Do Women Make More Credible Threats? Gender Stereotypes, Audience Costs, and Crisis Bargaining

Joshua A. Schwartz and Christopher W. Blair

Accepted at *International Organization*

Abstract

As more women attain executive office, it is important to understand how gender dynamics affect international politics. Toward this end, we present the first evidence that gender stereotypes affect leaders' abilities to generate audience costs. Using survey experiments, we show that female leaders have political incentives to combat gender stereotypes that women are weak by acting "tough" during international military crises. Most prominently, we find evidence that female leaders, and male leaders facing female opponents, pay greater inconsistency costs for backing down from threats than male leaders do against fellow men. These findings point to particular advantages and disadvantages women have in international crises. Namely, female leaders are better able to tie hands—an efficient mechanism for establishing credibility in crises. However, this bargaining advantage means female leaders will also have a harder time backing down from threats. Our findings have critical implications for debates over the effects of greater gender equality in executive offices worldwide.

During the 1984 presidential campaign, Geraldine Ferraro, the first major-party female vice-presidential candidate in US history, was asked at a debate, "do you think in any way that the Soviets might be tempted to try to take advantage of you simply because you are a woman?"[1] Similar questions dogged her entire campaign. In a follow-up on *Meet the Press*, hosts questioned whether Ferraro was "strong enough to push the [nuclear] button."[2] These statements reveal a pervasive gender stereotype: that men are better equipped to handle national security issues than women.[3] While gender stereotypes persist, the number of female political leaders has grown markedly over time. As Figure 1 indicates, women occupied nearly 10.5 percent of all executive offices worldwide in 2015, and have served as head of state in sixty-six countries since 1875.[4] The growing prevalence of women in high political office thus raises important questions about the role of leaders' gender in the conduct of war and peace. Toward this end, we investigate how common gender stereotypes affect crisis-bargaining dynamics. Specifically, we address a gap in the literature by presenting the first evidence of how gender stereotypes affect leaders' abilities to generate audience costs.

[Insert Figure 1 here]

Audience costs are the domestic political punishments leaders face for making a threat and then backing down.[5] Kertzer and Brutger identify two components of audience costs: inconsistency and belligerence.[6] Inconsistency costs, the traditional audience cost, are those leaders pay for making threats but failing to follow through. These threats

[1] Ferraro and Franke 2004, 40.

[2] Ibid., 306.

[3] Alexander and Andersen 1993; Dolan 2014; Falk and Kenski 2006; Holman, Merolla, and Zechmeister 2011; Huddy and Terkildsen 1993; Kahn 1992; Lawless 2004; Post and Sen 2020; Rosenwasser and Dean 1989; Sanbonmatsu 2002.

[4] See Table A.1 of the appendix for the full list of female leaders over time. Barnes and O'Brien 2018 show women's representation is also increasing in defense ministries worldwide.

[5] Fearon 1994.

[6] Kertzer and Brutger 2016.

tie leaders' hands because inconsistency costs are paid only if leaders back down.[7] Belligerence costs are those leaders pay for threatening force in the first place. These are sunk costs since leaders pay them immediately after issuing a threat.[8] Given that leaders always have an incentive to bluff, the benefit of being able to generate higher audience costs is greater credibility at the bargaining table, since only genuinely resolved leaders would be willing to tie their hands and sink costs.[9] Generating audience costs also allows leaders to better communicate their intentions, thereby reducing the chances that miscalculation will lead to war.[10] The disadvantage—especially with inconsistency costs—is that backing down from threats becomes more difficult as leaders become "locked into their position[s]," which can hamper efforts to de-escalate existing crises.[11] Drawing on insights from political science and psychology, we argue that female leaders pay greater inconsistency costs than male leaders facing male opponents. If female leaders demonstrate "weakness" by backing down from threats, they activate descriptive gender stereotypes about women's ill-preparedness for the demands of high office generally[12] and conflict in particular.[13] Male leaders who act inconsistently, by contrast, are judged less harshly because men's failures are more often attributed to situational factors beyond their control rather than dispositional factors related to their character.[14] In other words, female leaders are held to a higher standard than their male counterparts and are punished more for perceived policy failures, like inconsistency.

But gender stereotypes are not wholly irrelevant for male leaders. We also contend that male leaders pay greater inconsistency costs for backing down against women than they do for backing down against fellow men. Since gender stereotypes dictate that

---

[7] Fearon 1997.

[8] Fearon 1997; Kertzer and Brutger 2016.

[9] Schelling 1960.

[10] Fearon 1994, 587; Fearon 1995; Jervis 1978; Weeks 2008, 60.

[11] Fearon 1994, 577.

[12] Heilman 1995, 2001, 2012.

[13] Carlin, Carreras, and Love 2019.

[14] Swim and Sanna 1996.

women are less capable in the realm of national security, and that men should be strong and assertive, backing down against women is viewed as emasculating and seen as a negative signal of a male leader's competence. This kind of dynamic is evident even in schoolyard disputes, where "you lost to a girl" is a common pejorative.

Finally, given that female leaders may have political incentives to "act tough" during international crises to combat gender-stereotypical expectations of weakness, and male leaders have incentives to avoid appearing weak against female foes, we argue that female leaders will pay *lower* belligerence costs than male leaders facing fellow males, and the same is true for male leaders acting belligerently against female leaders.[15]

To isolate the effects of gender stereotypes on public evaluations of leaders in interstate disputes, we conducted two survey experiments. Experiments help overcome two related issues that plague observational studies on this topic: sample size and selection issues.[16] Because war and female leadership are historically rare, and since women both attain and perform in high political office nonrandomly, the feasibility of inference from observational data is limited. In an experimental setting, we can randomly vary leaders' genders and crisis behaviors, while holding other factors constant. Our primary experiment, which includes 2,342 subjects recruited through the Time-Sharing Experiments for the Social Sciences (TESS) panel conducted with the National Opinion Research Center (NORC) at the University of Chicago, reveals support for our theory. Female leaders pay greater inconsistency costs for backing down from threats than male leaders do against fellow men, and likewise for male leaders acting inconsistently against female leaders. These results also held in a pilot experiment we conducted on 1,607 Amazon Mechanical Turk (mTurk) subjects, lending further confidence in our findings.

Our results with respect to belligerence costs are somewhat more mixed, but also generally support our hypotheses. Results from our TESS experiment reveal that female leaders, and male leaders facing female leaders, pay lower belligerence costs than male

[15] Bauer 2017; Caprioli and Boyer 2001; Enloe 1990; Koch and Fulton 2011.

[16] Baturo and Gray 2018, Jalalzai 2013, and Reiter 2014 discuss the methodological challenges of identifying effects of female leadership with observational data.

leaders facing fellow men. A similar pattern emerges in our mTurk study, though the results are not statistically significant. Sentiment analysis conducted using open-ended responses from our TESS study corroborate our main findings on inconsistency and belligerence.

In sum, this study makes four principal contributions. First, we extend the bargaining literature by applying the logic of audience costs to an important empirical trend: the growing number of women in high political office. A large literature on audience costs has examined how these vary with regime type;[17] electoral structure;[18] media environment;[19] leaders' rhetoric;[20] and audience characteristics.[21] However, no study we are aware of has analyzed the impact of gender and gender stereotypes on leaders' abilities to generate audience costs.[22] More broadly, our study extends the burgeoning experimental literature on gender.[23]

Second, our findings extend those of Kertzer and Brutger and lend further support for the notion that it is essential to disaggregate audience costs into inconsistency and belligerence in order to draw appropriate inferences from audience-cost experiments.[24] Simply looking at overall audience costs obscures the key fact that female leaders generally pay greater inconsistency costs and lower belligerence costs. Because these two effects are

[17] Weeks 2008

[18] Schultz 1998.

[19] Potter and Baum 2010.

[20] Levendusky and Horowitz 2012; Trager and Vavreck 2011.

[21] Kertzer and Brutger 2016; Tomz 2007.

