Gender Differences in Economics Seminars*

Pascaline Dupas Amy Handlan Alicia Sasser Modestino Muriel Niederle
Mateo Seré Haoyu Sheng Justin Wolfers†
and the Seminar Dynamics Collective‡

May 8, 2025

Abstract

We assess whether men and women are treated differently when presenting their research in economics seminars. We collected data on every interaction between presenters and audience members across thousands of seminars, job market talks and conference presentations, leveraging both human judgment and audio processing algorithms to measure the number, tenor, tone and type of interruptions. Within a seminar series, women are interrupted more than men, and this finding holds when controlling for characteristics of the presenter and their paper topic and for audience size. Interruptions that may not be favorable to the presenter, such as those that are negative in tenor or tone, or cutoff the presenter mid-sentence, are common occurrences in economics seminars, and increase for women presenters. We also find greater engagement with female presenters in the form of larger, more diverse audiences, suggesting a potential role model effect.

JEL codes: A1, C8, C45, J4, J7

Keywords: academic environment, audio analysis, audio processing, differential treatment, economics profession, gender, interruptions, machine learning, seminar culture.

*Special thanks to Jim Poterba for providing access to all of the sessions at the 2019 NBER Summer Institute and to recordings of the 2022 NBER Summer Institute. Our 2019 human data collection protocols were reviewed by the Institutional Review Boards at Northeastern University and NBER and granted "exempt" status under category 2. Our 2022 protocols were reviewed by the Institutional Review Boards at Brown University and NBER and granted "exempt" status under category 4. The authors would like to acknowledge research funding from Stanford University's Department of Economics. We thank Stefano DellaVigna, Erzo Luttmer, anonymous referees, and many seminar audiences for highly valuable feedback and input. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

†Dupas: Princeton and NBER (pdupas@princeton.edu), Handlan: Brown (amy_handlan@brown.edu), Modestino: Northeastern (a.modestino@northeastern.edu), Sheng: Brown (haoyu_sheng@brown.edu), Niederle: Stanford and NBER (niederle@stanford.edu), Seré: University College London (m.sere@ucl.ac.uk), Wolfers: Michigan and NBER (jwolfers@umich.edu).

‡The Seminar Dynamics Collective is a group of 103 members of the economics community who collected and/or processed the human-coded data used in this study and its members are considered co-authors of this paper. These coauthors include: Corinne Andriola, Victoria Barone, Maryam Blooki, Stephanie Bonds, Nina Buchmann, Drew Burd, Anne Burton, Mrinmoyee Chatterjee, Vittoria Dicandia, Maria Dieci, Karl Dunkle Werner, Holly Dykstra, Luciana Etcheverry, John Fallon, Camille Falézan, Valeria Ferraro, Ellen Fu, Chelsea Garber, Shresth Garg, Anomita Ghosh, Laurie Hakes, Hyoyoung Han, Emma Harrington, Juan Herreño, Kelsi G. Hobbs, Lakshita Jain, Amna Javed, Michelle Jiang, Ariadna Jou, Catherine Michaud Leclerc, Domininkas Mockus, Erica Moszkowski, Philip Mulder, Tuan Nguyen, Urbashee Paul, Dev Patel, Grace Phillips, Xuechao Qian, Rizwanur Rob, Monica Rodriguez, Fernanda Rojas, Arvind Sharma, Rachel Schuh, Rachel Sederberg, Cory Smith, Rizki Nauli Siregar, Melissa Spencer, Anna Stansbury, Ishaana Talesara, Carly Trachtman, Francesca Truffa, Silvia Vannutelli, Joanna Venator, David N. Wasser, Melanie Wallskog, Ashley Wong, Alice Wu and 46 other coauthors who have chosen to remain anonymous.

1 Introduction

Are similarly situated men and women treated differently when presenting their economics research?[1] This question is important given both the distinctively interactive (some would say aggressive) culture of economics seminars, and the continuing under-representation of women among the senior ranks of the economics profession (Lundberg and Stearns, 2019; AEA, 2021).

