available surgeon (Shea et al., 1999). Physicians also look for surgeons with good communication skills so that they can easily follow up to manage the patient's post-surgical recovery (Kinchen et al., 2004; Forrest et al., 2006). Insurance status plays a role in a physician's referral choice, but the patients I look at all have Medicare coverage, making this factor less relevant. Surprisingly, risk attitudes and tolerance for uncertainty appear to be weak predictors of whom a physician will refer to (Forrest et al., 2006). While these variables do influence the physician's decision, they would have to be systematically correlated with surgeon gender to explain the results.

Patients can also request certain surgeons and these cases are not always identifiable in the data. In a paper studying how patients choose surgeons, however, Freedman et al. (2015) find that the most frequently reported reason a patient gives for choosing a particular surgeon is that their physician recommended that surgeon. In Section 1.5, I show that while the results could be partly driven by patient preferences, a large part is due to physicians changing their beliefs and referral behavior.

### 1.3.2 Data

The primary data source for this study is the Medicare Carrier file, a 20% random sample of fee-for-service claims of all Medicare beneficiaries in the U.S. between 2008 and 2012. Three features of the Medicare data make it ideal for understanding belief updating. First, the dataset includes detailed information on patients, including diagnoses, demographic information, medical

---

<sup>6</sup>See, for example, Barnett et al., 2012a; Barnett et al., 2012b

history, and procedure codes. I can therefore compare surgeons who are performing the same procedure on patients with similar demographic backgrounds, medical histories, and risk levels. Second, referrals are frequent, allowing me to document how a physician changes her behavior immediately following a surgery. Finally, the Medicare data lets me track surgeons over time and see how well they perform on surgeries that the referring physician might not witness. I can then calculate the career implications of biased updating and assess whether a physician's beliefs about a particular surgeon match the surgeon's actual skill.

I supplement the Carrier file with two other datasets: the Physician Compare National file and the Dartmouth Atlas of Health Care. The Physician Compare National file contains information on physicians and surgeons, such as the doctor's gender, specialty, medical school, and experience. It also has information on the hospital or group practice to which the doctor belongs. The Dartmouth Atlas of Health Care is a geographic dataset that lets me to match physicians and surgeons to their Hospital Referral Region (HRR). HRRs are geographic units representing regional health care markets and are the geographic unit within which physicians typically refer<sup>7</sup>. There are 306 HRRs in the U.S. The Atlas also has information on the number of specialists within each HRR so that I know how many surgeons a physician has the option of referring to.

I restrict the data in four ways. I first limit the sample to surgical procedures and specialties, as surgical procedures have clear outcomes, such as a patient death. Second, I only include physicians who have the option of referring to at least two specialists for the procedure they are referring for to ensure that their behavior is not constrained by the number of specialists in a region. I then limit the sample to cases in which one surgeon performed a procedure, rather than a team of surgeons. Although there would still be others in the operating room, such as an anesthesiologist and nurses, restricting to cases where there is a single lead surgeon means that, as much as possible, there is a clearly identified person responsible for the case. Finally, I include physician-surgeon pairs for which there was at least one referral prior to the patient event so that I can observe pre-trends.

### **Primary Variable Construction**

**Referrals** Referrals between a physician and a surgeon are identifiable in the Medicare data. For each patient diagnosis, I see whether the patient was referred to a surgeon. I use the term physician when describing any doctor who makes a referral to a surgeon. These physicians can be primary care doctors or specialists.

I do not see instances in which a physician refers a patient to a surgeon and the surgeon turns down the patient. As such, referrals are defined as those that were actually followed through: a physician refers a patient to a surgeon and the surgeon sees the patient. The results could therefore be driven by the surgeon's behavior if women are more likely to stop taking referrals after a bad outcome. However, in Section 1.5, I show that the results are not explained by surgeons changing their behavior. I exclude referrals from medical professionals other than doctors, such as nurse

---

<sup>7</sup>For more information, see <http://www.dartmouthatlas.org/data/region/>.

practitioners, as well as self-referrals in which a specialist refers a patient to him or herself.<sup>8</sup>

**Patient Risk** If women receive less risky patients than men, it would be natural for a physician to react more strongly to a patient death under a woman. I therefore construct and match on a measure of patient risk to ensure that I am comparing male and female surgeons who are receiving similar patients. I follow the medical literature to calculate patient risk, combining a comorbidity index with patient characteristics to predict patient mortality for a given procedure<sup>9</sup>. Specifically, I first use ICD-9 diagnosis codes to calculate the Elixhauser Index for each patient visit.<sup>10</sup> The Elixhauser Index uses information on patients' medical histories to categorize comorbidities, pre-existing medical conditions known to increase the risk of death. I then use the Elixhauser Index, along with other patient characteristics such as age, gender, and race, to predict in-hospital mortality. The higher is the probability of death, the riskier is the patient.

**Surgeon Ability** I construct a measure of surgeon ability for two reasons. First, I use it to estimate the true distribution of surgeon ability and compare it to what a Bayesian physician's prior distribution over ability would have to look like to explain the empirical results. Second, I track how the average ability of surgeons that a physician refers to changes when they switch surgeons.

I calculate surgeon ability as follows. For each patient a surgeon sees, I take the difference between the patient's risk of death for the procedure being performed (defined above) and an indicator variable that equals one if the patient died. I then take the average over all of the patients that a surgeon sees:

$$Ability_i = \frac{\sum p_i (Risk_p - \mathbb{1}(Death))}{n_i} \quad (1.1)$$

Here,  $n_i$  is the number of patients that surgeon  $i$  sees,  $Risk_p$  is patient  $p$ 's risk of death for the procedure being performed, and  $\mathbb{1}(Death)$  is an indicator if patient  $p$  died.

**Events (Signals)** I consider two signals that surgeons send to physicians: bad and good. Bad signals are defined as patient deaths that occur within 7 days of a procedure. I look within 7 days of a procedure to minimize the possibility that a patient died for reasons unrelated to the surgeon, such as a nurse making a mistake during post-operative recovery. In Appendix Figure A5, I show that the results do not change if I restrict bad events to be deaths that occur on the same day of a procedure.

To identify good signals, I look at the top percentile of the riskiest patient-procedure pairs,

---

<sup>8</sup>Self-referrals are often cases where a patient requests a particular surgeon.

<sup>9</sup>See, for example, Tsugawa et al. (2016), Pine et al. (2007).

<sup>10</sup>ICD is the International Classification of Diseases coding.

where patient risk is defined as before and procedure risk is a procedure’s 30-day mortality rate, controlling for patient characteristics. The top 1% of pairs are those for which there is a high probability of death, either due to the patient’s risk, the procedure risk, or a combination. I call a patient outcome “unexpectedly good” if the patient does not die and is not readmitted to the hospital within 30 days of the procedure. While good events are rare by construction, 26% of surgeries that could be categorized as good end up being so. In the majority of cases, the patient is re-hospitalized. As a robustness check, I show how the results change with a 5% risk cutoff in Appendix Figure A6.

**Propensity to Refer to Female Surgeons** In the analysis, I test whether physicians with a high propensity to refer to women respond differently to signals than physicians with a low propensity to refer to women. To do so, I construct a measure of whether a physician over- or under-refers to women relative to the average physician in her HRR:

$$\pi_{j,s} = \frac{\text{Referrals to Women}_{j,s}}{\text{Total Referrals}_{j,s}} - \sum_{j' \in J} \frac{\text{Referrals to Women}_{j',s}}{\text{Total Referrals}_{j',s}} \quad (1.2)$$

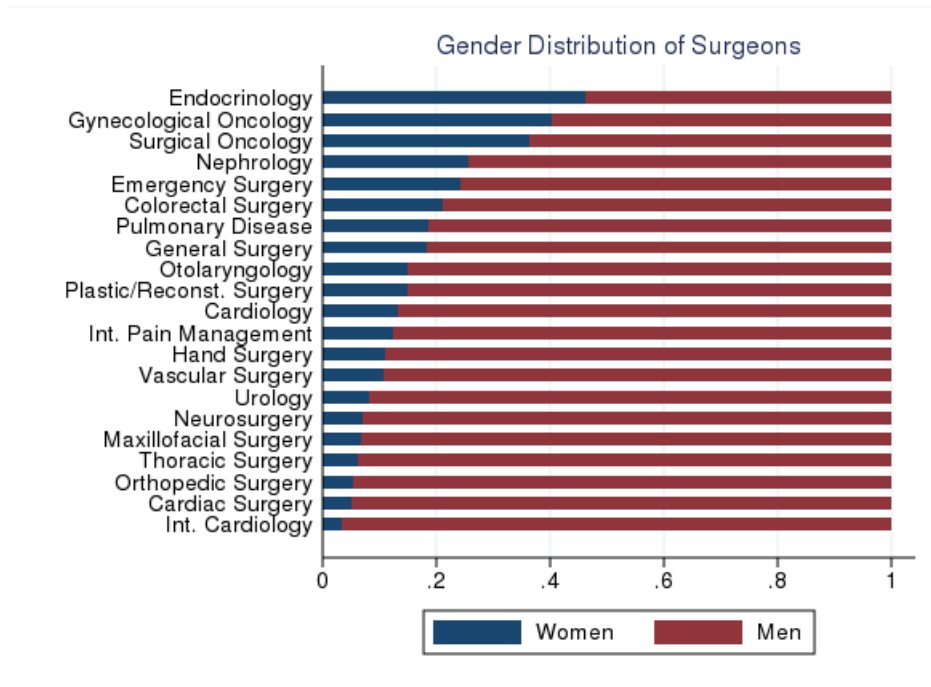
Here, the first term is the number of physician  $j$ ’s referrals going to women in specialty  $s$  divided by physician  $j$ ’s total number of referrals to surgeons in specialty  $s$ . The second term is the total number of referrals going to women in specialty  $s$  from other physicians ( $j'$ ) in the same HRR divided by the total number of referrals from other physicians to surgeons in specialty  $s$ . If  $\pi_{j,s} < 0$ , the physician under-refers to women relative to other physicians in the same referral region and is classified as a “low propensity” physician. If  $\pi_{j,s} > 0$ , the physician over-refers to women relative to other physicians and is classified as a “high propensity” physician.

## Summary Statistics

The list of surgical specialties considered in the paper and the gender distribution of each are presented in Figure ?? . Women are under-represented, making up 17% of the full sample, but with heterogeneity by specialty.

Summary statistics for the unmatched sample of specialists who perform a surgery are presented in Table 1.1. In constructing this table, I reweight the sample so that men and women have the same specialty distribution. Panel A displays surgeon characteristics. All of the statistics are the average over one year. For example, the total number of patients is the number of patients that a surgeon sees on average in one year. Several important differences emerge. Female surgeons receive fewer total patients than male surgeons (81 compared to 112). Women’s patients are less risky, younger, and more likely to be female or minority.

Patient deaths are relatively rare among the full set of surgeons. Approximately 0.85% of the patients that a male surgeon sees and 0.72% of patients a female surgeon sees in a given year die within 7 days of a procedure. Note that this does not account for differences in factors like patient risk. Good outcomes are, by construction, also quite rare when all referrals are taken as the base.



Notes: This figure displays the surgical specialties that could be included in the analysis (e.g. the specialties available to match on) and the gender composition of each. The gender distributions are calculated using the number of surgeons represented in each specialty in the Medicare dataset and therefore may be slightly different than the gender distribution when including all surgeons (such as those not accepting Medicare). Int. Cardiology is Internal Cardiology and Int. Pain Management is Internal Pain Management.

**Figure 1.1: Gender Breakdown of Specialties**

**Table 1.1: Summary Statistics: Unmatched Sample**

	Male Surgeons		Female Surgeons		p-value
	Mean	SD	Mean	SD	
<i>Panel A: Surgeons</i>					
Total Patients	112.4	170.0	81.2	144.1	0.001
Freq. of Bad Events (%)	0.85	4.4	0.72	4.3	0.001
Freq. of Good Events (%)	26.0	27.0	26.1	26.0	0.372
# Physicians Referring	18.6	27.1	14.7	25.2	0.001
Patient Risk (%)	0.48	0.35	0.40	0.33	0.001
Patient Female (%)	52.9	30.0	58.5	32.5	0.001
Patient Minority (%)	17.6	25.4	18.5	28.0	0.001
Patient Age	71.8	8.3	70.8	9.7	0.001
Surgeon Ability	0.0011	0.0053	0.0012	0.0052	0.001
Observations	152,237		30,603		
<i>Panel B: Referring Physicians</i>					
	Male Surgeons		Female Surgeons		p-value
	Mean	SD	Mean	SD	
# Surgeons Referred To	15.5	13.2	2.5	3.2	0.001
# Surgeons Referred to for a Proc.	1.41	1.24	0.17	0.51	0.001
Observations	181,237		50,603		

*Notes:* This table shows summary statistics for the full sample of surgeons. The variables are measured at the yearly so that Total Patients is the average number of patients a surgeon receives in a year, for example. Total Events is the number of good and bad events a surgeon experiences and # of Physicians Referring is the average number of physicians that a given surgeon receives referrals from in a year. Patient Risk and Surgeon Ability are calculated as in Section 1.3

In Table 1.1, I use the number of referrals a surgeon receives that qualify as risky procedures (i.e. the patients are in the top risk percentile) as the base. For both men and women, 26% of referrals they receive that qualify as risky will not result in a death or rehospitalization.

Mean surgeon ability is also reported. Women have a slightly higher mean ability and slightly lower variance of ability, but the differences are small. Figure 1.2 plots the probability distribution function and the cumulative distribution function of surgeon ability. The distributions for men and women largely line up with one another.

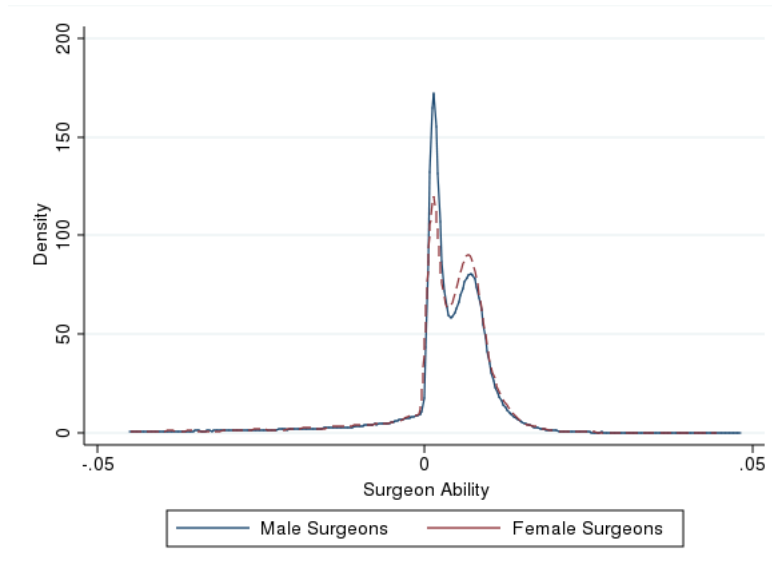
Panel B presents summary statistics for physicians, again measuring variables over the course of a year. Physicians refer to 15.5 different male surgeons and 2.5 different female surgeons per year. For any given procedure, physicians have relatively few surgeons whom they refer to, referring to 1.4 male surgeons per procedure and 0.2 female surgeons. Note that this is not the number of surgeons that are available to be referred to. Within the set of specialties that are included in the matched sample, physicians have an average of 7 available surgeons to refer to in a given specialty.

## 1.4 Empirical Strategy

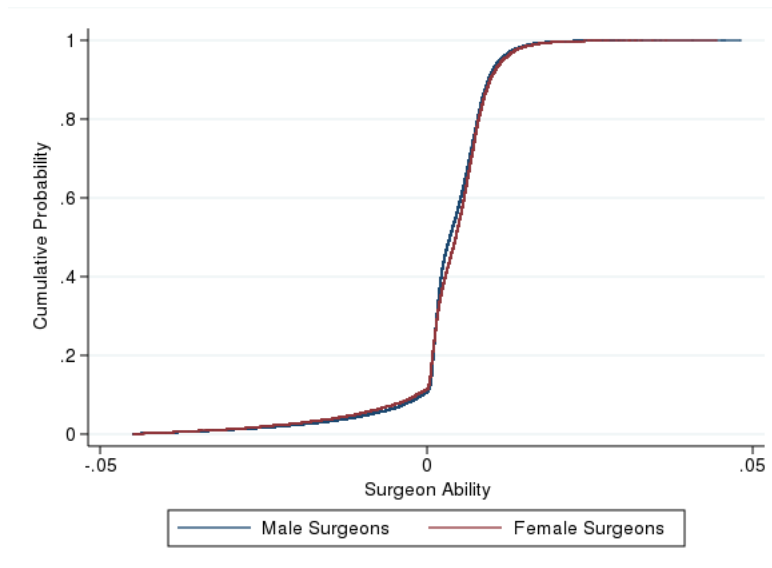
To understand how physicians update their beliefs, I use an event study design to compare how referrals to male and female surgeons change after good and bad patient outcomes. The previous

**Figure 1.2: Distribution of Surgeon Ability**

(a) Ability PDF



(b) Ability CDF



*Notes:* This figure shows the distributions of male and female surgeon ability, where ability is defined as described in Section 1.3.2. The top figure is the PDF and the bottom figure is the CDF. The full sample of surgical specialists is used to create these distributions.

section showed that male and female surgeons receive different types of patients. I therefore use a matching procedure to match men and women who have similar characteristics and are performing the same surgery on similar patients. This section first describes the matching procedure and shows that the resulting sample is balanced in terms of both surgeon and patient characteristics. It then describes the event study design, estimating equations, and identification assumptions.

### 1.4.1 Matching Procedure

I carry out a coarsened-exact match for male and female surgeons on a set of patient and surgeon characteristics.<sup>11</sup> For this procedure, I start with the sample of all events, meaning that surgeons can be included in the sample multiple times if they experience multiple events. If a surgeon experiences multiple events with the same physician, I only keep the first instance of an event. Before matching, I have a sample of 265,734 distinct surgeon-physician pairs with bad events and 302,294 distinct pairs with good events. I then match male and female surgeons on variables measured over the four quarters before an event. I match with replacement of individual surgeons and physicians but allow each surgeon-physician pair to be matched only once.

I match exactly on the surgeon's specialty, the procedure being performed, and the patient's gender and minority status (white vs. non-white). Individuals for whom there is no exact match for each of these categories are dropped. Procedures are identified via the Current Procedural Terminology (CPT) code that the American Medical Association maintains to standardize medical service reporting. There are multiple "layers" of coding with 19 broad parent groups of surgical procedures. Within a parent group, there are three additional layers, each identifying a procedure more uniquely. In Appendix A.1.1, I give an example of what the layers look like for a sample parent group. I match on the second-finest layer, denoted by an asterisk in the appendix. Because there are slightly different procedures within each of these groups, I also match on the average risk of death for each procedure. I am therefore either matching on the exact procedure or on very similar procedures that differ slightly in treatment (for example, treating a humeral shaft fracture with plates/screws versus treating a humeral shaft fracture with rods).

I match coarsely on the patient's age and risk; the number and fraction of physician  $j$ 's referrals going to surgeon  $i$  prior to the event; the total number of referrals surgeon  $i$  received from any physician prior to the event; and surgeon experience, measured in terms of the number of years since the surgeon finished medical school and the number of specific procedures the surgeon had completed prior to the event. Finally, I match on a physician's outside option, using the number of other surgeons available in surgeon  $i$ 's speciality in the same HRR. I use an average of 9 bins for each variable.

Matching on patient age, gender, race, and risk attempts to minimize the differences between patients that the surgeons see. Matching on past surgeries, referrals, and years of experience allows me to compare two surgeons with similar experience levels, both in performing the procedure at

---

<sup>11</sup>This procedure involves matching surgeons exactly on the surgeon's specialty, the procedure being performed, and the patient's gender and race. The other variables are then divided into bins and surgeons are matched across bins.

hand and in their overall experience. Matching on the number of surgeons in the same specialty as surgeon  $i$  in physician  $j$ 's referral region ensures that I am comparing physicians who have similar outside options. Note that I do not require the matched surgeons to have the same referring physician.

Tables 1.2 and 1.3 show that the final samples of matched male and female surgeons who experience bad and good events respectively are balanced. I end up with 7,757 surgeon matches for bad events and 6,979 matches for good events.

**Table 1.2: Balance Table: Matched Sample for Bad Outcomes**

	Male Surgeons		Female Surgeons		p-value
	Mean	SD	Mean	SD	
Patient Refs from Physician	24.9	33.7	25.2	34.3	0.639
Physician's Refs for Proc. (%)	50.6	14.0	50.7	14.1	0.454
Total Patient Refs	74.2	116.5	73.4	115.2	0.651
Patient Age	77.2	12.0	77.2	12.1	0.887
Patient Minority (%)	8.4	27.7	8.4	27.7	0.999
Patient Female (%)	49.6	50.0	49.6	50.0	0.999
Patient Risk (%)	0.81	0.40	0.81	0.40	0.915
Risk All Past Ptnts	0.007	0.004	0.007	0.004	0.830
Total Procs. Performed	10.5	15.5	10.6	15.5	0.593
Years of Experience	23.3	8.1	23.2	8.0	0.174
Available Surgeons	44.5	27.8	45.1	27.8	0.137
# of Matched Surgeons	7,757		7,757		
# Unique Surgeons	5,579		3,561		

*Notes:* The balance table shows summary statistics for the matched sample of surgeons who experience a death. The sample is matched exactly on patient gender and minority status as well as surgeon specialty and procedure (not shown above). The sample is matched coarsely on all other variables. Patient Refs from Physician is the number of referrals that surgeon  $i$  received from physician  $j$  in quarters  $t = -5$  through  $t = -2$  where patient  $p$  dies in  $t = 0$ . Physician's Refs for Proc. is the percent of physician  $j$ 's referrals for the procedure being performed that went to surgeon  $i$  before the event. Similarly, Total Patient Refs is the number of patients surgeon  $i$  received from any physician during this period. Patient Risk is the risk level of the patient who dies while Risk All Past Ptnts is the average risk of all past patients surgeon  $i$  received between  $t = -5$  and  $t = -2$ . Total Procs. Performed is the number times surgeon  $i$  performed the procedure that physician  $j$  refers for between  $t = -5$  and  $t = -2$ . Years of Experience is the number of years since the surgeon graduated from medical school. Available Surgeons is the number of surgeons who are in the same specialty as surgeon  $i$  that the referring physician has the option of referring to. This is measured using Hospital Referral Region as the geographical unit.

## 1.4.2 Estimating Equations and Identification Assumptions

I briefly overview the main estimating equations and identification assumptions that I use in the analysis. The remaining estimating equations are presented with the results in Section 1.5.

**Table 1.3: Balance Table: Matched Sample for Good Outcomes**

	Male Surgeons		Female Surgeons		p-value
	Mean	SD	Mean	SD	
Patient Refs from Physician	15.5	21.0	15.5	21.0	0.481
Physician's Refs for Proc. (%)	34.2	7.3	34.1	7.1	0.835
Total Patient Refs	45.5	68.5	45.0	68.0	0.645
Patient Age	91.3	4.7	91.3	4.6	0.353
Patient Minority (%)	4.8	21.3	4.8	21.3	0.999
Patient Female (%)	24.6	43.1	24.6	43.1	0.999
Patient Risk	0.017	0.002	0.017	0.002	0.650
Risk All Past Ptnts	0.015	0.004	0.015	0.004	0.850
Total Procs. Performed	8.0	15.5	8.0	15.3	0.861
Years of Experience	22.9	8.4	22.7	8.5	0.461
Available Surgeons	38.9	25.9	38.7	25.7	0.603
# Matched Surgeons	6,979		6,979		
# Unique Surgeons	5,554		2,882		

*Notes:* The balance table shows summary statistics for the matched sample of surgeons who experience an “unexpectedly good” outcome. The sample is matched exactly on patient gender and minority status as well as surgeon specialty and procedure (not shown above). The sample is matched coarsely on all other variables. Patient Refs from Physician is the number of referrals that surgeon  $i$  received from physician  $j$  in quarters  $t = -5$  through  $t = -2$  where patient  $p$  dies in  $t = 0$ . Physician's Refs for Proc. is the percent of physician  $j$ 's referrals for the procedure being performed that went to surgeon  $i$  before the event. Similarly, Total Patients Refs is the number of patients surgeon  $i$  received from any physician during this period. Patient Risk is the risk level of the patient who dies while Risk All Past Ptnts is the average risk of all past patients surgeon  $i$  received between  $t = -5$  and  $t = -2$ . Total Procs. Performed is the number times surgeon  $i$  performed the procedure that physician  $j$  refers for between  $t = -5$  and  $t = -2$ . Years of Experience is the number of years since the surgeon graduated from medical school. Available Surgeons is the number of surgeons who are in the same specialty as surgeon  $i$  that the referring physician has the option of referring to. This is measured using Hospital Referral Region as the geographical unit.

### Estimating Equations and Identification for Matched Male/Female Surgeon Pairs

To identify the impact of a good or bad event on referrals, I designate the quarter in which an event occurs as quarter 0. I then sum the number of referrals a surgeon received from the referring physician in each quarter, starting four quarters before the event and ending six quarters after the event, and leaving out the patient referred for surgery. I stack the events of all physician-surgeon pairs and estimate equation 1.3 below, where  $event_{ij,t-k}$  is a dummy variable indicating that an event occurred in quarter  $t$  and  $fem_i$  is a dummy variable indicating the surgeon is female:

$$R_{ijk} = \sum_{k=-4}^6 \beta_k event_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (event_{ij,t-k} \times fem_i) + \theta_{ij} + \epsilon_{ijk} \quad (1.3)$$

The outcome variable,  $R_{ijk}$ , is the number of referrals physician  $j$  sends to surgeon  $i$  in quarter  $k$ . The coefficient  $\hat{\gamma}_k$  tells us how a physician's reaction to an event changes when the surgeon is a woman.

I include physician-surgeon match fixed effects,  $\theta_{ij}$ , to absorb any initial differences between matches and cluster standard errors at the physician-surgeon match level in case there are idiosyncratic factors that are specific to a particular physician-surgeon pair. This assumes that each pair's errors are uncorrelated with the errors of other pairs.

The main identification assumption is that women do not systematically receive different types of patients or perform different procedures than men. For example, if women receive less risky patients, it would make sense that a physician would update more about them than about men after a bad event. Matching on patient characteristics, including patient risk, as well as on the procedure code should alleviate this concern. However, further robustness checks are presented in Section 1.6.

### Estimating Equations and Identification for "Placebo" Surgeons

To understand what referral patterns between the physician and the surgeon would have looked like in the absence of an event, I create a set of placebo surgeons who perform a surgery but do not experience a good or bad event (e.g. they experience the expected outcome). To do so, I use the matching procedure described in Section 1.4.1 to match female surgeons who are identical on all observables and who receive observably similar patients, but one of whom has a patient die while the other does not. I do the same for male surgeons. For bad events, I am thus comparing surgeons who had a patient die (the treated surgeons) to those who did not (the placebo surgeons) and for good events I am comparing surgeons whose patient lived (treated) to those whose patient dies or is re-hospitalized (placebo). Balance tables are presented in Appendix Table A.3.

To quantify the impact an event has on referrals to a surgeon, I estimate

$$R_{ijk} = \sum_{k=-4}^6 \beta_k \text{surgery}_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (\text{surgery}_{ij,t-k} \times \text{event}_k) + \theta_{ij} + \epsilon_{ijk} \quad (1.4)$$

on the matched female and matched male samples separately. The variable  $\text{surgery}_{ij,t-k}$  is a dummy variable that equals one during the quarter of the surgery and  $\text{event}_k$  is a dummy variable indicating that a good or bad event occurred. All other variables are defined as before.

### Estimating Equations for Spillovers

In the analysis I test whether one surgeon's performance affects how the physician updates about other surgeons. To do so, I estimate

$$f_{ijgsk} = \sum_{k=-4}^6 \beta_k \text{event}_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (\text{event}_{ij,t-k} \times \text{fem}_i) + \delta_{\text{available}}_{js} + \theta_{ij} + \epsilon_{ijgsk} \quad (1.5)$$

separately on the samples of male and female surgeons. The outcome variable,  $f_{ijgsk}$ , is the fraction of physician  $j$ 's new referrals going to surgeons in group  $g$  (men or women) in specialty  $s$  in quarter  $k$ , when surgeon  $i$  performs a surgery in quarter  $t$ . The measure leaves out referrals

going to the performing surgeon  $i$ . Other variables driving the physician to refer less to women therefore have to change at the time of the event to explain the changes in the physician's behavior. For example, if other women in the physician's HRR are less skilled or have full schedules, it is unlikely that the physician referred to them before the event occurred so the change in the fraction of referrals going to other women should not change.

I control for the fraction of available surgeons (*available<sub>js</sub>*) who are of the same gender and in the same specialty as the performing surgeon to ensure that the results are not driven by constraints in physicians' surgeon options. The result of this estimation tell us whether physicians change their beliefs about other female (male) surgeons after one of their patient lives or dies under the care of a female (male) surgeon.

### Variables Correlated with Physician's Reaction

To understand what drives a physician's reaction, I correlate several variables with the change in referrals from a physician to a surgeon. Specifically, I estimate

$$R_{ijt} = \beta_1 Fem_i + \beta_2 Post_t + \beta_3 Var + \beta_4 (Fem_i \times Post_t) + \beta_5 (Fem_i \times Var) + \beta_6 (Post_t \times Var) + \beta_7 (Fem_i \times Post_t \times Var) + \sum_{X^k \in \mathbb{X}} X_{ijt} + \theta_{ij} + \epsilon_{ijt} \quad (1.6)$$

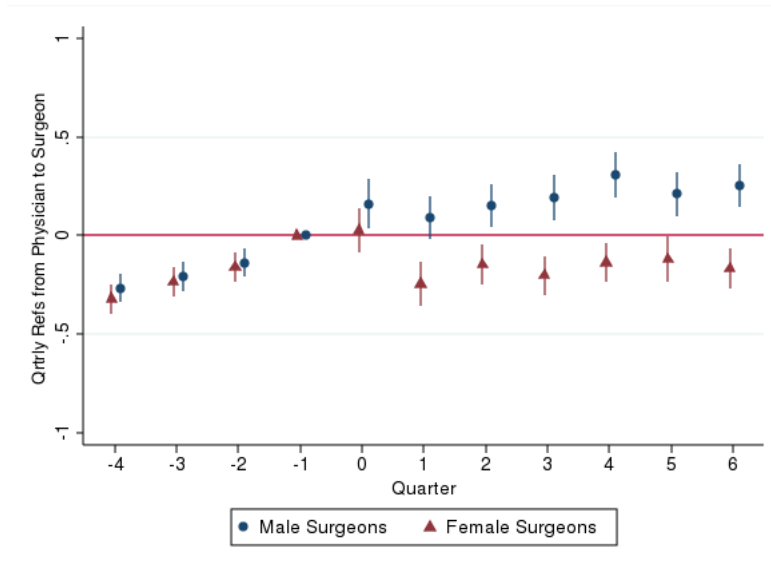
where  $Post_t$  is a dummy equalling one for quarters after the event occurs and  $Var$  is the variable being tested for correlation with the physician's response. A time trend and time trend interactions ( $\sum_{X^k \in \mathbb{X}} X_{ijt}$ ) with the variables  $\mathbb{X} = \{Post_t, Fem_i, Var\}$  are also included.

## 1.5 Results

The empirical analysis uncovers two asymmetries in how physicians react to signals. First, there is an individual-level asymmetry. Physicians respond more to good signals if the surgeon is male and more to bad signals if the surgeon is female. Comparing surgeons who experienced a good or bad outcome with the placebo surgeons who did not, I find that while the referral path for men and women would have been similar had the event not occurred, women receive far fewer referrals following a bad outcome than they otherwise would have. Conversely, men receive more referrals after a good outcome than they otherwise would have. Second, there is an asymmetry at the group level. Physicians update their beliefs about other women after receiving a signal from an individual woman but do not update about other men after receiving a signal from one man.

Physicians who just began a referral relationship with a particular surgeon react the strongest. The longer is a physician's referral relationship with a surgeon, the less that physician reacts to patient outcomes. Further, the gender gap in physicians' responses is decreasing in the length of the referral relationship. In what follows, I therefore focus primarily on physicians and surgeons who just began a referral relationship (i.e. where the physician had sent fewer than 10 referrals prior to the event), as the asymmetric response is concentrated among these physicians and

**Figure 1.3: Quarterly Estimates of Physician’s Reaction to Death**



*Notes:* This figure plots the quarterly regression coefficients and 95% confidence intervals from estimating equation 1.3 using the sample of matched male and female surgeons who experience a patient death. The coefficients are plotted relative to the number of referrals the surgeon was receiving from the physician in  $k = -1$ , which are normalized to zero. In  $k = -1$ , male and female surgeons both received an average of 0.65 referrals from the referring physician  $j$ . A patient that physician  $j$  referred to surgeon  $i$  dies in  $k = 0$ . The outcome variable is the total number of referrals that physician  $j$  sends to surgeon  $i$  each quarter. Standard errors are clustered at the physician-surgeon match level.

surgeons. However, I show how the results depend on the length of the referral relationship in Section 1.5.2.