[22] For a partial exception see Croco and Gartner 2014. They examine whether female politicians are punished more for "flip-flopping" on support for the Afghanistan War, a kind of inconsistency. Their approach differs from ours though because they: (1) do not examine audience costs since their focus is on inconsistent support for a war rather than backing down from a *threat*; (2) study gender monadically not dyadically; (3) focus on senators rather than the president; (4) utilize a nonrepresentative sample of college students; and (5) study a nine-year time gap in inconsistency, likely diluting its effect.

[23] See also Bauer 2015, 2017; Clayton, O'Brien, and Piscopo 2019; Holman et al. 2019; Karim et al. 2018; Klar 2018; Naurin, Naurin, and Alexander 2019; Teele, Kalla, and Rosenbluth 2018.

[24] Kertzer and Brutger 2016.

countervailing, a nondisaggregated replication of our study would miss critical nuances in the role of gender stereotypes during crises.

Third, our results strengthen the emerging consensus that leader attributes matter in important ways.[25] Research examines how factors like age,[26] post-tenure security,[27] and attitudinal dispositions[28] affect leaders' behavior, but pays less attention to gender. Melding the rich literature on gender and politics with scholarship on leaders, our findings highlight the importance of gender and gender stereotypes in international relations.[29] We hope future scholarship will pay closer attention to the roles of gender and gender stereotypes in shaping leader conduct.

Fourth, this study has implications for debates about whether increasing gender equality in executive office holding will lead to less belligerent foreign policies and more peace, or the reverse. Supporters of the "women-as-peacemakers" view, like Steven Pinker, argue that "over the long sweep of history, women have been and will be a pacifying force. Traditional war is a man's game."[30] This perspective implies that bioevolutionary factors[31] and socialization processes[32] incline women toward peace, so a world with more female leaders should be more pacific. Alternatively, supporters of the "iron ladies" view contend that more belligerent female leaders are selected into office,[33] and that once in office, female executives face incentives to combat gender stereotypes by adopting hawkish policies.[34] Our findings help reconcile these perspectives.

On one hand, our findings suggest that women's increasing roles in executive office

---

[25] Goemans, Gleditsch, and Chiozza 2009; Horowitz, Stam, and Ellis 2015.

[26] Horowitz, Stam, and Ellis 2015.

[27] Goemans, Gleditsch, and Chiozza 2009.

[28] Yarhi-Milo, Kertzer, and Renshon 2018.

[29] See also Naurin, Naurin, and Alexander 2019.

[30] Pinker 2011, 527.

[31] Fukuyama 1998; McDermott et al. 2007.

[32] Caprioli and Boyer 2001; Regan and Paskeviciute 2003.

[33] Enloe 1990.

[34] Bashevkin 2018; Koch and Fulton 2011; Schramm and Stark forthcoming.

may have a pacifying effect because female leaders have bargaining advantages. Since they are punished more for inconsistency, female leaders are better able to tie hands, which is the most efficient means for establishing credibility in crises.[35] Enhanced credibility should lead to more effective communication, reduced uncertainty, and a lower chance of international conflict. This is especially the case since male leaders competing against female leaders are also better able to generate inconsistency costs, facilitating clear communication. On the other hand, the mechanism driving this relationship is not that women are innately pacifistic or socialized to avoid aggression, but that they face political pressure to combat gender stereotypes by acting tough. Female leaders have political incentives to behave hawkishly, rendering their threats more credible, but also locking them into their positions and making it harder to de-escalate after threats have been made.[36]

Theory

Stereotypes are pervasive, durable, shared beliefs held about groups on the basis of certain (often ascriptive) characteristics. These biases typically incorporate both *descriptive* and *prescriptive* dimensions, meaning gender stereotypes influence beliefs about both what men and women are *perceived* to be like and what they *ought* to be like.[37] In complex environments like international crises, stereotypes serve as heuristic devices, guiding decision making on the basis of simplified categories.[38] Our intuition that evaluations of leaders' behavior are influenced by gender stereotypes and the normative expectations these biases conjure builds from these social psychological insights. Specifically, we draw on Heilman's Lack of Fit model.[39]

The Lack of Fit model suggests that individuals rely on stereotypes to form ex-

[35] Fearon 1997.

[36] Crisman-Cox and Gibilisco 2018.

[37] Burgess and Borgida 1999.

[38] McGarty, Yzerbyt, and Spears 2002.

[39] Heilman 1995, 2001, 2012.

pectations of performance when assessing leaders.[40] Even though the number of female executives has increased over time, descriptive stereotypes implying women are ill-suited for the realm of national security endure. Specifically, many studies find that men are viewed as tougher and better able to handle military crises than women.[41] For instance, Lawless finds that 61 percent of respondents believe that men are better prepared to respond to military crises than women; just 3 percent of respondents believe women are better able to handle military crises than men.[42] Likewise, those who consider national security as the top issue facing the country are significantly more likely to believe that a male president would do a better job than a female president,[43] and the public prefers male leadership during times of heightened terrorist threat.[44]

As the Lack of Fit model implies, these findings reflect a perceived discordance between the qualities women possess and the qualities necessary for success in foreign affairs. Particularly, gender stereotypical expectations that men are strong, aggressive protectors, and women are delicate and require protection, drive divergent beliefs about how male and female leaders will perform in military crises.[45] Because of female leaders' perceived "lack of fit" for the role of commander-in-chief, they face heightened scrutiny for their decisions, meaning women in power are often held to higher standards and have to outperform men in order to be evaluated equally highly.[46]

Perceptions of women's "lack of fit" for positions of leadership during crises are compounded by the fact that women's *failures* are more likely to be attributed to dispositional factors like incompetence, while men's *failures* are more likely to be attributed

---

[40] Ibid.

[41] Alexander and Andersen 1993; Dolan 2014; Huddy and Terkildsen 1993; Kahn 1992; Lawless 2004; Rosenwasser and Dean 1989; Sanbonmatsu 2002.

[42] Lawless 2004, 482.

[43] Falk and Kenski 2006.

[44] Holman, Merolla, and Zechmeister 2011.

[45] Goldstein 2001, 273; Sjoberg and Tickner 2011, 176.

[46] Bauer 2015, 2017; Carlin, Carreras, and Love 2019; Heilman 2001.

to situational factors beyond their control.[47] This means that observers will be likely to view female leaders' failures as confirming gender-stereotypical expectations about women's "lack of fit," while male leaders' failures may not shift expectations about male fitness for leadership.

Further, gender stereotypes may also operate as second-order beliefs, or beliefs about what others believe. This means that even if individuals do not personally subscribe to gender stereotypes—though many do—they may behave in accordance with the Lack of Fit model because they believe that other individuals and world leaders hold gender stereotypes. In the context of a military crisis, for example, a respondent might hold a female leader to a higher standard not because they personally believe women are ill-suited to the role of commander-in-chief, but because they believe foreign leaders subscribe to gender-stereotypical expectations about women's lack of fit, and so fear any misstep will cause the female leader to be viewed as an irresolute and incredible target.[48]

To combat gender-stereotypical expectations of weakness and minimize criticism, female leaders have political incentives to act tough during international crises.[49] For example, female chief executives are more likely to increase defense spending[50] and initiate militarized interstate disputes than male leaders.[51] Likewise, high-ranking female foreign policymakers—like Jeane Kirkpatrick, Madeline Albright, Condoleezza Rice, and Hillary Clinton—often advocate more aggressive foreign policies than their male counterparts.[52] In the medieval period, married queens were more likely than kings to be aggressors in interstate conflicts.[53] Examples of modern "iron ladies"—like Margaret Thatcher, Indira Gandhi, and Golda Meir—and ancient "warrior queens"—like Cleopatra, Boudica,

---

[47] Swim and Sanna 1996, 515.

[48] For more on reputation as a second-order belief, see Brutger and Kertzer 2018. We thank an anonymous reviewer for noting this possibility.

[49] Bauer 2017; Caprioli and Boyer 2001; Enloe 1990.

[50] Koch and Fulton 2011.

[51] Schramm and Stark forthcoming.

[52] Bashevkin 2018; McGlen and Sarkees 1993.