We report the findings of an ambitious data collection effort in which we coded data on every interaction between a presenter and their audience across thousands of seminars, workshops, conferences, and job market talks. Our rich microdata—some of which were hand coded in real time and some of which rely on machine learning algorithms applied to audio recordings—document the time, duration, type, tenor and tone of what we call "interruptions": questions or remarks made by audience members who may have raised their hand or simply interjected. We also record, for each interruption, the gender (and other characteristics) of the audience member. And we collect a rich set of controls that allow us to account for the characteristics of the presenter, the research they are presenting, their field, and their audience.

The data paints a comprehensive picture of economics seminar culture across five separate samples—each with its strengths (and idiosyncrasies). Overall, audience interruptions are very common (35 on average during job talks, i.e., one interruption every 2.5 minutes). They are often *remarks* (e.g., suggestions, comments, criticisms), rather than questions. They often interrupt the presenter mid-sentence. They are rated by a machine-learning algorithm as having a negative tone of voice about half the time. And when we used graduate students to code the tenor of the interruption, we found that seminars typically have a couple of openly hostile interruptions from audience members.

In this context, we test for differences in how men and women are treated when presenting their research. Our analyses show that women typically receive about 10 to 20 percent more interruptions than men in economics seminars, and that this holds true even when accounting for a range of control variables. Importantly, this disparity is evident during recruitment seminars (job talks), which shape the future composition of the economics profession. This holds for in-person as well as online seminars, and we observe similar gender gaps in data collected by fellow economists and data coded by computer-based audio processing algorithms.[2] We also show that

[1] The present paper draws together research previously circulated in working paper form by Dupas et al. (2021), Handlan and Sheng (2023), and Seré (2023).

[2] We separately analyze talks as seminar presentations or conference presentations. We find that women do not receive significantly more interruptions than men at NBER SI conferences or online conferences.

female speakers tend to have larger seminar audiences, and in particular, more women—suggesting a potential role-model effect. This would suggest that women speakers generate more interest and engagement, which could be a benefit both to them and to other women in the profession. However, greater and more diverse audiences do not explain the greater number of interruptions, since the results hold when we control for audience size and composition.

The finding that women are interrupted more often than men in economics seminars raises two questions. First, is this evidence of disparate treatment, or does it reflect differences in what they present or other confounders correlated with gender? We argue that the latter explanation seems unlikely, since we are able to control for a host of covariates, including the paper's topic (JEL codes) and type (whether it includes theory, empirical analysis, experimental design, or policy evaluation), and since female presenters in the sample appear neither positively nor negatively selected.

Second, what explains this disparate treatment? There are two potential explanations. On the one hand, it could indicate a bias *in favor* of women: one could argue that more engagement is a sign of greater interest (Simonsohn, 2021), and too little engagement is a concern. Research seminars provide an opportunity to receive valuable feedback, and interruptions in the form of suggestions can be welcome. Alternatively, it could indicate a bias *against* women: one could argue that the decision to interrupt to make suggestions or question the analysis reflects a more negative pre-judgment of the presenter's credibility or authority. The latter perspective aligns with findings in the literature on conversational dynamics, where interruptions—especially when frequent or overlapping—are interpreted as signals of dominance or reduced conversational status (e.g., Kendrick and Torreira, 2015; Tannen, 1993). It is also echoed in the defense of economics' "interrupting culture" in a 2009 *The Economist* article: "The difference between interrupting and non-interrupting cultures is not a simple and arbitrary choice of social norm, but instead reflects a judgment about whose words are likely to be most valuable to hear."[3] According to this view, the fact that female economists are interrupted more often would imply that their words are not judged as valuable as those of male economists. Note that the two biases may coexist within an audience, both contributing to women receiving more interruptions. To shed light on the relative importance of these two potential effects, we move beyond analyzing the number of interruptions to a broader assessment of their type, tenor, tone and tempo.

Although there are mixed indicators, the data highlights some concerning patterns. First, in

[3]This quote is en economist's response to the following comment by a sociologist: "The economist engaged in the usual norms for his own department's culture: interrupting at pretty much every slide."