## 1.5.1 Responses to Signals

### Updating after a Bad Event

This section documents the referring physician’s reaction to a patient death. Because I match male and female surgeons on both the number and fraction of referrals they received from a physician before an event occurs, I am comparing two surgeons for whom physicians have roughly the same beliefs.<sup>12</sup>

Figure 1.3 documents how physicians respond to a patient death, plotting the quarterly coefficients  $\hat{\beta}_k$  and  $\hat{\beta}_k + \hat{\gamma}_k$  for  $k \in \{-4, 6\}$  from estimating equation 1.3. In this figure, physician  $j$  refers patient  $p$  to surgeon  $i$ . Surgeon  $i$  performs the surgery in quarter 0 and the patient dies within 7 days of the surgery. Coefficients are plotted relative to the number of referrals the surgeon received from physician  $j$  in the quarter before the event ( $k = -1$ ). Both male and female surgeons were receiving 0.65 referrals in  $k = -1$  from their referring physician.

<sup>12</sup>This is assuming that referrals are a function of a physician’s belief about a surgeon’s average ability, a claim that I further explore in Section 1.8

The figure shows that before the patient death, the surgeon receives an increasing number of referrals from the physician as they have just started their referrals relationship. When the death occurs in quarter 0, the physician refers less to the surgeon, with a marked drop in referrals occurring if the surgeon is female. Controlling for time trends, men receive 0.101 more referrals in the quarters after the death relative to the quarter before the death, whereas women receive 0.222 fewer referrals after the death relative to the quarter before.<sup>13</sup> The gap in referrals that emerges between male and female surgeons is significant at the 1% level and persists up to a year and a half following the death. To put these numbers into perspective, a male surgeon would have to have three patient deaths to be treated the same as a female surgeon. Given the rarity of patient deaths, this difference is substantial. The results are summarized in column 1 of Table 1.4.

**Table 1.4: Impact of Event on Referrals to Performing Surgeon**

Event	Referrals		Medicare Pay (\$)	
	(1) Bad	(2) Good	(3) Bad	(4) Good
Post	0.006 (0.058)	0.509*** (0.069)	-80.05*** (25.09)	145.29*** (45.21)
Female × Post	-0.291*** (0.087)	-0.222** (0.100)	-138.97*** (32.54)	-22.88 (64.14)
Time Trend	0.099*** (0.011)	0.073*** (0.011)	31.84*** (5.84)	27.58*** (8.07)
Female × Time Trend	-0.009 (0.013)	-0.010 (0.015)	13.67*** (4.65)	-6.08 (8.59)
Post × Time Trend	-0.072*** (0.012)	-0.046*** (0.012)	-35.51*** (6.00)	-41.83*** (8.82)
<i>Average Post Effect On:</i>				
Male Surgeons	0.101	0.604	-92.90	95.42
Female Surgeons	-0.222	0.346	-184.02	51.26
Mean of Outcome Var.	0.65	0.48	309.17	264.86
Observations	34,053	29,214	34,053	29,214
Clusters	3,425	2,948	3,425	2,948
R-Squared	0.265	0.325	0.237	0.249

*Notes:* This table displays the effect of a bad event (columns 1 and 3) and a good event (columns 2 and 3) on referrals and Medicare payments to the performing surgeon. In columns 1 and 2, the outcome variable is the number of referrals from the referring physician to the performing surgeon in a quarter. In columns 3 and 4, the outcome variable is the total Medicare pay that the surgeon receives from referrals from the physician. All regressions include surgeon-physician match fixed effects and standard errors are clustered at the surgeon-physician match level. Columns 1 and 3 are estimated on the sample of matched male and female surgeons who experience a patient death. Columns 2 and 4 are estimated on the sample of matched male and female surgeons who experience a good patient outcome. Both samples are limited to physician-surgeon pairs in which the physician had referred between 1 and 10 patients in the past. Physician-surgeon match fixed effects are included in all regressions and standard errors are clustered at the physician-surgeon match level. Levels of significance: \*10%, \*\* 5%, and \*\*\* 1% level.

<sup>13</sup>See Table 1.4, column 1.

I also find a change in the types of patients and procedures that the physician refers in the future. In Figure 1.4, I re-estimate equation 1.3 but use the average risk of patients that physician  $j$  refers to surgeon  $i$  in quarter  $k$  as the dependent variable. There is no significant change in the riskiness of patients that physicians send to men. However, women start to receive patients that are 0.28 standard deviations less risky after a death. Similarly, women are referred procedures that are 0.041 standard deviations less risky than what they received before the death. The impact that this has on surgeon pay and skill accumulation is further discussed in Section 1.7

To further quantify the impact that a death has on male and female surgeons, I show what the referral path would have looked like in the absence of a death using the matched sample of treated/placebo surgeons described in Section 1.4. The matched surgeons are identical on observables and receive patients with similar risk levels and demographics, but one surgeon experiences a patient death while the other does not. It is worth noting that the absence of a death is not equivalent to the physician receiving no signal from a surgeon. The fact that the patient lives is in itself a good signal. I am therefore not comparing a surgeon who sends a bad signal to one who does not send a signal.

Figure 1.5 plots the results from estimating equation 1.4 separately for male and female surgeons. In both figures, the estimates for the placebo surgeons indicate that the number of referrals to male and female surgeons would have continued to grow at a similar rate had the patient not died. However, the gap between the actual and “projected” number of referrals is smaller for men (top panel). Men who experience a patient death receive 0.22 fewer referrals each quarter than they would have if the patient lived. The F-statistic for the joint significance of  $\hat{\delta}_1 - \hat{\delta}_6$  is 2.48. Women who experience a patient death, shown in the bottom panel, receive 0.6 fewer referrals each quarter (with an F-statistic of 15.88). Physicians thus seem to update their beliefs downward about both male and female surgeons, but by a larger amount when the surgeon is female.

The difference in outcomes between women and men who do and do not experience a patient death is summarized in a difference-in-differences plot shown in panel (c) of Figure 1.5. There is no difference in the referral path between male and female surgeons who do not experience a patient death (grey circles) but women who experience a patient death receive approximately 0.3 fewer referrals per quarter than men who experience a death (green triangles).

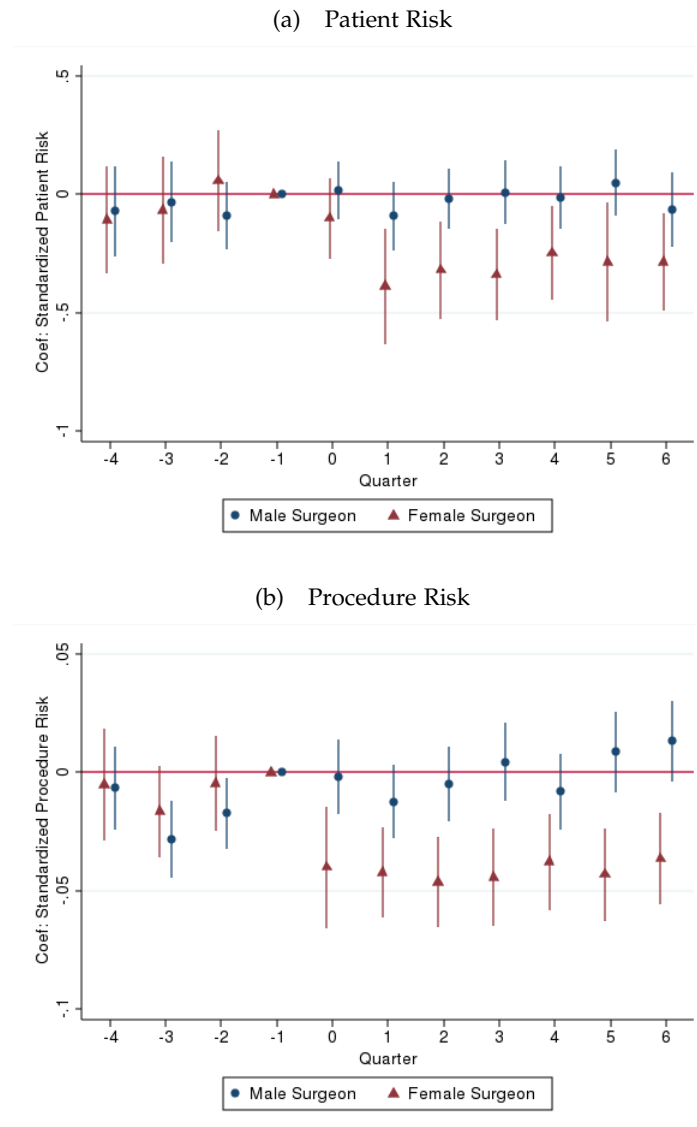
### Updating after a Good Event

It is possible that physicians react more to bad signals from women because the variance of women’s ability is larger. If this is true, physicians should react strongly to good signals from women as well.<sup>14</sup> I now test how physicians react to “unexpectedly good” outcomes, defined as the top 1% of the riskiest patient-procedures pairs for which no death or re-hospitalization

---

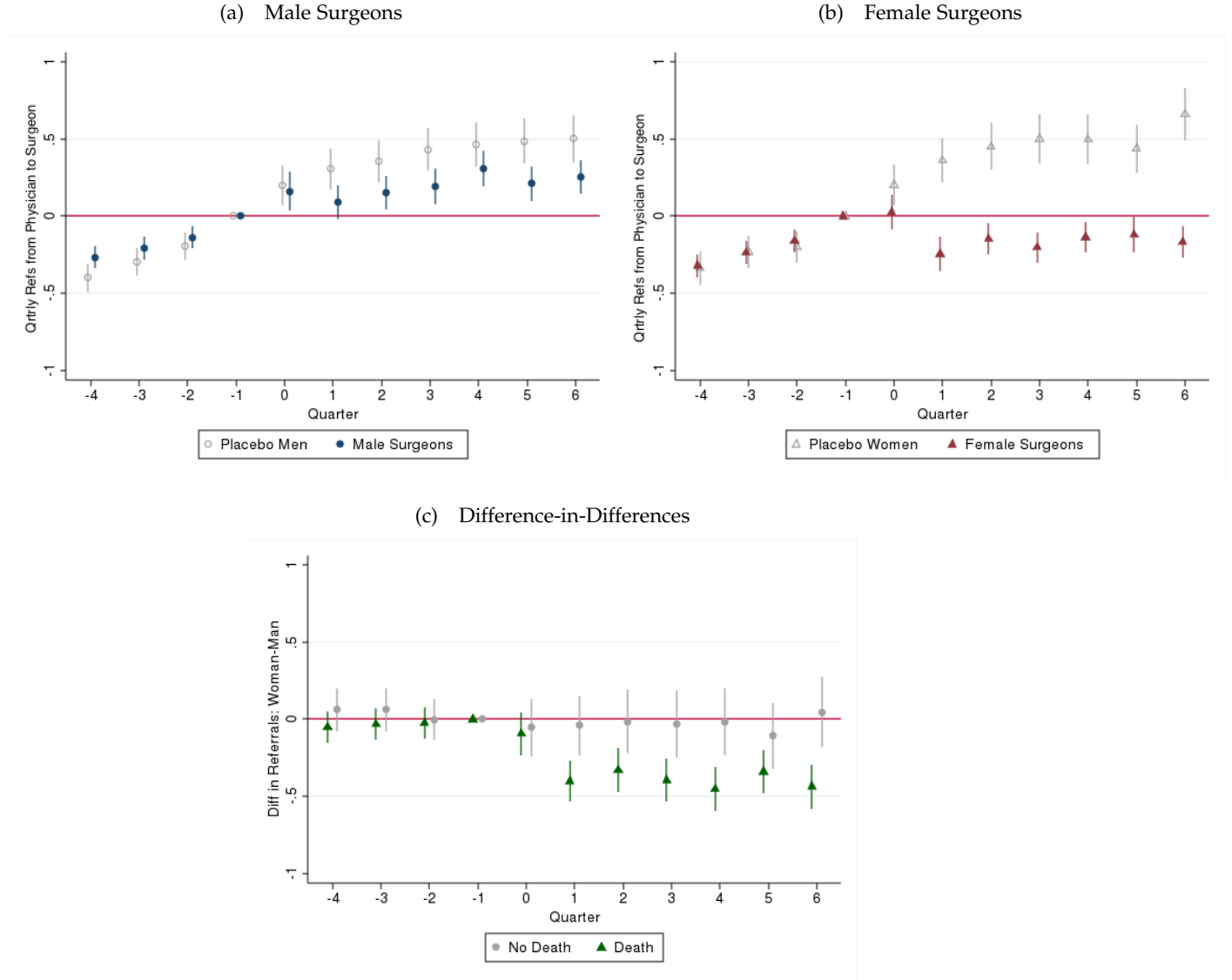
<sup>14</sup>This is assuming that ability is distributed according to a symmetric distribution. Such assumptions are formally analyzed in Section 1.8

**Figure 1.4: Riskiness of Future Procedures and Patients**



*Notes:* Panels (a) and (b) show how the riskiness of future patients and procedures change after a patient death. The outcome variable in Panel (a) is the average patient risk of the patients that physician  $j$  sends to surgeon  $i$  in quarter  $k$ , excluding the risk level of the patient who died. The outcome variable in Panel (b) is the average death rate of the procedures that physician  $j$  sends to surgeon  $i$  in quarter  $k$ , again excluding the riskiness of the procedure that the patient who died was referred for. The coefficients are plotted relative to the average patient risk or procedure death rate that the surgeon received in  $k = -1$ , which is normalized to zero. Standard errors are clustered at the physician-surgeon match level. See Section 1.3 for the definition and calculation of patient risk.

**Figure 1.5: Comparison with Placebo Surgeons**



*Notes:* This figure shows three plots that compare the impact of a bad event to “placebo” outcomes. Panels (a) and (b) plot the quarterly regression coefficients and 95% confidence intervals from estimating equation 1.4 on the matched sample of male (panel a) and female (panel b) surgeons who did not experience a patient death. These coefficients are represented by the hollow circles and triangles. For the placebo outcomes in each panel, physician  $j$  refers a patient to surgeon  $i$  and the surgery does not result in a patient death. I then plot the quarterly coefficients from Figure 3. Panel (c) shows difference-in-differences estimates and 95% confidence intervals. The estimates show the difference between referrals to a female and a male surgeon in quarter  $k$ . The grey circles show the estimates for the sample of placebo surgeons and the green triangles show the estimates for the sample of matched surgeons who experience a patient death. All coefficients are plotted relative to the number of referrals the surgeon was receiving from the physician in  $k = -1$ , which are normalized to zero. Standard errors are clustered at the physician-surgeon match level.

occurred. The quarterly coefficients from estimating equation 1.3 as well as the placebo outcomes are presented in Figure 1.6.

Opposite to their reaction to a death, physicians respond more strongly when the surgeon is male than when the surgeon is female. While referrals to both male and female surgeons increase, they increase by a significantly larger amount for male surgeons. Controlling for time trends, men receive roughly 0.6 more referrals than they did in the period before the good event whereas women receive 0.35 more referrals, a difference that is significant at the 5% level (see Table 1.4, column 2 for a summary of effects). A physician's response to good signals stands in stark contrast to how physicians react to bad signals, where their response to male surgeons is muted.

Relative to the sample of surgeons who did not experience a good outcome, men receive 0.25 more referrals per quarter than they otherwise would have while women receive 0.15 more. Note that a difference in the number of referrals to men and women who do not experience a good outcome also emerges as the expected outcome in this case is a hospital readmission or death. Although a re-hospitalization is expected, women still receive slightly fewer referrals than men do after the surgery.<sup>15</sup> The difference-in-differences coefficients are plotted in the bottom panel of Figure 1.6.

### Spillovers: Updating about Other Surgeons

I now turn to the question of whether a physician's experience with one surgeon influences that physician's beliefs about other surgeons of the same gender. Figure 1.7 plots the quarterly coefficients from estimating equation 1.5 where the event is a patient death.

The outcome in the top figure is the fraction of a physician's referrals going to surgeons the physician hasn't referred to before who are the same gender and in the same specialty as the performing surgeon. I focus on new referral relationships as a physician's experience with one surgeon does not significantly impact her beliefs about another surgeon she has been referring to for some time. This is shown in Appendix Figure A1 where physicians who have a long referral history with a particular surgeon do not change their referrals to that surgeon after having a bad experience with surgeon  $i$ .

In cases where a female surgeon has a patient die (the red triangles), the physician becomes less likely to form referral relationships with female surgeons in the future. However, physicians do not change their propensity to form new referral relationships with male surgeons after a patient dies under a man (the blue circles). Men appear to be treated as individuals while information about one woman in specialty  $s$  affects the physician's beliefs about other women in specialty  $s$ .

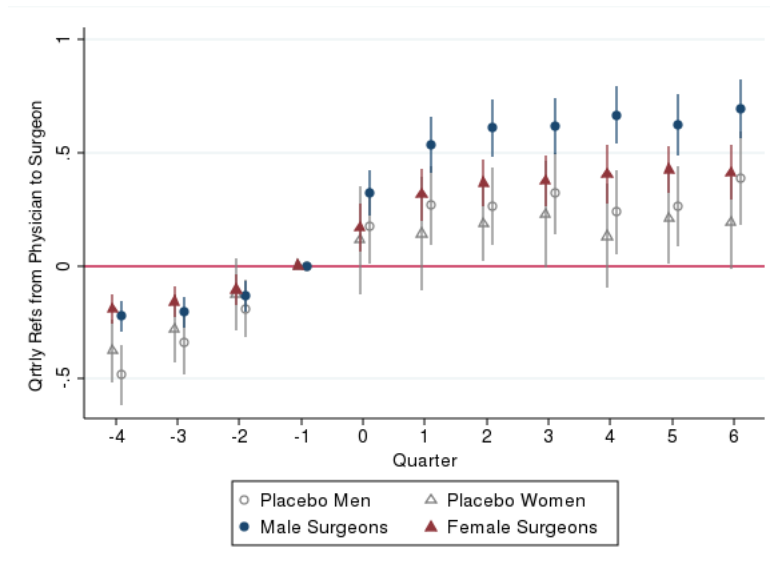
The bottom figure shows the change in a physician's referrals to new surgeons of the same gender but in different specialties than the performing surgeon. Physicians slightly reduce the fraction of their referrals going to female surgeons in other specialties but the post-death coefficients are jointly insignificant.

---

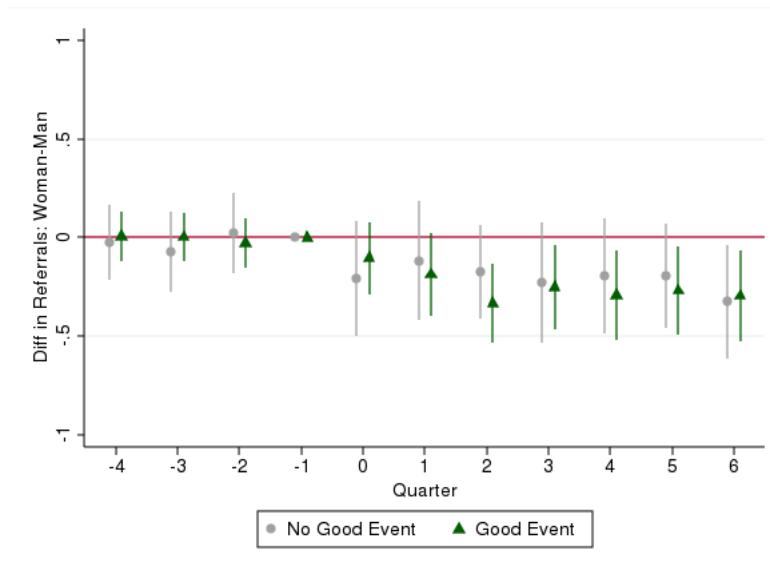
<sup>15</sup>In Section 1.5.2, I show how physicians' responses to deaths depend on patient risk as well as several other factors.

**Figure 1.6: Quarterly Estimates for Unexpectedly Good Outcomes**

(a) Treated and Placebo Surgeons

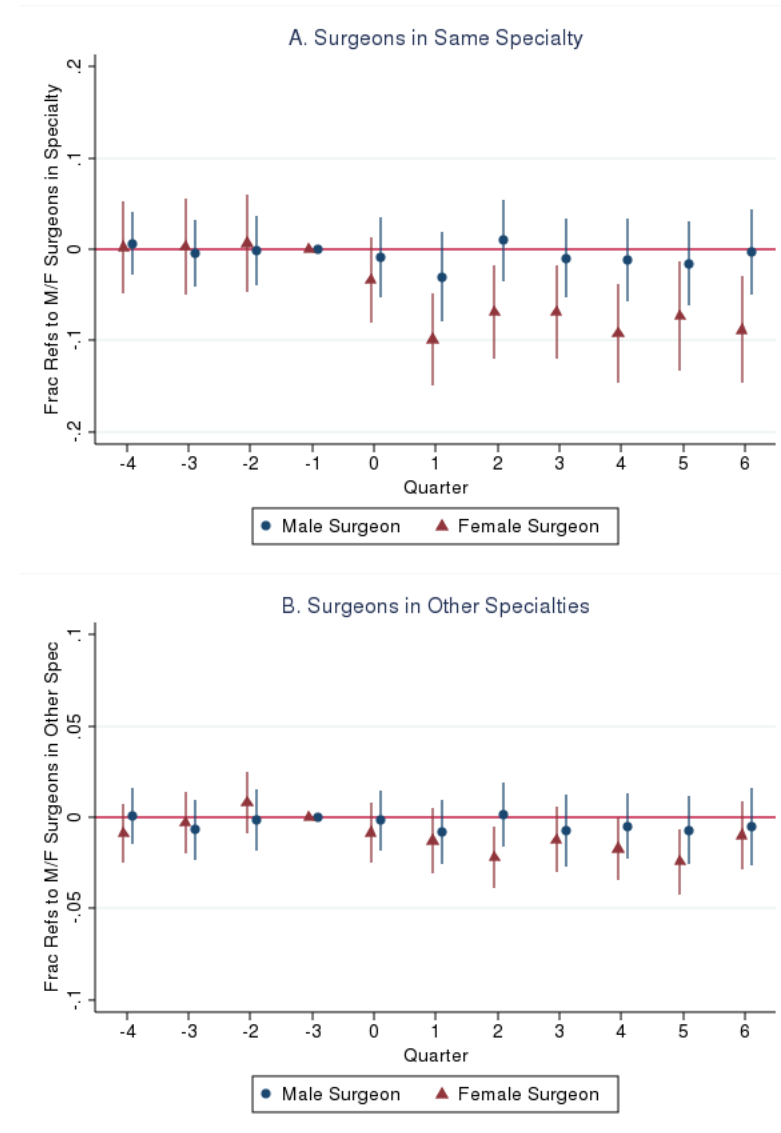


(b) Difference-in-Differences



Notes: Panel (a) in this figure shows the quarterly coefficients and 95% confidence intervals from estimating equations 1.3 and 1.4. The blue circles and red triangles show the impact that an unanticipated patient survival has on referrals to the male and female surgeon respectively. The grey circles and triangles show what would have happened in the absence of a survival. Panel (b) shows the difference-in-differences estimates and corresponding 95% confidence intervals. The grey circles show the estimates for the sample of placebo surgeons and the green triangles show the estimates for the sample of matched surgeons who have an unanticipated patient survival. All coefficients are plotted relative to the number of referrals the surgeon was receiving from the physician in  $k = -1$ , which are normalized to zero. Standard errors are clustered at the physician-surgeon match level.

**Figure 1.7: Spillovers to Other Surgeons after Bad Outcome**



*Notes:* This figure shows how physicians change their behaviour toward other surgeons after a patient death. In  $k = 0$ , a patient that physician  $j$  sent to surgeon  $i$  dies. The outcome variable is the fraction of new referrals going to female or male surgeons (whom physician  $j$  hasn't referred to before) in the same specialty as surgeon  $i$  (Panel A) or a different specialty (Panel B). Both outcomes variables exclude the performing surgeon. The blue circles are estimated from a regression using the fraction of referrals going to men as the outcome and the red triangles are estimated from a regression using the fraction of referrals going to women as the outcome. The coefficients are plotted relative to the number of referrals that physician  $j$  was sending to surgeons she had not previously referred to in quarter  $k = -1$ . I control for the fraction of available surgeon who are male or female and also include physician-surgeon match fixed effects. Standard errors are clustered at the physician-surgeon match level.

The spillover results are summarized in columns 3 and 4 of Table 1.5. Relative to the mean fraction of referrals going to new female surgeons in the same specialty, the decline in referrals is substantial. The fraction of a physician’s referrals going to new women in the same specialty as the performing surgeon declines by 53% (column 1). The fraction going to new women in other specialties (column 3) declines by 20% (but again, this result is insignificant).

**Table 1.5: Impact of an Event on Referrals to Others**

Fraction of New Referrals to Women in:	Same Specialty		Other Specialty
	(1)	(2)	(3)
Event	Bad	Good	Bad
Post	0.011 (0.025)	-0.022 (0.028)	0.007 (0.012)
Female × Post	-0.097** (0.039)	0.034 (0.044)	-0.028 (0.018)
Time Trend	0.003 (0.006)	0.003 (0.008)	0.004 (0.003)
Female × Time Trend	-0.003 (0.010)	0.002 (0.012)	0.001 (0.004)
Post × Time Trend	-0.004 (0.007)	-0.001 (0.009)	-0.005 (0.004)
Female × Post × Time Trend	0.006 (0.012)	0.001 (0.014)	-0.001 (0.005)
Avg. Post Effect	-0.079	0.029	-0.025
Mean of Outcome Var.	0.15	0.12	0.13
Observations	34,053	29,214	34,053
Clusters	3,425	2,948	3,425
R-Squared	0.417	0.350	0.206

*Notes:* This table displays the effect of a bad (columns 1 and 3) or a good (column 2) event by a performing surgeon on referrals to other surgeons of the same gender. The outcome variable is the fraction of a physician’s new referrals that go to men or women, excluding the performing surgeon. Columns 1 and 2 look at the fraction of new referrals going to women/men in the same specialty as the performing surgeon while Column 3 looks at the fraction of new referrals going to women/men in other specialties. I control for the fraction of available of surgeons who are women or men within the same specialty (Columns 1 and 2) or in any specialty (Columns 3). Regressions are estimated using the sample of matched male and female surgeons who experience a patient death or an unexpectedly good outcome. Levels of significance: \*10%, \*\* 5%, and \*\*\* 1% level.

Interestingly, physicians do not appear to treat women as a group when a woman performs well. Column 2 of Table 1.5 shows the spillovers to other women after a female surgeon has a good patient outcome. The coefficient on the female×post interaction variable is positive but insignificant.<sup>16</sup> It is possible, though, that a physician updates her beliefs about other female surgeons upward but continues to only refer to the performing surgeon.

These results also provide evidence that the drop in referrals is not simply due to female

<sup>16</sup>Appendix Figure A2 plots the coefficients from estimating equation 1.5 using good events. There are no significant positive spillovers to other women or men in the same specialty as the performing surgeon.

surgeons changing their behavior. A large literature shows that women are less likely to be overconfident than men (Lichtenstein, Fischhoff, and Phillips, 1982; Beyer, 1990; Barber and Odean, 2001). It is possible that female surgeons turn away more referrals after a patient death if the event hurts their confidence. The fact that physicians are changing their behavior toward other female surgeons, though, suggests that at least part of the change is due to physicians.

### Information Spillovers to Other Physicians

The implications that asymmetric updating has on a surgeon's career depends in part on whether information about an event spreads to other physicians. I test whether other physicians react to a bad event in Figure 1.8. I plot the coefficients from estimating

$$R_{i,-j,k} = \sum_{k=-4}^6 \beta_k event_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (event_{ij',t-k} \times fem_i) + \theta_{ij} + \epsilon_{ijk} \quad (1.7)$$

where  $event_{ij,t-k}$  is still a dummy variable indicating that a bad event occurred between physician  $j$  and surgeon  $i$ , but the outcome variable is the number of referrals that surgeon  $i$  receives from other physicians (excluding  $j$ ). I consider two outcome variables: the number of referrals going from other members of physician  $j$ 's group practice and the number of referrals from physicians outside of physician  $j$ 's practice.

Panel (a) plots the coefficients when the outcome variable is the number of referrals to the performing surgeon from other members of physician  $j$ 's group. I restrict the group practices to be those containing at most 20 members. This is done to account for the fact that many group practices defined in the Medicare data cover groups with branches in multiple regions. It is therefore unclear whether group practices with many members are large practices in one geographic location or normal-sized practices with multiple branches. There is a small decline in the number of referrals from other members of small group practices. Female physicians receive on average 0.5 fewer referrals per quarter after a bad event than before, but the coefficients in quarters 1 through 6 are jointly insignificantly different from zero. Male surgeons, on the other hand, continue to receive referrals from the referring physician's practice.

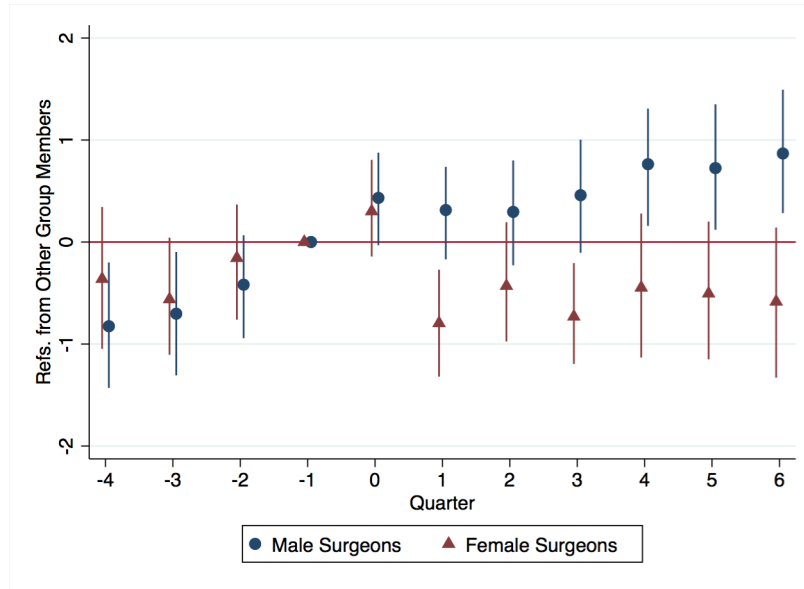
Panel (b) shows the change in referrals from physicians outside of physician  $j$ 's practice but in the same HRR. Here we see no impact on the number of referrals the surgeon receives. Both women and men continue to receive referrals from physicians outside of the referring physician's practice, suggesting that information does not spread or that physicians do not act on this information.

### 1.5.2 What Influences a Physician's Reaction to a Signal?

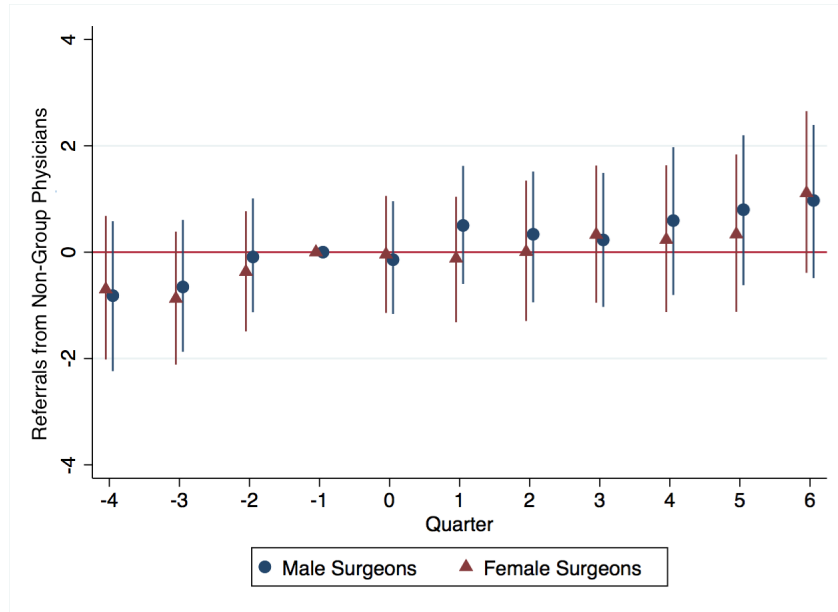
I have shown that physicians respond differently to a given patient outcome depending on the surgeon's gender. Physicians lower referrals to female surgeons more than male surgeons after a bad outcome and increase referrals more to male surgeons than to female surgeons after a good outcome. I now explore other factors that influence a physician's reaction, finding two main

**Figure 1.8: Information Spillovers to Other Physicians**

(a) Referrals from Physicians in Group Practice



(b) Referrals from Physicians Outside of Practice



Notes: This figure shows how referrals from other physicians to a performing surgeon change after a bad event. The outcome variable in Panel (a) is the number of quarterly referrals from physicians in the referring physician’s group practice. The outcome variable in Panel (b) is the number of quarterly referrals from physicians outside of the referring physician’s group practice but in the same Hospital Referral Region. Both panels are estimated on the sample of matched male and female surgeons who experience a patient death. All coefficients are plotted relative to the number of referrals the surgeon was receiving from the physician in  $k = -1$ , which are normalized to zero. Standard errors are clustered at the physician-surgeon match level.

drivers. First, physicians' reactions to signals are weaker the more signals they have received from a surgeon in the past. Physicians who have received many signals from a woman prior to the event are also more likely to treat her the same as a man. Second, physicians with a high propensity to refer to women before the event treat men and women more equally - e.g. exhibit less asymmetric updating - than physicians with a low propensity to refer to women.