[53] Dube and Harish forthcoming.

and Isabella of Spain—lend further credence to the view that female leaders have political motivations to pursue relatively hard-line policies to combat gender stereotypes.[54] Mark Penn, Hillary Clinton's chief strategist in 2008, argued that Clinton had political incentives to portray strength:

> Regardless of the sex of the candidates, most voters in essence see the presidents as the "father" of the country. They do not want someone who would be the first mama, especially in this kind of world ... [Thatcher] represents the most successful elected woman leader in this century—and the adjectives that were used about her (Iron Lady) were not of good humor or warmth, they were of smart, tough leadership.[55]

The Lack of Fit model suggests that if women demonstrate weakness by, for example, acting inconsistently, support will wane. Because audiences are stereotypically inclined to believe women will fare worse in conflicts, a female leader's failure to follow through will confirm mass suspicions about her "lack of fit" for executive office, and the public will respond punitively, attributing her perceived failures more to dispositional than situational factors. Even individuals who do not themselves believe women are ill-suited for leadership may believe that foreign leaders believe gender stereotypes and will view female leaders as incredible; these individuals will punish female inconsistency because of second-order gender-stereotypical beliefs and extrinsic concerns about reputation. In short, when female leaders perform poorly in international crises by making a threat and then backing down, gender stereotypes are likely to be activated, leading to greater disapproval from the general population than when male leaders behave identically.

From the Lack of Fit model's logic, we derive a number of testable implications about how gender stereotypes affect leaders' abilities to generate audience costs. In any potential conflict dyad, there are four possible gender combinations: (1) the most common male-male (MM) dyad, involving two male leaders; (2) the female-male (FM) dyad, where the domestic leader is a female and the foreign leader is a male; (3) the male-female (MF) dyad; and (4) the presently rare female-female (FF) dyad.[56] The male-male dyad, the

[54] Fraser 1990.

[55] Joshua Green, "Penn's 'Launch Strategy' Ideas, December 21, 2006" *The Atlantic*, 11 August 2008, 2.

[56] While recognizing the spectrum of gender identity, we presume a gender dichotomy for the purpose of

most common historical combination by far, can be thought of as the baseline group against which we are comparing other dyads.[57] Our first two hypotheses compare the FM and FF dyads to the MM baseline:

> *H*1*a: Female leaders pay greater inconsistency costs compared to the MM dyad.*
>
> *H*1*b: Female leaders pay lower belligerence costs compared to the MM dyad.*

While there may be a strategic logic to bluffing, the public typically perceives acting inconsistently by making a threat and then backing down as a policy failure.[58] Indeed, inconsistency is what scholars commonly think of when they discuss audience costs.[59] The Lack of Fit model predicts that gender stereotypes will be activated when female leaders behave inconsistently, leading to greater disapproval from the general population than when male leaders behave the same way against fellow men. Thus, female leaders in mixed (FM) and same-gender (FF) dyads should face higher inconsistency costs than male counterparts in same-gender (MM) dyads. Because female executives' failures are often perceived as dispositional,[60] women in general are more likely to be perceived as incompetent for acting inconsistently or failing to respond forcefully to aggression. Essentially, when female leaders perform poorly in international crises by backing down, gender stereotypes are activated regardless of the gender of the rival leader, leading to greater disapproval from the general population. There is empirical support for this argument, as Carlin, Carreras, and Love find that increases in terrorism—a clear policy failure—reduce the public approval of female but not male leaders.[61]

---

analytical simplicity and because individuals still overwhelmingly think about gender in binary terms. Ellemers 2018, 277.

[57] Observational data support our decision to treat the male-male crisis dyad as the baseline or "control" group. Looking at bilateral militarized interstate disputes (MIDs), almost 96 percent since 1816 have occurred between men.

[58] Fearon 1994.

[59] Kertzer and Brutger 2016.

[60] Swim and Sanna 1996.

[61] Carlin, Carreras, and Love 2019.

We also expect that female executives will pay lower belligerence costs compared to the male-male baseline. In traditional audience-cost experiments, including ours, domestic leaders are faced with a clear case of foreign aggression: the invasion of a third country by an adversary. In this context, the Lack of Fit model implies that female heads of state will have political incentives to act belligerently to combat *descriptive* gender stereotypes that they are weak.[62] To understand this intuition, think of the inverse of belligerence costs: "inaction costs." These are the costs that leaders pay for doing nothing in response to the invasion of a third country, relative to making a threat in response and following through on it.[63] We expect that female leaders will pay greater inaction costs—and consequently lower belligerence costs—because, according to the Lack of Fit model, doing nothing in response to foreign aggression will activate *descriptive* gender stereotypes of perceived female weakness in military affairs.

We now turn to situations where male leaders face female opponents. Comparing the mixed-gender MF dyad to the male-male baseline, we hypothesize:

> *H2a: Male leaders facing female opponents pay greater inconsistency costs compared to the MM dyad.*
>
> *H2b: Male leaders facing female opponents pay lower belligerence costs compared to the MM dyad.*

In this situation, relational stereotypes are relevant. As Ellemers describes, gender stereotypes do not merely prescribe how individuals of different genders are expected to perform in general, but also how they are expected to perform in relation to one another.[64]

[62] Granted, *prescriptive* gender stereotypes imply that women should not be too assertive in foreign affairs (Rudman and Glick 2001), and so may cut in the opposite direction. Since war is often viewed as a male domain (Goldstein 2001), the public may punish female leaders for violating gender norms (Goodyear-Grant 2013). However, prescriptive gender stereotypes likely have the strongest effect on public opinion when female leaders clearly initiate conflicts and are the aggressors, which is not the case in our experiment.

[63] As we discuss in more detail later, belligerence costs are equal to disapproval in the engage condition minus disapproval in the stay out condition. Inaction costs are the opposite: disapproval in the stay out condition minus disapproval in the engage condition.

[64] Ellemers 2018.

Building from the Lack of Fit model's expectation that men are perceived as better equipped to handle national security affairs than women, the logic of relational stereotypes suggests that backing down against a female leader will be viewed as emasculating and a particularly negative sign of a male leader's competence. Put simply, for male targets of female-initiated threats, backing down should be perceived as a sign of weakness, defying expectations about masculine strength and "fit" for leadership according to the Lack of Fit model. Consequently, male leaders have political incentives to act tough against female leaders to avoid perceptions that they backed down against an opponent people expect to be weaker. Anecdotal evidence corroborates this expectation. In 60 CE, Boudica, a Celtic queen, led an uprising against Rome. Cassius Dio, a Roman historian, wrote of Roman losses to Boudica: "all this ruin was brought upon them by a woman, a fact which in itself caused them the greatest shame."[65]

This logic also extends to our expectations regarding belligerence costs. We predict that, on balance, male leaders facing female opponents will pay lower belligerence costs compared to the MM baseline. According to the Lack of Fit model, male leaders are likely to be viewed as better suited than women for military crises. *Descriptive* stereotypes that men are stronger and more capable in military affairs mean male leaders will have political incentives to act belligerently against female leaders to avoid the perception that they feared fighting a weaker opponent. Returning to the hypothetical inverse of belligerence costs, inaction costs, our logic suggests that male leaders should face greater inaction costs—and thus lower belligerence costs—in a crisis against a female initiator because inaction against a female adversary could signal surprising "lack of fit" for the role of commander-in-chief.[66]

By way of illustration, consider Yahya Khan's eagerness to fight Indira Gandhi

[65] Quoted in Gillespie 2018, 105.

[66] *Prescriptive* gender stereotypes suggest that men should protect women, not fight them. Stiehm 1982. While Naurin, Naurin, and Alexander 2019 find a chivalry effect consistent with prescriptive stereotypes, their experimental scenario involved cooperation between European Union allies. Our scenario, by contrast, involves conflict rather than cooperation, and the public is unlikely to sympathize with foreign leaders credibly accused of aggression, regardless of gender.

during the Bangladesh crisis of 1970–71. As he noted, "If that woman [Indira Gandhi] thinks she is going to cow me down, I refuse to take it. If she wants to fight, I'll fight her!"[67] Clearly, Khan was not afraid of fighting a female leader, as prescriptive stereotypes might suggest. Rather, documentary evidence suggests Khan was motivated by the fear that he would be perceived as weak if he refused to fight Gandhi in the first place, or failed to follow through on his threats once made.

Tables 1 and 2 summarize these hypotheses. In our experimental framework, the domestic leader is the leader whose cost-generating capacities we measure.