the human-coded samples, we find that the additional interruptions women receive are disproportionately more likely to be rated as *negative* in tenor (e.g., patronizing or hostile): while the total number of interruptions increases by 11 percent for women presenters, they see a 43 percent increase in the number of interruptions rated by human coders as negative in tenor (from just above 1 per seminar, to 1.5), and no increase in the number of interruptions rated as supportive. Importantly, this is driven by both men and women in the audience. Additionally, our results for the online sample indicate that female presenters tend to experience relatively more interruptions characterized by overlapping speech—where the audience talks at the same time the presenter—than their male counterparts. This result is driven entirely by male audience members. These potentially disquieting results coexist with results whose implications are less clear-cut. For example, the tone results in the computer-coded sample indicate that women receive the same balance of positive and negative interruptions as men do—i.e., they receive more of both. Finally, women receive more interruptions of all types (e.g., comments, clarifying questions, suggestions) in equal proportions. Given that many interruptions are suggestions, women receive *more* suggestions than men. However, it is unclear if these are beneficial or not because suggestions may include helpful feedback but they could also be counterproductive distractions.

Whether this differential treatment is favorable for women or not is something our observational data cannot answer definitively, but we offer a few reflections. First, our data reveal that, for both male and female presenters, economics audiences are quite challenging to manage, with a large number of less-than-collegial interruptions and a few clearly hostile ones. Even just one hostile interruption in a seminar early on in one's career (e.g., on the job market) could have a demoralizing effect, and if so, then our findings that women receive more such interruptions could be detrimental. A similar argument could be made for mid-sentence interruptions. The literature on conversation dynamics shows how the prevalence of interruptions during spoken interactions can hinder effective communication and be perceived as impolite or disrespectful (Kendrick and Torreira, 2015; Tannen, 1993). While some degree of overlapping speech is a natural aspect of group discussions, the excessive or prolonged nature of such interruptions, particularly when disproportionately directed towards a specific group, can adversely affect the quality and flow of academic discourse.

Taken together, these patterns suggest that the increased engagement observed in seminars may not be entirely benign. Yet, our data also show that women presenters encourage higher attendance of *female* participants. And some of the additional interruptions of women presenters come from women. This suggests a potential role model or encouragement effect similar to the additional

service women take on when mentoring junior women: it may be easier for women to engage in professional discourse in the presence of other women during economics seminars.

Overall, our results suggest that audience members may react differently to female presenters, in ways that are both positive and negative, making it useful to distinguish between two related research questions: disparate treatment versus disparate impact. Our analysis focuses on the question of *disparate treatment*, asking whether women are treated differently than men within otherwise similar academic contexts. As such, this research can also be read as something of a progress report on whether economists are living up to the ideals laid out in the recently adopted Code of Professional Conduct requiring "civil and respective discourse ... [where] each idea is considered on its own merits ... [in] a professional environment with equal opportunity and fair treatment for all." We distinguish this from an alternative but equally important question, which is whether the distinctive seminar culture within economics—even if applied with equal force to men and women—has a *disparate impact* on women economists. Boustan and Langan (2019) address this latter question, arguing that even a seminar culture that treats men and women equally aggressively may have a disparate impact, if women find this aggressive culture less welcoming than men do.

Finally, we want to acknowledge that part of our data reflects the efforts of an unusual collaboration with the Seminar Dynamics Collective, a group of (mainly) graduate students who volunteered to study seminar dynamics by collecting and coding relevant data in real-time. It was logistically infeasible to recruit "blinded" coders, and so we note that our collaborators are a self-selected group of experts, able to follow economic discourse, but who may also care about gender equity to some degree. To allay concern that this may bias our findings, we also explored the unique research opportunity offered by the COVID-19 pandemic, when many seminars and conferences were moved online. We collected separate samples of these online seminars and used sophisticated audio processing algorithms to analyze audio files from a large number of presentations. These algorithms provided a scalable approach to coding seminar dynamics that could both enforce a uniform coding scheme across interactions and also be implemented by agents that were blind to our research question. Yet machines are not yet sophisticated enough to assess the *intent* of what was said, which depends on context beyond the audio recording or from outside the seminar setting. Thus, our computer- and human-based coding each have different strengths and weaknesses, but together they help us paint a useful picture of how presenters are treated in economics seminars, and how that differs by presenter gender.

The rest of this paper is organized as follows. Section 2 situates our contribution within the

existing literature. Section 3 describes how the various datasets were created and Section 4 describes the analytical framework. The heart of the paper follows, where Section 5 analyzes gender differentials in the number of interruptions and Section 6 explores both *who* interrupts and *how*. Section 7 concludes. Online appendices provide further detail on our samples, control variables, and robustness checks.