### Referral History with Surgeon

So far I have restricted the analysis to physician-surgeon pairs in which the physician has referred at most 10 patients to the surgeon before the event. This concentrates the analysis on physicians who are presumably learning about a particular surgeon. If physicians learn about surgeons over time, they will become more certain in their beliefs about a surgeon's ability the longer they have been referring to one. To test how a physician's reaction changes with the length of her relationship with a surgeon, I estimate equation 1.6 where the independent variable of interest is the number of referrals that physician  $j$  sent to surgeon  $i$  in the year before the event.<sup>17</sup> I estimate equation 1.6 on the set of matched surgeons who have had up to 50 prior referrals from a given physician. I am still comparing male and female surgeons who receive the same number and fraction of referrals before an event occurs, but am looking at how a physician's reaction changes the more a physician has worked with a surgeon.

The results are presented in columns 1 and 3 of Table 1.6. *Pre-Referrals* is the number of patient referrals that physician  $j$  sent to surgeon  $i$  before the bad (column 1) or good (column 3) event. These "pre-referrals" are referrals that ended well: the patient did not die and was not readmitted to the hospital. I am therefore measuring how an additional referral that ended in a good patient outcome affects the physician's response. An additional referral from a physician to a surgeon reduces the physician's negative response to a death by 0.014 referrals per quarter for male surgeons and 0.022 referrals per quarter for female surgeons, thereby diminishing the gender gap in physician response. Female surgeons need to have sent 22 good signals (performed 22 procedures that went well) to a physician in the year before the event for there to be no gender gap in the physician's response.

An additional referral mitigates a physician's positive reaction to a good outcome when the surgeon is male but does not change the positive response when the surgeon is female, again leading to more equal responses. These results suggest that physicians learn about surgeons, becoming more certain in their beliefs about a surgeon's ability over time. Any bias that they exhibit comes out only when working with new surgeons.

---

<sup>17</sup>I do not look at the full referral history as the data is cut off in 2008. However, these two measures should be correlated.

**Table 1.6: Experience with Surgeons and Physician Response**

Outcome Var:	Referrals from Physician to Performing Surgeon			
	(1)	(2)	(3)	(4)
	Bad Outcome		Good Outcome	
Post	-0.084 (0.061)	0.247*** (0.075)	0.379*** (0.064)	0.619*** (0.089)
Female × Post	-0.475*** (0.068)	-0.548*** (0.102)	-0.285*** (0.061)	-0.511*** (0.128)
Post × Pre-Referrals	0.014** (0.005)		-0.015** (0.007)	
Female × Post × Pre-Referrals	0.008* (0.004)		0.013** (0.006)	
Post × High Propensity		-0.145 (0.090)		-0.147 (0.135)
Female × Post × High Propensity		0.210** (0.103)		0.373** (0.159)
Mean of Outcome Var.	1.30	0.54	1.04	0.54
Observations	48,610	26,387	39,640	20,245
Clusters	4,972	2,378	4,025	1,745

*Notes:* This table looks at how a physician’s relationship with the performing surgeon and with other surgeons changes her response. The outcome variable is the number of patients that the referring physician sends to the performing surgeon in a quarter. “Pre-Referrals” is the number of referrals that surgeon  $i$  received from physician  $j$  before the patient death. “High Propensity” is an dummy variable that equals one when a physician sends a larger fraction of her referrals to female surgeons than the average physician in her HRR. All regressions include a time trend and time trend interactions (with Female Surgeon, Post Event and the Pre-Referrals/High Propensity) as well as physician-surgeon match fixed effects. Standard errors are clustered at the physician-surgeon match level. Levels of significance: \*10%, \*\* 5%, and \*\*\* 1% level.

### Physician Response and Referral History with Other Surgeons

The previous section showed that the longer the referral relationship between a physician and a surgeon, the more muted the physician’s response to an event is. It is unclear, though, how a physician’s response is influenced by her referral history with other surgeons of the same gender as the performing surgeon. For example, a physician who has favorable beliefs about women and who sends a large volume of referrals to them presumably has more information about women. That physician may not react as much to new signals. Depending on the shape of the physician’s priors, though, a surgeon who thought women were very good and was consequently surprised to see a patient die under a woman might drastically revise her beliefs.

To understand how a physician’s response varies with her referral history with other surgeons, I again estimate equation 1.6 and use the “propensity to refer to women” variable described in Section 1.3 as the main explanatory variable. A physician’s propensity to refer to women is measured as the difference between the fraction of her referrals going to women in a particular specialty and the average fraction going to women of other physicians in the same referral area. I then define a variable, *High Propensity*, that equals one if a physician sends a greater fraction of her referrals to women within a given specialty than the average physician in her referral region.

The results are presented in columns 2 and 4 of Table 1.6. Physicians with a high propensity to refer to women react less negatively after bad outcomes and more positively after good outcomes. The gender gap in the physician’s reaction shrinks by 22% if a physician has a high propensity to refer to women.

### Other Variables Influencing Physician’s Reaction

I also test whether a physician’s outside option, surgeon and physician experience, and physician gender influence the physician’s reaction to a signal.

**Outside Options** It is plausible that a physician who has many possible surgeons she can refer to would have a stronger response to a patient death as it is easier for her to shift to a new surgeon. To test whether a physician’s outside option matters, I use the number of surgeons in the same specialty as the performing surgeon in the referring physician’s HRR. The results from estimating equation 1.6 are presented in Column 1 of Table 1.7, the x-variable of interest is physician  $j$ ’s outside option. There is a small impact of having more outside options but the result is statistically insignificant.

**Table 1.7: Variables Correlated with Surgeon Response**

X-Var:	Referrals to Surgeon			
	Num. Surgeons (1)	Surgeon Exp. (2)	Physician Exp. (3)	Fem. Physician (4)
Post	-0.256* (0.129)	-0.043 (0.131)	0.177 (0.160)	0.109 (0.058)
Female × Post	-0.522*** (0.119)	-0.306** (0.130)	-0.518** (0.199)	-0.434*** (0.056)
Post × X-Var	0.004 (0.002)	0.002 (0.006)	-0.004 (0.006)	0.058 (0.140)
Female × Post × X-Var	0.004 (0.002)	0.003 (0.010)	0.008 (0.008)	0.185 (0.123)
Observations	17,691	34,054	28,989	28,989
Clusters	1,781	2,945	2,578	2,914

*Notes:* This table tests whether the four variables listed at the top of each column are correlated with a physician’s response. Each variable is interacted with the Post and Post×Female variables. “Num. Surgeons” is the number of surgeons in the same specialty as the performing surgeon who work in the physician’s Hospital Referral Region (i.e. the physician’s outside options). “Surgeon Exp.” and “Physician Exp.” are the surgeon and physician’s experience, measured as the number of years that each doctor has been out of medical school. “Physician Fem.” is a dummy variable indicating that the physician is a woman. A time trend, time trend interactions with all variables, and a physician-surgeon match fixed effect are included in each regression. Columns 2 and 3 also include surgeon experience squared and physician experience squared respectively. Standard errors are clustered at the physician-surgeon match level. Levels of significance: \*10%, \*\* 5%, and \*\*\* 1% level.

**Experience** I look at both the physician’s and the surgeon’s experience, measured as the number of years since they graduated from medical school, to understand whether experience influences

a physician’s reaction to signal. I again estimate equation 1.6 but include  $Exper_{i,t}$  (the surgeon’s experience at the time of the event) or  $Exper_{j,t}$  (the physician’s experience) as the main independent variable. I also include squared experience terms to allow for non-linearities in the doctor’s response. The results are presented in columns 2 and 3 of Table 1.7. There is no significant relationship between either doctor’s experience and the physician’s reaction. A physician’s experience working with a particular surgeon seems to matter more than absolute years of experience.

**Physician Gender** The evidence on how women evaluate other women is mixed. For example, Casadevall and Handelsman (2014) find that having a woman on a convening team for scientific conferences increases the proportion of invited women. Bagues et al. (2017), however, find that women evaluating tenure cases do not favor female candidate and that men in fact become harsher in their evaluations in the presence of a female evaluator.

I find no significant evidence that male and female physicians treat female surgeons differently. In column 4 of Table 1.7, a female dummy indicating that the physician is female is included as the independent variable of interest. The point estimate on the triple interaction between *Female*, *Post*, and *Female physician* is positive, suggesting that female physicians might be easier on female surgeons than male physicians, but the result is noisy and I cannot rule out no or negative effects.

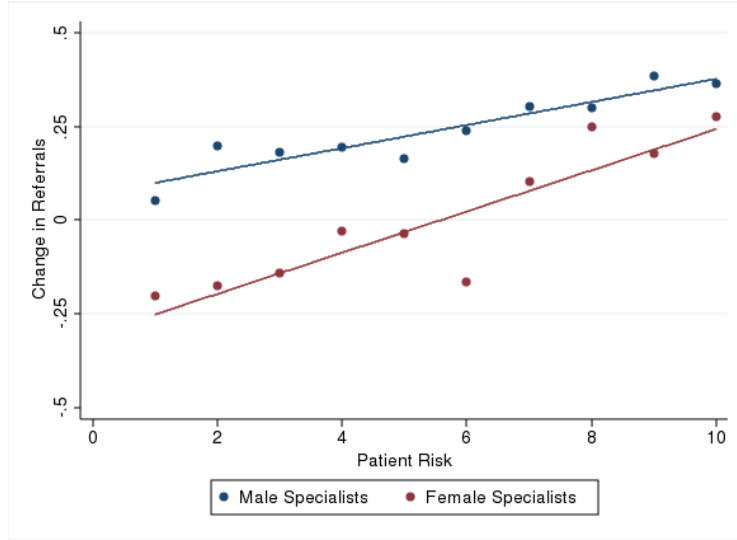
## Summary of Results

In sum, I look at the impact of two signals, good and bad, on two sets of outcomes for male and female surgeons: referrals to the performing surgeon and referrals to other surgeons of the same gender. The following table, which draws from Tables 1.4 and 1.5, shows the average post effect of an event on referrals to individual surgeons and to other surgeons.

	Performing Surgeon		Other Surgeons	
	Male	Female	Male	Female
Bad Outcome	0.101	-0.222	∅	-0.079
Good Outcome	0.604	0.346	∅	∅

## 1.6 Alternative Interpretations

Before discussing what the empirical findings tell us about how physicians update their beliefs, I explore three alternative interpretations of the results. I show that differences in unobservable risk, differences in the predictiveness of events for future events, and changes in the physician’s behavior cannot account for the findings.



Notes: This binned scatterplot shows the relationship between the risk of the patient that died and the change in referrals from physician  $j$  to surgeon  $i$ , using the matched sample of male and female surgeons who experience a patient death. Patient risk is calculated as in Section 1.3 and is then binned into deciles. Both variables are residualized on a physician-surgeon match pair effect. The line of best fit using OLS is shown separately for male and female surgeons. The lines of best fit have slopes of 0.060 (s.e. = 0.004) for female surgeon and 0.026 (s.e. = 0.002) for male surgeons.

**Figure 1.9: Physician Response by Patient Risk**

### 1.6.1 Differences in Patient Risk

Although I match on observable patient risk, there could be unobservable factors influencing risk. If female surgeons receive less risky patients, deaths are surprising and survivals are unsurprising, meaning that physicians respond strongly to deaths and weakly to survivals. Here, I put bounds on what the unobservable risk difference between male and female surgeons' patients would have to be to justify the degree of differential updating.

In Figure ??, I plot the regression coefficients  $\hat{\beta}_1$  and  $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$  from estimating

$$\Delta R_{ijp} = \beta_1 Risk_p + \beta_2 (Risk_p \times Fem_i) + \beta_3 Fem_i + \theta_i + \epsilon_{ijp} \quad (1.8)$$

where  $\Delta R_{ijp}$  is the change in referrals from physician  $j$  to surgeon  $i$  before and after patient  $p$ 's death, and  $Risk_p$  is the patient  $p$ 's observed risk level. The coefficient  $\hat{\beta}_2$  tells us what the difference in unobserved risk between a female surgeon's and a male surgeon's patient would have to be to account for the gender difference in the physician's reaction.

While physicians respond to patient risk regardless of surgeon gender, they respond much more when the surgeon is female. Figure ?? shows that a male surgeon with a patient in the bottom risk decile experiences a drop in referrals equivalent to what a female surgeon with a patient in the 7th decile receives. Thus, any differences in unobserved risk would have to be large enough to move a patient that a male surgeon sees from being in the bottom 10th percentile of patient risk to the top 70th percentile of observed risk, a large difference.

### 1.6.2 Are Outcomes Differentially Predictive of Future Outcomes?

A physician's behavior is also be justified if bad events are predictive of future bad events for female surgeons and good events are predictive of future good events for male surgeons. To test this hypothesis, I estimate

$$\mathbb{P}(Event_i = 1 | X_i, X_{pt}, X_{i,p'}) = \beta_1 Fem_i + \beta_2 FutRefs_i + X_i' \gamma + \alpha \log(PastRisk_p) + \delta \log(FutRisk_{p'}) + \epsilon_{ip}$$

on the matched samples of surgeons who experienced a bad or good event. The outcome variable is the probability that surgeon  $i$  has another event in the future under any physician (not just physician  $j$ ). I condition on the number of referrals the surgeon receives in the future from all physicians ( $FutRefs_i$ ) as well as the log of future patient risk ( $\log(FutRisk_{p'})$ ). I also control for the performing surgeon's characteristics ( $X_i'$ ), including experience, specialty, and work history<sup>18</sup>, as well as the log of the surgeon's past patients' risk levels ( $\log(PastRisk_p)$ ).

The results are presented in Table 1.8. I find no evidence that patient deaths are more predictive of future deaths for female surgeons than for male surgeons. In fact, women are less likely to have future patients die even conditional on the risk of future patients. If physicians are responding to an unobservable factor, it is not something that influences the future performance of surgeons.

### 1.6.3 Do Physicians Stop Referring for Certain Procedures?

An additional concern is that physicians stop referring patients for a particular surgery after a patient dies. For example, Keating et al. (2017) find that doctors who refer a patient for a colonoscopy are less likely to refer patients for that procedure if the patient has an extreme adverse outcome. However, the effect is short-lived as physicians begin referring for the procedure again, and at the same rate as before, one quarter after the adverse outcome. Nevertheless, if women tend to specialize in one type of surgery while men perform many different surgeries, the drop in referrals to women could be due to the physician changing the types of procedures she refers.

I test this by looking at how referrals for the surgery that was performed change after a patient death. Specifically, I estimate

$$S_{jk} = \sum_{k=-4}^6 \beta_k event_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (event_{ij,t-k} \times Fem_i) + \theta_{ij} + \epsilon_{ij} \quad (1.9)$$

where  $S$  is the surgery that was being performed on the patient who died and  $S_{jk}$  is the number of those surgeries that physician  $j$  refers to any surgeon in quarter  $k$ .

The results are shown in Appendix Figure A3. For female surgeons, the coefficients on all quarters before and after the event are precise zeros, indicating that the physician does not change her referral patterns for the surgery in question. For male surgeons, there is a small drop in the

---

<sup>18</sup>This includes the number of patients seen before the death.

**Table 1.8: Are Events Predictive of Future Events?**

	# Bad Events (1)	Any Bad Event (2)	# Good Events (3)
Female Surgeon	-0.792** (0.366)	-0.031*** (0.011)	-0.045 (0.177)
Log Future Ptnt Risk	0.283 (0.197)	0.223*** (0.017)	0.859** (0.321)
Future Referrals	0.006*** (0.001)	0.001** (0.001)	0.005*** (0.001)
Log Past Ptnt Risk	-1.173*** (0.380)	-0.038*** (0.005)	0.243*** (0.070)
Past Referrals	-0.001* (0.001)	0.001*** (0.000)	-0.002*** (0.000)
Surgeon Experience	0.001 (0.013)	-0.001 (0.001)	0.012*** (0.004)
Observations	24,459	24,459	24,478
R-Squared	0.594	0.373	0.418

*Notes:* This table shows the results from testing whether events are differentially predictive of future events for male and female surgeons. The sample consists of the full sample of surgeons who every experienced a bad or good event. Log Future Patient Risk is the average risk of the future patients. Future Referrals is the number of referrals that the surgeon receives from any physician in the future. Similarly, # Past Referrals and Log Patient Past Risk are the number of referrals and average risk of those patients that the surgeon received before the patient death. Surgeon specialty fixed effects are also included. Standard errors are clustered at the specialty level. Levels of significance: \*10%, \*\* 5%, and \*\*\* 1% level.

number of referrals the physician gives for the procedure in question in  $k = 1$  but the physician's referrals revert back to the mean shortly thereafter.

I also show that the physician decreases referrals to surgeon  $i$  for all procedures, not just the one that was being performed on the patient who died. In Appendix Figure A4, I plot the coefficients from estimating

$$OtherRef_{ijk} = \sum_{k=-4}^6 \beta_k event_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (event_{ij,t-k} \times Fem_i) + \theta_{ij} + \epsilon_{ij} \quad (1.10)$$

where the outcome variable,  $OtherRef_{ijk}$ , is the number of referrals that the physician sends to the surgeon aside from the procedure that was being performed on the patient who died. The results are noisy as physicians typically only refer to a surgeon for a particular procedure, but the patterns are the same. Referrals for other procedures drop for both men and women but by a greater amount for women. The coefficients for female surgeons are significantly different from those for male surgeons at the 10% level.

## 1.7 Welfare Analysis and Career Effects

### 1.7.1 Surgeon Ability

If asymmetric updating distorts a physician's belief about a surgeon's ability, physicians may switch away from high ability female surgeons after receiving negative signals. For example, if physicians have some cutoff ability, below which they do not refer to a surgeon, physicians will stop referring to female surgeons earlier than similar men. The average ability of male surgeons they refer to will eventually be lower than that of the female surgeons they refer to.

To test whether asymmetric updating affects the average quality of surgeons a physician refers to, I use the definition of surgeon ability described in Section 1.3. I then calculate the average ability of all surgeons a physician refers to in each quarter and plot how it changes after a bad event under one surgeon. If physicians give male surgeons too many chances to make mistakes and female surgeons too few chances, average surgeon ability will decline after a patient death as physicians move away from potentially qualified female surgeons to potentially less qualified male surgeons.

Figure 1.10 plots the quarterly coefficients from estimating

$$\bar{a}_{jk} = \sum_{k=-4}^6 \beta_k event_{ij,t-k} + \sum_{k=-4}^6 \beta_k (event_{ij,t-k} \times fem_i) + \theta_{ij} + \epsilon_{ijk}$$

where  $\bar{a}_{jk}$  is the average ability of the surgeons that physician  $j$  refers to in quarter  $k$ .

There is no significant change in the quality of surgeons that physicians refer to when the performing surgeon is a man. However, when the performing surgeon is a woman, the average surgeon quality falls by approximately 0.1 standard deviations in the year following a death, although these results are only marginally significant at the 10% level. However, they provide suggestive evidence that physicians may be misestimating female surgeons' abilities after deaths and switch from some high ability female surgeons to other lower ability surgeons.

### 1.7.2 Surgeon Pay and Skill Accumulation

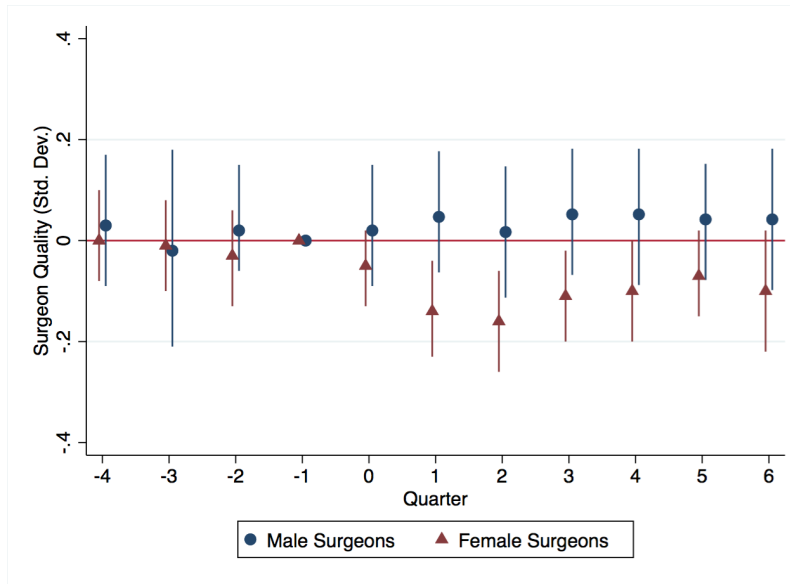
I now turn to the impact that asymmetric updating has on surgeons' career trajectories. In Section 1.5, I find limited evidence that information about a patient death spreads to other physicians. There is a small decline in the number of referrals that women receive from other members of the referring physician's group practice, provided the practice is small. However, this decline is statistically indistinguishable from zero. As such, I abstract from changes in other physicians' behavior and focus on the impact that asymmetric updating by the referring physician has on pay and skill accumulation.

Numerous papers have documented a pay gap between male and female physicians.<sup>19</sup> Closely related to this paper, Zeltzer (2017) decomposes the Medicare earnings gap, showing that in the

---

<sup>19</sup>See, for example, Ly et al. (2016), Lo Sasso et al. (2011), and Sasser (2005)

**Figure 1.10: Change in Surgeon Quality**



Notes: This figure shows how the quality of surgeon that a physician refers to changes after a patient death. In the figure, a patient that surgeon  $i$  received from physician  $j$  dies in quarter  $k = 0$ . The outcome variable is the standard deviation of surgeon ability from the mean. The coefficients are plotted relative to the standard deviation of surgeon quality that the physician was referring to in  $k = -1$ . For example, if a physician was referring to surgeons who were an average of 1 standard deviation above the mean in  $k = -1$ , all coefficients are plotted relative to 1. Standard errors are clustered at the physician level.

raw data, women earn 48% less than men. About a third of the gap can then be explained by difference in the specialties men and women select into. Controlling for career interruptions and differences in experience and education, Zeltzer shows that gender homophily in referrals explains an additional 15% of the gap. However, the remainder of the gap remains unaccounted for.

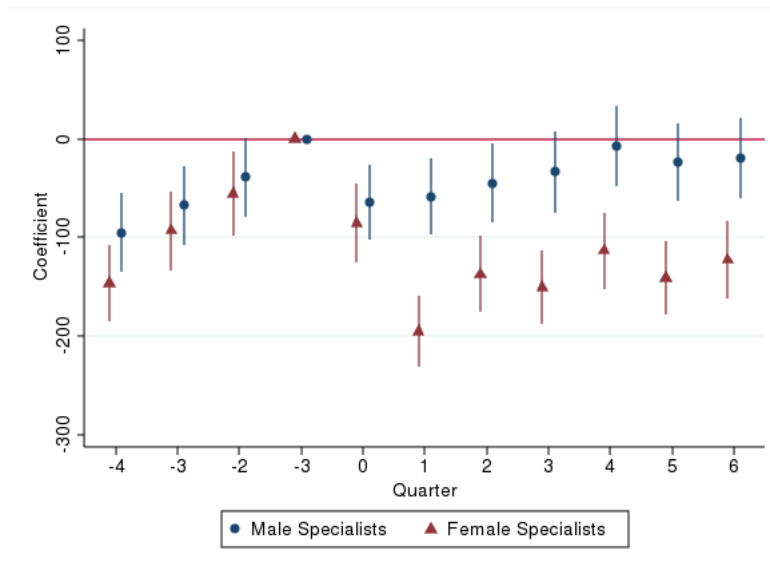
Here, I provide an additional mechanism that contributes to the gap, but do not quantify the contribution. It is difficult to estimate the full impact that asymmetric updating has on the surgeon pay gap as I look at two specific and relatively infrequent events. Surgeons experience less than one bad outcome per year, for example. Furthermore, physicians exhibit asymmetric updating when they are starting new referral relationships, so the relevant baseline pay gap for comparison is the gap that exists between men and women at the beginning of their careers or who have moved. I therefore show that asymmetric updating creates a wedge between male and female pay but do not speculate on its relative importance in explaining the overall surgeon pay gap.

Figure 1.11 shows the change in Medicare payments following a patient death, plotting the quarterly coefficients from estimating

$$Pay_{ijk} = \sum_{k=-4}^6 \beta_k Death_{ij,t-k} + \sum_{k=-4}^6 \gamma_k (Death_{ij,t-k} \times fem_i) + \theta_{ij} + \epsilon_{ijk} \quad (1.11)$$

on the sample of matched surgeons who experience a patient death. The outcome variable,  $Pay_{ijk}$ , is the total quarterly Medicare pay that surgeon  $i$  receives from physician  $j$ 's referrals in quarter  $k$ .

Figure 1.11: Medicare Payments



Notes: This figure shows plots quarterly Medicare payments using the sample of matched surgeons who experience a patient death. The outcome variable is the total quarterly Medicare payments that surgeon  $i$  received from  $j$ 's referrals. The coefficients are plotted relative to the total Medicare payments that the surgeon was receiving from the physician in  $k = -1$ . Payments in  $k = -1$  are normalized to zero. A patient that physician  $j$  referred to surgeon  $i$  dies in  $k = 0$ . Standard errors are clustered at the physician-surgeon match level.

A patient death occurs in  $k = 0$ .

Because I have matched on a number of variables that influence the gender pay gap (such as volume of referrals, experience, and specialty), there is a small but statistically insignificant difference in male and female surgeon pay before the death. After the death, a gap of approximately \$140 per quarter emerges. Column 3 of Table 1.4 summarizes the effect, showing that women lose approximately 60% of their Medicare billings from the referring physician while men lose 30%. Women incur a substantial pay penalty from the referring physician but given that other physicians do not change their behavior much, women do not experience a substantial pay penalty overall.

Column 4 of Table 1.4 shows the impact of a good patient outcome on Medicare pay. Men receive a 36% increase in quarterly Medicare billings while women receive a 19% increase, although the difference is statistically insignificant.

A second channel through which asymmetric updating can influence women's careers is through skill acquisition. In Section 1.5, I showed that female surgeons who still receive referrals after a patient death receive easier cases, either with less risky patients or less risky procedures. Since learning-by-doing is important for surgeon learning<sup>20</sup>, asymmetric updating might impact women's skill accumulation which can also influence future pay as well as their career trajectory.

<sup>20</sup>See, for example, Hughes (1991), Keehner et al. (2006).

## 1.8 Theoretical Framework

This section sets up a theoretical framework to link the empirical results to belief updating, answering whether and under what conditions the observed behavior is in line with Bayesian updating.

To be consistent with the main results, a model must have two key features:

1. *Asymmetry in Updating about the Individual*: Physicians must update their beliefs more about men after a good signal and more about women after a bad signal.
2. *Asymmetry in Updating about Groups*: Physicians update their beliefs about other women upon receiving a signal from one woman but do not update beliefs about other men after receiving a signal from one man.

I do not argue that one particular model explains the results, but rather outline the assumptions about a physician's beliefs that are needed for the behavior to be consistent with Bayesian updating. To derive these assumptions, I model the physician's decision problem to map referrals to beliefs. I then show that the behavior is consistent with Bayesian updating if (1) physicians believe that women are higher ability than men, or (2) the difference in the average variance of women's and men's abilities increases as physicians receive more signals. I discuss the empirical validity of these assumptions and show that if physicians hold such beliefs, they are inconsistent with the data on the distribution of surgeon ability, allowing me to reject rational expectations. I then discuss an alternative model in which physicians exhibit bias dynamically rather than through their priors.<sup>21</sup>

### 1.8.1 Physician's Decision Problem

#### Setup

I follow the setup in Zeltzer (2017). There are two types of agents, physicians and surgeons. Physicians, denoted  $j \in J$ , decide which surgeon,  $i \in I_j$ , to refer a patient to where  $I_j$  is the pool of surgeons available to  $j$ . Surgeons belong to an identifiable group  $g \in \{m, w\}$  (men or women) and their ability,  $a_i$ , is unknown to the physician. In period  $t = 0$ , physicians have a prior probability distribution over a surgeon's ability  $a_{i,t} \sim f(\bar{a}_i, \sigma_i^2)$  where  $\bar{a}_i$  is the physician's prior about surgeon  $i$ 's average ability. In time 0, the physician's prior is based on her beliefs about the group:  $\bar{a}_{i,0} = \bar{a}_g$ . Similarly,  $\sigma_i^2$  is the variance and  $\sigma_i^2 = \sigma_g^2$  in  $t = 0$ . I assume that  $f$  has a defined mean and variance but do not place restrictions on higher-level moments.

After receiving a patient, surgeons draw and send a signal (a patient outcome). To match the data, I assume that signals can be either good or bad:  $s \in \{s_G, s_B\}$ .<sup>22</sup> The probability of drawing each type of signal depends on a surgeon's ability, with higher ability surgeons being more likely

---

<sup>21</sup>Under some assumptions, various other models also fit the data. See, for example, Bordalo et al.'s (2016) model of stereotypes.

<sup>22</sup>The model can be extended to include a set of finitely many ordered signals and the results do not change.

to draw a good signal. Specifically, let the probability that a member from group  $g$  draws a signal  $s$  be  $\mathbb{P}(s|g) = \sum_{i \in g} \mathbb{P}(s|a_i) f(a_i) da$ .

Physicians want to maximize patient utility by referring to the best available surgeon subject to idiosyncratic factors like patient preferences and wait times. That is, physician  $j$  chooses a surgeon  $i$  to maximize

$$\operatorname{argmax}_{i \in I_j} U_p(a) = \beta \mathbb{E}[a_{i,t} | \bar{a}_i, g, s] - \lambda \sigma_{i,t}^2 + \epsilon_{ipt} \quad (1.12)$$

where  $\bar{a}_i$  is the physician's prior about surgeon  $i$ 's mean ability, and where  $i$  belongs to group  $g \in \{m, w\}$ . The constant  $\lambda$  represents risk preferences and  $\epsilon_{ipt}$  represents the idiosyncratic factors discussed above. The physician is thus trying to choose the surgeon with the highest expected ability, the first term, while trading off the variance of ability, the second term.

Assuming  $\epsilon_{ipt}$  is independently and identically distributed according to the extreme value distribution, we can write equation 1.12 in terms of the logit probability,

$$\mathbb{P}(R_{i,p,t}^j = 1 | a_{i,g,t-1}) = \frac{e^{v_{ij}}}{\sum_{i' \in I} e^{v_{i'j}}} \quad (1.13)$$

where  $R_{i,p,t}^j = 1$  if physician  $j$  refers patient  $p$  to surgeon  $i$  in time  $t$ , and  $v_{ij} = \beta \mathbb{E}[a_{i,t} | \bar{a}_i, s] - \lambda \sigma_{i,t}^2$ . Physician  $j$ 's total number of referrals to surgeon  $i$  in period  $t$  can then be written as

$$R_{ijt}^{total} = n_t \cdot \frac{e^{v_{ij}}}{\sum_{i' \in I} e^{v_{i'j}}}$$

where  $n_t$  is the number of patients that the physician refers to surgeon  $i$  in time  $t$  if the surgeon's capacity constraint has not been reached.

Aside from idiosyncratic factors like patient preferences, two main variables influence the physician's referral choice in this model: beliefs about ability ( $\mathbb{E}[a_{i,t}]$ ) and the variance of ability ( $\sigma_{i,t}^2$ ). Referrals are increasing in the surgeon's expected ability and decreasing in the surgeon's expected variance of ability.<sup>23</sup>

I consider each case in turn. I first look at how Bayesian physicians react to signals if they only care about mean ability, placing no restrictions on higher-order moments of the ability distribution. I then consider how physicians react if they have risk preferences and care about the variance of ability.

## 1.8.2 Bayesian Updating: Physician Cares about Mean Ability

Recall that physicians have the prior probability distribution  $a_{i,t} \sim f(\bar{a}_i, \sigma_i^2)$  over a surgeon's ability. If a physician only cares about mean ability, she chooses  $i$  to maximize  $U_p(a) = \beta \mathbb{E}[a_{i,t} | \bar{a}_i, g, s] + \epsilon_{ip}$ ,

---

<sup>23</sup>This assumes risk aversion in which a physician prefers to refer to a surgeon who produces average outcomes for sure than one who produces either very good or very bad outcomes with some probability. There may be cases in which a physician would prefer to refer to a surgeon with a high expected  $\sigma^2$  instead of a "safe" option, but I abstract from this case here.

which is equation 1.12 setting  $\lambda = 0$ . Holding the physician's beliefs about other surgeons constant, the change in referrals after a signal is then

$$\Delta \bar{R}_{ijt} = \frac{(n_t - n_{t-1}) \cdot e^{\beta(\mathbb{E}[a_{i,t}|\bar{a}_{i,g,s}] - \bar{a}_i)}}{\sum_{i' \in S} e^{\beta \bar{a}_{i',t-1}}} \quad (1.14)$$

Empirically, the change in referrals to women is larger after a bad signal and smaller after a good signal compared to the change in referrals to men. For this to be true, it must be that  $\mathbb{E}[a_{i,t}|s, w] - \bar{a}_{i,w} \leq \mathbb{E}[a_{i,t}|s, m] - \bar{a}_{i,m}$  in equation 1.14. This is possible under Bayesian updating when physicians believe that women have higher average ability than men.