[Insert Table 1 here]

[Insert Table 2 here]

Experimental Design

To test our hypotheses, we designed and administered a 3 x 2 x 2 x 2 between-subjects experiment fielded in collaboration with TESS on a pool of 2,342 subjects recruited from NORC's nationally representative AmeriSpeak panel.[68] Our design and hypotheses were pre-registered with Evidence in Governance and Politics (EGAP).[69] To maximize comparability, the design and wording of the experiment closely follow that of seminal audience-cost experiments conducted by Tomz and Kertzer and Brutger.[70] The factors we varied are the United States' crisis action (stay out, not engage, and engage); the US president's gender; the foreign leader's gender; and the US president's partisan affiliation. We blocked on respondent party identification to ensure approximately equal numbers of Democrats, Independents, and Republicans in each experimental cell. Every respondent was presented with the following introduction:

[67] Malhotra 1991, 137.

[68] AmeriSpeak is a representative, probability-based panel with households selected from a sample frame based on the NORC National Frame and address-based sample.

[69] The design is registered under EGAP ID # 20190731AB. An exploratory pilot study was fielded on Amazon's mTurk platform prior to pre-registration and fielding on the AmeriSpeak panel.

[70] Kertzer and Brutger 2016; Tomz 2007.

> The following questions are about US relations with other countries around the world. You will read about a situation our country has faced many times in the past and will likely face again. Different leaders have handled the situation in different ways. We will describe one approach US leaders could take in the future and ask whether you approve or disapprove.

The only difference between this introduction and the one utilized by Tomz and Kertzer and Brutger is that instead of telling respondents that "we will describe one approach US leaders *have taken*," we told them that "we will describe one approach US leaders *could take* in the future."[71] The reason for this difference is that there have not been any female US presidents in the past and so, to be realistic, our scenario had to be forward looking. With this caveat in mind, we were sanguine about the prospect that respondents would approach scenarios describing female presidents seriously. In three of the last four US presidential elections, a woman has served as a major party presidential or vice presidential nominee, and in all four of the last US presidential elections, female candidates have made serious primary bids.[72] Further, we fielded our study in August and September 2019, in a period when six women—Elizabeth Warren, Amy Klobuchar, Kamala Harris, Kirsten Gillibrand, Tulsi Gabbard, and Marianne Williamson—were Democratic primary candidates for the 2020 presidential election.[73] Despite the fact that the US has never had a female president, we think concerns that respondents did not take our prompt realistically are mitigated because of the realistic possibility of a female president.

After the introduction, we presented respondents with information about a hypothetical international crisis scenario:

> A country sends its military to take over a neighboring country. The attacking country is controlled by a [female/male] leader.

[71] Tomz 2007; Kertzer and Brutger 2016.

[72] In the 2008 election, Sarah Palin was the Republican vice presidential nominee, and Hillary Clinton was a Democratic primary candidate. In the 2012 election, Michele Bachmann was a Republican primary candidate. In the 2016 election, Hillary Clinton was the Democratic presidential nominee, and Carly Fiorina was a Republican primary candidate. In the 2020 election, presumptive Democratic nominee Joe Biden committed to choosing a woman as his running mate, and a historic number of women ran for the Democratic nomination.

[73] Gillibrand ended her candidacy during our study period.

Next, we presented respondents with the identity of the US president:

> The [Republican/Democratic] US President, [Erica/Eric, Stephanie/Steven] Smith...

Following Trager and Vavreck, we randomized the party of the US president.[74] This is particularly important for analyzing the effects of gender since women are often perceived as more liberal than men.[75] The name combinations we utilized are similar, but clearly primed gender.[76] They should not, however, have primed any notable politician because no former US presidents or vice presidents share any of the names we employed. Although Hillary Clinton is the most prominent female politician in US history, an advantage of fielding this study during the 2020 campaign cycle is that the large number of female candidates running should reduce the extent to which respondents thought solely about Clinton when evaluating the crisis scenario. Research by Kromer and Parry also demonstrates that priming Hillary Clinton does not aggravate or diminish gendered expectations.[77] We randomized name assignment within the US president's gender condition to mitigate any effects of name choice.

After presenting respondents with the identity of the US president, we randomly assigned them to one of three different scenarios for how the United States responds. To distinguish between inconsistency and belligerence costs, we employed the same three categories that Kertzer and Brutger used: stay out, not engage, and engage.[78] In the stay-out scenario, the US president promises to refrain from intervening in the crisis and abides by this promise:

> ...says the United States will stay out of the conflict. The attacking country continues to invade. In the end, [Erica/Eric, Stephanie/Steven] Smith decides not to send troops, and the attacking country gains 20 percent of the contested territory.

---

[74] Trager and Vavreck 2011.

[75] Hayes 2011; Koch 2000; McDermott 1997.

[76] On the use of names to prime gender, see MacNell, Driscoll, and Hunt 2015.

[77] Kromer and Parry 2019.

[78] Kertzer and Brutger 2016.

In the not-engage scenario, the US president promises to deploy troops to resolve the crisis, but fails to do so:

> ...says that if the attack continues, the United States military will push out the invaders. The attacking country continues to invade. In the end, [Erica/Eric, Stephanie/Steven] Smith does not send troops, and the attacking country gains 20 percent of the contested territory.

In the engage scenario, the US president promises to deploy troops to resolve the crisis and follows through:

> ...says that if the attack continues, the United States military will push out the invaders. The attacking country continues to invade. In the end, [Erica/Eric, Stephanie/Steven] Smith orders the US military to engage. The attacking country gains 20 percent of the contested territory and the US experiences zero casualties.

Note that following Kertzer and Brutger, we hold constant outcomes in all three conditions to isolate the effect of inconsistency and belligerence.[79] Like previous studies, our outcome measures are binary and seven-point Likert scales to measure approval or disapproval of the US president's handling of the crisis. Within this framework, inconsistency costs equal disapproval in the not-engage condition minus disapproval in the engage condition. Belligerence costs equal disapproval in the engage condition minus disapproval in the stay-out condition. Audience costs equal inconsistency plus belligerence costs.

Experimental Results

Table 3 displays the percentage point difference in mean disapproval for the FM, FF, and MF dyads compared to the MM baseline.[80] Positive values indicate that audience, inconsistency, or belligerence costs are greater for the respective dyad relative to the MM baseline, and negative values indicate that these costs are lower. In accordance with previous studies, Table 3 collapses the seven-point measure of approval or

[79] Kertzer and Brutger 2016.

[80] Table 3 excludes respondents that failed the attention check, leaving us 1,816 respondents. Results are substantively similar with the full sample. See Table A.2.

disapproval into a binary measure of disapproval to more clearly illustrate substantive effects.[81] Substantively identical results emerge with the full seven-point measure.[82]

We begin by examining *H1a* and *H2a*, which hold that inconsistency costs should be greater in the FM, FF, and MF dyads than in the MM baseline. Column 2 in Table 3 demonstrates statistical support for these hypotheses, as well as substantively large effects. Disapproval is 20.7 percentage points greater for a female president acting inconsistently against a foreign male leader compared to a male president acting inconsistently against a fellow male ($p \approx 0.001$; 95% bootstrapped CI: 6.7, 34.2). Similarly, disapproval is 18.2 percentage points greater for a female president acting inconsistently against a foreign female leader than the MM baseline ($p \approx 0.008$; 95% bootstrapped CI: 3.4, 32.2). Further, male presidents who act inconsistently against foreign female leaders face disapproval rates that are 15.4 percentage points greater compared to when they act inconsistently against male leaders ($p \approx 0.018$; 95% bootstrapped CI: 1.2, 29.7).

[Insert Table 3 here]

Our findings with respect to belligerence costs also comport with our hypotheses. Recall that *H1b* and *H2b* predict that belligerence costs will be lower in the FM, FF, and MF dyads compared to the MM baseline. In accordance with this expectation, disapproval is 14.4 percentage points lower for a female president acting belligerently against a foreign male leader compared to a male president acting belligerently against a fellow male ($p \approx 0.026$; 95% bootstrapped CI: -29.0, 1.0). For a female president acting belligerently against a fellow female, disapproval is 13.6 percentage points lower than the baseline ($p \approx 0.037$; 95% bootstrapped CI: -28.9, 1.4). Finally, disapproval is 10.8 percentage points lower for a male president acting belligerently against a foreign female leader compared to a male president acting belligerently against a fellow male ($p \approx 0.079$; 95% bootstrapped CI: -25.7, 3.7).