2 What Do We Know About Gender in Academic Settings?

2.1 On the Differential Treatment of Female Economists

Our research adds to an emerging literature that has documented the many margins on which women economists are treated differently—and often worse—than similarly situated men. Sarsons (2017) and Sarsons et al. (2021) find that women economists receive less credit than their male co-authors when assessed for tenure and promotion. Koffi (2021) finds papers published in top economic journals by women are less likely to be cited by top-tier journals and less likely to be cited by men.[4] Card et al. (2022) find that journal editors and referees are more likely to reject papers written by women economists than if they were seeking to maximize citations. Hengel (2022) finds that women economists experience longer turn-around times and more stringent writing requirements from journal reviewers. Costello et al. (2022) compare working papers with published manuscripts, finding that the publication process pushes women economists to add more hedging words and to express greater uncertainty about their findings. Finally, Wu (2020) finds that women economists have been systematically trivialized and even sexualized in online forums.

Each of these factors likely contributes to the finding by Chen et al. (2022) that women economists are less likely to be promoted than men, and also to Ginther and Kahn (2004, 2021) finding that women are less likely to be promoted in economics than in other academic disciplines. But not all gender differences disfavor women, and elections provide a useful counterpoint. For example, Donald and Hamermesh (2006) find that women are more likely to be elected to be office holders of the American Economic Association. Card et al. (2022) document that for over 40 years women were much less likely to be elected as Fellows of the Econometric Society, but this gap disappeared in the 1980s, and women are now significantly more likely to be elected (a pattern

[4] In a subsequent paper, Koffi (2025) finds that female-authored papers (at 16 economics journals) are less likely to be cited, but this effect goes away when publication history is controlled for. Koffi (2025) also finds that citation omissions are less pronounced for women with names perceived as masculine rather than feminine, and more pronounced for men with names perceived as feminine rather than masculine, suggesting that citations may be influenced by perceived author gender.

that became even more pronounced after the Society's Nominating Committee was given an explicit mandate to nominate highly qualified women in 2012). A very similar pattern was found for elections to prestigious scientific bodies like the American Association for the Advancement of Science (AAAS) and National Academy of Sciences (NAS)—something the authors of the study conjecture could stem from recognition within these institutions that female scientists could face higher publication barriers and receive less credit for their work (Card et al., 2023).

What is missing from this brief review of the literature is any systematic analysis of the academic seminars where reputations are often forged, and of the job talks that launch fledgling careers. Our paper seeks to fill that gap. The recent AEA climate survey (Allgood et al., 2019) suggests that this is a fruitful area of inquiry: nearly half (47 percent) of female respondents reported that they had not presented their question, idea, or view at their school or place of work to avoid possible harassment, discrimination, or unfair or disrespectful treatment, compared to 24 percent of male respondents. Similarly, 46 percent of women versus 18 percent of men said they had "not spoken at a conference or during a seminar presentation" to avoid those types of experiences. Boustan and Langan (2019) confirm that economics seminar culture does indeed matter. Their structured interviews at top economic departments lead them to conclude that "departments with better relative outcomes for women are reported to have a less aggressive and more constructive climate in their research seminars."

Other recent studies have focused more closely on seminars as a potential source of gender disparity in economics. Doleac et al. (2021) document the share of invited seminar speakers by gender and under-represented minority status across a balanced panel of 66 leading economics departments between 2015 and 2019, revealing that less than one quarter of talks were given by women. Chari and Goldsmith-Pinkham (2017) document a similar disparity in the share of papers presented at the NBER Summer Institute that were (co-)authored by women, with greater under-representation within finance and macroeconomics. While some of this literature looks beyond gender to also highlight differences in the experience of other under-represented groups, the present study only analyzes gender.[5]

[5]Beyond gender, most under-represented groups—whether by race, ethnicity, sexuality, or disability status—are so under-represented at top economics departments that we lack the statistical power in our samples to say anything of interest, much less draw any substantive conclusions.

2.2 On the Dynamics of Academic Discourse

Studies of fields outside of economics have also examined how women and men are treated differently when presenting their research, although none appear to be as large or systematic as the present study. Blair-Loy et al. (2017) analyze videotapes of job talks across five engineering departments at two R1-designated universities, finding that women received more questions, more follow-ups, and that more of their presentations were consumed by audience speech. Davenport et al. (2014) analyze presentations at the annual meetings of the American Astronomical Society, finding that women were asked slightly more questions than men.