Consider the following proposition:

**Proposition 1.** *If physicians are Bayesian and the following statements are true:*

1.  $\mathbb{E}[a_i|s_G, g] > \mathbb{E}[a_i|s_B, g]$
2.  $\mathbb{E}[a_{i,t}|s, w] - \bar{a}_{i,w} \leq \mathbb{E}[a_{i,t}|s, m] - \bar{a}_{i,m}$

then it must be that  $\frac{\mathbb{P}(s_G|w)}{\mathbb{P}(s_G|m)} > 1$ .

*Proof.* See Appendix A.2. □

Proposition 1 states that if expected ability is increasing in the signal and that the change in beliefs about women's ability are always more negative than the change in beliefs about men's ability, then women must have a higher average ability than men, which is equivalent to women having to send more good signals. Intuitively, if a Bayesian physician believes that women are more likely to send good signals, any good signal is going to be close to the physician's prior and will not move her beliefs much. In contrast, when a man sends a good signal, the signal is far from the physician's prior, leading her to update more. Similarly, if a woman sends a bad signal, the physician will update much more than she does about a man, whom the physician already believes to be low ability.

## Discussion of Assumptions

If physicians believe that women are higher ability than men and physicians only care about mean ability, equation 1.12 implies that women should receive at least as many referrals as men do. Empirically, female surgeons receive 10% of referrals while they make up 17% of the surgeon population. Adjusting for the fact that women work fewer hours than men, they should receive at least 15% of referrals if physicians refer to men and woman at equal rates.<sup>24</sup> While the difference is small, female surgeons are under-referred to even after adjusting for work hours.

Theoretically, a cost of referring to women could result in men and women receiving an equal number of referrals but women being higher ability than men. A female surgeon would have to be

---

<sup>24</sup>See Appendix A.1.1 for description and calculation of work hours adjustment

sufficiently high ability to offset any cost associated with referring to her. Such costs could arise for a variety of reasons. For example, in a model of taste-based discrimination, physicians pay a utility cost to refer to women so they will only refer to women whose ability outweighs this cost.

25

Two empirical observations suggest that physicians do not believe that women are higher ability than men. First, in Table 1.6, I showed that physicians who send a larger fraction of their referrals to women react less to bad events and more to good events. If physicians who refer more to women than average have higher beliefs about women's ability, we should see the opposite result. Beliefs about women's ability should be negative correlated with referrals after a bad event as these are the physicians who should be the most surprised to see a bad outcome. However, these are empirically the physicians who react strongly to good signals and weakly to bad signals.

Second, while I cannot observe beliefs directly, I can look at the ability distributions of men and women to test for differences in average ability and to test whether physicians have rational expectations. Women are slightly higher ability than men. Figure 1.2 shows the ability PDF and CDF for male and female surgeons in the unmatched sample. Table 1.1 reports the means. Women have higher average ability but the difference is small: women's ability is 0.0001 points larger than men's. It is unlikely that such a small difference in ability leads to such a large reaction from physicians. Nevertheless, because I do not directly observe physicians' beliefs, it is possible that they believe that there is a large difference in surgeon ability, which produces the observed asymmetries in referrals. If this is the case, physicians' actions are in line with Bayesian updating, but are inconsistent with physicians having rational expectations as their beliefs would not match the actual surgeon ability distribution.

### 1.8.3 Bayesian Updating: Physician Cares about Variance

If physicians also care about the variance of ability, they might choose to refer to a surgeon with a lower average ability but also a lower variance as this would mean less uncertainty in the patient's outcome. For example, if physicians are risk averse and want to avoid very bad outcomes like deaths, they might refer to a surgeon with lower expected ability, but whose ability she is certain of, over a surgeon with higher expected ability but for whom she is less certain about. The opposite could be true if physicians are willing to take a risk to try to get a good outcome.

The following proposition states the assumption under which the empirical results are consistent with Bayesian updating when the physician has risk preferences. It holds regardless of which group has higher variance but assume for now that the variance of beliefs for female surgeons' ability is larger than that for male surgeons.

**Proposition 2.** *If there are asymmetric changes in beliefs about male and female surgeons, it must be that the difference in the expected variance for female and male surgeons increases as the physician receives more*

---

<sup>25</sup>Similarly, women might work less convenient hours or be geographically distant from patients, making them more costly to refer to. Physicians would only refer to surgeons if their ability was sufficiently high to offset these costs

signals:

$$\sigma_w^2 - \sigma_m^2 < \mathbb{E}[\sigma_w^2|s] - \mathbb{E}[\sigma_m^2|s] \quad \forall s \in \mathcal{S}$$

*Proof.* See Appendix A.2. □

The proof in Appendix A.2 shows that if  $\sigma_w^2 - \sigma_m^2 < \mathbb{E}[\sigma_w^2|s] - \mathbb{E}[\sigma_m^2|s]$  does not hold, the physician must have a larger spread in beliefs about the group with the larger initial variance. Put differently, if the prior variance is larger for women, there is a larger spread in possible outcomes under female surgeons than male surgeons. This means that the physician should update more about women after *both* good and bad signals. Physicians will only update asymmetrically if the difference in posterior variance between women and men ( $\mathbb{E}[\sigma_w^2|s] - \mathbb{E}[\sigma_m^2|s]$ ) is larger than the difference between the prior variances ( $\sigma_w^2 - \sigma_m^2$ ) as the physician receives signals. This is a fairly strong assumption and does not hold for many of the distributions commonly used to model updating (for example, with normal or binomial distributions). Proposition 2 also holds for non-symmetric distributions and does not require the means of the distributions to be the same.

### Discussion of Assumptions

The assumption that the difference in variance increases with signals is strong and there are few cases in which it would be violated. One such case is if good signals are informative for male surgeons but uninformative for female surgeons. In this case, the posterior for men could be much narrower than the posterior for women even if the prior on men is only slightly narrower than the prior on women, meaning that  $\sigma_w^2 - \sigma_m^2 < \mathbb{E}[\sigma_w^2|s] - \mathbb{E}[\sigma_m^2|s]$  would hold. In Table 1.8, I test whether events are differentially predictive of future events for men and women and do not find any such evidence. However, it is possible that physicians hold these beliefs even if they do not match the data.

Distributions with odd higher-level moments can also violate the assumption.<sup>26</sup> Again, while I do not observe physicians' beliefs, I can measure the variance of surgeons' abilities. Table 1.1 gives the standard deviation of the ability measure for both men and women using the unmatched sample. The variance of women's abilities is slightly smaller than that of men. Figure 1.2 also shows that higher-level moments are roughly the same (there is no difference in skewness, for example). The difference in the variance of priors for male and female surgeons would thus have to be weakly increasing as the physician receives signals despite the actual variance in abilities being roughly the same. This again rejects rational expectations.

---

<sup>26</sup>For example, if a physician has received few signals from women and the female ability distribution is strongly left-skewed relative to the male ability distribution, the observed asymmetries will appear.

#### 1.8.4 Alternative Models

Many existing discrimination models in labor economics start with the assumption that employers hold different beliefs about two groups<sup>27</sup> For example, if employers think that women are on average less skilled than men, they will be reluctant to hire women. However, after receiving signals from individual workers, employers correctly update based on these signals and can eventually end up with the same beliefs about a male and female worker. Once an employer holds the same beliefs about a man and a woman's abilities, the employer will form the same posterior about the man and the woman conditional on receiving the same signal from each.<sup>28</sup>

My empirical results suggest that physicians may exhibit bias dynamically rather than through their priors. The surgeons I compare received the same number and fraction of a physician's referrals before an event. If referrals are a proxy of beliefs, the physicians should have the same priors about these surgeons, and the surgeons should be treated similarly after an event. Yet we see a male and a female surgeon for whom a physician holds the same prior being treated differently post-event.

Further, physicians' responses are strongest when they receive extreme or surprising signals. When the "placebo" surgeons produce the expected outcome, physicians do not treat men and women differently.

Several behavioral models exhibit agents who respond to signals in a biased way, either ignoring or misattributing signals depending on their priors.<sup>29</sup> Here, I discuss a simple example of physicians ignoring some signals, drawing on the psychology literature on attribution bias. I do not set up the model in full, but rather use features of attribution bias to show how it can lead to asymmetric updating.

#### Attribution Bias

Attribution theory studies how agents use and understand information to arrive at a conclusion, especially regarding an unexpected event (Fiske and Taylor, 1991). Typically, agents attribute information either to internal or external factors. Attribution bias occurs when there are systematic differences in how agents interpret information. For example, in the medical setting, physicians can attribute patient outcomes to the surgeon's ability (an internal factor) or to noise (an external

---

<sup>27</sup>For examples, see Coate and Loury (1993), Altonji and Pierret (2001), and Lange (2007)

<sup>28</sup>There are also cases where beliefs do not converge to the worker's true ability, specifically if employees alter their behavior. Coate and Loury (1993), for example, show that potential workers may under-invest in skills depending on the standards that employers set. Glover et al. (2017) provide experimental evidence that grocery store clerks do change their behavior based on the degree of bias that managers exhibit.

<sup>29</sup>Confirmation bias, for example, assumes that agents misread signals to fit with their prior (Rabin and Schrag, 1999; Fryer et al., 2017). It assumes, though, that an agent's bias depends on her belief about the individual. I find that physicians update different about two surgeons for whom they hold the same prior. Further, physicians should misread all signals to confirm their prior about an individual surgeon. However, I do not find evidence that physicians update after mundane events, such as general surgeries going as expected, as evidenced by the lack of change in referrals to placebo surgeons.

factor). A physician exhibits attribution bias if she systematically interprets unexpectedly good or bad events differently depending on the surgeon’s gender.

Under attribution bias, agents are unbiased when the actual outcome matches their expected outcome. However, if the outcome is far from an agent’s expected outcome, the agent needs to decide whether her beliefs were wrong or whether the outcome was an anomaly. In doing so, the agent, either consciously or unconsciously, relies on stereotypes or deep-seated biases. For example, physicians might generally treat male and female surgeons equally, but hold a deep-seated belief that men are better surgeons. When a physician receives a signal far from her prior, she rationalizes it using this stereotype. Here, I briefly show how attribution bias produces asymmetries in physicians’ responses to signals.

### Updating about the Individual

As in the Bayesian case, physicians have the prior probability distribution  $a_{i,t} \sim f(\bar{a}_i, \sigma_i^2)$  over a surgeon’s ability. For simplicity, I consider the case in which physicians only care about a surgeon’s mean ability. Physicians thus choose a surgeon  $i \in I_j$  to maximize patient utility:

$$\operatorname{argmax}_{i \in I_j} U_p(a) = \beta \mathbb{E}[a_{i,t} | \bar{a}_i, g, s] + \epsilon_{ipt} \quad (1.15)$$

All variables are defined as before but I consider the case in which the physician sees a continuous signal,  $s \sim f$ . Physicians have an expected outcome of each patient referral,  $\mathbb{E}[s|a_i]$ .

When an actual outcome matches the physician’s expected outcome, physicians update as Bayesians, relying on their prior about the individual surgeon to update. Specifically, the physician’s posterior when updating as a Bayesian, denoted  $f^b(a_i|s)$ , is

$$f^b(a_i|s) \propto f(s|a_i) \cdot f(a_i)$$

If a physician has the same belief about a man and a woman’s ability, the physician will update about them in the same way (e.g. form the same posterior about each) conditional on  $s$ . Empirically, this is what is seen in the “placebo” surgeon outcomes, where the surgical outcome is the expected outcome.

When the actual outcome does not match the expected outcome, physicians rationalize it, attributing the outcome to the surgeon’s ability or to noise. If the physician attributes the outcome to the surgeon’s ability, she updates as a Bayesian. If the physician attributes the outcome to noise, she throws away the signal and does not update. Assume that even if a physician believes that a given male and female surgeon are equally qualified, there is a societal stereotype that women make worse surgeons. As such, the physician attributes unexpectedly good signals ( $s > \mathbb{E}[s]$ ) to ability if the surgeon is male. If the surgeon is female, however, the physician attributes it to noise with some probability  $q$ . Similarly, the physician attributes unexpectedly bad signals ( $s < \mathbb{E}[s]$ ) to ability when the surgeon is female but attributes it to noise with probability  $q$  when the surgeon is male.

Compare this with a physician who exhibits attribution bias. Because the physician only accounts for a fraction  $q$  of a female surgeon's unexpectedly good outcomes and a fraction  $1 - q$  of a male surgeon's unexpectedly bad outcomes, her posterior will be distributed

$$f^c(a_i | s > \mathbb{E}[s]) \propto [1 - (1 - q) \cdot \mathbb{1}(g = w)] \cdot f^b(a_i | s) + (1 - q) \cdot \mathbb{1}(g = w) f(a_i) \quad (1.16)$$

$$f^c(a_i | s < \mathbb{E}[s]) \propto [1 - q \cdot \mathbb{1}(g = m)] \cdot f^b(a_i | s) + q \cdot \mathbb{1}(g = m) f(a_i) \quad (1.17)$$

after a good and bad signal respectively. In the above equations,  $\mathbb{1}(g = w)$  is an indicator taking the value one if the surgeon is female. Similarly,  $\mathbb{1}(g = m)$  takes the value one when the surgeon is male.

To help interpret these posteriors, consider a female surgeon. In equation 1.16,  $\mathbb{1}(g = w) = 1$  and in equation 1.17,  $\mathbb{1}(g = m) = 0$ . If a female surgeon sends a good signal, the physician updates as a Bayesian with probability  $q$  and does not update with probability  $1 - q$ . In the latter case, the physician's beliefs do not change from the prior period. The physician's posterior is thus a weighted average of the Bayesian posterior and the physician's prior. However, if a male surgeon sends a good signal, the physician updates as a Bayesian. After a bad signal we see the opposite: the physician updates as a Bayesian if the surgeon is female and has a posterior that is a weighted average if the surgeon is male, producing the asymmetry observed in the data.

### Updating about the Group

Attribution bias also helps explain the group-level results. Assume that physicians update iteratively about groups after seeing signals from different surgeons. Because women are underrepresented among surgeons, physicians will see fewer patient outcomes under women each quarter even if they are referring to them at their population share. If the variance of  $a_i$  is decreasing in the number of outcomes a physician sees, a common assumption, physicians' beliefs will be more diffuse for women than for men. That is,  $\sigma_{i,m} < \sigma_{i,w}$ , leading physicians to react more strongly to patient outcomes when the surgeon is female. However, if physicians exhibit attribution bias, the distribution about women's ability will shift more after bad outcomes than after good outcomes as physicians attribute female surgeons' good outcomes to noise with some probability. A physician's beliefs about a group can change, but it will take many more signals than it would if the physician was Bayesian.

## 1.9 Conclusion

Gender gaps in hiring, promotion, and pay persist in many industries. This paper identifies a mechanism that contributes to these gaps. Using referral volume to proxy for a physician's beliefs about a surgeon's ability, I show that physicians exhibit asymmetric updating, lowering their referrals more to women than to men after bad outcomes and increasing them more to men than to women after good outcomes. In addition, physicians use their experience with one woman

to infer the ability of other female surgeons. After a bad experience with one female surgeon, physicians become less likely to form new referral connections with women.

The results are consistent with Bayesian updating under specific assumptions on physicians' priors about surgeon ability. Physicians would have to believe that women are higher ability than men or that the variance of men and women's abilities shrinks differentially following signals. However the set of priors required do not match the actual distribution of surgeon ability. Therefore, although the results are reconcilable with Bayesian updating, they are inconsistent with physicians having rational expectations. Behavioral models in which physicians selectively ignore some signals, such as attribution bias, help to explain the empirical patterns.

Regardless of the model, the results have implications for how we think about employment decisions. Bayesian updating suggests that two individuals of equal ability and who perform equally well on a set of tasks will be evaluated in the same way. That is, an employer will hold the same posterior about each individual conditional on having the same prior and seeing the same signal from each. However, if employers exhibit asymmetric updating, two employees with the same objective performance could end up on different career tracks as the employer treats their performances differently. In this setting, women are punished for one mistake in the same way that men are punished for three mistakes, leading to lower skill accumulation and pay.

The implications of these results are especially important in occupations in which women are underrepresented. Because employers see relatively few signals from women, the performance of one woman influences the employers' beliefs about other women. If employers update asymmetrically, they might be too quick to let a woman go after a mistake and then be unwilling to hire more women in the future, preventing them from learning about the true distribution of women's abilities.

## Chapter 2

# Gender Differences in Recognition for Group Work

### 2.1 Introduction

How do employers infer workers' productivity under uncertainty? In many workplaces, employers must make hiring and promotion decisions under incomplete information or in the face of subjective productivity measures (Prendergast, 1999; MacLeod, 2003). Employees working in teams is an example of this. Unless employers can perfectly observe each worker's contribution to the team's output, employers must decide how to allocate credit without having full information as to what each worker did. This is of increasing relevance as many organizations rely predominantly on group work for production<sup>1</sup>. Yet there is little empirical evidence on how credit for group work is allocated and whether demographic characteristics such as gender play a role.

In this paper, I test whether uncertainty over an individual's contribution to a project leads to differential attribution of credit that contributes to the gender promotion gap. In many industries, women are not only hired at lower rates than men are, they are also promoted at lower rates. Blau and DeVaro (2007), for example, find that across jobs, women are less likely to be promoted than men conditional on worker's performance and ability ratings. In the UK, female managers are nearly 40% less likely to be promoted than male managers (Elmins et al. 2016). A significant portion of the promotion gap remains unexplained even after accounting for factors such as productivity, personality and behavioural differences (such as competition aversion), and fertility preferences<sup>2</sup>.

I specifically look at tenure decisions within academia, an ideal setting for two reasons. First,

---

<sup>1</sup>Lazear and Shaw (2007), for example, report that the share of large U.S. firms whose workers predominantly work in teams rose from 27% to 78% between 1987 and 1996

<sup>2</sup>There is a large literature documenting gender differences in productivity, attitudes toward different types of work, and family choices. See, for example, Niederle and Vesterlund (2007), Antecol et al. (2016), Ceci et al. (2014), and Ginther and Kahn (2004).

there is a large tenure gap between men and women, over 30% of which can not be explained by observable productivity differences or family commitments (Ginther and Kahn, 2004). Second, disciplines within academia differ in their treatment of authorship. This provides variation in the clarity of signals that employers receive. For example, in economics, coauthors are listed alphabetically, making it difficult to discern who did what on a coauthored paper. Solo-authored papers, on the other hand, provide a clear signal of the author's ability. In sociology, authors are listed in order of contribution, making both coauthored and solo-authored papers clear signals.

I primarily use data from economists' CVs to track individuals' career trajectories and compare whether the trajectory is different for individuals who coauthor versus solo-author, and whether there is a difference by gender. I then contrast economics with sociology to test whether there is a smaller unexplained promotion gap in academic disciplines where it is easier to discern an author's specific contribution.

Within economics, I find that men and women who solo-author most of their work have similar tenure rates conditional on a proxy for the quality of papers. However, an additional coauthored paper is correlated with an 8% increase in tenure probability for men but only a 2% increase for women. This gap is significantly less pronounced for women who coauthor with women, suggesting that the attribution of credit is related to the gender mix of coauthors. Furthermore, a man who coauthors is no less likely to receive tenure than a comparable man who solo-authors even though there is presumably more uncertainty as to how much work he did. In sociology, women who coauthor are as likely to receive tenure as men who coauthor, suggesting that it is the uncertainty contained in the alphabetical ordering of authors that disproportionately hurts women.

To ensure that I am not picking up on ability differences between men and women, I control for the quality of papers using both journal rankings and citations, allowing for a comparison of men and women with similar research portfolios. The results are also robust to including other individual-level controls such as length of time to tenure and the seniority of one's coauthors, as well as tenure year, tenure institution, and primary field fixed effects.

While I cannot pinpoint a specific mechanism that explains why coauthoring has lower returns for women, I rule out several standard explanations. The difference in credit does not appear to be due to women's own behaviour: they are not especially likely to coauthor with senior or high-ability faculty, nor do they present their coauthored work less. The empirical patterns are also inconsistent with taste-basted discrimination.

The remainder of the paper is organized as follows. Section 2.2 describes the data and shows that a tenure gap exists between male and female economists. In Section 2.3, I show that the tenure gap is driven by women having a lower probability of tenure for each additional coauthored paper than men. I then contrast economics with sociology to see how results change based on the quality and clarity of the signal. I show that the results are robust to using different journal rankings and definitions of tenure. In Section 2.4, I test four theories that might explain this relationship and argue that none can fully explain the observed empirical patterns. I look at how the relationship

between coauthoring and tenure has changed over time and by school in Section 2.5 and conclude in Section 2.6.

## 2.2 Data

To examine the relationship between paper composition and tenure, I construct a dataset using the CVs of economists who came up for tenure between 1985 and 2014 at one of the top 30 U.S. PhD-granting universities<sup>3</sup>. The academic progression documented in the CVs makes it possible to evaluate the relationship between the individual's research output and career progression. I can then compare the degree of collaborative work and reward for that work, and compare these results for men versus women.

### 2.2.1 Sample Selection and Data Overview

I include only PhD-granting institutions in the sample so that, as much as possible, I am comparing individuals facing similar tenure requirements. For example, liberal arts colleges place greater weight on teaching ability for tenure, something that I cannot measure. I exclude business and public policy schools for similar reasons. Many business schools require professors to write cases, for example, and it is not clear how this factors into tenure decisions. It is reasonable to assume that the top 30 economics departments in the U.S. emphasize research which is measured by the number and quality of papers one produces.

One problem in collecting tenure information is that the CVs of individuals who went up for tenure, were denied it, and left to industry or government are difficult to find, leading to a sample selection problem. To deal with this issue, I was able to collect historical faculty lists from 17 of the 30 schools and locate over 90% of faculty who had ever gone up for tenure at these 17 institutions. For the remaining 13 schools, I looked at the top 75 U.S. institutions, the top 5 Canadian institutions, and the top 5 European institutions to locate anyone who went up for tenure at a top 30 U.S. school and then moved to another institution. I also checked economists' CVs at the major Federal Reserve Boards and other large research institutes, such as Mathematica, in the U.S. While there might still be a sample selection problem, I show in Section 2.3.3 that the results are robust to using only the sample for which I have historical faculty lists.

From individuals' CVs, I code where and when they received their PhDs, their employment and publication history, and their primary and secondary fields. When looking at the relationship between publications and tenure in the main analysis, I only include papers that were published up to and including the year an individual goes up for tenure. Forthcoming papers and book chapters are not included in the paper count, but I later include such papers as a robustness check.

I primarily use the 2015 RePEc/IDEAS ranking of economics journals to control for the quality of a person's publications. For this, I take the top 85 journals and give the top journal a score of 86.

---

<sup>3</sup>The list of institutions included can be found in Appendix A

The lowest quality journal has a score of two. Any journals below this are given a score of one<sup>4</sup>. Finally, I include current citations of pre-tenure papers as a control variable. These citations were scraped from Google Scholar.

I supplement this dataset with results from a survey designed to measure individuals' beliefs about the returns to various types of papers. The survey also contains information on how frequently individuals present their papers. The exact questions and nature of the survey are discussed in greater detail in Section 2.4.

### 2.2.2 Construction of Tenure

To determine whether someone received tenure, I follow the guidelines on each school's website (as of 2015) as to when tenure decisions are made. The majority of schools require faculty to apply for tenure 7 years after their initial appointment. I therefore consider years 6-8 to be the "tenure window" in which someone applies for tenure to account for people who go up for tenure early or late (because of a leave of absence, for example). I assume that an individual is denied tenure if s/he moves to a university ranked 5 positions below the initial institution during the tenure window. Similarly, I assume that an individual is denied tenure if he moves from academia to industry during the tenure window. Defining tenure in this way accounts for the fact that some people switch institutions 2-3 years after their initial appointment, not because they were denied tenure but for personal preferences, and that some people might choose to move to a comparable school around the time of tenure even though they were offered tenure at their original institution. For example, someone who moves from MIT to Harvard after 7 years was presumably offered tenure at MIT but chose to move to Harvard for other reasons.

As mentioned, a person who moves 5 or fewer years after his or her initial appointment is not assumed to have been denied tenure since s/he moved before the tenure window starts. If someone moves before the tenure window, I use the second institution they were at to determine tenure. For example, if a person's first job is at University A but s/he moves to University B after three years, I use University B as the tenure institution but do not start the tenure clock over. In this example, I would look 3-5 years after the individual starts at University B (that is, 6-8 years after the initial position) to determine whether the individual received tenure. If a person moves from an academic institution into industry before the tenure window, I exclude them from the sample.

### 2.2.3 Summary Statistics

Table 2.1 presents summary statistics of the data. Approximately 70% of the full sample received tenure, but this masks a stark difference between men and women. Only 52% of women received tenure while 75% of men did.

---

<sup>4</sup>Less than 10% of papers are published in journals not included in the RePEc/IDEAS ranking.

**Table 2.1: Summary Statistics**

	Full	Male	Female	p-value
<i>Panel A:</i>				
Tenure	0.69 (0.46)	0.74 (0.44)	0.52 (0.50)	0.001
Years to tenure	6.7 (1.7)	6.6 (1.7)	7.2 (1.9)	0.001
Total papers	8.5 (4.0)	8.6 (4.1)	8.1 (3.4)	0.221
Solo-authored	3.1 (2.3)	3.1 (2.4)	3.1 (2.3)	0.912
Coauthored	5.4 (3.6)	5.5 (3.7)	5.1 (3.2)	0.159
<i>Panel B:</i>				
Top 5 Solo	0.68 (0.98)	0.67 (0.98)	0.70 (0.93)	0.811
Top 5 Coauthored	1.4 (1.5)	1.4 (1.5)	1.2 (1.4)	0.161
<i>Avg. Journal Rank:</i>				
All Pubs.	45.5 (17.2)	46.3 (17.2)	42.8 (17.3)	0.037
Solo Pubs.	467.2 (23.4)	47.8 (23.5)	45.5 (23.1)	0.348
Coauthored Pubs.	45.1 (20.2)	46.0 (20.3)	42.0 (19.5)	0.052
<i>Panel C</i>				
Total Unique CAs	5.0 (3.1)	5.0 (3.0)	4.9 (3.2)	0.607
Frac. Sr Coauthors	0.30 (0.32)	0.30 (0.30)	0.31 (0.32)	0.567
Frac. Fem Coauthors	0.14 (0.23)	0.11 (0.19)	0.27 (0.30)	0.001
Observations	597	459	138	

This table presents summary statistics for the full sample and separately for men and women. All paper count variables (*Total Papers*, *Solo-authored*, *Coauthored*, and *Top 5s*) are the number of papers an individual had published at the time of tenure.

*Total Papers*, *Solo-authored*, and *Coauthored* are the number of papers in each group that an individual had published by the time of tenure. These publication counts do not include books or book chapters. Papers published in non-economics journals (such as a political science journal) are included but receive a ranking of 1 (the lowest ranking). The results are robust to excluding publications in non-economics journals.

There is no statistically significant difference in the number of papers that men and women produce. Panel B looks at differences in the quality of papers. Men are no more likely to publish their papers in "Top 5" journals (American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, and The Review of Economic Studies) than women are. The only statistically significant productivity difference is that men tend to publish their coauthored papers in slightly higher-ranking journals. I therefore control for the quality of papers, measured as the average journal ranking and average citations, throughout the analysis.

Panel C displays differences in coauthoring patterns between men and women. *Total Unique CAs* is the number of unique coauthors an individual has had by tenure. Men and women have roughly the same number of coauthors and women are no more likely to coauthor with senior faculty than men are. The only significant difference in coauthorship is that 26% of women's coauthors are other women whereas only 11% of men's coauthors are women.

## 2.3 Empirical Strategy and Results

To understand how credit for group work is allocated, I correlate paper composition with tenure while controlling for individual-level characteristics as well as school, year, and primary field fixed effects. While an additional solo-authored paper is associated with the same increase in tenure probability for men and women, an additional coauthored paper is correlated with a larger increase in tenure probability for men than for women.

After establishing this result, I contrast economics with sociology, a discipline in which authors are listed according to contribution rather than alphabetically. In this case, an author's contribution is clear and there is no longer a gender difference in the relationship between coauthored work and tenure. Taken together, these results suggest that the uncertainty created by alphabetically ordering authors leads to women receiving less credit and contributes to the gender promotion gap.

### 2.3.1 Main Results

I show three main results. I first establish that a significant tenure gap exists between men and women. I then show that the gap becomes more pronounced the more women coauthor, and that women who solo-author all of their papers have comparable tenure rates to men. Finally, I show that the gender of a woman's coauthor matters. Women who coauthor with other women do not suffer a coauthor penalty, providing evidence that women receive less credit for joint work than men do, whether warranted or not.

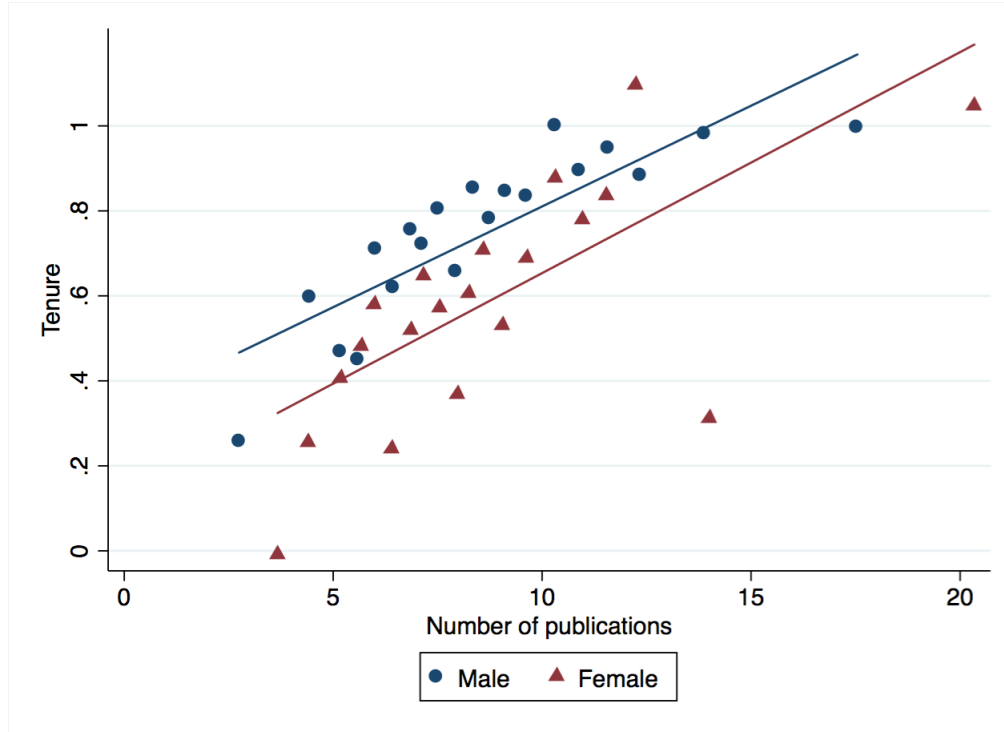
## The Tenure Gap

Figure 2.1 plots the lines of best fit from estimating

$$T_{ifst} = \beta_1 \text{TotPapers}_i + \gamma' Z_i + \theta_f + \theta_s + \theta_t + \epsilon_{ifst} \quad (2.1)$$

separately for men and women using OLS. The dependent variable,  $T_{ifst}$ , is an indicator that individual  $i$  in field  $f$  at school  $s$  receives tenure in year  $t$ .  $\text{TotPapers}_i$  is the number of papers (both coauthored and solo-authored) individual  $i$  has at the time he or she went up for tenure. The vector of individual-level controls,  $Z_i$ , includes average journal rank, total citations, the number of years it took  $i$  to go up for tenure, and the number of coauthors on papers. Tenure institution ( $\theta_s$ ), tenure year ( $\theta_t$ ), and field fixed effects ( $\theta_f$ ) are also included as tenure standards likely vary over time and by field and department.

Figure 2.1: Total Papers and Tenure



Notes: This figure is a binned scatterplot of the correlation between the total number of publications an individual has at the time they go up for tenure and the probability of receiving tenure. Both variables are residualized on the following controls before plotting: number of years it took to go up for tenure, average journal rank of pre-tenure publications, log citations, total coauthors, and tenure school, tenure year, and field fixed effects. The line of best fit using OLS is shown separately for men and women. The lines of best fit are estimated using the full sample (N=587) and have slopes of  $\beta = 0.050$  (s.e. = 0.014) for women and  $\beta = 0.047$  (s.e. = 0.005) for men. The y-variable is a binary variable indicating whether an individual received tenure. Each dot represents the mean of approximately 30 observations along both dimensions.