We did not hypothesize about total audience cost effects because we anticipated the effects of inconsistency and belligerence costs to countervail one another. Specifically,

[81] Tomz 2007.

[82] See Tables A.3 and A.4.

because we expected that inconsistency costs would be greater in the FM, FF, and MF dyads compared to the MM baseline, while belligerence costs would be lower, our theory predicts null or small aggregate effects. These expectations bear out. In column 1 of Table 3, we examine whether there are any differences in total audience costs across dyads. Consistent with our expectations, no statistically significant differences emerge when we analyze total audience costs. This null, however, masks critical heterogeneity. Thus, our results provide additional support for Kertzer and Brutger's argument that it is essential to disaggregate audience costs.[83] Simply looking at overall audience costs obscures the fact that female leaders pay greater inconsistency costs and lower belligerence costs because these two effects work against one another.

To ensure the robustness of our core findings, we take a number of steps. First, we verify that results are substantively similar when we use the full sample of respondents, rather than only those who passed the attention check.[84] Second, we show that substantively identical results emerge when we employ the full seven-point measure of approval or disapproval.[85] Third, we show that results hold in a regression that controls for factors like the partisan identity of the US president in the scenario; the respondents' gender, age, education, partisanship, level of sexism, and level of militant assertiveness; and whether our sexism battery was administered pre- or post-treatment.[86] Fourth, we present results from our exploratory mTurk pilot study fielded in February 2019, which are substantively similar, though yield more modest support on belligerence costs.[87] The robustness of our results across these tests builds confidence in our main findings.

[83] Kertzer and Brutger 2016.

[84] See Tables A.2 and A.4

[85] See Tables A.3 and A.4

[86] See Table A.5.

[87] See Table A.10.

Sentiment Analysis

To further probe the robustness of our findings, we asked respondents (after presenting each crisis scenario) to provide four words that they believe best described the US president.[88] Open-ended questions can help provide a more direct view into a survey subject's beliefs.[89] Using the `tidytext` package in `R`, and a dictionary developed by Liu, we classified respondents' word answers as positive or negative.[90] As an alternative to our primary measurement strategy, which relies on a forced-choice Likert item, we use the average sentiment score for each respondent calculated from the mean of the four words given. Each respondent's sentiment score about the president in the crisis scenario serves as an alternative way to operationalize their disapproval of the president's crisis action. Table 4 presents the results from our sentiment-analysis exercise. Positive values indicate that audience, inconsistency, or belligerence costs are greater for the relevant gender dyad compared to the MM baseline, and negative values indicate that these costs are lower. Results in Table 4 are substantively identical to our estimates in Table 3, lending further confidence in the robustness of our main results.

[Insert Table 4 here]

Internal Validity

Experiments are the gold standard for causal identification, but they are not entirely immune from confounding. In our context, the most likely source of confounding is a lack of information equivalence, where manipulating one factor (e.g., gender) leads respondents to update their beliefs about other relevant, but not experimentally manipulated, dimensions.[91] Our experimental design explicitly controlled for one possible

[88] This approach follows Rothschild et al. 2019.

[89] Roberts et al. 2014.

[90] Liu 2015. Words not captured by the original dictionary were hand-coded as either positive, negative, or missing (neither positive nor negative) according to guidelines in Liu 2015. See our replication files for the list of our hand-coded words.

[91] Dafoe, Zhang, and Caughey 2018.

confounding factor—the party of the US president—but two other possibilities stand out. First, it is possible respondents will think that female presidents are more likely to be nonwhite than male presidents. If this is the case, then it could be racial stereotypes that drive higher inconsistency costs for female leaders rather than gender. Second, survey subjects might infer that foreign countries led by a woman are more likely to be democratic. To rule out these possibilities, we asked respondents placebo questions at the end of the survey to gauge their perceptions about the US president's race and the foreign country's regime type. Promisingly, we find no systematic evidence of confounding. Female US presidents were only marginally more likely to be perceived as nonwhite ($\rho \approx 0.05$), and foreign countries led by women were only slightly more likely to be perceived as democratic ($\rho \approx 0.11$). These correlations demonstrate that there is no widespread association between the gender of US presidents and race, or the gender of foreign leaders and regime type. More importantly, our results are robust to the inclusion of controls for these variables in a regression.[92]

Three other potential concerns also warrant mention. First, it is possible that respondents intuited from our experiment that our focus was on gender. This possibility raises the specter of experimenter demand effects, which occur if respondents surmise researchers' hypotheses and adjust their behavior to validate those expectations. Recent work, however, suggests respondents are often unable to adjust behaviors to conform with researchers' expectations, so demand effects are unlikely to bias our results.[93] A second, related concern stems from social desirability. Respondents could have intuited our focus on gender stereotypes, and adjusted their behavior to appear less sexist. While possible, this would bias *against* our inconsistency cost results because respondents seeking to appear less sexist would be *more approving* of women's crisis actions. Order effects are a third potential concern because some respondents received a battery of questions designed to measure sexism before treatment, while others received the battery after treatment. However, assignment to the order of the sexism battery was randomized, and our results

[92] See Table A.5.

[93] Mummolo and Peterson 2019.

hold when the order is controlled for in a regression.[94]

Heterogeneous Effects

In the appendix, we analyze whether the effects of gender stereotypes on audience costs vary across respondent subgroups, focusing on five respondent characteristics: militant assertiveness, partisanship, sexism, age, and respondent gender. Contrary to our expectations, we find no evidence that our hypotheses are stronger among Republican, more sexist, older, or male respondents.[95] These null results, especially with respect to sexism, are consistent with gender stereotypes mattering more as second-order beliefs. We cannot test this contention directly, but it is a ripe avenue for future research. We also replicate Kertzer and Brutger's findings: Democrats and individuals low in militant assertiveness impose higher belligerence costs, and Republicans and individuals high in militant assertiveness impose higher inconsistency costs.[96] By replicating Kertzer and Brutger's well-known findings about partisanship and militant assertiveness in the context of disaggregated audience costs, we build confidence in our design.

Conclusion

As the number of women in executive office grows, it is imperative to consider how gender dynamics impact international politics. This study provides the first causal evidence that gender stereotypes affect leaders' abilities to generate audience costs. Our most important finding is that female leaders, and male leaders facing female leaders, pay greater inconsistency costs for backing down from threats than male leaders do against fellow men. These results have critical implications for theory and policy, and speak to calls for more nuance in understanding the reasons men and women have for fighting.[97]

---

[94] See Table A.5.

[95] See Tables A.7 and A.8.

[96] Kertzer and Brutger 2016; See Table A.6.

[97] McDermott 2015.

The evidence in this article suggests that female leaders hold important advantages and disadvantages in bargaining situations. On one hand, their greater ability to generate inconsistency costs means women should find it easier to tie their hands in crises, and in turn are better able to establish credibility and signal resolve. As a result, female leadership may facilitate peace by making it easier to communicate intentions ex ante. On the other hand, because women face higher costs for backing down from threats, and lower costs for initiating in the first place, gender stereotypes may contribute to military adventurism and conflict risk because female leaders will find it tempting to make threats and difficult not to escalate once threats have been made.

As far as theory, these findings build on the rich literature on feminist approaches to international relations, and bear critically on the debate over the peace-inducing effects of female leadership in world politics. While some scholars contend that greater equality in holding executive office will facilitate peace because women are innately less belligerent than men for bioevolutionary[98] or social reasons,[99] our work in this piece points to a more complicated view. Because female leaders hold bargaining advantages, more women holding executive office may indeed lead to peace, but not because women are less willing to fight than men. In fact, our results suggest women may actually be *more willing* to fight. The peace-inducing effects of female attainment of high office, rather, stem from the fact that women make more credible threats, and can communicate their intentions and resolve more effectively. In sum, our empirical results may help unify extant theoretical[100] and empirical critiques[101] of the women-as-peacemakers view that Fukuyama and Pinker, among others, espouse.[102] In this way, our theoretical framework and results can account for the seemingly disparate facts that female leadership is associated with peace,[103] and

[98] Fukuyama 1998; McDermott et al. 2007; Pinker 2011.