Observational studies and self-reported data also reveal gender disparities in the motivation of the *interrupter*. Across a diverse set of academic disciplines and geographies, multiple studies have found that men in the audience ask more questions than women, whether in absolute numbers, on a per capita basis, or even with a gender-balanced audience (Davenport et al., 2014; Hinsley et al., 2017; Salem et al., 2021; Jarvis et al., 2022). Carter et al. (2018) surveyed a convenience sample of the academic community across a range of fields, finding that men were twice as likely as women to report being motivated to ask a question because they felt they spotted an error. Jarvis et al. (2022) conducted a six-month follow-up survey of conference attendees, revealing that women reported less comfort in asking questions, citing a fear of backlash and feelings of anxiety.

Other studies suggest that seminar dynamics are somewhat path-dependent and can be shaped. Carter et al. (2018) and Salem et al. (2021) found that women asked relatively fewer questions when a man asked the first question, or when there were fewer questions overall. Salem et al. (2021) also report that an intervention designed to increase female participation—encouraging the selection of female chairs and instructing all session chairs to offer the opening question to female audience members—led to greater participation from women.

While there have been few studies on the occurrence of interruptions outside of conference settings, an analysis of Congressional hearing transcripts by Miller and Sutherland (2022) revealed that women were more likely to be interrupted than men (including by other women). Jacobi and Schweers (2017) reviewed oral arguments at the Supreme Court and found that female justices were interrupted more frequently by both their fellow justices and male lawyers. Finally, Bisbee et al. (2025) showed that when Janet Yellen was Chair of the Federal Reserve, legislators interrupted her more often and more aggressively than they had interrupted the male Fed Chairs who preceded or succeeded her.

3 Data collection

The present research reflects the efforts of three independent data collection efforts by different subsets of the authors who were initially working independently. As with any data collection effort, each team faced a trade-off between convenience and representativeness, and each faced an array of logistical, legal, and ethical constraints. That, in turn, led each team to collect different samples using different methods, each of which reflects a good-faith effort to represent the discourse among academic economists. The result is a collection of separate and individually informative samples, and we present our findings from each sample separately, as each sample has different strengths and weaknesses. Collectively, these multiple samples speak with more power than any individual sample.

3.1 Logistical, Ethical, and Regulatory Issues

We initially considered videotaping or audio recording in-person seminars ourselves, but quickly learned that this would require gathering explicit (opt-in) permission from those whose comments would be recorded. This presented formidable logistical constraints. The first few economics departments we approached for permission perceived only downside risk for their individual departments from cooperating with such a study, even as they conceded that such research would be valuable for the broader profession.

This led us to pursue two strategies. Our first strategy was to collect data in real-time using human coders during in-person seminars pre-COVID in 2019. Because seminar attendance was only open to invitees, this required recruiting collaborators who would be invited as regular attendees. As such, our human-coded data was collected as part of an unusual collaboration with the Seminar Dynamics Collective, a group of (mainly) economics doctoral students who volunteered to collect and code data from the department seminars and job talks that they were invited to attend. The advantage of these data is that they were coded by people who are part of the economics profession and understand the idiosyncrasies of economic discourse. The disadvantage is that these data were coded by collaborators who were aware of (and indeed, volunteered to help explore) the research question, raising some concern about potential coding biases.

Our second strategy became possible during the COVID-19 pandemic when much of the economics discourse was pushed online in 2020 and 2021, solving the prior logistical challenge of recording seminars. That disruption created a ready stream of online and hybrid seminars and

conferences—and hence audio files—for our analysis. In addition, we could now rely on purely objective computer algorithms, rather than human coders, by exploiting recent advances in machine learning and audio processing to transform these audio files into a dataset of interruptions during economics seminars.

We describe both approaches in much greater detail below. For now, we note that we pursued these approaches because they met our goals of providing useful data, being feasible, and also meeting ethical and regulatory obligations.[6] Because of our additional commitment to privacy protection, we did not record the identities of audience members, only their gender. We also did not reveal the identities of individual departments or programs when reporting results. As an added precaution, we protected the identities of our human coders (even from each other) unless authorized to reduce the likelihood of retaliation (which was a concern several expressed).