The figure shows that a significant tenure gap exists between men and women even after

controlling for productivity, primary field, tenure institution, and tenure year. While an additional paper is correlated with a roughly 4 percentage point increase in tenure probability for both men and women, women are consistently 17 percentage points less likely to receive tenure than men conditional on having written the same number of papers of similar quality. The corresponding estimates for equation 2.1 using a probit model are presented in Column 1 of Table 2.2.

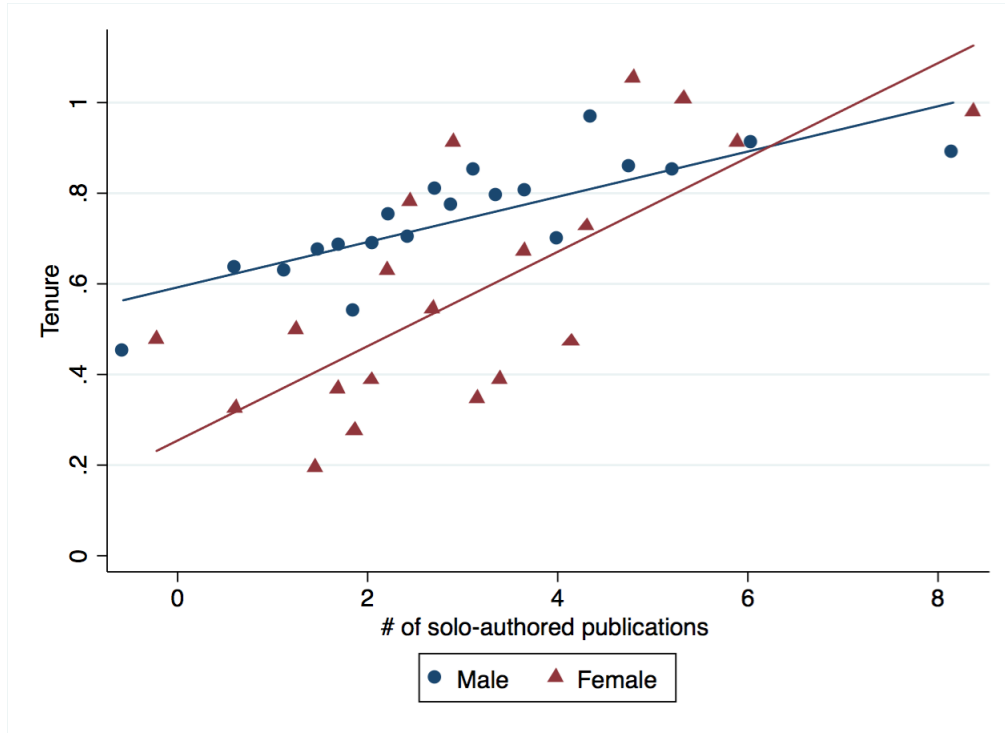
### The Tenure Gap and Paper Composition

If solo-authored papers provide a clear signal of ability, we would expect men and women to benefit similarly from an additional solo-authored paper, conditional on paper quality. Figure 2.2 plots the coefficients  $\hat{\beta}_1$  and  $\hat{\beta}_1 + \hat{\beta}_2$  from

$$T_{ifst} = \beta_1 S_i + \beta_2 (fem_i \times S_i) + \beta_3 CA_i + \beta_4 (fem_i \times CA_i) + \delta_1 fem_i + \gamma' Z_i + \theta_f + \theta_s + \theta_t + \epsilon_{ifst} \quad (2.2)$$

where  $S_i$  and  $CA_i$  are the number of solo-authored and coauthored papers an individual has at the time of tenure.

Figure 2.2: Solo Authored Papers and Tenure



Notes: This figure is a binned scatterplot of the correlation between the number of solo-authored publications an individual has at the time they go up for tenure and the probability of receiving tenure. Both variables are residualized on the same controls in Figure 2.1. The lines of best fit are estimated using the sample of individuals who have at least one solo-authored publication (N=534) and have slopes of  $\beta = 0.106$  (s.e. = 0.020) for women and  $\beta = 0.050$  (s.e. = 0.009) for men. Each dot represents the mean of approximately 27 observations.

**Table 2.2: Relationship Between Papers & Tenure**

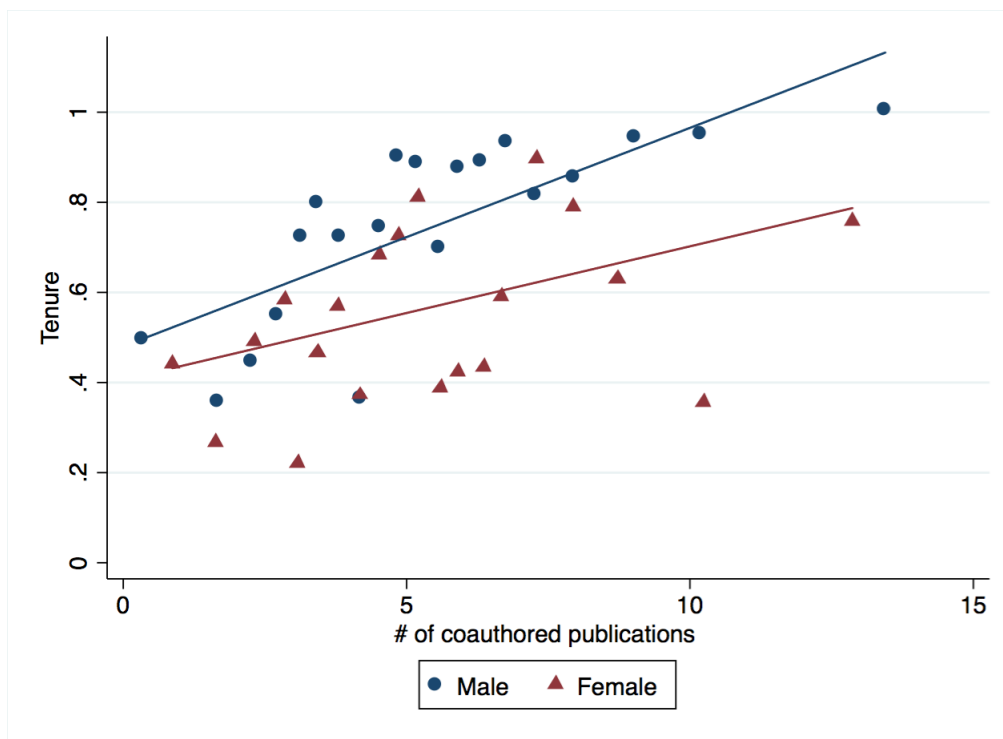
	Tenure		
	(1)	(2)	(3)
Total papers	0.115*** (0.015)		
Solo-authored		0.065** (0.027)	0.103** (0.046)
Fem x Solo		-0.012 (0.069)	-0.052 (0.081)
Coauthored		0.095*** (0.015)	0.117*** (0.032)
Fem x Coauthored		-0.075* (0.038)	-0.108* (0.064)
Total coauthors			0.014 (0.012)
Log Citations			
Avg. Journal Rank	1.340*** (0.341)		
Avg. Solo Rank		-0.055 (0.071)	0.100 (0.072)
Avg. Coauthored Rank		0.023*** (0.009)	0.037*** (0.011)
Female	-0.163*** (0.036)	0.159 (0.177)	0.279 (0.253)
School FE	Yes	No	Yes
Tenure Year FE	Yes	No	Yes
Field FE	Yes	No	Yes
Observations	572	495	463
Pseudo R-Sq.	0.393	0.275	0.406

This table shows the relationship between the number and types of papers an individual publishes and tenure. The dependent variable is a binary variable indicating whether the individual received tenure 6-7 years after being hired at the initial tenure institution. Total papers is the number of papers an individual had published by the time s/he went up for tenure. Solo-authored and Coauthored are the number of solo or coauthored papers s/he had published at the time of tenure. (XX) Avg. Journal Rank is the average journal rank of these pre-tenure publications, measured using the RePEc/IDEAS ranking. Tenure length is the number of years it took the individual to go up for tenure. Citations are from Google Scholar and measured at present. The equations are estimated using a probit model and the marginal probabilities calculated at the mean are displayed. All regressions include a quadratic term in the number of papers (coauthored and solo-authored) and number of papers interacted with female. Standard errors, reported in parentheses, are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

Here we see a large tenure gap between men and women who have few solo-authored papers. However, the gap converges as women write more solo-authored papers. It seems that the signal from the solo papers begins to outweigh the employer’s prior, which is consistent with a model in which employers start with lower beliefs about women and update as they receive clear signals about a woman’s ability.

If coauthored papers are an unclear signal of ability, an employer must make a judgment call as to how much each coauthor contributed to the paper which could lead to differential attribution of credit. Indeed, we see tenure rates diverging in Figure 2.3, which plots the relationship between an additional coauthored paper and tenure ( $\hat{\beta}_3$  and  $\hat{\beta}_3 + \hat{\beta}_4$  from equation 2.2). While an additional coauthored paper helps both men and women, men benefit much more than women, suggesting that coauthored work is typically attributed to men. Columns 2 and 3 of Table 2.2 show the corresponding coefficients for equation 2.2, estimated using a probit model. From these estimates, men’s tenure rates increase by 7.7 percentage points when they produce a coauthored paper whereas women’s increase by 2 percentage points.

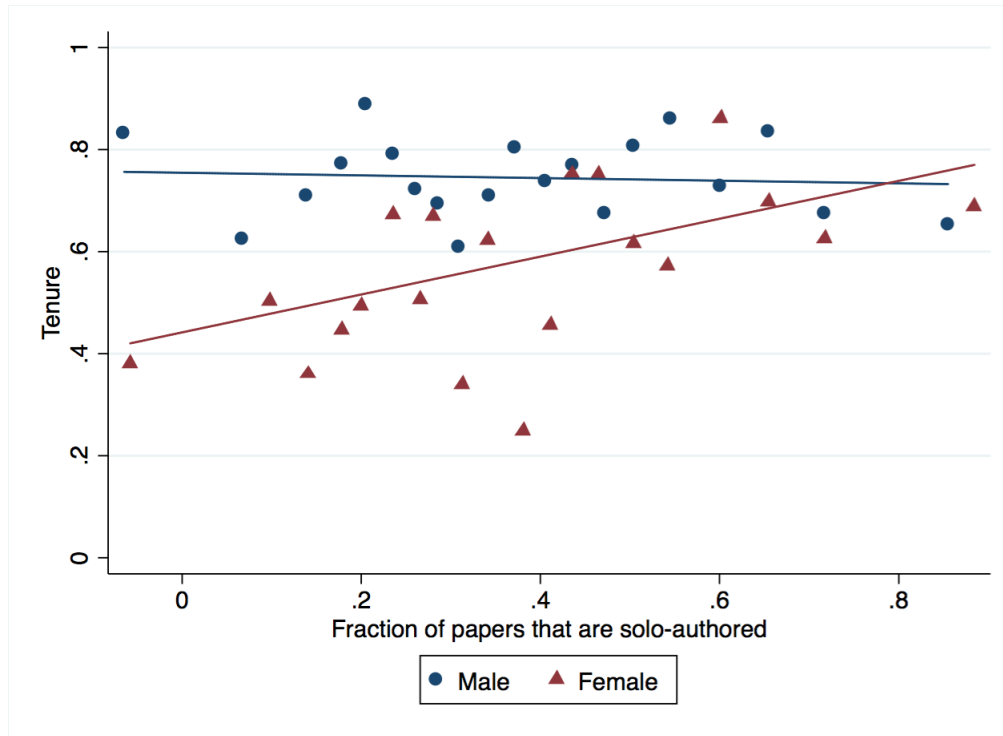
**Figure 2.3: Coauthored Papers and Tenure**



Notes: This is a binned scatterplot of the correlation between the number of coauthored publications an individual has at the time they go up for tenure and the probability of receiving tenure. Both variables are residualized on the same controls in Figure 2.1. The lines of best fit are estimated using the sample of individuals who have at least one solo-authored publication (N=529) and have slopes of  $\beta = 0.358$  (s.e. = 0.014) for women and  $\beta = 0.050$  (s.e. = 0.006) for men. Each dot represents the mean of approximately 26 observations.

The results are summarized in Figure 2.4.

Figure 2.4: Relationship Between Paper Composition and Tenure



Notes: This figure is a binned scatterplot of the correlation between tenure and the fraction of an individual's papers that are solo-authored, split by gender. Both variables are residualized on the same controls in Figure 2.1. The line of best fit using OLS is shown separately for men and women. The lines of best fit are estimated using the full sample (N=587) and have slopes of  $\beta = 0.036$  (s.e. = 0.167) for women and  $\beta = -0.028$  (s.e. = 0.078) for men. The y-variable is a binary variable indicating whether an individual received tenure. Each dot represents the mean of approximately 30 observations along both dimensions.

This figure plots the relationship between the fraction of an individual's papers that are solo-authored, controlling for the total number of papers, citations, journal quality, number of coauthors, and tenure institution, year, and field fixed effects. For men, it does not matter if one coauthors or solo-authors: tenure rates are comparable conditional on the quality of papers. Women who write all of their papers alone have similar tenure rates to men. However, women who coauthor all of their papers have an approximately 40% tenure rate, substantially lower than that of men who coauthor all of their papers (75%). The slope for women is 0.37 and is statistically significant at the 5% level (s.e.=0.167).

### Does Coauthor Gender Matter?

The probability of receiving tenure is not lower for all women who coauthor. In Table 2.3, I categorize coauthored papers into those written with only men, only women, or a mix of men and women:

$$\begin{aligned}
T_{ifst} = & \beta_1 S_i + \beta_2 (fem_i \times S_i) + \beta_3 CA_{male}_i + \beta_4 (fem \times CA_{male}_i) + \beta_5 CA_{mix}_i \\
& + \beta_6 (fem \times CA_{mix}_i) + \beta_7 CA_{fem}_i + \beta_8 (fem_i \times CA_{fem}_i) + \beta_9 fem_i \\
& + \gamma' Z_i + \theta_f + \theta_s + \theta_t + \epsilon_{ifst}
\end{aligned} \tag{2.3}$$

As before,  $S_i$  is the number of solo-authored papers individual  $i$  has.  $CA_{fem}_i$  is the number of coauthored papers individual  $i$  has in which all of the coauthors are female. Similarly,  $CA_{male}_i$  is the number of papers  $i$  has in which all of the coauthors are male and  $CA_{mix}_i$  is the number of papers  $i$  has in which the coauthors consist of men and women.

**Table 2.3: Coauthor Gender**

Dep Var: Tenure	(1)	
	Probit	Probit x Female
Solo-authored	0.064*** (0.009)	0.010 (0.014)
Pubs. with only Fem CAs	0.092*** (0.026)	0.001 (0.027)
Pubs. with only Male CAs	0.067*** (0.013)	-0.066*** (0.018)
Pubs. with M and F CAs	0.084** (0.031)	-0.048 (0.035)
Avg Journal Rank	0.004 (0.001)	
Total coauthors	-0.001 (0.009)	
Log Citations	-0.001 (0.018)	
Female	0.058 (0.096)	
Observations	548	

This table presents the results of one regression where the interaction terms are displayed in the right-hand column. *Pubs. with only Fem CAs* is the number of publications an individual has in which all coauthors are female. Similarly, *Pubs with only Male CAs* and *Pubs with M and F CAs* are the number of publications with only male coauthors and with a mix of male and female coauthors respectively. The equations is estimated using a probit model and the marginal probabilities calculated at the mean are displayed. Standard errors, reported in parentheses, are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

The estimated coefficients on the interaction terms show that women receive almost no marginal benefit from producing a coauthored paper with a man but receive some benefit when there is either a mix of male and female coauthors or only female coauthors. The point estimates suggest that an additional paper that a woman writes with a mix of men and women improves tenure

chances by 3.6% and an additional paper with other women improves tenure probability by 9.2%, the same increase a man receives. While the estimates are imprecise due to sample size, I can rule out that an additional coauthored paper with a woman has the same effect as an additional coauthored paper with a man. The estimates are robust to including all of the control variables discussed earlier.

These results imply that the differential attribution of credit arises when signals are noisy and that men tend to disproportionately benefit from this noise. If a woman coauthors with a woman, credit is given to both of them. However, when a woman coauthors with a man, the man receives the bulk of the credit, suggesting that employers rely on gender in some way to infer ability or effort. However, this does not mean that the employer is acting sub-optimally or is biased against women. I explore possible explanations in Section 2.4. First, though, I test whether removing the noise from a signal reduces the gender gap in credit.

### 2.3.2 Testing Against Other Coauthoring Conventions

If the uncertainty contained in a “coauthored” signal is contributing to the gender gap, clearly stating an individual’s role on a project would alleviate this problem. I test whether this is true by looking at sociology, where authors are listed by order of contribution. This removes uncertainty over each author’s contribution to the paper. Of course it should be noted that sociology is not the perfect comparison as many other factors differ across the two disciplines, including the fraction of women in each discipline. Still, the results shed light on whether the uncertainty contained in coauthored papers in economics contributes to a promotion gap.

The sociology sample consists of randomly sampled faculty at the top 20 sociology PhD-granting departments in the U.S.<sup>5</sup>. There are 250 sociologists in the sample, 40% of whom are female. Summary statistics are presented in Table 2.4. There is no statistically significant difference between men and women’s tenure rates (with the mean tenure rate being 76%) although men seem to publish more solo-authored articles than women.

To test whether men and women are treated differently, I reestimate equation 2.2 but include measures of the number of papers that researcher  $i$  is first author on. The results are presented in Table 2.5. I include the number and fraction of papers a researcher is first author on in Columns 1 and 2 respectively, along with female dummy interaction terms.

Being first author on a paper is correlated with a 5% increase in tenure probability for both men and women. Importantly, women are not penalized for coauthoring. Results are somewhat noisy due to the small sample, but the coefficient on the female/total coauthored papers interaction term is close to zero, indicating that women receive credit for their work when their contribution to a project is clear.

---

<sup>5</sup>Ranking from U.S. News Education

**Table 2.4: Sociology Summary Statistics**

	Men	Women	p-value
Tenure	0.752 (0.433)	0.776 (0.419)	0.547
Total papers	12.15 (7.808)	10.18 (5.726)	0.033
Total coauthored	6.409 (6.641)	5.959 (4.999)	0.567
Solo papers	5.745 (4.451)	4.224 (2.892)	0.003
Time to tenure	7.584 (1.607)	7.520 (1.724)	0.686
Books	0.779 (1.185)	0.571 (0.799)	0.139
Observations	150	100	

This table presents summary statistics for the full sample of sociologists and separately for men and women. All paper and book count variables (*Total Papers*, *Solo-authored*, *Coauthored*, and *Top 5s*) are the number of papers or books an individual had published at the time of tenure.

### 2.3.3 Robustness Checks

One may be concerned that the results are a product of the types of productivity measures used or of missing data. In this section, I show that the results are robust to using only the sample for which I have no missing observations, to using different journal rankings, and to accounting for papers published shortly after tenure.

#### Attrition

The results might be biased if the sample excludes individuals who are denied tenure and go into industry, government, or other institutions where I do not observe them. This would be particularly problematic if men who go to industry after being denied tenure disproportionately coauthored their papers. If this is true, I would be overestimating the benefit of coauthoring for men. I would have a similar problem if women who go to industry after being denied tenure typically wrote solo-authored papers.

As discussed in Section 2.2.1, I attempted to find such individuals by searching institutions outside of the top 30 U.S. schools, federal reserves, and other research institutes. Doing so certainly does not guarantee I found everyone who went up for tenure, though. To allay concerns about sample selection, I run the analysis on the sample for which I received historical faculty lists. These lists allow me to track who went up for tenure and find them even if they left academia. The results, presented in Column 1 of Table 2.6, do not change when run on the sample for which there should be very few missing observations.

**Table 2.5: Sociology: Papers and Tenure**

Dep Var: Tenure	Probit (1)	Probit (2)
Total first author	0.050** (0.017)	
Fem x First Author	0.026 (0.040)	
Fraction first author		0.403*** (0.043)
Fem x Frac. First Author		-0.042 (0.172)
Solo papers	0.008 (0.006)	0.000 (0.006)
Fem x Total Solo	0.002 (0.011)	0.007 (0.011)
Total Coauthored	-0.010* (0.004)	0.009 (0.007)
Fem x Total CA	-0.020 (0.017)	0.001 (0.015)
Books	0.063* (0.032)	0.058 (0.035)
Book chapters	0.007 (0.013)	0.005 (0.012)
Female	0.026 (0.114)	0.010 (0.163)
School FE	Yes	Yes
Tenure Year FE	Yes	Yes
Observations	237	209

This table shows the relationship between the number and types of papers an individual publishes and tenure for a sample of sociologists. The dependent variable is a binary variable indicating whether the individual received tenure 6-7 years after being hired at the initial tenure institution. *Total first author* is the number of papers an individual is first author on while *Fraction first author* is the fraction of an individual's papers that s/he was first author on. The equations are estimated using a probit model and the marginal probabilities calculated at the mean are displayed. Standard errors, reported in parentheses, are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

**Table 2.6: Robustness Checks**

	Attrition	Journal Rankings		Publication Count	
	Fac. List Sample	AER Equiv.	Over Time	Tenure +1	Tenure +2
	(1)	(2)	(3)	(4)	(5)
Solo-authored	0.058*** (0.011)	0.057*** (0.010)	0.072*** (0.010)	0.070*** (0.010)	0.069*** (0.011)
Coauthored	0.056** (0.017)	0.059*** (0.013)	0.077*** (0.014)	0.063*** (0.010)	0.053*** (0.008)
Fem x Coauthored	-0.047*** (0.010)	-0.060*** (0.013)	-0.055*** (0.014)	-0.052*** (0.013)	-0.045** (0.014)
Fem x Solo	0.019 (0.015)	0.015 (0.013)	0.011 (0.015)	0.009 (0.014)	0.016 (0.014)
Years to tenure	-0.019* (0.008)	-0.053*** (0.012)	-0.053*** (0.010)	-0.050*** (0.011)	-0.050*** (0.011)
Total coauthors	0.005 (0.011)	0.009 (0.010)	0.002 (0.011)	0.011 (0.009)	0.016* (0.008)
Log Citations	-0.012 (0.015)	0.004 (0.019)	0.007 (0.021)	-0.004 (0.021)	-0.006 (0.022)
Avg Journal Rank	0.002* (0.001)		0.005** (0.001)	0.004** (0.001)	0.004** (0.001)
AER Equiv. Solo		0.028 (0.063)			
AER Equiv. CA		0.298** (0.111)			
Female	0.058 (0.067)	0.088 (0.104)	0.056 (0.100)	0.087 (0.106)	0.040 (0.122)
School FE	Yes	Yes	Yes	Yes	Yes
Tenure Year FE	Yes	Yes	Yes	Yes	Yes
Field FE	Yes	Yes	Yes	Yes	Yes
Observations	278	467	551	549	543

The dependent variable is an indicator for receiving tenure. Column 1 restricts the sample to those schools I received a historical faculty list from. Column 2 uses the number of AER-equivalents as individual has as the paper quality measure. This is broken up into the number of solo-authored AER-equivalents and coauthored AER-equivalents. Column 3 uses historical journal rankings to allow for rankings to change over time and to account for new journals entering. In Columns 4 and 5, I include papers that were published one and two years after an individual went up for tenure in the paper counts. Standard errors are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

## Journal Rankings

While the economics profession largely agrees on what the “top” journals are, rankings of field journals or lower-tier journals have changed over time and might be disputed. Furthermore, including an average journal ranking could mask differences in the rankings of solo and coauthored papers between men and women. In Columns 2 and 3 of Table 2.6, I attempt to account for this by using two alternative journal ranking metrics.

In Column 2, I separately include the average ranking of one’s solo and coauthored papers. In addition, I convert each journal ranking into its “AER equivalent” where each raw publication is converted into its number of American Economic Review-equivalent papers. This has the advantage of standardizing the journal quality measure, and for allowing different distances between paper ranks. For example, in the RePEc/IDEAS ranking, *Econometrica* is one ranking below the AER. Using the AER-equivalent measure, a paper in *Econometrica* is nearly equivalent to having a paper in the AER. A paper in many of the top field journals is equivalent to having published 0.15-0.2 papers in the AER. For details on the methodology and the ranking, see Kalaitzidakis, Mamuneas, and Stengos (2003). Again, the results are unchanged.

In Column 3, I allow journal rankings to change over time. I use historical rankings of economics journals (drawn from Laband and Piette, 1994, and combined with current rankings) and match each paper a person publishes with its journal ranking at the time it was published. In Column 3, the variable *Avg Journal Rank* is then the average rank of an individual’s papers, measured using the historical rankings. Using these rankings accounts for journals moving in rank over time as well as new journals being added. Again, the results do not qualitatively change. An additional coauthored paper is associated with a 7.7 percentage point increase in tenure probability for men but only a 2.2 percentage point increase for women. In section 2.4, I also separate papers into “Top 5s” and “non-Top 5s”.

## Tenure Definition

In the main analysis, I only consider papers that were published up to and including the year that an individual goes up for tenure. If an individual goes up for tenure in 1995, for example, papers published in 1996 are not included in the paper count even though they may have been “revise and resubmits” at the time of tenure. This could affect the results if men who coauthor have several promising unpublished papers at the time of tenure but women who coauthor do not, in which case I am not actually comparing people with similar publication records. In Columns 4 and 5 of Table 2.6, I include papers that are published one and two years after a person’s tenure year in the paper count variables. The results do not change: women continue to benefit less from coauthored papers than men do.

## 2.4 Channels

There are many possible explanations for the above findings, not all of which can be tested with these particular data. Here I shed light on four standard and testable channels: ability-based sorting, preference-based sorting, women not claiming credit for their work, and taste-based discrimination. The empirical patterns are inconsistent with all of the proposed explanations, most of which would have suggested that a woman's own behaviour leads her to receive less credit for joint work.

### 2.4.1 Ability-Based Sorting

Employers might rationally deny women who coauthor tenure if individuals sort such that only lower ability women coauthor with men. This could arise for several reasons. For example, if coauthoring lowers the cost of producing a paper, but women know that they receive less credit for papers, high ability women might forego the cost savings and choose to work alone. They know they can produce high quality papers by themselves and send the employer a clearer signal of their ability. However, if low ability women can only produce high quality papers with the help of a high ability man, they might coauthor even if they receive less credit. High ability men will agree to coauthor with them if it reduces the cost of the paper without reducing the quality. Employers would then know that any woman coauthoring with a man is lower ability.

A simple sample of worker sorting and promotion is presented in Appendix B. The framework provides intuition for when women of different ability types would choose to coauthor with men. If women do not know that they receive less credit for coauthoring, we should see assortative matching. Men and women of the same ability level will collaborate, making the decision to deny coauthoring women tenure sub-optimal. If women do know that they receive less credit, high ability women will opt to work alone while, under certain conditions, low ability women will coauthor. In this case, it is optimal for an employer to tenure coauthoring women at lower rates than men.

In what follows, I test whether women anticipate receiving less credit and consequently sort accordingly. To do so, I first present survey evidence suggesting that women do not know that the returns to coauthoring are lower than solo-authoring. I then show that women do receive some credit for papers that publish well, suggesting that employers believe that there might be some assortative matching. Finally, I provide evidence that even when women tend to work with men who are slightly higher ability than themselves this unequal match does not explain the gender gap in tenure.

### Survey Evidence on Knowledge of Returns to Coauthoring

If women know that their returns to coauthoring with men are low, it is plausible that high ability women would choose to solo-author or only work with other women. Here I test whether women anticipate receiving less credit for collaborative work. However, it is first worth noting that the

results presented in Section 2.3.1 show that women have a nearly 0% return to coauthoring with a man. If women know the true returns to coauthoring, *any* woman, regardless of ability, should be hesitant to collaborate (see Appendix B for details). The fact that we see women coauthoring with men already implies that women misjudge the true returns to coauthoring.

I test whether women know the true returns to coauthoring using a survey conducted with economists currently working at the top 35 U.S. economics departments. The survey was sent to all professors, regardless of rank, at these institutions and received an 32% response rate. The gender composition of the sample is representative of the profession today, with 89 respondents being female and 300 being male. In the survey, economists were asked the following question:

*Suppose a solo-authored AER increases your chance of receiving tenure by 15%. For each of the following, please give an estimate of how much you think the described paper would increase your chance of receiving tenure.*

Respondents then go through five types of papers (coauthored AER, coauthored AER with senior faculty, coauthored AER with junior faculty, solo-authored top field, and coauthored top field) and record their beliefs about the returns to these papers<sup>6</sup>.

In Table 2.7, I test the difference in the mean beliefs of men and women<sup>7</sup>.

**Table 2.7: Survey Results**

	(1) Men	(2) Women	(3) p-value
<i>Panel A: Beliefs about Returns to Papers</i>			
Coauthored AER	12.1	12.2	0.939
Coauthored AER, Sr. Faculty	9.1	8.8	0.528
Coauthored AER, Jr. Faculty	13.3	13.4	0.796
Solo Top Field	8.0	8.2	0.669
Coauthored Top Field	6.3	6.8	0.223
<i>Panel B: Frequency of Presenting Papers</i>			
Times Presented	3.1	2.2	0.07
Present More Freq. than CA	0.37	0.44	0.20
Observations	300	89	

This table presents the mean responses for men and women to the following survey questions: Panel A: "Suppose a solo authored AER increases your chance of receiving tenure by 15 percent. By how much do you think each of the following increases your change of receiving tenure?" Panel B: "How many times per year do you typically present your solo-authored papers? Are you more or less likely than your coauthors to present a joint paper?" The survey was conducted with a sample of academic economists currently working at a top 35 U.S. economics department.

<sup>6</sup>I did not ask respondents about paper coauthored with men/women so that they would not be primed to think about gender

<sup>7</sup>Because the survey was anonymous, the answers can not be linked to the CV data. I can therefore only test for differences in means without controls.

There is no statistically significant difference in the beliefs of men and women for any type of paper. Men believe that a coauthored AER will increase their chance of receiving tenure by 12.1%, and women by 12.2%. Women believe that there are slightly lower returns to AER papers coauthored with senior faculty (8.8% versus 9.1% for men), but the difference is not statistically significant. These results suggest that, in this context, women are unaware of the true returns to coauthoring.

### Evidence on Sorting by Ability

A second test of whether women know that they will receive less credit for papers and sort accordingly is to look at the correlation between propensity to coauthor and ability. I first test whether high ability women are less likely to coauthor than low ability women and then test for assortative matching among coauthors. I proxy for ability using the quality of journal that an individual's job market paper was published in. I assume that the job market paper is the first solo-authored paper an individual publishes after he or she graduates.

If women anticipate discrimination, ability and the fraction of one's papers that are coauthored will be negatively correlated. High ability women should be less likely to coauthor. In Figure 2.5.A I plot the coefficients  $\hat{\beta}_1$  and  $\hat{\beta}_2$  from estimating

$$FracCA_{ifst} = \beta_1 a_i + \beta_2 (fem_i \times a_i) + \beta_3 fem_i + \beta_4 TotPapers_i + \theta_f + \theta_s + \theta_t + \epsilon_{ifst} \quad (2.4)$$

where  $FracCA_{ifst}$  is the fraction of person  $i$ 's papers that are coauthored and  $a_i$  is person  $i$ 's ability (job market paper rank). If high ability women anticipate receiving less credit, we expect  $\hat{\beta}_2 < 0$ . In Figure 2.5.A, however, we see that ability is uncorrelated with the fraction of papers that are coauthored for both men and women: both estimates are precise zeros. There is no evidence that women along the ability distribution act strategically in their choice to coauthor versus solo author.

I also find no evidence that high ability women strategically coauthor with other women rather than men. Figure 2.5.B plots the results from equation 2.4 using the fraction of papers that are coauthored with women as the dependent variable. Women are more likely to coauthor with other women than men are but there is no sorting by ability.

While women do not seem to be sorting according to ability, it is possible that women tend to work with higher-ability coauthors who then receive more credit for a paper. I can test for this by correlating a person's ability with that of his or her coauthors. While I do not have the job market paper information for all coauthors in the dataset, I can see where the coauthors were working at the time the individual went up for tenure. As a measure of average coauthor ability, I take the average school rank of all of an individual's pre-tenure coauthors. For example, if  $i$  coauthors with  $j$  and  $k$  and  $j$  works at the 5th-ranked institution and  $k$  works at the 15th-ranked institution, the average ability of  $i$ 's coauthors is 10. I correlate  $i$ 's ability with the average ability of her coauthors in Figure 2.6. The line of best fit is plotted controlling for number of coauthored and solo-authored publications, time until tenure, and field, institution, and tenure year fixed effects.