[99] Caprioli and Boyer 2001; Enloe 1990; Goldstein 2001.

[100] Tickner 1992, 1994.

[101] Dube and Harish forthcoming; Schramm and Stark forthcoming.

[102] Fukuyama 1998; Pinker 2011.

[103] Caprioli and Boyer 2001.

that women are as or more likely than men to initiate conflicts.[104]

Our results also highlight a number of promising avenues for future research. First, new work suggests that apart from inconsistency and belligerence costs, incompetency costs also weigh in the public's mind during international crises.[105] These are costs that leaders pay for failing to achieve their audiences' desired outcomes. While beyond the scope of this project, it would be interesting to extend our argument about gender stereotypes to an analysis of incompetency costs to determine whether women are also held to higher standards than men in evaluations of policy success, as some scholars imply.[106] Second, more research is needed to unpack the diverse ways gender stereotypes matter, ranging from chivalry reactions in cooperative scenarios[107] to the costs we identify in interstate crises. Third, our findings speak to the need for more research on whether gender stereotypes operate primarily as first- or second-order beliefs among members of the public. Fourth, and relatedly, what are *leaders'* first- and second-order beliefs about how gender stereotypes affect rival leaders' credibility? Future research could fruitfully tackle this question with elite surveys.[108] Finally, our results raise questions about how other pervasive biases, such as racial stereotypes, affect international policymaking. Greater appreciation for the role of gender and other stereotypes in international relations can help scholars understand the likely implications of greater diversity in the world's executive offices.

References

Alexander, Deborah, and Kristi Andersen. 1993. Gender as a Factor in the Attribution of Leadership Traits. *Political Research Quarterly* 46 (3):527–45.

Barnes, Tiffany D., and Diana Z. O'Brien. 2018. Defending the Realm: The Appoint-

[104] Dube and Harish forthcoming; Schramm and Stark forthcoming.

[105] Nomikos and Sambanis 2019.

[106] Carlin, Carreras, and Love 2019.

[107] Naurin, Naurin, and Alexander 2019.

[108] See Naurin, Naurin, and Alexander 2019 for a prominent study of gender stereotypes in an elite sample.

ment of Female Defense Ministers Worldwide. *American Journal of Political Science* 62 (2):355–68.

Bashevkin, Sylvia. 2018. *Women as Foreign Policy Leaders: National Security and Gender Politics in Superpower America.* Oxford University Press.

Baturo, Alexander, and Julia Gray. 2018. When Do Family Ties Matter? The Duration of Female Suffrage and Women's Path to High Political Office. *Political Research Quarterly* 71 (3):695–709.

Bauer, Nichole M. 2015. Who Stereotypes Female Candidates? Identifying Individual Differences in Feminine Stereotype Reliance. *Politics, Groups, and Identities* 3 (1):94–110.

Bauer, Nichole M. 2017. The Effects of Counterstereotypic Gender Strategies on Candidate Evaluations. *Political Psychology* 38 (2):279–95.

Brutger, Ryan, and Joshua D. Kertzer. 2018. A Dispositional Theory of Reputation Costs. *International Organization* 72 (3):693–724.

Burgess, Diana, and Eugene Borgida. 1999. Who Women Are, Who Women Should Be: Descriptive and Prescriptive Gender Stereotyping in Sex Discrimination. *Psychology, Public Policy, and Law* 5 (3):665–92.

Caprioli, Mary, and Mark A. Boyer. 2001. Gender, Violence, and International Crisis. *Journal of Conflict Resolution* 45 (4):503–18.

Carlin, Ryan E., Miguel Carreras, and Gregory J. Love. 2019. Presidents' Sex and Popularity: Baselines, Dynamics and Policy Performance. *British Journal of Political Science*: 1–21.

Cialdini, Robert B., and Melanie R. Trost. 1992. Social Influence: Social Norms, Conformity, and Compliance. In *The Handbook of Social Psychology* edited by Daniel Todd Gilbert, Sustain T. Fiske, Gardner Lindzey, 151–92. McGraw-Hill.

Clayton, Amanda, Diana Z. O'Brien, and Jennifer M. Piscopo. 2019. All Male Panels? Representation and Democratic Legitimacy. *American Journal of Political Science* 62 (1):113–29.

Crisman-Cox, Casey, and Michael Gibilisco. 2018. Audience Costs and the Dynamics of War and Peace. *American Journal of Political Science* 62 (3):566–80.

Croco, Sarah E., and Scott Sigmund Gartner. 2014. Flip-Flops and High Heels: An Experimental Analysis of Elite Position Change and Gender on Wartime Public Support. *International Interactions* 40 (1):1–24.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. Information Equivalence in Survey Experiments. *Political Analysis* 26 (4):399–416.

Dolan, Kathleen. 2014. *When Does Gender Matter?* Oxford University Press.

Dube, Oeindrila, and S.P. Harish. Forthcoming. Queens. *Journal of Political Economy*.

Ellemers, Naomi. 2018. Gender Stereotypes. *Annual Review of Psychology* 69 (1):275–98.

Enloe, Cynthia. 1990. *Bananas, Beaches, and Bases: Making Feminist Sense of International Politics*. University of California Press.

Falk, Erika, and Kate Kenski. 2006. Issue Saliency and Gender Stereotypes: Support for Women as Presidents in Times of War and Terrorism. *Social Science Quarterly* 87 (1):1–18.

Fearon, James D. 1994. Domestic Political Audiences and the Escalation of International Disputes. *American Political Science Review* 88 (3):577–92.

Fearon, James D. 1995. Rationalist Explanations for War. *International Organization* 49 (3):379–414.

Fearon, James D. 1997. Signaling Foreign Policy Interests: Tying Hands versus Sinking Costs. *Journal of Conflict Resolution* 41 (1):68–90.

Ferraro, Geraldine, and Linda Bird Francke. 2004. *My Story*. Northwestern University Press.

Fraser, Antonia. 1990. *Warrior Queens: The Legends and the Lives of the Women Who Have Led their Nations in War*. Random House.

Fukuyama, Francis. 1998. Women and the Evolution of World Politics. *Foreign Affairs* 77 (5):24–40.

Gillespie, Caitlin C. 2018. *Boudica: Warrior Woman of Roman Britain.* Oxford University Press.

Goemans, Henk E., Kristian Skrede Gleditsch, and Giacomo Chiozza. 2009. Introducing Archigos: A Dataset of Political Leaders. *Journal of Peace Research* 46 (2):269–83.

Goldstein, Joshua S. 2001. *War and Gender.* Cambridge University Press.

Goodyear-Grant, Elizabeth. 2013. *Gendered News: Media Coverage and Electoral Politics in Canada.* University of British Columbia Press.

Hayes, Danny. 2011. When Gender and Party Collide: Stereotyping in Candidate Trait Attribution. *Politics and Gender* 7 (2):133–65.

Heilman, Madeline E. 1995. Sex Stereotypes and Their Effects in the Workplace: What We Know and What We Don't Know. *Journal of Social Behavior and Personality* 10 (6):3–26.

Heilman, Madeline E. 2001. Description and Prescription: How Gender Stereotypes Prevent Women's Ascent Up the Organizational Ladder. *Journal of Social Issues* 57 (4):657–74.

Heilman, Madeline E. 2012. Gender Stereotypes and Workplace Bias. *Research in Organizational Behavior* 32:113–35.

Holman, Mirya R., Jennifer Merolla, and Elizabeth Zechmeister. 2011. Sex, Stereotypes, and Security: A Study of the Effects of Terrorist Threat on Assessments of Female Leadership. *Journal of Women, Politics and Policy* 32 (3):173–92.

Holman, Mirya R., Jennifer L. Merolla, Elizabeth J. Zechmeister, and Ding Wang. 2019. Terrorism, Gender, and the 2016 US Presidential Election. *Electoral Studies* 61:1–8.