For regular seminars and job market talks, the only personally identifiable information we collected was that of the presenter, including their name and the title of their talk, both of which were publicly available online. Even then, we only used this information to create variables that allow us to account for the influence of presenter and paper characteristics (such as home institution of the presenter and JEL code of the paper), and to link to other public information (such as author citation counts and paper publication outcomes). Because a speaker's voice may also be identifiable, we constructed measures of the auditory patterns for the machine learning approach that cannot be used to identifiably reverse-engineer the voice recording.

For the NBER Summer Institute, our data were collected under two separate cooperative agreements with the NBER in 2019 and 2022. The NBER required even stronger protections than the ones we followed for the regular seminar and job market talks. In addition to not recording the identities of audience members, workshop participants were given advance notice that their participation was being studied. Although we had pledged not to reveal the identities of the presenters or the specific NBER "program" (effectively "conference track"), we were not allowed to collect any individually-identifiable data, even to generate control variables. For the 2019 in-person conference, the NBER additionally asked that our human coders not record potentially sensitive information

[6] Our data collection for the in-person and hybrid seminars was considered "exempt" under Institutional Review Board (IRB) guidelines—with no need to obtain informed consent from seminar participants. This is because there is typically no expectation of privacy among seminar attendees, who often take notes—including on what other participants say, and our research is simply a more structured form of such data collection. Our data from online seminars were collected in Europe and simply involved downloading publicly available content from YouTube. Because this research was conducted for academic purposes, it complies with Article 89 of the General Data Protection Regulation (GDPR), which permits the processing of personal data for scientific research, provided that appropriate safeguards, such as data anonymization, are implemented.

such as whether a comment was deemed negative ("patronizing," "disruptive," "demeaning," or "hostile"). For the 2022 hybrid conference, the NBER additionally asked that our audio recordings be deleted two weeks after their creation to align with when the Summer Institute recordings were removed from YouTube. While we are grateful to the NBER for granting us permission to collect the Summer Institute data, these limitations constrained our ability to harmonize our approach across the conference and department seminar samples.

3.2 Samples

3.2.1 U.S. In-Person Job Market Seminars (Human Coded)

In the first half of 2019, members of the Seminar Dynamics Collective collected data on a sample of job talks at leading U.S. universities. Our sample of economics departments is probably best described as a directed convenience sample: our search for collaborators was driven by a desire to cover the "top thirty" economics departments, but wherever we found willing collaborators, we happily worked with them. And so this sample includes (but is not restricted to) twenty of the "top thirty" economics departments. The resulting sample includes 176 job market talks involving 80 job applicants (31 women and 49 men) across 26 universities.

3.2.2 U.S. In-Person Department Seminars (Human Coded)

In addition to collecting data on job market seminars, the Seminar Dynamics Collective also collected data from regular departmental seminar series. In a typical case, an individual coder would volunteer to code a particular seminar series of interest for the whole semester (such as the Stanford Development Seminar). We collected data on 245 presentations involving 220 unique external speakers (64 women and 156 men) across 46 seminar series from 25 universities.

3.2.3 NBER Summer Institute Presentations, 2019 (Human Coded)

This sample is the near-universe of all presentations made at the 2019 NBER Summer Institute, a leading invitation-only economics conference. In 2019, the conference was held in person, with no remote option. As such, data were collected by members of the Seminar Dynamics Collective who were recruited specifically for this effort. The Summer Institute includes a broad array of (mostly applied) fields, with a robust representation of macroeconomics (and very little pure economic or econometric theory). This sample includes 443 talks, collected across 48 of the 51 NBER program

meetings held that year.

3.2.4 Online Talks (Machine Coded)

This sample reflects an exhaustive attempt at coding data in a broad set of online seminars, workshops and conference presentations that occurred in the wake of the COVID-19 pandemic, which shifted a lot of economic discourse online. The key sample inclusion criteria were that the talk occurred online, and that it be organized or sponsored by an American or European economics department or highly regarded research institution (such as the AEA or CEPR) (Seré, 2023). To ensure accurate measurement, we exclude hybrid talks, talks where a moderator aggregates questions, talks with discussants, and talks of non-standard duration (under 20 minutes or over 110 minutes). This process resulted in a final sample of 994 talks across 86 distinct online series.