Men and women both sort positively on ability but women are more likely to collaborate with

individuals at more highly-ranked institutions than men are. To see whether this can explain the results, I estimate

$$\begin{aligned}
T_{ifst} = & \beta_1 S_i + \beta_2 (fem_i \times S_i) + \beta_3 CA_i + \beta_4 (fem_i \times CA_i) + \beta_5 rank_{ij} \\
& + \beta_6 (CA_i \times rank_{ij}) + \beta_7 (fem_i \times CA_i \times rank_{ij}) + \beta_8 (fem_i \times rank_{ij}) \\
& + \beta_9 fem_i + \gamma' Z_i + \theta_f + \theta_s + \theta_t + \epsilon_{ifst}
\end{aligned} \tag{2.5}$$

where  $rank_{ij}$  is the average institution rank of  $i$ 's coauthors and all other variables are defined as before. The results are reported in Table 2.8. If men receive more credit because they are coauthoring with lower ability women,  $\hat{\beta}_7$  should be negative. However,  $\hat{\beta}_7$  is close to zero, indicating that the ability of one's coauthor is not driving the tenure gap for coauthoring women.

**Table 2.8: Accounting for Sorting on Ability**

Dep Var: Tenure	(1)	(2)
		x Avg Coauthor Rank
Solo-authored	0.081*** (0.015)	
Fem x Solo	0.012 (0.019)	
Coauthored	0.093*** (0.017)	-0.0003 (0.0002)
Fem x Coauthored	-0.053* (0.021)	-0.0003 (0.0002)
Female	0.108 (0.128)	
Avg Journal Rank	0.004** (0.001)	
Tenure Length	-0.056*** (0.012)	
Log Citations	-0.031 (0.024)	
Constant		0.001 (0.001)
School FE	Yes	
Tenure Year FE	Yes	
Field FE	Yes	
Observations	415	

This table presents the results of one regression where the coefficients on the interaction terms (with *Avg. Coauthor Rank*) are displayed in Column 2. *Avg Coauthor Rank* is calculated by taking the mean of the school rank that an individual's coauthors are at when they wrote their joint paper. For example, if a person has two coauthors, one who is at the 5th-ranked school and one who is at the 10th-ranked school, *Avg Coauthor Rank* would be 7.5. Individuals who have no coauthored papers are not included in the sample. Standard errors are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

## Returns to Top Papers

For high ability women to receive no credit for their coauthored papers, employers would have to believe that there is no assortative matching by ability. Otherwise, employers would receive a signal that women who coauthor with high ability men are also high ability, and be more likely to promote them (see Appendix B for details). Figure 2.6 shows that assortative matching does occur, but it is possible that employers do not recognize this. I test for this by looking at how credit for top 5 publications is allocated. If employers know that there is assortative matching, they should believe that women coauthoring with high-ability men are also likely to be high ability.

Table 2.9 shows the results from estimating

$$\begin{aligned} T_{ifst} = & \beta_1 TopS_i + \beta_2 (fem_i \times TopS_i) + \beta_3 TopCA_i + \beta_4 (fem_i \times TopCA_i) + \beta_5 NonTopS_i \\ & + \beta_6 NonTopCA_i + \beta_7 (fem_i \times NonTopS_i) + \beta_8 (fem_i \times NonTopCA_i) + \beta_9 fem_i + \gamma' Z_i \\ & + \theta_f + \theta_s + \theta_t + \epsilon_{ifst} \end{aligned} \quad (2.6)$$

The female interaction terms are presented in the second column.  $TopS_i$  and  $TopCA_i$  are the number of solo and coauthored papers that individual  $i$  has published in a top 5 journal. Similarly,  $NonTopS_i$  and  $NonTopCA_i$  are the number of solo and coauthored papers the individual has published in non-top 5 journals.

Note that power becomes an issue as (1) there are relatively few people publishing in the top 5 journals, and (2) cutting by gender means that there are even fewer women in each category. Still, Table 2.9 shows that coauthored papers published in a top 5 journal help women but still by less than they help men. The point estimate on the interaction term is negative and insignificant due to large standard errors, but it suggests that a coauthored publication in a top journal is associated with a 10% increase in tenure probability for women and a 15% increase for men. Non-top 5 coauthored papers do not have any positive influence on women's tenure probability. It seems that employers receive some signal when a woman publishes her coauthored papers in top journals which is at odds with the hypothesis that only low ability women coauthor with men.

Overall, there is little evidence that ability-based sorting is driving the results. If anything, employers seem to recognize that high ability men and women might work together and are therefore more likely to grant these women tenure. However, their tenure rate is still lower than that of high ability men.

### 2.4.2 Preference-Based Sorting

If women prefer to coauthor with senior faculty, we could reasonably expect that women would have lower tenure rates. Assuming senior faculty are more likely to be credited for a paper, the fact that most senior faculty are men would drive the correlation between coauthoring with a man and tenure. That is, women receive less credit because they enjoy coauthoring with senior faculty and these senior faculty are predominantly male.

To test whether women are more likely to coauthor with senior faculty and whether this can

**Table 2.9: Paper Split by Top 5**

Dep Var: Tenure	(1)	
	Top 5	Non-Top 5
Solo	0.135*** (0.027)	0.057*** (0.011)
Coauthored	0.127*** (0.025)	0.055*** (0.013)
Female x Solo	-0.033 (0.039)	0.028 (0.023)
Female x Coauthored	-0.036 (0.032)	-0.063*** (0.015)
Female	0.047 (0.111)	
Total coauthors	0.008 (0.010)	
Log Citations	-0.010 (0.017)	
Years to tenure	-0.052*** (0.010)	
School FE	Yes	
Tenure Year FE	Yes	
Field FE	Yes	
Observations	535	

This table presents the results of one regression where Column 1 shows the coefficients on the number of top 5 solo and coauthored papers and Column 2 shows the coefficients on the number of non-top 5 solo and coauthored papers. Top 5 papers are those published in the American Economic Review, Econometrica, the Journal of Political Economy, Quarterly Journal of Economics, or the Review of Economic Studies. Standard errors are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

explain the results, I first estimate

$$MaleSr_{ifst} = \beta_1 fem_i + \beta_2 TotCA_i + \beta_3 CA_i + \beta_4 S_i + \gamma' Z_i + \theta_f + \theta_s + \theta_t + \epsilon_{ifst} \quad (2.7)$$

where  $MaleSr_{ifst}$  is either the number of fraction of male senior coauthors an individual has. The independent variables are defined as before with  $TotCA_i$  being the number of coauthors an individual has worked with by the time s/he goes up for tenure. The results are presented in Column 1 of Table 2.10. Women are not more likely to coauthor with senior men: the point estimate on  $fem_i$  in Column 1 is positive but small and insignificant.

**Table 2.10: Coauthor Seniority**

Dep. Variable:	Frac. Sr Coauthors	Tenure
	(1) OLS	(2) Probit
Female	0.043 (0.040)	0.112 (0.124)
Coauthored	0.019 (0.014)	0.078*** (0.015)
Solo-authored	-0.023** (0.008)	0.070*** (0.010)
Fem x Solo		0.005 (0.018)
Fem x Coauthored		-0.059*** (0.014)
Frac. Sr Coauthors		-0.131 (0.072)
Fem x Frac. Sr Coauthors		0.040 (0.102)
Total coauthors	-0.014 (0.011)	0.005 (0.012)
Years to tenure	0.005 (0.009)	-0.038*** (0.010)
Avg Journal Rank	0.001 (0.001)	0.005*** (0.001)
Log Citations	-0.016 (0.013)	-0.010 (0.022)
School FE	Yes	Yes
Tenure Year FE	Yes	Yes
Field FE	Yes	Yes
Observations	507	531

The dependent variable in Column 1 is the fraction of an individual's coauthors that are senior (full professors). The dependent variable in Column 2 is the binary tenure outcome. Column 1 is estimated using OLS and Column 2 uses probit. Standard errors are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

As an additional check, I reestimate equation 2.2 but control for the fraction of a person's coauthors who are senior. The results are presented in Column 2 of Table 2.10. The seniority

of women's coauthors does not explain the results. Controlling for coauthor composition, an additional coauthored paper increases a man's probability of tenure by 8 percentage points but a woman's by only 2 percentage points.

### 2.4.3 Women Not Claiming Credit for Papers

Women might be given less credit for their work if they are less likely to claim it as their own. For example, if women present less frequently than men, people might associate a paper with the male coauthor who presents it more. The survey discussed in Section 2.4.1 also asked individuals how many times per year they present their work and whether they are more or less likely to present their coauthored papers than their coauthor. Panel B of Table 2.7 shows that women do not report presenting their coauthored papers less frequently than their coauthors. Interestingly, though, women present their solo-authored papers fewer times per year than men do. It is possible that women do not "advertise" their work as much as men do and this leads to women receiving less recognition for their work in general. If this were true, though, women who solo author should also be less likely to receive tenure.

### 2.4.4 Taste-Based Discrimination

If some employers have a distaste for tenuring women, as in Becker (1971), we should see women who write solo-authored papers being denied tenure as well. If employers cannot plausibly deny a woman who solo-authored several well-published papers, however, they might be constrained to deny tenure only to those for whom they can make a reasonable case. If it can be argued that a woman who coauthors did little of the work, taste-based discrimination could help to explain the results as employers have an excuse for denying tenure to coauthoring women. However, as shown in Table 2.3, only women who coauthor with men have lower tenure rates. This would imply that employers have a particular distaste for tenuring women who coauthor with men, which seems unlikely.

## 2.5 Further Discussion: Are Things Improving and Where?

As more women enter the economics profession, we would hope that any bias against women would begin to dissipate. As there are more examples of women doing high quality research, people might be less prone to innate biases. In Table 2.11, I reestimate equation 2.2 but interact the publication count variables with a linear time trend for the year an individual is expected to go up for tenure. The interaction coefficients are presented in column 2.

Women continue to receive less credit for coauthored work throughout the time period but the interaction on  $Fem_i \times Coauthored_i \times TimeTrend$  is positive although insignificant. This provides some evidence that as more women enter the profession, they may be starting to receive more credit for joint work. Interestingly, the coefficient on the interaction term on  $Coauthored_i \times TimeTrend$  is

**Table 2.11: Results Over Time**

	(1)	(2)
		x Time Trend
Solo-authored	0.099** (0.037)	-0.001 (0.001)
Fem x Solo	0.066 (0.066)	-0.001 (0.002)
Coauthored	0.130*** (0.032)	-0.002 (0.001)
Fem x Coauthored	-0.118* (0.051)	0.002 (0.001)
Female	-0.003 (0.100)	
Total Coauthors	0.001 (0.009)	
Log Citations	0.017 (0.023)	
Constant		0.009 (0.006)
Observations	553	

This table shows the results from one regression where solo and coauthored papers are interacted with a linear time trend in the year an individual is expected to go up for tenure. The coefficients in Column 2 show are the interactions between paper types, author gender, and the time trend. Standard errors are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

**Table 2.12: Interaction w/ School Rank**

	(1) Probit x School Rank	
Solo-authored	0.091*	-0.0002
	(0.042)	(0.001)
Coauthored	0.172***	-0.001*
	(0.041)	(0.001)
Fem x Solo	0.070	-0.001
	(0.063)	(0.001)
Fem x Coauthored	-0.154**	0.001*
	(0.048)	(0.001)
Avg Journal Rank	0.003**	
	(0.001)	
Female	0.112	
	(0.106)	
Years to tenure	-0.048***	
	(0.010)	
Total coauthors	0.004	
	(0.010)	
Log Citations	0.001	
	(0.022)	
Constant	-0.001	
	(0.003)	
Observations	523	

This table shows the results from one regression where solo and coauthored papers are interacted with the ranking of an individual's tenure institution. *SchoolRank* is defined such that 30 is the highest-ranked school and 1 is the lowest-ranked. The coefficients in Column 2 show are the interactions between paper types, author gender, and school rank. Standard errors are clustered by tenure institution. (\*=p<0.10, \*\*=p<0.05, \*\*\*=p<0.01)

negative (but again insignificant), meaning that men might begin to suffer a coauthor penalty in later years.

An additional question is whether some institutions are better than others. For example, women in the top schools might be viewed more favourably if people think that anyone who makes it to a top school must be good. On the other hand, if people think that there is affirmative action for women, especially at the top schools, women in such institutions might be viewed less favourably.

In Table 2.12, I report the results from estimating equation 2.2 interacting all paper count and gender variables with an individual's tenure institution ranking. I have inverted the rankings so that 30 is the highest rank and 1 the lowest.

The interaction terms in column 2 show that women in higher-ranked schools receive slightly more credit for their coauthored papers while men receive slightly less credit. For women,

moving up one spot in rankings improves the correlation between coauthoring and tenure by 0.001 percentage points. Moving from the lowest to the highest-ranked school is not enough to eliminate the negative correlation between coauthoring and tenure for women, but it does reduce the gap. These results could suggest that women at high-ranking schools are viewed more favourably and given more credit for joint work. Of course, I can not rule out the possibility that there is less gender bias at high-ranking schools, etc.

The results also show that men at lower-ranking schools benefit immensely from coauthoring: an additional coauthored paper is correlated with a 17.2 percentage point increase in tenure. This is in line with Ginther and Kahn's (2004) finding that men have many more publications, often coauthored, in lower ranked journals that they receive credit for. These publications are unlikely to count for much at the top schools.

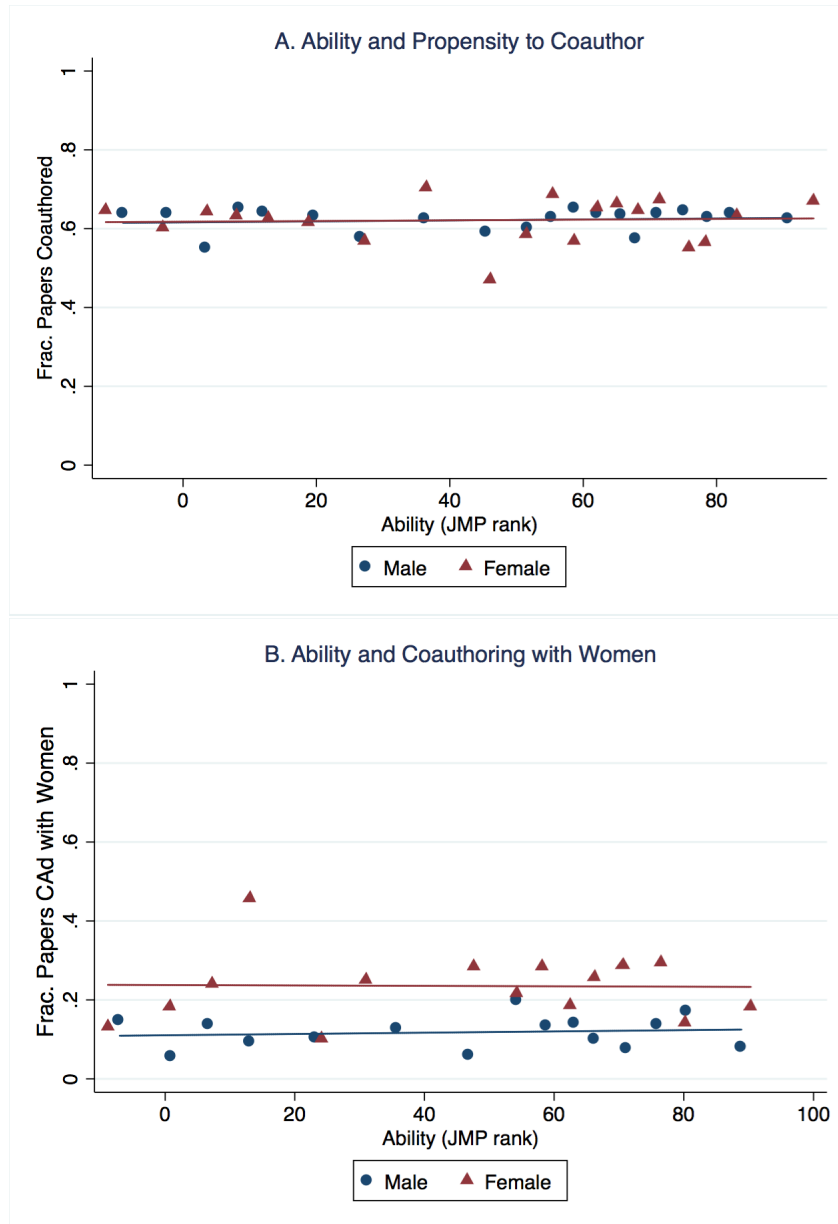
## 2.6 Conclusion

Women receive tenure at significantly lower rates than men in many academic fields. As discussed in the introduction, this phenomenon is not exclusive to academia. Several explanations have been put forward for the gap, but it persists even after accounting for observable characteristics such as fertility preferences and productivity.

This paper proposes an alternative explanation. I argue that women receive less credit for group work when employers can not perfectly observe their contribution. When signals are noisy, employers have to infer each worker's ability or productivity. Coauthored papers provide employers with a noisy signal. The fact that women who work specifically with men receive tenure at lower rates than comparable women who work alone or with other women suggests that gender enters into the employer's inference process. However, when employers receive clear signals, men and women are treated similarly. For example, men and women receive the same amount of credit for solo-authored papers, which provide a clear signal of ability. Furthermore, when the uncertainty in a coauthored paper is resolved, as in sociology, women and men again receive the same amount of credit for joint work. I show that these results are not explained by sorting or women presenting their work less. I also argue that it is not due to blatant taste-based discrimination.

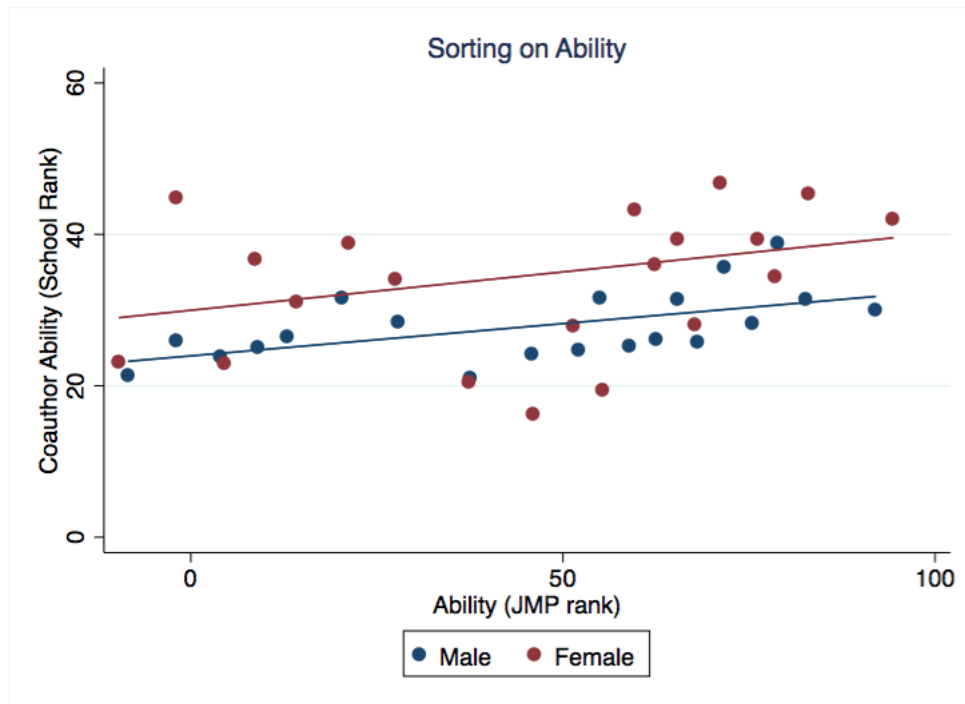
While a specific channel cannot be established with these data, being aware of this phenomenon is important in a world that is increasingly relying on group work for production. The tech industry, for example, prides itself on collaboration. In such male-dominated fields, however, group work could result in fewer women moving up the career ladder if credit is not properly attributed. Further work is needed to determine why women seem to receive less credit for group work.

Figure 2.5: Ability and Sorting



Notes: This binned scatterplot shows the correlation between an individual’s ability and the propensity to coauthor (Fig. 5A) and the propensity to coauthor with women (Fig. 5B). I proxy for an individual’s ability using the journal in which his or her job market paper is published in. Both variables are residualized on the following controls before plotting: total solo and coauthored papers, the number of years it took to go up for tenure, log citations, and tenure school, tenure year, and field fixed effects. The lines of best fit using OLS are shown separately for men and women. The estimates for Fig. 5A are  $\beta = 0.00002$  (s.e. = 0.0003) for women and  $\beta = -0.0001$  (s.e. = 0.0002) for men. The estimates for Fig. 5B are  $\beta = -0.00005$  (s.e. = 0.0008) for women and  $\beta = 0.00016$  (s.e. = 0.0003) for men.

Figure 2.6: Assortative Matching



Notes: This binned scatterplot shows the correlation between an individual's ability, proxied by the journal in which their job market paper is published in, and their coauthor's ability, proxied by the average school rank of their coauthors. Both variables are residualized on the following controls before plotting: total solo and coauthored papers, the number of years it took to go up for tenure, log citations, and tenure school, tenure year, and field fixed effects. The line of best fit using OLS is shown separately for men and women. The lines of best fit are estimated on the full sample and have slopes of  $\beta = -0.00005$  (s.e. = 0.0008) for women and  $\beta = 0.0002$  (s.e. = 0.0003) for men.

# Chapter 3

## Confidence Men<sup>1</sup>

### 3.1 Introduction

Gender gaps in labour market outcomes have remained large despite continued efforts to promote equality. While the sources of these gaps have been traditionally ascribed to differences in human capital accumulation and discrimination, a growing body of literature has attributed the gender gap to psychological factors, in particular the role of confidence (Bertrand 2011). From a young age, women appear less confident than men. This confidence gap has been argued to play a key role in explaining differences in academic success, occupational choices, and career progression. Samek (2015), for example, uses data on college graduates to show that women are less likely to apply for jobs with competitive compensation schemes. This boxes them out of many high-paying careers, such as those in finance, that have been traditionally dominated by men.

While confidence is often treated as a pre-determined personality trait, it can in itself be a result of labour market discrimination: Are women less confident because they are discriminated against on the labour market? Or do they progress less in their careers because they feel less confident? Making progress towards understanding the nature of the confidence gap is an important step to removing remaining barriers to gender equality.

This paper asks whether a confidence gap exists between men and women who have made it to the very top of their careers. While the existing literature explains why we see fewer women in upper-level management and STEM positions (Bertrand et al. 2010; Goldin 2014), few have looked at how the confidence gap changes when women break the glass ceiling. It could be that women who rise to the top of their profession do so because they are confident, meaning that the confidence gap should disappear. Similarly, rising through the ranks could heighten a woman's sense of confidence and beliefs about her ability.

Using data from a select group of economists working in top U.S. universities, we find that women are still less confident than men along two margins. First, when asked about their level of agreement on survey questions about the economy, women are 7.6% points less likely to give

---

<sup>1</sup>Coauthored with Guo Xu

“extreme” answers in which they strongly agree or disagree. Second, women are less confident in the accuracy of their answer. Women express a level of confidence that is 0.340 points lower than a comparable man, as measured on a scale of 1 (unconfident) to 10 (very confident). The results persist after controlling for the year the PhD was granted, the PhD awarding institution, the current institution, and the number of solo and co-authored publications up to the point of tenure. We also provide suggestive evidence that the confidence gap is largely driven by women being less confident when asked questions that are outside their field of expertise.

Overall, the paper provides evidence of a confidence gap that persists even among the most successful academics and distinguishes between two types of confidence along which men and women differ. The results hence complement existing experimental evidence on the role of confidence in economic choices. Mobius et al. (2015) show in a lab experiment that individuals update their beliefs in a biased way when they receive signals about their ability. Women, however, are more conservative than men when updating their beliefs in the face of positive signals. This leads high ability women to be less confident than high ability men which could explain why we see fewer women in high ranking positions. Niederle and Vesterlund (2007) also conduct a lab experiment to show that men are twice as likely to choose competitive over piece-rate (non-competitive) compensation than women. They argue that these preferences are explained by gender differences in preferences for competition and by gender differences in confidence, with men being overconfident.

Our finding that the confidence gap is driven by women being less confident about questions outside their field of expertise is closely aligned with two papers: Coffman (2014) and Lundeberg et al. (1994). Coffman (2014) finds in a lab experiment that women are hesitant to contribute their ideas to a group. She has students answer questions about a series of subjects, ranging from pop culture to science, and then assigns students to groups. Students can choose to contribute their answer to the group’s answer. She finds that even when women are the group’s expert on a subject, they are unwilling to share their ideas. This is especially true when questions are stereotyped as male (e.g. math and science). Similarly, men are less likely to contribute ideas when tasks are stereotypically female.

Lundeberg et al. (1994) present undergraduates with a series of exam questions and ask them to state their confidence in their answer. They find that both men and women are overconfident in their answers but, in line with our findings, men are on average more confident. While we cannot test whether men are wrongly confident as the questions asked in our survey do not have definitive answers, Lundeberg et al. find that men are more confident even when their answer is wrong.

The remainder of this paper is organized as follows. Section 2 describes the data used in our analysis. In Section 3, we present the main results and explore mechanisms that might explain the findings. Section 4 concludes.

## 3.2 Data

Our main data source is the Initiative of Global Markets (IGM) survey. The survey asks a select group of 51 economists, all at top U.S. institutions, questions related to the current state of the economy as well as other policy issues. The economists are chosen to represent a range of political views, ages, and research interests. All are reported to have a “keen interest in public policy”.<sup>2</sup> The set of economists answering questions only changes in that new members have been added to the panel over time.

A total of 166 questions<sup>3</sup> were asked between September 2011 and May 2015, leaving us with 7,026 responses. The questions cover a range of economic questions, from education to the minimum wage to monetary policy. Answers are given privately but displayed publicly on the IGM website.

Respondents answer each question along two dimensions. First, they are asked the degree to which they agree with a given statement (e.g. “Some Americans who work in the production of competing goods, such as clothing and furniture, are made worse off by trade with China.”), measured on a Likert scale of 1 (strongly disagree) to 5 (strongly agree). Second, they are asked how much confidence they have in their answer which is measured on a scale of 1 (lowest) to 10 (highest).

Each response captures a different aspect of confidence. The first captures whether respondents are overly confident in their level of dis/agreement with the statement, known as overestimation in the psychology literature (Moore and Healy 2008). The second elicits confidence in the accuracy of an individual’s response, known as overprecision. We look at whether the confidence gap is primarily a result in differences in extremeness of opinions or confidence in views. As shown in Figures ?? and ??, women are less likely to hold strong views than men and are more dispersed in their self-reported confidence.

We combine the IGM data with pre-tenure characteristics for the 51 economists. This is the period all respondents have already completed and for which more comparable data is obtainable from the public CVs. These data include each economist’s primary field of specialization, the year the PhD was received, the number of published papers broken down into solo and coauthored papers, and whether the individual received tenure at his or her initial placement school. These data are collected and coded as described in Sarsons (2015). Due to missing values, the final sample consists of 47 economists.

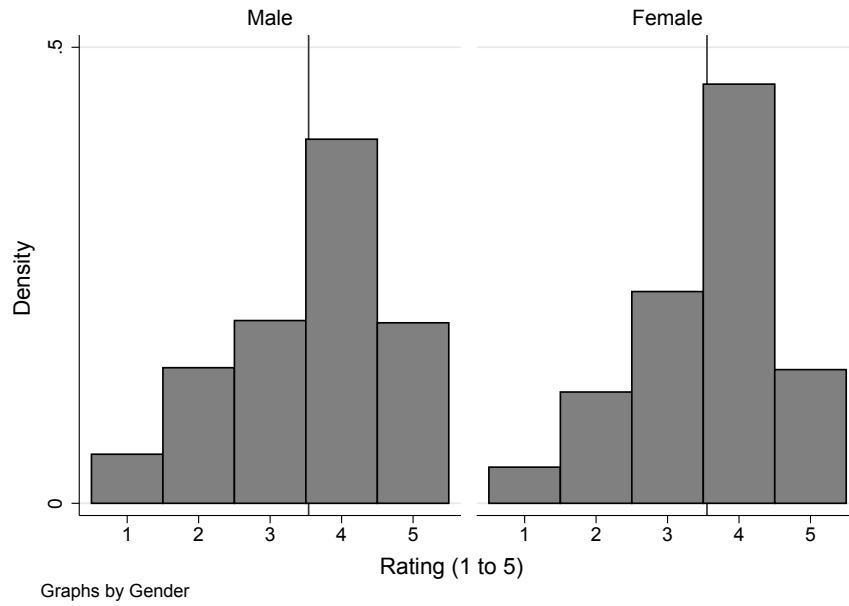
Table 3.1 summarizes the differences among male and female economists of the IGM panel. Given the highly selected nature of a sample of top economists, male and female economists are, in terms of their average characteristics, comparable (Panel A). However, given the small sample

---

<sup>2</sup>Refer to <http://www.igmchicago.org/igm-economic-experts-panel/>, retrieved 11 June 2015 and Gordon and Dahl (2013) for a description and alternative use of the data.

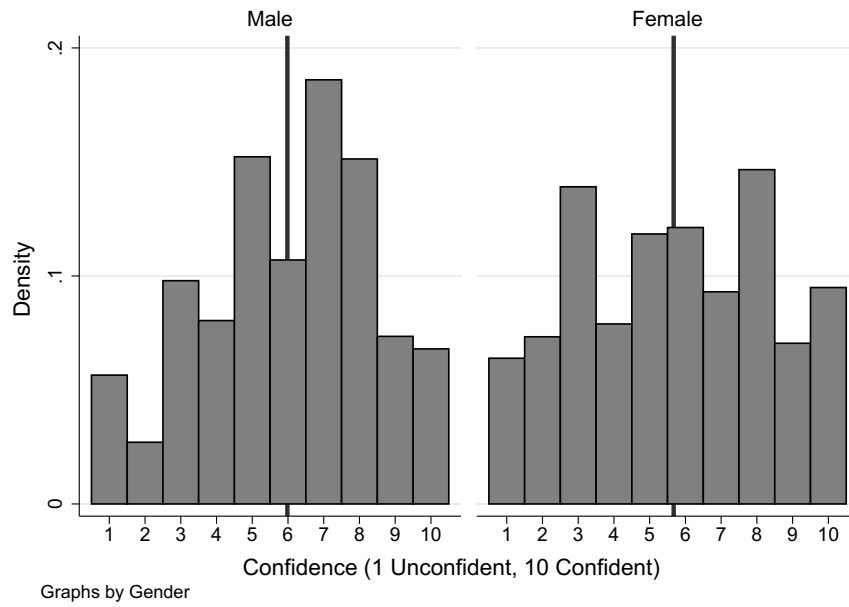
<sup>3</sup>The full list of questions can be viewed at [http://www.igmchicago.org/igm-economic-experts-panel/poll-results?SurveyID=SV\\_72JJHkpH4FvJb9j](http://www.igmchicago.org/igm-economic-experts-panel/poll-results?SurveyID=SV_72JJHkpH4FvJb9j).

**Figure 3.1: Distribution of Responses (Likert scale)**



Note: N=5241. Vertical line marks the mean.

**Figure 3.2: Distribution of Confidence**



Note: N=5241. Vertical line marks the mean

size, some differences may be insignificant due to lack of power. Women appear, for example, to be younger than their male counterparts, as indicated by the later average year of PhD award, although the difference is statistically insignificant. The share of PhDs awarded from Harvard and MIT are comparable at roughly 60%. While there are differences in fields, these are, with the exception of International/Trade, insignificant. The number of publications is comparable across gender.

**Table 3.1: Comparison of background characteristics and confidence by gender**

	(1)	(2)	(3)	(4)	(5)	(6)
	Male		Female		p-value	
	Mean	SD	Mean	SD	Diff mean	KS-test
<b>Panel A</b>						
Individual characteristics						
Year PhD award	1985.6	10.222	1990.6	2.887	0.174	0.135
Harvard/MIT PhD	0.567	0.502	0.600	0.516	0.857	1.000
Field: IO	0.027	0.164	0.100	0.316	0.320	1.000
Field: Labour	0.108	0.314	0.300	0.483	0.136	0.882
Field: Dev/Hist/Pol.	0.162	0.373	0	0	0.180	0.969
Field: Behavioural/Exp.	0.027	0.164	0	0	0.608	1.000
Field: Public/Health/Env	0.189	0.397	0.400	0.516	0.169	0.797
Field: Finance	0.081	0.276	0	0	0.362	1.000
Field: International	0	0	0.100	0.316	0.053	1.000
Field: Macro	0.216	0.417	0.100	0.316	0.418	1.000
Field: Microtheory	0.162	0.373	0	0	0.180	0.969
Field: Econometrics	0.027	0.164	0	0	0.608	1.000
Solo pubs. until tenure	5.540	3.870	4.600	2.319	0.469	0.448
Co-authored pubs. until tenure	6.405	4.312	6.100	1.462	0.840	0.958
Tenured at first job	0.972	0.164	0.700	0.152	0.005	0.474
Observations	37		10			
<b>Panel B</b>						
Voting behaviour						
Vote (Likert 1-5)	3.612	1.116	3.606	0.982	0.873	0.002
Extreme answers (1 or 5)	0.268	0.443	0.187	0.390	0.001	0.001
Confidence (1-10)	5.983	2.396	5.671	2.681	0.001	0.001
Observations	4177		1064			

Column (5) shows the p-value of the simple t-test for equality of means between male and female. Column (6) shows the p-value for the Kolmogorov-Smirnov test for equality of distributions. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

Even with the highly selected sample, we observe stark differences in confidence (Panel B). While the average level of confidence is the same, the distributions for men and women are very different. Men hold more extreme views, as measured by the share of extreme answers, and believe these views to be more accurate, as measured by the self-reported confidence level. These differences are significant at the 1% level.