Horowitz, Michael C., Allan C. Stam, and Cali M. Ellis. 2015. *Why Leaders Fight.* Cambridge University Press.

Huddy, Leonie, and Nayda Terkildsen. 1993. Gender Stereotypes and the Perception of Male and Female Candidates. *American Journal of Political Science* 37 (1):119–47.

Jalalzai, Farida. 2013. *Shattered, Cracked, or Firmly Intact? Women and the Executive Glass Ceiling Worldwide*. Oxford University Press.

Jervis, Robert 1978. Cooperation Under the Security Dilemma. *World Politics* 30 (2):167–214.

Kahn, Kim Fridkin. 1992. Does Being Male Help? An Investigation of the Effects of Candidate Gender and Campaign Coverage on Evaluations of US Senate Candidates. *Journal of Politics* 54 (2):497–517.

Karim, Sabrina, Michael J. Gilligan, Robert Blair, and Kyle Beardsley. 2018. International Gender Balancing Reforms in Postconflict Countries: Lab-in-the-Field Evidence from the Liberian National Police. *International Studies Quarterly* 62 (3):618–31.

Kertzer, Joshua D., and Ryan Brutger. 2016. Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory. *American Journal of Political Science* 60 (1):234–49.

Klar, Samara. 2018. When Common Identities Decrease Trust: An Experimental Study of Partisan Women. *American Journal of Political Science* 62 (3):610–22.

Koch, Jeffrey W. 2000. Do Citizens Apply Gender Stereotypes to Infer Candidates? Ideological Orientations? *Journal of Politics* 62 (2):414–29.

Koch, Michael T., and Sarah A. Fulton. 2011. In the Defense of Women: Gender, Office Holding, and National Security Policy in Established Democracies. *Journal of Politics* 73 (1):1–16.

Kromer, Mileah, and Janine A. Parry. 2019. The Clinton Effect? The (Non)Impact of a High-Profile Candidate on Gender Stereotypes. *Social Science Quarterly* 100 (6):2134–47.

Lawless, Jennifer. 2004. Women, War, and Winning Elections: Gender Stereotyping in the Post-September 11th Era. *Political Research Quarterly* 57 (3):479–90.

Levendusky, Matthew S., and Michael C. Horowitz. 2012. When Backing Down Is the Right Decision. *Journal of Politics* 74 (2):323–38.

Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. *Innovative Higher Education* 40 (4):291–303.

Malhotra, Inder. 1991. *Indira Gandhi: A Personal and Political Biography.* Northeastern University Press.

McDermott, Monika L. 1997. Voting Cues in Low-Information Elections: Candidate Gender as a Social Information Variable in Contemporary US Elections. *American Journal of Political Science* 41 (1):270–83.

McDermott, Rose, Dominic Johnson, Jonathan Cowden, and Stephen Rosen. 2007. Testosterone and Aggression in a Simulated Crisis Game. *Annals of the American Academy of Political and Social Science* 614 (1):15–31.

McDermott, Rose. 2015. Sex and Death: Gender Differences in Aggression and Motivations for Violence. *International Organization* 69 (3):753–75.

McGarty, Craig, Vincent Y. Yzerbyt, and Russell Spears. (2002). *Stereotypes as Explanations: The Formation of Meaningful Beliefs About Social Groups.* Cambridge University Press.

McGlen, Nancy E., and Meredith Reid Sarkees. 1993. *Women in Foreign Policy: The Insiders.* Routledge.

Mummolo, Jonathan, and Erik Petersen. 2019. Demand Effects in Survey Experiments: An Empirical Assessment. *American Political Science Review* 113 (2):517–29.

Naurin, Daniel, Elin Naurin, and Amy Alexander. 2019. Gender Stereotyping and Chivalry in International Negotiations: A Survey Experiment in the Council of the European Union. *International Organization* 73 (2):469–88.

Nomikos, William G., and Nicholas Sambanis. 2019. What Is the Mechanism Underlying Audience Costs? Incompetence, Belligerence, and Inconsistency. *Journal of Peace Research* 56 (4):575–88.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined.* Penguin.

Post, Abigail S., and Paromita Sen. 2020. Why Can't a Woman Be More Like a Man? Female Leaders in Crisis Bargaining. *International Interactions* 46 (1):1–27.

Potter, Phillip B.K., and Matthew A. Baum. 2010. Democratic Peace, Domestic Audience Costs, and Political Communication. *Political Communication* 27 (4):453–70.

Regan, Patrick M., and Aida Paskeviciute. 2003. Women's Access to Politics and Peaceful States. *Journal of Peace Research* 40 (3):287–302.

Reiter, Dan. 2014. The Positivist Study of Gender and International Relations. *Journal of Conflict Resolution* 59 (7):1301–26.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G, Rand. 2014. Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science* 58 (4):1064–82.

Rosenwasser, Shirley Miller, and Norma G. Dean. 1989. Gender Role and Political Office: Effects of Perceived Masculinity/Femininity of Candidate and Political Office. *Psychology of Women Quarterly* 13 (1):77–85.

Rothschild, Jacob E., Adam J. Howat, Richard M. Shafranek, and Ethan C. Busby. 2019. Pigeonholing Partisans: Stereotypes of Party Supporters and Partisan Polarization. *Political Behavior* 41 (2):423–43.

Rudman, Laurie A., and Peter Glick. 2001. Prescriptive Gender Stereotypes and Backlash Toward Agentic Women. *Journal of Social Issues* 57 (4):743–62.

Sanbonmatsu, Kira. 2002. Gender Stereotypes and Vote Choice. *American Journal of Political Science* 46 (1):20–34.

Schelling, Thomas. 1960. *The Strategy of Conflict.* Harvard University Press.

Schramm, Madison, and Alexandra Stark. Forthcoming. Peacemakers or Iron Ladies? A Cross National Study of Gender and International Conflict. *Security Studies.*

Schultz, Kenneth A. 1998. Domestic Opposition and Signaling in International Crises. *American Political Science Review* 92 (4):829–44.

Sjoberg, Laura, and J. Ann Tickner. 2011. *Feminism and International Relations: Conversations About the Past, Present and Future*. Routledge.

Stiehm, Judith Hicks. 1982. The Protector, The Protected, The Defender. *Women's Studies International Forum* 5 (3–4):367–76.

Swim, Janet K., and Lawrence J. Sanna. 1996. He's Skilled, She's Lucky: A Meta-Analysis of Observers' Attributions for Women's and Men's Successes and Failures. *Personality and Social Psychology Bulletin* 22 (5):507–19.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth. 2018. The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics. *American Political Science Review* 112 (3):525–41.

Tickner, J. Ann. 1992. *Gender in International Relations: Feminist Perspectives on Achieving Global Security*. Columbia University Press.

Tickner, J. Ann. 1994. Why Women Can't Run the World: International Politics According to Francis Fukuyama. *International Studies Review* 1 (3):3–11.

Tomz, Michael. 2007. Domestic Audience Costs in International Relations: An Experimental Approach. *International Organization* 61 (4):821–40.

Trager, Robert F., and Lynn Vavreck. 2011. The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of the Party. *American Journal of Political Science* 55 (3):526–45.

Weeks, Jessica L. 2008. Autocratic Audience Costs: Regime Type and Signaling Resolve. *International Organization* 62 (1):35–64.