## 3.3 Results

### 3.3.1 Main Results

To test whether women are less confident than men, we estimate:

$$y_{ijs} = \beta_1 fem_i + x_i' \gamma + \theta_j + \theta_s + \epsilon_{ijs} \quad (3.1)$$

where  $fem_i$  is a dummy indicating whether respondent  $i$  is female,  $x_i$  is a vector of individual-level control discussed in Section 2, and  $\theta_j$  and  $\theta_s$  are question and school fixed effects. The question fixed effect is included to confine the identifying variation to within-question comparisons. As a conservative specification and to ensure the results are not driven by differences across schools, we also include current school fixed effects to compare gender differences only among academics at the same school.<sup>4</sup>

As discussed in the data section, we measure confidence,  $y_{ijs}$ , in two ways: (i) the propensity to provide extreme judgements, as measured by a dummy for whether the respondent strongly agreed or disagreed, and (ii) the self-reported confidence level on an integer scale of 1 to 10. If a confidence gap exists, with women being less confident than men, we expect  $\beta_1$  to be negative. We cluster the standard errors on the question-level.<sup>5</sup>

The results are reported in Tables 3.2 and Table 3.3 and confirm that a gap exists for both measures of confidence. Women are less confident in both the level and precision of their answers (Column 1).

This gap is not driven by gender-specific differences across questions or schools, as it is robust to question and school fixed effects (Column 2-3). The inclusion of pre-tenure individual controls does not substantially affect the gap (Column 4).

In terms of magnitude, the confidence gap is economically large. Women are 7.6% points less likely to provide extreme judgements (Table 3.2, Column 4).<sup>6</sup> Compared to the mean of the dependent variable (25.2%), this corresponds to a gap of 30%. The gap is somewhat smaller for the confidence level (Table 3.3). On average, women tend to report a confidence score that is 0.340 point lower than men. This corresponds to a gap of 6% when evaluated against the mean. Interestingly, those who were awarded PhDs from Harvard or MIT are 0.604 points more confident than respondents who received their PhD elsewhere (Table 3.3, Column 4). The gender confidence gap is about half of the size.<sup>7</sup>

---

<sup>4</sup>The same university may both have a business school and an economics department. We distinguish between both institutions within the same university by including separate school fixed effects.

<sup>5</sup>We also cluster at the individual level. However, this leaves us with only 46 clusters which greatly reduces the significance of our results. More data is therefore needed to determine whether the results are robust to individual-level clustering.

<sup>6</sup>This finding is consistent with Mondak and Anderson (2004), who document that women are more likely to report “I don’t know” in surveys of political knowledge than men.

<sup>7</sup>The reader can therefore expect the confidence gap between the co-authors of this paper to be more than closed

**Table 3.2: Propensity to Provide Extreme Judgements and Gender**

	(1)	(2)	(3)	(4)
	Dependent variable: Strongly agree or disagree			
Mean of dep. var.	0.252	0.252	0.252	0.252
Female	-0.082*** (0.01)	-0.084*** (0.01)	-0.092*** (0.01)	-0.076*** (0.01)
Year PhD award				-0.002*** (0.00)
PhD Harvard/MIT				0.001 (0.01)
Total solo pubs.				-0.001 (0.00)
Total co-authored pubs.				-0.006*** (0.00)
Question FE		X	X	X
School FE			X	X
Observations	5,241	5,241	5,241	5,241

The unit of observation is the economist-question pair. The dependent variable is a dummy that is 1 if the respondent replied either strongly disagree or strongly agree to a question. *Female* is a dummy that is 1 if the respondent is female. *Year of PhD award* is the year the respondent was awarded the PhD. *PhD Harvard/MIT* is a dummy that is 1 if the respondent was awarded a PhD from either Harvard or MIT. *Total solo pubs.* is the number of total single authored publication. *Total co-authored pubs.* is the total number of co-authored publications. Robust standard errors in parentheses, clustered at the question level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

**Table 3.3: Self-reported Confidence Level and Gender**

	(1)	(2)	(3)	(4)
	Dependent variable: Level of confidence (1-10)			
Mean of dep. var.	5.920	5.920	5.920	5.920
Female	-0.082*** (0.01)	-0.330*** (0.09)	-0.399*** (0.09)	-0.340*** (0.09)
Year PhD award				-0.011*** (0.00)
PhD Harvard/MIT				0.604*** (0.08)
Total solo pubs				0.019* (0.01)
Total co-authored pubs				-0.038*** (0.01)
Question FE		X	X	X
School FE			X	X
Observations	5,241	5,241	5,241	5,241

The unit of observation is the economist-question pair. The dependent variable is the self-reported measure of confidence vis-a-vis a given question's reply (1 lowest, 10 highest on integer scale). *Female* is a dummy that is 1 if the respondent is female. *Year of PhD award* is the year the respondent was awarded the PhD. *PhD Harvard/MIT* is a dummy that is 1 if the respondent was awarded a PhD from either Harvard or MIT. *Total solo pubs.* is the number of total single authored publication. *Total co-authored pubs.* is the total number of co-authored publications. Robust standard errors in parentheses, clustered at the question level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

### 3.3.2 Mechanisms

The persistence of a confidence gap among highly selected economists is striking. Having confirmed that a confidence gap exists, we now attempt to disentangle mechanisms that could be driving the results. In particular, we explore the role of differential confidence in answering questions outside one’s primary field of expertise in driving the gender gap. We explore this mechanism by exploiting variation in the questions asked. Since respondents are asked about a variety of questions related to many aspects of economics, some of the questions fall outside their field of expertise. We estimate the following equation to test whether women’s lack of confidence appears when they are asked questions about topics outside of their fields:

$$y_{ijs} = \beta_1 fem_i + \beta_2 foreign_{ij} + \beta_3 (fem_i \times foreign_{ij}) + x_i' \gamma + \theta_j + \theta_s + \epsilon_{ijs} \quad (3.2)$$

Here,  $foreign_{ij}$  equals one when the question  $j$  is outside of respondent  $i$ ’s primary field and all other variables are defined as in Section 3.1.

The results are presented in Table 3.4 for both measures of confidence. In Column 1 and 3, we report the results of Section 3.1 with an added dummy for whether a question lies outside the respondent’s own field. Respondents are less confident when answering questions outside their fields. In particular, respondents are 6.1% points less likely to provide an extreme judgement (Column 1) and are 0.857 points less confident (Column 3) when answering a foreign field question compared to a question in their own field.

In Column 2 and 4 we report the results when  $foreign_{ij}$  is interacted with  $fem_i$ . The results show that women’s confidence falls more when being asked questions outside of their field. Men are also less confident when being asked questions outside of their field but women suffer from an additional lack of confidence. This result is statistically insignificant for the propensity to provide extreme judgements (Column 2, with  $p = 0.124$  for the interaction) and significant on the 10% level for the self-reported level of confidence (Column 4). More importantly, accounting for the differential confidence when moving beyond one’s own field “explains away” the level effect of gender.

These findings are somewhat in line with Coffman (2014)’s work showing that both men and women are less likely to voice their opinion while working on group projects that are outside of their expertise. However, Coffman finds that even when women find out that they are the “expert” on a topic, they are still unwilling to contribute their ideas. We do not find this to be the case: women who are asked a question related to their area of expertise are not significantly less confident than men are. It could be, then, that successful women are not underconfident but rather are more aware of the bounds of their expertise. Given the data constraints, however, we are unable to firmly corroborate this interpretation.

---

upon completion of their PhDs.

**Table 3.4: Differential Confidence in Answering Foreign Field Questions and Gender**

	(1)	(2)	(3)	(4)
	Extreme answer		Confidence	
Mean of dep. var.	0.252	0.252	5.920	5.920
Female	-0.077*** (0.01)	-0.031 (0.03)	-0.354*** (0.09)	-0.054 (0.18)
Year PhD award	-0.002*** (0.00)	-0.002*** (0.00)	-0.011*** (0.00)	-0.011*** (0.00)
PhD Harvard/MIT	-0.004 (0.01)	-0.003 (0.01)	0.533*** (0.08)	0.537*** (0.08)
Total solo pubs	-0.001 (0.00)	-0.001 (0.00)	0.022** (0.01)	0.022** (0.01)
Total co-authored pubs	-0.005*** (0.00)	-0.005*** (0.00)	-0.036*** (0.01)	-0.036*** (0.01)
Foreign field question	-0.061*** (0.02)	-0.047** (0.02)	-0.857*** (0.09)	-0.767*** (0.11)
Female × Foreign field		-0.058 (0.04)		-0.374* (0.20)
Question FE	X	X	X	X
School FE	X	X	X	X
Observations	5,241	5,241	5,241	5,241

The unit of observation is the economist-question pair. For Column 1-2 the dependent variable is a dummy that is 1 if the respondent replied either strongly disagree or strongly agree to a question. For column 3-4 the dependent variable is the self-reported measure of confidence vis-a-vis a given question's reply (1 lowest, 10 highest on integer scale). *Female* is a dummy that is 1 if the respondent is female. *Year of PhD award* is the year the respondent was awarded the PhD. *PhD Harvard/MIT* is a dummy that is 1 if the respondent was awarded a PhD from either Harvard or MIT. *Total solo pubs.* is the number of total single authored publication. *Total co-authored pubs.* is the total number of co-authored publications. Robust standard errors in parentheses, clustered at the question level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

## Breadth of expertise

If women actually have a narrower range of expertise, we would expect them to be less confident than men when answering questions outside of their field. For example, women might choose to specialize in topics related to a single field while men work on topics in a range of areas. Our data only distinguishes two primary fields but it is possible that economists work in several fields during their careers. If men do work in more fields than women, the confidence gap is justified.

In Table 3.5 we control for an economist's breadth of expertise using data from RePEc.<sup>8</sup> RePEc creates a measure of "breadth" using the number of distinct fields in which there are one or more papers citing an economist's work. This proxy should be closely related to an individual's breadth of expertise since being cited in a given field suggests that the author has worked in that field or does work closely related to that field. We estimate equation (3.3) below and present the results in Table 3.5:

$$y_{ijs} = \beta_1 fem_i + \beta_2 foreign_{ij} + \beta_3 (fem_i \times foreign_{ij}) + \beta_4 breadth_i + \beta_5 (fem_i \times breadth_i) + x'_i \gamma + \theta_j + \theta_s + \epsilon_{ijs}. \quad (3.3)$$

Correlating breadth and gender reveals that women do have a narrower breadth of expertise<sup>9</sup> but this does not affect our results. The gender gap in confidence remains significant: conditional on an individual's breadth of expertise, women are still less confident in their views than men are.

## Are women responding to disagreement?

Some of the questions on the IGM panel are more controversial than others. Opinion differs widely for such questions while there is much more consensus for other questions. For example, all economists agree that the benefits of free trade and NAFTA outweigh the costs. There is much disagreement, though, as to whether using tax incentives to affect a firm's location choice is beneficial. In Table 3.6 we test whether women are less confident on questions that have a greater spread of answers.

Specifically, we estimate

$$y_{ijs} = \beta_1 SDrating_j + \beta_2 fem_i + \beta_3 (fem_i \times SDrating_j) + x'_i \gamma + \theta_s + \epsilon_{ijs} \quad (3.4)$$

where  $SDrating_j$  is the standard deviation in responses for question  $j$ . Column 3 includes question fixed effects which absorb the  $SDrating$  term. The results show that while both men and women are less confident when responding to questions on which there is less consensus, a marginal increase in the dispersion of answers matters less for women's confidence than it does for men's.

Figure 3.3 plots the regression lines from equation (3.4). Here we see that there is a confidence gap when there is more of a consensus on a question with men being significantly more confident

---

<sup>8</sup>See <https://ideas.repec.org/top/top.person.nepcites.html> for more details.

<sup>9</sup>Difference in means between male - female is 4.128 (standard errors: 2.284 in a two-sided t-test).

**Table 3.5: Confidence in Answering Foreign Field Questions**

	(1)	(2)	(3)	(4)
	Extreme answer		Confidence	
Mean of dep. var.	0.252	0.252	5.920	5.920
Female	-0.019 (0.04)	-0.017 (0.04)	0.208 (0.20)	0.207 (0.20)
Year PhD award	-0.002** (0.00)	-0.002*** (0.00)	-0.006* (0.00)	-0.006* (0.00)
PhD Harvard/MIT	0.002 (0.01)	0.001 (0.01)	0.481*** (0.08)	0.481*** (0.08)
Total solo pubs	-0.001 (0.00)	-0.001 (0.00)	0.032*** (0.01)	0.032*** (0.01)
Total co-authored pubs	-0.005*** (0.00)	-0.005*** (0.00)	-0.032*** (0.01)	-0.032*** (0.01)
Foreign field question	-0.048** (0.02)	0.186 (0.18)	-0.752*** (0.11)	-0.851 (1.12)
Female × Foreign field	-0.070* (0.04)	-0.073* (0.04)	-0.525** (0.22)	-0.524** (0.22)
Breadth	-0.000 (0.00)	0.002 (0.00)	-0.029*** (0.01)	-0.030*** (0.01)
Female × Breadth		-0.003 (0.00)		0.001 (0.01)
Question FE	X	X	X	X
School FE	X	X	X	X
Observations	5,044	5,044	5,044	5,044

The unit of observation is the economist-question pair. For Column 1-2 the dependent variable is a dummy that is 1 if the respondent replied either strongly disagree or strongly agree to a question. For column 3-4 the dependent variable is the self-reported measure of confidence vis-a-vis a given question's reply (1 lowest, 10 highest on integer scale). *Female* is a dummy that is 1 if the respondent is female. *Year of PhD award* is the year the respondent was awarded the PhD. *PhD Harvard/MIT* is a dummy that is 1 if the respondent was awarded a PhD from either Harvard or MIT. *Total solo pubs.* is the number of total single authored publication. *Total co-authored pubs.* is the total number of co-authored publications. *Breadth* measures the number of distinct fields in which there are one or more papers citing an economist's work. Robust standard errors in parentheses, clustered at the question level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

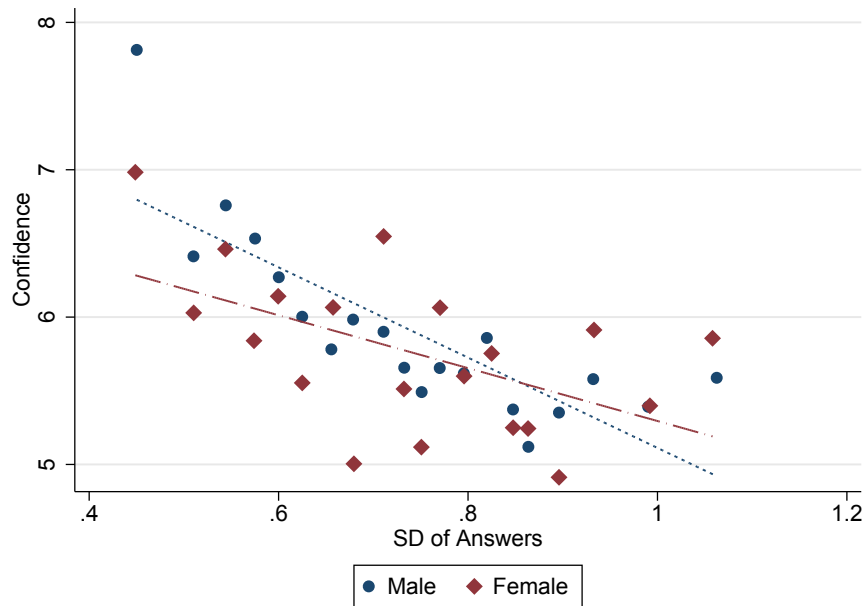
**Table 3.6: Confidence and Disagreement by Gender**

	(1)	(2)	(3)
		Confidence	
Mean of dep. var.	5.920	5.920	5.920
Female	0.052 (0.18)	-0.960*** (0.33)	-1.053*** (0.32)
Year PhD award	-0.010*** (0.00)	-0.010*** (0.00)	-0.011*** (0.00)
PhD Harvard/MIT	0.539*** (0.07)	0.539*** (0.07)	0.535*** (0.08)
Total solo pubs	0.018* (0.01)	0.018* (0.01)	0.022** (0.01)
Total co-authored pubs	-0.037*** (0.01)	-0.037*** (0.01)	-0.036*** (0.01)
Foreign field question	-0.636*** (0.11)	-0.635*** (0.11)	-0.766*** (0.11)
Female × Foreign field	-0.496** (0.20)	-0.484** (0.20)	-0.362* (0.20)
SD rating	-2.752*** (0.45)	-3.029*** (0.45)	
Female × SD rating		1.355*** (0.38)	1.339*** (0.38)
Question FE			X
School FE	X	X	X
Observations	5,241	5,241	5,241

The unit of observation is the economist-question pair. For Column 1-2 the dependent variable is a dummy that is 1 if the respondent replied either strongly disagree or strongly agree to a question. For column 3-4 the dependent variable is the self-reported measure of confidence vis-a-vis a given question's reply (1 lowest, 10 highest on integer scale). *Female* is a dummy that is 1 if the respondent is female. *Year of PhD award* is the year the respondent was awarded the PhD. *PhD Harvard/MIT* is a dummy that is 1 if the respondent was awarded a PhD from either Harvard or MIT. *Total solo pubs.* is the number of total single authored publication. *Total co-authored pubs.* is the total number of co-authored publications. *SD rating* measures the standard deviation of ratings submitted, capturing the level of disagreement in a given question. Robust standard errors in parentheses, clustered at the question level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

than women.

Figure 3.3: Confidence and disagreement



Self-reported confidence (1-10) and level of disagreement (as measured by the standard deviation in ratings) for a question, broken down by gender.

On questions with more disagreement, though, men’s confidence converges that of women. Therefore, on questions in which there is broad disagreement, both men and women recognize that disagreement and adjust their confidence accordingly. However, women do not seem to take others’ agreement with their view as a signal to be confident. We cannot say, though, whether men are being overconfident in these situations or if women are being underconfident. It also could be that while there is a consensus within the economics community on a topic, there is broad disagreement in other communities and women take this into account more than men. From this analysis, we can only say that women seem to be “sticky” in their confidence as they do not adjust it as much as men do. Finally, the *female × foreign field* remains negative and significant, providing further evidence for the robustness of our results.

### The Wallflower Effect

A complementary explanation is the existence of a “wallflower effect” in which women do not want to stand out and therefore give answers that are closer to the mean. Researchers have documented this phenomenon experimentally. Linardi and Jones (2014), for example, show that individuals whose charitable donations are made public are more likely to donate the mean amount of past donations. Women are particularly likely to condition their behaviour on what others have done.

In our context, women may avoid giving extreme answers to avoid standing out. This might in part have to do with confidence but there could also be other aspects, such as being given a hard time if they are wrong, that drive women to stick to the mean. In Appendix Table 8, we look at whether women are less likely to both strongly disagree or strongly agree with a statement. Women are 6 percentage points less likely to strongly agree with an answer but are not statistically less likely to strongly disagree with an answer. The wallflower effect would imply that women shade their answers on both ends, making them less likely to both strongly agree and strongly disagree. That women are no less likely to strongly disagree than men are suggests that the results are not driven by a desire to blend in.

### **Robustness**

We provide robustness checks to rule out alternative mechanisms (Table 3.7). An alternative explanation for the gender gap in confidence is that women sort into fields in which people are generally less confident or less extreme in their answers. It could be, for example, that macroeconomists are especially confident. Since there are few women in macroeconomics, the effects could be picking up this sorting rather than measuring an overall lack of confidence among women. However, the confidence gap persists, and is in fact larger, when controlling for field of study (Column 2) suggesting that sorting is not at play.

We include answer fixed effects in Column 3 of Table 3.7. In this sense, we are holding constant one margin of confidence and asking, for example, for all individuals who strongly agree with a statement, are there differences in how confident they are about strongly agreeing? The fixed effect takes out any correlation between the extremeness of the answer and the confidence in the answer. The fact that the size of the coefficient decreases, suggests that strongly agreeing with an answer is correlated with being more confident and that women are less likely to strongly agree with answers. Further, within each answer type, women are still less confident than men. Some of the confidence gap is thus driven by men having more extreme stances in addition to being more confident in their stance.<sup>10</sup>

## **3.4 Conclusion**

Several papers have found that women are less confident than men. We test whether the confidence gap persists for women who have reached the top of their careers. While we do find that a confidence gap persists, it is primarily driven by women being less confident when asked about topics they are not an expert on. It is therefore difficult to state that women being less confident is always negative. It could be that women have an optimal level of confidence and adjust it

---

<sup>10</sup>For brevity, we report the results for the explicit confidence-level measure but the results are similar for the propensity to provide extreme judgements. We also estimate the equations controlling for tenure. Tenure is negatively correlated with confidence, however, because only 4 people in our sample did not receive tenure and 3 of them are women, we refrain from drawing any conclusions from this result.

**Table 3.7: Robustness Checks**

	(1)	(2)	(3)
Dependent variable: Level of confidence (1-10)			
Mean of dep. var.	5.920	5.920	5.920
Female	-0.054 (0.18)	-0.357** (0.18)	0.020 (0.16)
Year PhD award	-0.011*** (0.00)	-0.022*** (0.00)	-0.004 (0.00)
PhD Harvard/MIT	0.537*** (0.08)	0.319*** (0.10)	0.553*** (0.07)
Total solo pubs	0.022** (0.01)	0.023** (0.01)	0.029*** (0.01)
Total co-authored pubs	-0.036*** (0.01)	-0.048*** (0.01)	-0.024*** (0.01)
Foreign field question	-0.767*** (0.11)	-0.827*** (0.11)	-0.697*** (0.09)
Female × Foreign field	-0.374* (0.20)	-0.509** (0.21)	-0.190 (0.18)
Question FE	X	X	X
School FE	X	X	X
Field FE		X	
Answer FE			X
Observations	5,241	5,241	5,144

The unit of observation is the economist-question pair. The dependent variable is the self-reported measure of confidence vis-a-vis a given question's reply (1 lowest, 10 highest on integer scale). *Female* is a dummy that is 1 if the respondent is female. *Year of PhD award* is the year the respondent was awarded the PhD. *PhD Harvard/MIT* is a dummy that is 1 if the respondent was awarded a PhD from either Harvard or MIT. *Total solo pubs.* is the number of total single authored publication. *Total co-authored pubs.* is the total number of co-authored publications. Robust standard errors in parentheses, clustered at the question level. \*\*\* p<0.01, \*\* p<0.05, \* p<0.1

depending on whether they are an expert whereas men are consistently overconfident. We also look at measures of confidence and find that the confidence gap has two components. Women hold less extreme views and are also less confident in their views.

Our paper helps explain why women can still be less confident than men even after breaking through the glass ceiling. Further research would help to understand why this gap exists and the implications that holding less extreme views and being less confident in their views has for women. For example, are women penalized for holding extreme views and does this contribute to the low number of women reaching upper-level positions? We leave such questions for future research.

# References

- [1] Altonji, Joseph and Charles Pierret. 2001. "Employer Learning and Statistical Discrimination." *The Quarterly Journal of Economics*, 116(1): 313-350.
- [2] Antecol, Heather, Kelly Bedard, and Jenna Stearns. 2016. "Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies?" IZA Discussion Paper No. 9904.
- [3] Arrow, Kenneth. 1973. "The Theory of Discrimination." In Orley Ashenfelter and Albert Rees, eds. *Discrimination in Labor Markets*. Princeton: Princeton University Press.
- [4] Azmat, Ghazala and Rosa Ferrer. 2017. "Gender Gaps in Performance: Evidence from Young Lawyers." *Journal of Political Economy*, 125(5): 1306-1355.
- [5] Babcock, Linda and Sara Laschever. 2004. *Women Don't Ask*. Princeton University Press.
- [6] Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva. 2017. "Does the Gender Composition of Scientific Committees Matter?" *American Economic Review*, 107(4): 1207-38.
- [7] Barber, Brad and Terrance Odean. 2001. "Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment." *The Quarterly Journal of Economics*, 116(1): 261-292.
- [8] Barnett, Michael, Nancy Keating, Nicholas Christakis, James O'Malley, and Bruce Landon. 2012a. "Reasons for Choice of Referral Physician Among Primary Care and Specialist Physicians." *Journal of General Internal Medicine*, 27(5): 506-512.
- [9] Barnett, Michael, Ziruir Song, and Bruce Landon. 2012b. "Trends in Physician Referrals in the United States, 1999-2009." *Archives of Internal Medicine*, 172(2): 163-170.
- [10] Becker, Gary. 1957. *The Economics of Discrimination*. Chicago: University of Chicago Press.
- [11] Bertrand, Marianne (2011). "New Perspectives on Gender" Chapter 17 in Ashenfelter, Orley and Card, David "Handbook of Labor Economics", Elsevier, edition 1, vol. 4 (5)
- [12] Bertrand, Marianne and Goldin, Claudia and Katz, Lawrence F. (2013). "Dynamics of the Gender Gap for Young Professionals in the Corporate and Financial Sectors", *American Economic Journal: Applied Economics*, 2 (3), pp. 228-255.
- [13] Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review*, 94(4): 991-1013.
- [14] Beyer, Sylvia. 1990. "Gender Differences in the Accuracy of Self-Evaluations of Performance." *Journal of Personality and Social Psychology*, 59: 960-970.

- [15] Blau, Francine and Lawrence Kahn. 2016. "The Gender Wage Gap: Extend, Trends, and Explanations." NBER Working Paper No. 21913.
- [16] Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics*, 131(4): 1753-1794.
- [17] Bursztyn, Leonardo, Thomas Fujiwara, and Amanda Pallais. "Acting Wife': Marriage Market Incentives and Labor Market Investments." *American Economic Review*, forthcoming.
- [18] Casadevall, Arturo and Jo Handelsman. 2014. "The Presence of Female Conveners Correlates with a Higher Proportion of Female Speakers at Scientific Symposia." *mBio*, 5(1): e00846-13.
- [19] Card, David, Ana Rute Cardoso, and Patrick Kline. 2016. "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women." *The Quarterly Journal of Economics*, 131(2): 633-686.
- [20] Ceci, Stephen, Donna Ginther, Shulamit Kahn, and Wendy Williams. 2014. "Women in Academic Science: A Changing Landscape." *Psychological Science in the Public Interest*, 15(3): 1-67.
- [21] Choudhry, Niteesh K., Joshua M. Liao, and Allan S. Detsky. 2014. "Selecting a Specialist: Adding Evidence to the Clinical Practice of Making Referrals." *Journal of the American Medical Association*, 312(18): 1861-1862.
- [22] Coate, Stephen and Glenn Loury. 1993. "Will Affirmative Action Policies Eliminate Negative Stereotypes?" *American Economic Review*, 83(5): 1220-12042.
- [23] Coffman, Katherine Baldiga (2014). "Evidence on Self-Stereotyping and the Contribution of Ideas" *Quarterly Journal of Economics*, 129 (4), pp. 1625-1660.
- [24] Eil, David and Justin Rao. 2011. "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself." *American Economic Journal: Microeconomics*, 3(2): 114-38.
- [25] Egan, Mark L., Gregor Matvos, and Amit Seru. 2017. "When Harry Fired Sally: The Double Standard in Punishing Misconduct." NBER Working Paper No. 23242.
- [26] Elmins, William, Robert Joyce, and Monica Costa Bias. 2016. "The Gender Wage Gap." Institute for Fiscal Studies Briefing Note 186.
- [27] Farber, Henry and Robert Gibbons. 1996. "Learning and Wage Dynamics." *The Quarterly Journal of Economics*, 111(4): 1007-1047.
- [28] Fiske, Susan and Shelley Taylor. 1991. *Social Cognition* (2nd ed.). New York: McGraw-Hill.
- [29] Forrest, Christopher, Paul Nutting, Sarah von Schrader, Charles Rohde, and Barbara Starfield. 2006. "Primary Care Physician Specialty Referral Decision Making: Patient, Physician, and Health Care System Determinants." *Medical Decision Making*, Jan-Feb.
- [30] Freedman, Rachel A., Elena M. Kouri, Dee W. West, and Nancy L. Keating. 2015. "Racial/Ethnic Differences in Patients' Selection of Surgeons and Hospitals for Breast Cancer Surgery." *JAMA Oncology*, 1(2): 222-230.

- [31] Fryer, Roland. 2007. "Belief Flipping in a Dynamic Model of Statistical Discrimination." *Journal of Public Economics*, 91(5-6): 1151-1166.
- [32] Fryer, Roland, Philipp Harms, and Matthew Jackson. 2017. "Updating Beliefs when Evidence is Open to Interpretation: Implications for Bias and Polarization." Working paper.
- [33] Ginther, Donna and Shulamit Kahn. 2004. "Women in Economics: Moving Up or Falling Off the Academic Career Ladder?" *Journal of Economics Perspectives*, 18(3): 193-214.
- [34] Glover, Dylan, Amanda Pallais, and William Pariente. 2017. "Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores." *The Quarterly Journal of Economics*, 132(3): 1219-1260.
- [35] Goldin, Claudia (2014a). "A Grand Gender Convergence: Its Last Chapter", 104 (4), pp. 1091-1119.
- [36] Goldin, Claudia. 2014b. "A Pollution Theory of Discrimination: Male and Female Differences in Occupations and Earnings." in L. Boustan, C. Frydman, and R. Margo, *Human Capital and History: The American Record* (Chicago: University of Chicago Press), pp. 313-48.
- [37] Goldin, Claudia, and Cecilia Rouse. 2000. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians," *American Economic Review*, 90(4): 715-741.
- [38] Gordon, Roger and Dahl, Gordon B. (2013). "Views among Economists: Professional Consensus or Point-Counterpoint?", *American Economic Review: Papers & Proceedings*, 103 (3), pp. 629-635.
- [39] Gregory, A., R.J. Skiba, and P.A. Noguera. 2010. "The Achievement Gap and the Discipline Gap: Two Sides of the Same Coin?" *Educational Researcher*, 39: 59-68.
- [40] Hughes, Gordon. 1991. "The Learning Curve in Staples Surgery." *The Laryngoscope*, 101(12): 1280-1284.
- [41] Jena, Anupam, Andrew Olenski, and Daniel Blumental. 2016. "Sex Differences in Physician Salary in US Public Medical Schools." *JAMA Internal Medicine*, 176(9): 1294-1304.
- [42] Kahn, Lisa and Fabian Lange. 2014. "Employer Learning, Productivity, and the Earnings Distribution: Evidence from Performance Measures." *The Review of Economic Studies*, 81(4): 1575-1613.
- [43] Keating, Nancy L., A. James O'Malley, Jukka-Pekka Onnela, and Bruce E. Landon. 2017. "Assessing the Impact of Colonoscopy Complications on Use of Colonoscopy Among Primary Care Physicians and Other Connected Physicians: An Observational Study of Older Americans." *BMJ Open*, 7.
- [44] Keehner, Madeleine, Yvonne Lippa, Daniel Montello, Frank Tendick, and Mary Hegarty. 2006. "Learning a Spatial Skill for Surgery: How the Contributions of Abilities Change with Practice." *Applied Cognitive Psychology*, 20(4): 487-503.
- [45] Kinchen, Kraig, Lisa Cooper, David Levine, Nae Yuh Wang, and Neil Powe. 2004. "Referral of Patients to Specialists: Factors Affecting Choice of Specialist by Primary Care Physicians." *Annals of Family Medicine*, 2(3): 245-252.

- [46] Kleven, Henrik, Camille Landais, and Jakob Egholt Sogaard. 2017. "Children and Gender Inequality: Evidence from Denmark." Working paper.
- [47] Lautenberger, Diana, Valerie Dandar, Claudia Raezer, and Rae Anne Sloane. 204. "The State of Women in Academic Medicine." Association of American Medical Colleges Annual Report.
- [48] Lazear, Edward P. and Kathryn L. Shaw. 2007. "Personnel Economics: The Economist's View of Human Resources." *Journal of Economic Perspectives*, 21(4): 91-114.
- [49] Lichtenstein, S., B. Fischhoff, and L.D. Phillips. 1982. Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases*, (pp. 306-334). Cambridge, England: Cambridge University Press.
- [50] Lo Sasso, Anthony, Michael Richards, Chiu-Fang Chou, and Susan Gerber. 2011. "The \$16,819 Pay Gap for Newly Trained Physicians: The Unexplained Trend of Men Earning More Than Women." *Health Affairs*, 30(2): 193-201.
- [51] Ly, Dan, Seth Seabury, and Anupam Jena. 2016. "Differences in Incomes of Physicians in the United States by Race and Sex: Observational Study." *BMJ*, 353: i2921.
- [52] MacLeod, W. Bentley. 2003. "Optimal Contracting with Subjective Evaluation." *American Economic Review*, 93(1): 216-240.
- [53] Medscape Physician Compensation Report. 2017. Sarah Grisham, Senior Editor. Art Science Code LLC. <<http://www.medscape.com/slideshow/compensation-2017-overview-6008547>>
- [54] Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat (2015). "Managing Self-Confidence: Theory and Experimental Evidence", mimeo
- [55] Mondak, Jeffery and Anderson, Mary (2004). "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge", *Journal of Politics*, 66 (2), pp. 492-512.
- [56] Moore, Don and Paul Healy (2008). "The Trouble with Overconfidence", *Psychological Review*, 115 (2), pp. 502-17.
- [57] Niederle, Muriel and Lise Vesterlund (2007). "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, 122 (3), pp. 1067-1101.
- [58] Noonan, Mary, Mary Corcoran, and Paul Courant. 2005. "Pay Differences Among the Highly Trained: Cohort Differences in the Sex Gap in Lawyers' Earnings." *Social Forces*, 84(2): 851-870.
- [59] Pantelis Kalaitzidakis, Theofanis P. Mamuneas, and Thanasis Stengos. 2003. "Rankings of Academic Journals and Institutions in Economics." *Journal of the European Economic Association*, 1(6): 1346-1366.
- [60] Phelps, Edmond. "The Statistical Theory of Racism and Sexism." *American Economic Review*, 62(4): 659-661.
- [61] Pinkston, Joshua. 2009. "A Model of Asymmetric Employer Learning with Testable Implications." *The Review of Economic Studies*, 76(1): 367-394.
- [62] Prendergast, Canice. 1999. "The Provision of Incentives in Firms." *Journal of Economic Literature*, 104(5): 7-63.

- [63] Rabin, Matthew and Joel Schragg. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics*, 114(1): 37-82.
- [64] Samek, Anya (2015). "A University-Wide Field Experiment on Gender Differences in Job Entry Decisions", mimeo
- [65] Sarsons, Heather (2015). "Gender Differences in Recognition for Group Work", mimeo.
- [66] Sasser, Alicia. 2005. "Gender Differences in Physician Pay: Tradeoffs Between Career and Family." *The Journal of Human Resources*, 40(2): 477-504.
- [67] Shea, Dennis, Bruce Stuart, Joseph Vasey, and Soma Ng. 1999. "Medicare Physician Referral Patterns." *Health Services Research*, 34(1): 331-348.
- [68] Skiba, Russell J., Robert S. Michael, Abra Carroll Nardo, and Reece L. Peterson. 2002. "The Color of Discipline: Sources of Racial and Gender Disproportionality in School Punishment." *The Urban Review*, 34(3): 317-342.
- [69] Zeltzer, Dan. 2017. "Gender Homophily in Referral Networks: Consequences for the Medicare Physician Earnings Gap." Working paper. Available at SSRN: <https://ssrn.com/abstract=2921482>

# Appendix A

## Appendix to Chapter 1

### A.1 Data Appendix

#### A.1.1 Example of Procedure Code Groupings

**Parent Group:** Surgical Procedures on the Cardiovascular System

- Surgical Procedures on the Heart and Pericardium
  - Surgical Procedures on the Pericardium
  - Excision Procedures of Cardiac Tumor
  - Transmyocardial Revascularization Procedures
  - Pacemaker or Pacing Cardioverter-Defibrillator Procedures
  - Electrophysiologic Operative Procedures on the Heart and Pericardium
  - Introduction or Removal of Patient-activated Cardiac Event Recorder
  - Surgical Procedures on the Heart (Including Valves) and Great Vessels
  - Surgical Procedures on Cardiac Valves
    - \* Surgical Procedures on the Aortic Valve (22 procedures)
    - \* Surgical Procedures on the Mitral Valve (8 procedures)
    - \* Surgical Procedures on the Tricuspid Valve (5 procedures)
    - \* Surgical Procedures on the Pulmonary Valve (7 procedures)
  - Other Cardiac Valvular Procedures
  - Coronary Artery Anomaly Procedures
  - Endoscopy Procedures on the Heart and Pericardium
  - Venous Grafting Only for Coronary Artery Bypass
  - Combined Arterial-Venous Grafting for Coronary Bypass
  - Arterial Grafting for Coronary Artery Bypass

- Coronary Endarterectomy Procedures
  - Repair Procedures for Single Ventricle and Other Complex Cardiac Anomalies
  - Repair Procedures for Septal Defect
  - Repair Procedures for the Sinus of Valsalva
  - Repair Procedures for Venous Anomalies
  - Shunting Procedures on the Heart and Pericardium
  - Repair Procedures for Transposition of the Great Vessels
  - Repair Procedures for Truncus Arteriosus
  - Repair Procedures for Aortic Anomalies
  - Repair Procedures for Thoracic Aortic Aneurysm
  - Endovascular Repair Procedures of the Descending Thoracic Aorta
  - Surgical Procedures on the Pulmonary Artery
  - Heart/Lung Transplantation Procedures
  - Extracorporeal Membrane Oxygenation or Extracorporeal Life Support Services and Procedures
  - Cardiac Assist Procedures
  - Other Cardiac Surgery Procedures
- Surgical Procedures on the Arteries and Veins

### **A.1.2 Specialties included in matched sample**

Cardiac surgery/cardiology, emergency medicine, general surgery, interventional cardiology, interventional radiology, nephrology, neurosurgery, orthopedic surgery, otolaryngology, plastic and reconstructive surgery, pulmonary disease, surgical oncology, urology, vascular surgery.

## **A.2 Appendix to Theoretical Framework**

### **A.2.1 Calculation of Work Hours Adjustment**

On average, female physicians work fewer hours than male physicians. As such, women see approximately 0.67 patients for every patient a man sees (Medscape Physician Compensation Report, 2016). In the Medicare data, female surgeons receive 10% of referrals, meaning that they receive 0.1 referrals when working 2/3 of a “man day” (number of hours men work on average). If we increase a woman’s work hours to be equivalent to a man’s, women should receive 0.05 additional referrals (since they receive 0.1 referrals per 2/3 day worked, they receive 0.05 referrals per 1/3 day worked).

Women should thus receive 15% of total referrals.

## A.2.2 Proofs

### Proof of Proposition 1

Assume that statements 1 and 2 in Proposition 1 hold and that  $\frac{\mathbb{P}(s_G|w)}{\mathbb{P}(s_G|m)} < 1$ . Statement 2 says that

$$\bar{a}_w - \bar{a}_m > \mathbb{E}(a|s_G, w) - \mathbb{E}(a|s_G, m) \quad (\text{A.1})$$

$$\bar{a}_w - \bar{a}_m > \mathbb{E}(a|s_B, w) - \mathbb{E}(a|s_B, m) \quad (\text{A.2})$$

Multiplying equation A.1 by  $\mathbb{P}(s_G|w)$  and equation A.2 by  $\mathbb{P}(s_B|w)$  and adding the two inequalities together gives

$$\begin{aligned} \mathbb{P}(s_G|w)(\bar{a}_w - \bar{a}_m) + \mathbb{P}(s_B|w)(\bar{a}_w - \bar{a}_m) > \\ \mathbb{P}(s_G|w)(\mathbb{E}(a|s_G, w) - \mathbb{E}(a|s_G, m)) + \mathbb{P}(s_B|w)(\mathbb{E}(a|s_B, w) - \mathbb{E}(a|s_B, m)) \end{aligned}$$

Since  $\mathbb{P}(s_G|w) + \mathbb{P}(s_B|w) = 1$ , we have

$$\begin{aligned} \bar{a}_w - \bar{a}_m &> \mathbb{P}(s_G|w)(\mathbb{E}(a|s_G, w) - \mathbb{E}(a|s_G, m)) + \mathbb{P}(s_B|w)(\mathbb{E}(a|s_B, w) - \mathbb{E}(a|s_B, m)) \\ \bar{a}_w - \bar{a}_m &> \mathbb{P}(s_G|w)\mathbb{E}(a|s_G, w) + \mathbb{P}(s_B|w)\mathbb{E}(a|s_B, w) - \mathbb{P}(s_G|w)\mathbb{E}(a|s_G, m) - \mathbb{P}(s_B|w)\mathbb{E}(a|s_B, m) \\ \bar{a}_w - \bar{a}_m &> \bar{a}_w - \mathbb{P}(s_G|w)\mathbb{E}(a|s_G, m) - \mathbb{P}(s_B|w)\mathbb{E}(a|s_B, m) \\ \bar{a}_m &< \mathbb{P}(s_G|w)\mathbb{E}(a|s_G, m) + \mathbb{P}(s_B|w)\mathbb{E}(a|s_B, m) \end{aligned} \quad (\text{A.3})$$

where the law of iterated expectations is used in lines 3-4. Substituting in for  $\bar{a}_m$ , using the fact that  $\mathbb{P}(s_B|m) = 1 - \mathbb{P}(s_G|m)$ , and rearranging gives

$$\begin{aligned} \mathbb{P}(s_G|m)\mathbb{E}(a|s_G, m) + \mathbb{P}(s_B|m)\mathbb{E}(a|s_B, m) &< \mathbb{P}(s_G|w)\mathbb{E}(a|s_G, m) + \mathbb{P}(s_B|w)\mathbb{E}(a|s_B, m) \\ \mathbb{P}(s_G|m)\mathbb{E}(a|s_G, m) + (1 - \mathbb{P}(s_G|m))\mathbb{E}(a|s_B, m) &< \mathbb{P}(s_G|w)\mathbb{E}(a|s_G, m) + (1 - \mathbb{P}(s_G|w))\mathbb{E}(a|s_B, m) \\ \mathbb{P}(s_G|m) \cdot (\mathbb{E}(a|s_G, m) - \mathbb{E}(a|s_B, m)) &< \mathbb{P}(s_G|w) \cdot (\mathbb{E}(a|s_G, m) - \mathbb{E}(a|s_B, m)) \\ (\mathbb{P}(s_G|m) - \mathbb{P}(s_G|w)) \cdot (\mathbb{E}(a|s_G, m) - \mathbb{E}(a|s_B, m)) &< 0 \end{aligned} \quad (\text{A.4})$$

For equation A.4 to be negative, one of the terms must be negative and the other positive. Note that  $(\mathbb{E}(a|s_G, m) - \mathbb{E}(a|s_B, m))$  must be positive by the definition of a good and bad event, so it must be that  $\mathbb{P}(s_G|m) < \mathbb{P}(s_G|w)$  which violates our initial assumption that  $\frac{\mathbb{P}(s_G|w)}{\mathbb{P}(s_G|m)} < 1$ .

### Proof of Proposition 2

Assume without loss of generality that the variance over priors is larger for female surgeons than male surgeons:  $\sigma_w^2 > \sigma_m^2$ . The law of total variance states that the prior variance is the sum of the

variance of the posterior and the (expected) posterior variance:

$$\sigma_g^2 = \text{Var}(\mathbb{E}[a|s, g]) + \mathbb{E}[\sigma_g^2|s]$$

Taking the difference in the variance of the prior between men and women gives

$$\sigma_w^2 - \sigma_m^2 = \mathbb{E}[\sigma_w|s] - \mathbb{E}[\sigma_m|s] + \text{Var}(\mathbb{E}[a|s, w]) - \text{Var}(\mathbb{E}[a|s, m]) \quad (\text{A.5})$$

We have assumed that  $\sigma_w^2 - \sigma_m^2 > \mathbb{E}[\sigma_w|s] - \mathbb{E}[\sigma_m|s]$ . Substituting [A.5](#) into this inequality, we get

$$\begin{aligned} \mathbb{E}[\sigma_w|s] - \mathbb{E}[\sigma_m|s] + \text{Var}(\mathbb{E}[a|s, w]) - \text{Var}(\mathbb{E}[a|s, m]) \\ > \mathbb{E}[\sigma_w|s] - \mathbb{E}[\sigma_m|s] \\ \text{Var}(\mathbb{E}[a|s, w]) - \text{Var}(\mathbb{E}[a|s, m]) > 0 \end{aligned} \quad (\text{A.6})$$

Equation [A.6](#) shows that the spread over beliefs about a female surgeon after a signal is larger than the spread in beliefs about a male surgeon after the equivalent signal, meaning that the physician must have updated more about the female surgeon after both good and bad signals.

### A.3 Additional Tables

**Table A1: Balance for Placebo Matched Samples**

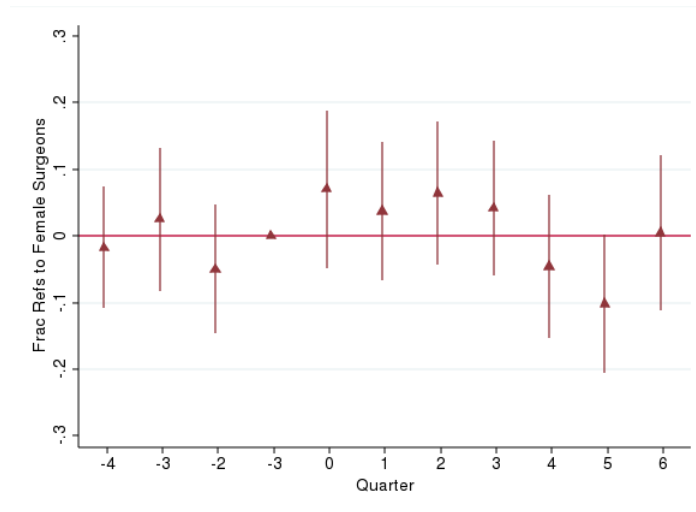
<i>Panel A: Male Surgeons</i>	No Death		Death		p-value
	Mean	SD	Mean	SD	
Patient Refs from Physician	25.6	33.4	25.4	34.2	0.654
Total Patient Refs	69.3	106.0	69.0	105.9	0.899
Patient Age	77.8	9.0	77.9	9.1	0.451
Patient Minority (%)	6.8	25.2	6.8	25.2	0.999
Patient Female (%)	59.0	49.2	59.0	49.2	0.999
Patient Risk (%)	0.85	0.34	0.85	0.35	0.156
Risk All Past Ptnts	0.010	0.007	0.010	0.006	0.194
Experience (Yrs)	24.9	10.4	24.9	10.3	0.997
Available Surgeons	48.7	29.7	49.0	29.6	0.123
Observations	15,012		15,012		

<i>Panel B: Female Surgeons</i>	No Death		Death		p-value
	Mean	SD	Mean	SD	
Patient Refs from Physician	25.4	44.4	25.1	44.2	0.664
Total Patients Refs	72.3	110.0	72.0	107.7	0.821
Patient Age	79.7	9.8	79.8	9.8	0.317
Patient Minority (%)	5.7	23.2	5.7	23.2	0.999
Patient Female (%)	61.2	48.7	61.2	48.7	0.999
Patient Risk (%)	0.010	0.004	0.010	0.005	0.162
Risk All Past Ptnts	0.009	0.004	0.009	0.004	0.284
Experience (Yrs)	22.2	8.3	22.2	8.5	0.640
Available Surgeons	48.8	29.7	48.7	29.4	0.254
Observations	4,658		4,658		

Notes: This balance table shows summary statistics for the matched sample of surgeons who did and did not experience a patient death. The surgeons who did not experience a death are called “placebos” in the paper. To create each sample, I take the set of male or female surgeons who ever experienced a bad event and match the male surgeons to other men who did not experience a death, and match the female surgeons to other women who did not experience a death. The sample is matched exactly on patient gender and minority status as well as surgeon specialty and procedure (not shown above). The sample is matched coarsely on all other variables. All variables are defined as in Table 1.2.

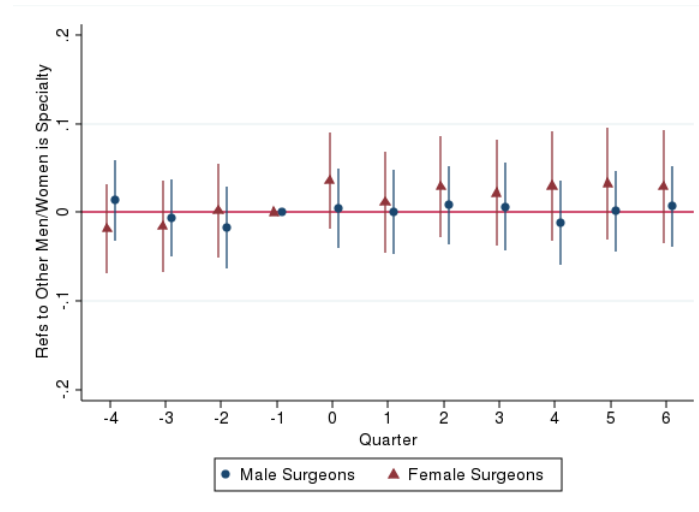
## A.4 Additional Figures

Figure A1: Spillovers to Female Surgeons, >10 Prior Refs



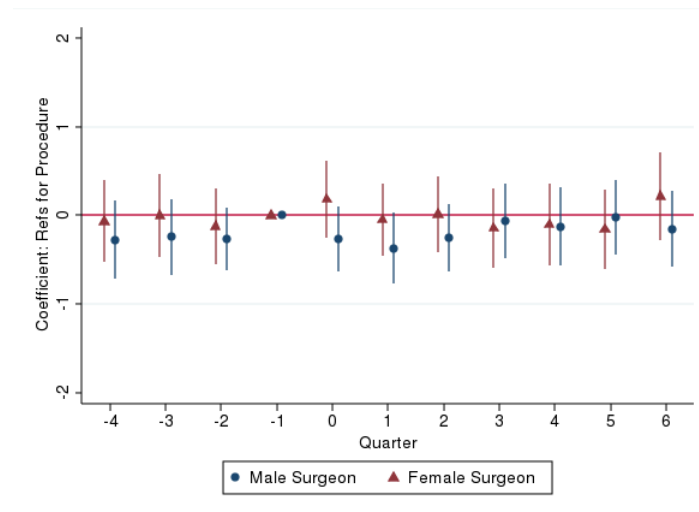
*Notes:* This figure shows how physicians change their behaviour toward other female surgeons after a patient death. In  $k = 0$ , a patient that physician  $j$  sent to surgeon  $i$  dies. The outcome variable is the fraction of referrals going to female surgeon whom physician  $j$  had referred at least 10 patients to prior to the death. The estimation is done on the sample of female surgeons who experience a patient death. The coefficients are plotted relative to the fraction of referrals that physician  $j$  was sending to these surgeons in quarter  $k = -1$ . I control for the fraction of available surgeon who are male or female and also include physician-surgeon match fixed effects. Standard errors are clustered at the physician-surgeon match level.

**Figure A2: Spillovers to Other Surgeons after Good Outcome**



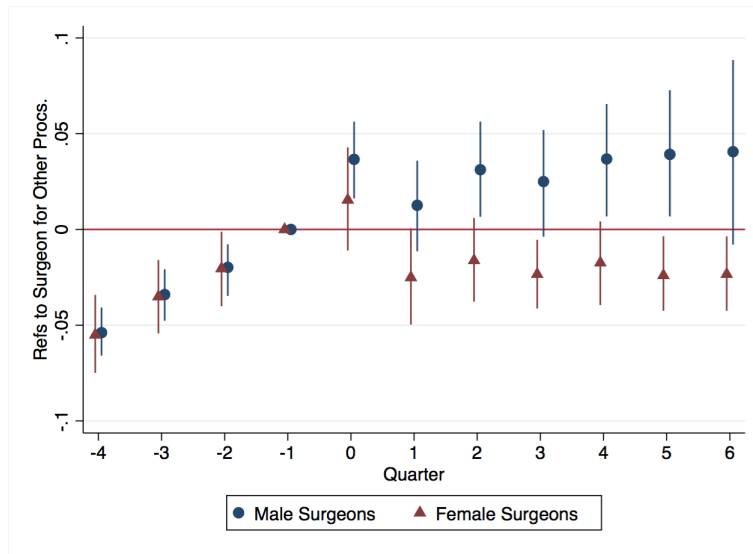
*Notes:* This figure shows how physicians change their behaviour toward other surgeons after an unexpectedly good outcome. In  $k = 0$ , a risky patient that physician  $j$  sent to surgeon  $i$  is not rehospitalized within 30 days of surgery. The outcome variable is the number of referrals going to female (Figure A) or male (Figure B) surgeons whom physician  $j$  hasn't referred to before. The coefficients are plotted relative to the number of referrals that physician  $j$  was sending to surgeons she had not previously referred to in quarter  $k = -1$ . Physician-surgeon match fixed effects are included in the regression and standard errors are clustered at the physician-surgeon match level.

**Figure A3: Referrals for Procedure**



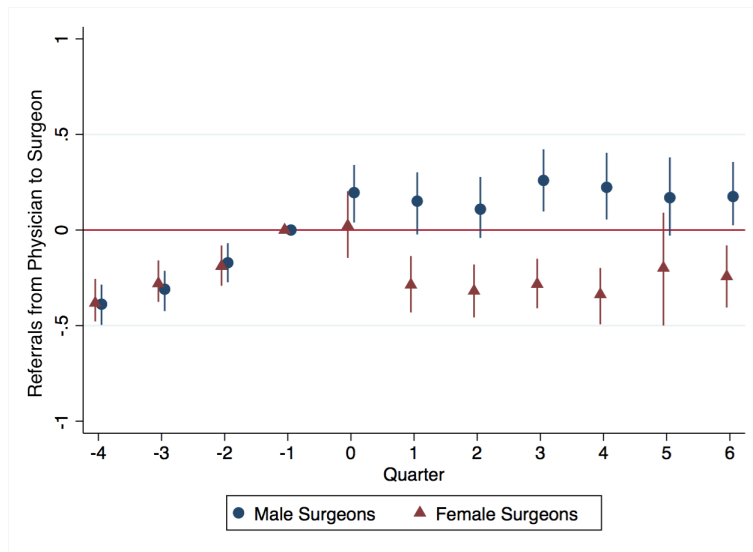
*Notes:* This figure plots the coefficients from a regression estimating how many patient referrals a physician gives for the procedure that a patient dies from in quarter  $k = 0$ . The outcome variable is the quarterly number of referrals for the particular procedure that physician sends to surgeons other than the performing surgeon. The estimation is done on the sample of matched male and female surgeons who experience a patient death. All coefficients are plotted relative to the number of procedural referrals the physician was sending in  $k = -1$ , which is normalized to zero. Surgeon-physician match fixed effects are included and standard errors are clustered at the physician-surgeon match level.

**Figure A4: Referrals for Other Procs. to Performing Surgeon**



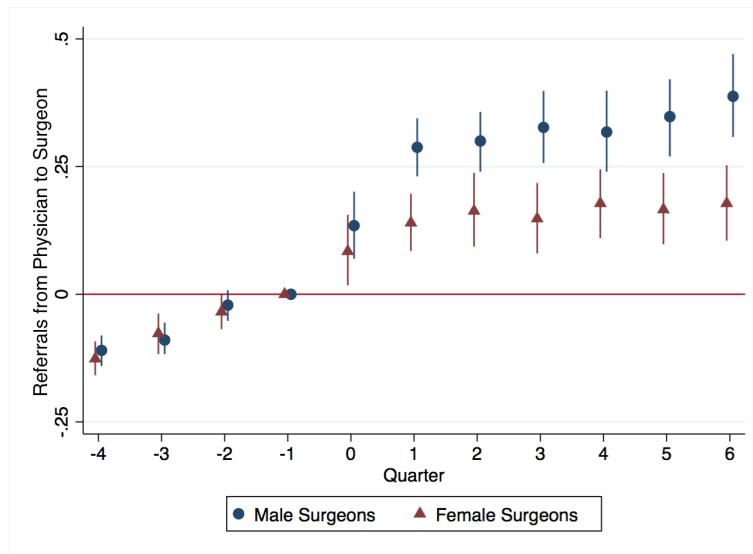
Notes: This figure shows the number of patients that physician  $j$  refers to surgeon  $i$  for a procedure other than the patient who died was referred for. In  $k = 0$ , a patient that  $j$  referred to  $i$  for procedure  $n$  dies. The outcome variable is the number of referrals that  $j$  refers to  $i$  in each quarter for any procedure other than  $n$ . The coefficients are plotted relative to  $k = -1$  and 95% confidence intervals are also shown. The coefficients for  $k = 1$  through  $k = 6$  are jointly significantly different from zero at the 10% level for both male and female surgeons. The sample used is the sample of matched surgeons who experienced a patient death. Surgeon-physician match fixed effects are included. Standard errors are clustered at the physician-surgeon match level.

**Figure A5: Deaths that Occur on Day of Surgery**



Notes: This figure plots the quarterly regression coefficients and 95% confidence intervals from estimating equation 1.3 using the sample of matched male and female surgeons who experience a patient death that occurs within 24 hours of surgery. The coefficients are plotted relative to the number of referrals the surgeon was receiving from the physician in  $k = -1$ , which are normalized to zero. In  $k = -1$ , male and female surgeons both received an average of 0.65 referrals from physician  $j$ . A patient that physician  $j$  referred to surgeon  $i$  dies in  $k = 0$ . The outcome variable is the total number of referrals that physician  $j$  sends to surgeon  $i$  each quarter. Standard errors are clustered at the physician-surgeon match level.

**Figure A6: Unexpectedly Good Outcomes: Top 5% Risk Level**



Notes: This figure plots the quarterly coefficients and 95% confidence intervals from estimating equation 1.3, looking at good events. Here, good events are defined using a 5% risk cutoff rather than a 1% risk cutoff. All coefficients are plotted relative to the number of referrals the surgeon was receiving from the physician in  $k = -1$ , which are normalized to zero. Standard errors are clustered at the physician-surgeon match level.

## Appendix B

# Appendix to Chapter 2

## Appendix A

### List of institutions included in analysis

**Received faculty list:** Brown, Columbia, Duke, Michigan State University, NYU, Northwestern, Ohio State University, Penn State, UC Berkeley, UC San Diego, UCLA, University of Virginia, University of Maryland, University of Michigan, University of Minnesota, University of Pennsylvania, University of Wisconsin-Madison

**No faculty list:** Boston College, Boston University, California Institute of Technology, Cornell, Harvard, MIT, Princeton, Stanford, University of Southern California, University of Chicago, University of Texas - Austin, University of Rochester, Yale

## Appendix B

This simple matching example illustrates how women's knowledge of the returns to coauthoring affects their authorship decisions. It describes the contexts in which we would (1) expect high ability men and women to collaborate, making the decision not to promote collaborating women sub-optimal, and (2) expect low ability women and high ability men to collaborate, making the decision not to tenure collaborating women optimal. The model abstracts from the employer's problem, assuming that employers want to promote anyone who is high ability and not promote anyone who is low ability, and focuses on how knowledge of the true returns to coauthoring could lead to sorting among workers.

### Setup

There are two types of agents: employers and workers. Workers produce papers and can choose whether to work alone or to collaborate. The quality of the paper depends on the worker's ability,

as well as that of the collaborator if the worker collaborates. The employer uses the quality of the paper as a signal to infer the worker's type so the worker wants to maximize the quality of the paper subject to production costs.

Workers are either high or low ability, ( $a_i \in \{h, l\}$ ), and belong to an identifiable group, men or women ( $g_i \in \{m, w\}$ ). I make the simplifying assumption that workers know each other's ability but the employer does not. As mentioned earlier, workers choose whether to work alone or collaborate,  $c_i \in \{S, CA\}$ , and are trying to maximize the quality of the paper. Paper quality is a function of each collaborator's type:  $f(a_{i,g})$  if  $c = S$  and  $f(a_{i,g}, a_{j,g})$  if  $c = CA$ . I assume that high ability men and women produce the same quality of papers ( $f(h_w) = f(h_m)$ ) as do low ability men and women.

The payoffs to producing solo and coauthored papers respectively are

$$\pi_{i,S} = f(a_{i,g}) - \kappa_S \quad (\text{B.1})$$

$$\pi_{i,CA} = f(a_{i,g}, a_{j,g}) - \kappa_{CA} \quad (\text{B.2})$$

where the costs of producing solo and coauthored papers,  $\kappa_S$  and  $\kappa_{CA}$ , are assumed to be constants with  $\kappa_S \neq \kappa_{CA}$ .

### Assortative Matching with Equal Credit for Papers

Consider a woman's final payoff to producing a solo-authored paper and a paper coauthored with a man:

$$\pi_S = f(a_w) - \kappa_S \quad (\text{B.3})$$

$$\pi_{CA} = f(a_w, a_m) - \kappa_{CA} \quad (\text{B.4})$$

If types are complimentary, the quality of a paper is higher when high ability women work with high ability men:

$$f(h_w, h_m) - f(l_w, h_m) > f(h_w, l_m) - f(l_w, l_m) \quad (\text{B.5})$$

In this case, a high ability woman is willing to offer a high ability man  $f(h_w, h_m) - f(h_w, l_m)$  to collaborate, while a low ability woman is willing to offer  $f(l_w, h_m) - f(l_w, l_m)$ . Inequality (B.5) states that  $f(l_w, h_m) - f(l_w, l_m) < f(h_w, h_m) - f(h_w, l_m)$ , meaning that high ability women can outbid low types. Therefore, high types will coauthor and low types will coauthor as long as the costs of coauthoring do not outweigh the benefits. For high types to work together, the following conditions must be true for both men and women:

$$f(h_w, h_m) - f(h_w) \geq \kappa_{CA} - \kappa_S \quad (\text{B.6})$$

$$f(h_w, h_m) - f(h_m) \geq \kappa_{CA} - \kappa_S \quad (\text{B.7})$$

If at least one of (B.6) or (B.7) does not hold, there will be no coauthoring between men and women. Because we assume that  $f(h_w) = f(h_m)$ , if (B.6) does not hold, neither does (B.7) and vice versa.

Since  $f(h_w, h_m) > f(h_w, l_m) = f(l_w, h_m)$ , high ability workers who do not want to collaborate with other high ability workers will not want to collaborate with low ability workers either.

Thus, in the case of assortative matching in which men and women have the same returns to papers, men and women will only coauthor with the same ability type. A woman coauthoring with a high ability man thus provides the employer with a signal that she too is a high type, and women who coauthor with men should be no less likely to receive tenure than high ability men. Note that this would also be the case if women received less credit for papers but did not know it. That is, if  $f_w(h_w, h_m) < f_m(h_w, h_m)$  but women believe these to be equal.

### Assortative Matching with Unequal Credit for Papers

Now assume that women receive less credit for their collaborative work and that they know this. Specifically, let the payoff to a woman who coauthors with a man be  $\hat{\pi}_{CA} = \beta f(m_a, w_a) - \kappa_{CA}$  where  $\beta < 1$ . The payoffs to coauthoring for high types are now

$$\text{Woman : } \beta f(h_w, h_m) - f(h_w) > \kappa_{CA} - \kappa_S \quad (\text{B.8})$$

$$\text{Man : } f(h_w, h_m) - f(h_m) > \kappa_{CA} - \kappa_S \quad (\text{B.9})$$

If both ((B.8)) and (B.9)) hold, we are back in the case of assortative matching. Both groups are willing to collaborate and high type women are able to outbid low type women. However, if  $\beta$  is sufficiently small, ((B.8)) becomes less likely to hold and high ability women will choose to solo author.

Since payoffs between men and women are now different, there are some cases in which high-ability men and low-ability women might collaborate. In particular, high-ability men will be willing to coauthor with low-ability women if

$$f(l_w, h_m) - f(m_h) > \kappa_{CA} - \kappa_S \quad (\text{B.10})$$

and low-ability women will coauthor if

$$\beta f(l_w, h_m) - f(w_l) > \kappa_{CA} - \kappa_S \quad (\text{B.11})$$

Note that  $\beta f(h_w, h_m) - f(h_w) < \kappa_{CA} - \kappa_S$  does not imply that  $\beta f(l_w, h_m) - f(l_w) < \kappa_{CA} - \kappa_S$  since  $f(h_w) > f(l_w)$ . Therefore, if  $\beta f(l_w, h_m)$  is sufficiently larger than  $f(l_w)$  for low-ability women (that is, the paper is greatly improved with the help of a coauthor), then we might still see low-type women wanting to coauthor with men. If the cost savings of coauthoring are large enough and if coauthoring with a low type does not reduce the quality of the paper, men will still be willing to coauthor with low ability women. Of course, many high-ability men will choose to coauthor with other high-ability men rather than a low-ability woman. The purpose of this example is to illustrate the circumstances under which we would see women sorting into solo and coauthoring by ability. For this, we need to assume that at least some high-ability men must coauthor with a woman.

Overall, if women do not know that they have lower returns to coauthoring, we will see assortative matching. In this case, employers who believe a man to be high ability should also believe his female collaborator is high ability. Denial of tenure in this case is sub-optimal. However, if women know the returns to coauthoring, high ability women will solo-author or work with other high-ability women and tenure denial is rational.

This example only considered two types of workers, high and low ability. Extending it to more types does not change the results. Assortative matching will ensure that workers will always match with workers of their type. When women know the true returns to coauthoring, the highest ability women will work alone and any male/female matching will require that the woman is slightly lower ability than the man.