Table 1: *Inconsistency cost predictions vs. male-male dyad*

		Foreign Leader	
		Female	Male
Domestic Leader	Female	Greater Inconsistency Costs	Greater Inconsistency Costs
	Male	Greater Inconsistency Costs	Baseline

Table 2: *Belligerence cost predictions vs. male-male dyad*

		Foreign Leader	
		Female	Male
Domestic Leader	Female	Lower Belligerence Costs	Lower Belligerence Costs
	Male	Lower Belligerence Costs	Baseline

Table 3: *Percentage point difference in mean disapproval compared to the male-male baseline*

Dyad (Baseline = Male-Male)	*Audience Cost (%) (1)*	*Inconsistency Cost (%) (2)*	*Belligerence Cost (%) (3)*
Female-Male	+6.3	+20.7***	-14.4**
	(31.6 – 25.3)	(61.9 – 41.2)	(-30.3 – -15.9)
Female-Female	+4.6	+18.2***	-13.6**
	(29.9 – 25.3)	(59.4 – 41.2)	(-29.5 – -15.9)
Male-Female	+4.6	+15.4**	-10.8*
	(29.9 – 25.3)	(56.6 – 41.2)	(-26.7 – -15.9)

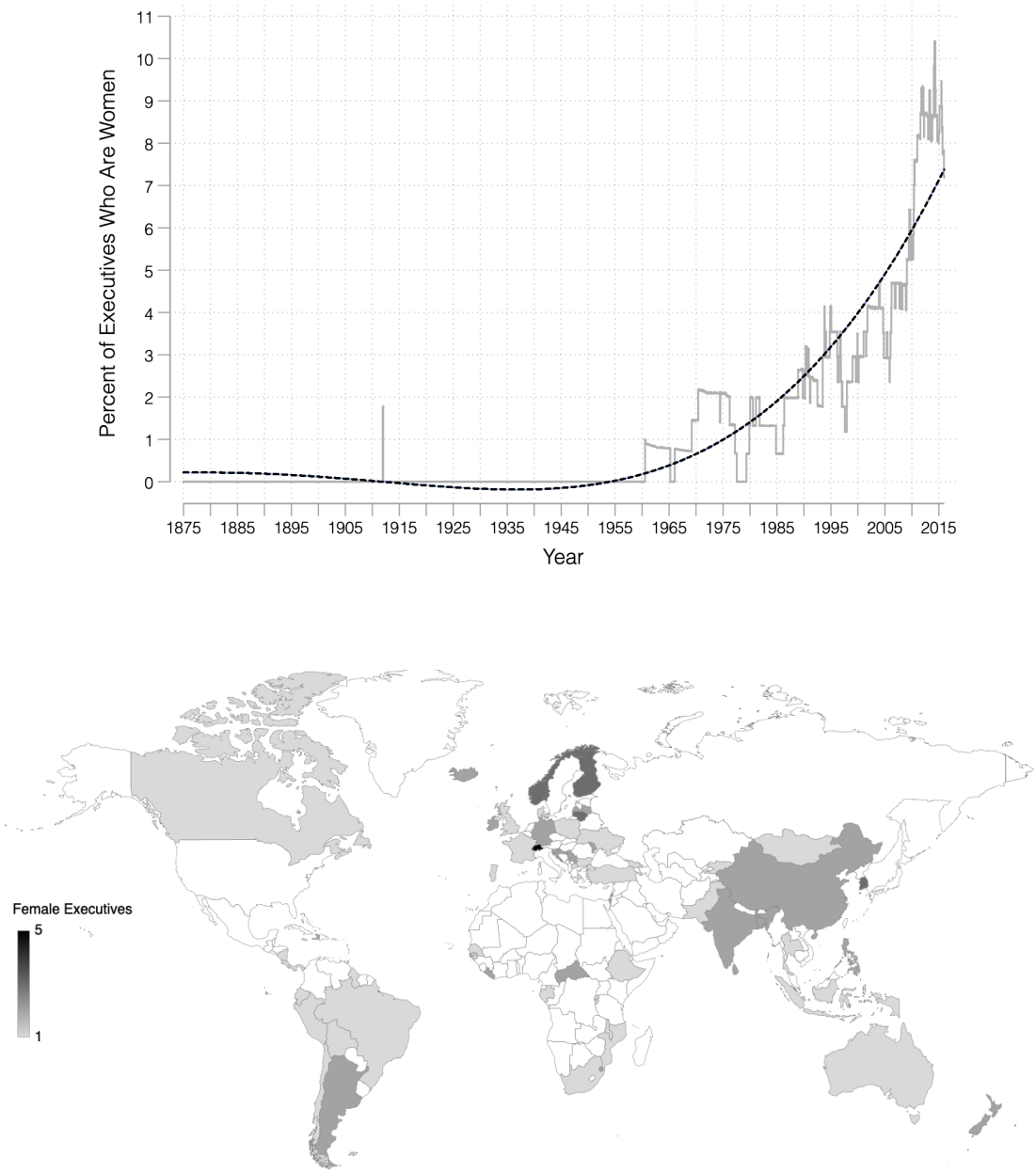
Notes: Results depict average treatment effects (ATE) for a binary measure of disapproval calculated from 2,000 bootstraps. The main quantities reflect the average percentage point difference in disapproval for the respective dyad in the left column compared to the male-male baseline. For example, 20.7 percentage points more respondents disapprove of a female president acting inconsistently against a foreign male leader than a male president acting inconsistently against a foreign male leader. Mean disapproval for the two experimental groups used to calculate ATE are in parentheses. For example, average disapproval of a female president behaving inconsistently against a foreign male leader was 61.9%, while average disapproval of a male president behaving inconsistently against a foreign male leader was 41.2%. * $p<.10$; ** $p<.05$; *** $p<.01$.

Table 4: *Percentage point difference in mean negative sentiment compared to the male-male baseline*

Dyad (Baseline = Male-Male)	*Audience Cost (%) (1)*	*Inconsistency Cost (%) (2)*	*Belligerence Cost (%) (3)*
Female-Male	+5.1	+14.0***	-8.9*
	(27.9 – 22.9)	(56.8 – 42.8)	(-28.9 – -20.0)
Female-Female	-1.9	+8.2*	-10.1*
	(21.0 – 22.9)	(51.0 – 42.8)	(-30.0 – -20.0)
Male-Female	-3.3	+8.9*	-12.2**
	(19.6 – 22.9)	(51.7 – 42.8)	(-32.1 – -20.0)

Notes: Results depict average treatment effects (ATE) calculated from 2,000 bootstraps. The main quantities reflect the average percentage point difference in negative sentiment for the respective dyad in the left column compared to the male-male baseline. For example, negative sentiment was 14 percentage points higher for a female president acting inconsistently against a foreign male leader than a male president acting inconsistently against a foreign male leader. Mean negative sentiment for the two experimental groups used to calculate ATE are in parentheses. For example, average negative sentiment of a female president behaving inconsistently against a foreign male leader was 56.8%, while average negative sentiment of a male president behaving inconsistently against a foreign male leader was 42.8%. * $p<.10$; ** $p<.05$; and *** $p<.01$.

Figure 1: *Female leadership is becoming more common over time and across countries*



Notes: The top panel shows the daily share of female executive officeholders worldwide between 1875 and 2015 (in gray), along with a fractional polynomial trend in the share of female executives (dashed in black). The bottom panel depicts a heat map of countries that were led by women from 1875 to 2015, shaded by the number of female executive officeholders a country had in that period. Data on female executive officeholding come from Archigos (Goemans, Gleditsch, and Chiozza 2009), LEAD (Horowitz, Stam, and Ellis 2015), and Jalalzai (2013).

Authors

Joshua A. Schwartz is a PhD candidate in Political Science at the University of Pennsylvania in Philadelphia. He can be reached at josha@sas.upenn.edu.

Christopher W. Blair is a PhD candidate in Political Science at the University of Pennsylvania in Philadelphia. He can be reached at cwblair@sas.upenn.edu.

Acknowledgments

This is one of several joint articles by the authors; the ordering of names reflects a principle of rotation with equal authorship implied. We thank Diana Mutz, Shira Pindyck, Dawn Teele, Ryan Brutger, Jonathan Chu, Jamie Druckman, Jeremy Freese, Michael Horowitz, Nicholas Sambanis, Dustin Tingley, Alex Weisiger, participants at the 2019 Harvard Experimental Political Science Conference, two anonymous TESS reviewers, two anonymous *International Organization* reviewers, and the editors and staff of *International Organization* for helpful comments and advice.

Funding

Generous support for this research was provided by the Christopher H. Browne Center for International Politics at the University of Pennsylvania and Time-Sharing Experiments in the Social Sciences (TESS). Some of the data were collected by Time-Sharing Experiments for the Social Sciences, NSF Grant 0818839, Jeremy Freese and James Druckman, principal investigators. This research was approved by the University of Pennsylvania Institutional Review Board (IRB Protocol #832589).

Key Words

Audience costs; gender; female leaders; crisis bargaining; credibility; inconsistency costs