3.2.5 NBER Summer Institute Presentations, 2022 (Machine Coded)

This sample includes presentations made at the 2022 NBER Summer Institute, which was held in hybrid format. All presentations were recorded and broadcast over YouTube (and were then subsequently deleted two weeks later). These recordings were not publicly downloadable, but were made available to us under a special arrangement with the NBER (Handlan and Sheng, 2023). A small number of sessions opted out of the study and we did not receive recordings for presentations in those sessions. This sample yields a total of 467 paper presentations, across a broad range of (predominantly applied) fields. For some analysis we need sufficiently good audio quality of the audience. In those specifications we use a smaller sample of 257 presentations which we verified to have sufficiently high-quality audio from audience members. Finally, our agreement with the NBER required that we delete all of these audio files within two weeks.

3.2.6 Summarizing the Five Samples

Table 1 summarizes the characteristics of our five (non-overlapping) samples arranged chronologically. Note first that the share of talks with female presenters in each sample ranges from 28 percent to 49 percent, larger than the share of women in the profession. We are therefore well-powered to study differences by presenter gender.

Beyond differences in coding (human vs. machine), there are four other important dimensions of heterogeneity across our samples. First, there is the type of presentation, which could be either a seminar (where questions and answers are free-flowing) or a conference presentation (which tends

to be both shorter and more formal). Second, there is the format, which varies between in-person seminars (with no online option), hybrid presentations (presented simultaneously to both an in-person and online audience), and online-only seminars. Third, there is the professional risk, which is likely considerably higher for job market talks. Fourth, there are the data collection limitations imposed by the NBER that create differences in the control variables that could be constructed.

Table 1: Summary Statistics

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022
Panel A: Descriptive details					
Coded by:	Humans	Humans	Humans	Computer	Computer
Type	Seminars	Seminars	Conf. talks	Seminar & conf. talks	Conf. talks
Mode	In person	In person	In person	Online only talks	Hybrid talks
Timing	Spring 2019	Spring 2019	Summer 2019	2020-2023	Summer 2022
Location	U.S.	U.S.	U.S.	U.S. & Europe	U.S.
Unique talks	176	245	442	994	467
Unique presenters	80	220	n.a.	904	n.a.
Field info	Yes	Yes	Coarse	Yes	Coarse
Paper type	Yes	Yes	No	Yes	No
Panel B: Summary statistics, by sample					
No. of interruptions	35.08	26.32	14.38	11.60	13.56
Avg. durations (mins)	84.41	79.73	38.65	62.67	45.75
No. of interruptions per hour	24.98	19.90	22.01	10.67	17.80
Female presenter	0.49	0.32	0.28	0.32	0.34
Observations	240	292	447	994	467

Notes: See Appendix C for detailed definitions of variables. Column 2: "Regular seminars" are presentations by external speakers in regular, departmental seminar series.

The average presentation duration varies quite substantially across these different samples representing quite different formats. The average in-person seminar is around 80-90 minutes, while NBER presentations are much briefer (and more variable), at 40 minutes. Online seminars are in the middle, at around an hour. This, in turn, shapes the number of interruptions. While this number varies quite sharply across our samples, when considered as a rate—say, interruptions per hour—the samples are markedly more similar.

Finally, the period over which these talks were given spans five years—from early 2019 to late 2023. Analyses of the first three (human-coded) samples were released in 2021, and audience members were possibly familiar with those when the later talks took place. It is therefore possible that the treatment of women is heterogeneous across our samples because it *changed* over time (we test this formally in Section 6.3), or because it changes when the issue is made salient. This is particularly relevant for the latest sample, the NBER SI 2022. Not only had the earlier results been

released by then (as an NBER working paper in 2021 and presented at NBER SI 2022), audience members at NBER SI 2022 were informed that talks were being recorded so that researchers could study audience behavior. Specifically, Jim Poterba included the following information in the email that he sent to all NBER participants on July 7, 2022: "*Audio recordings from the YouTube postings are also going to be transformed to anonymized data by a research team that is carrying out statistical analysis of conference dynamics. Please remember that all participants in NBER meetings, in-person or virtual, are expected to comply with the NBER's conference code of conduct.*" This may have created a tempering effect.

3.2.7 Presenter and Paper Characteristics by Gender

For three of the five samples (job market talks, regular seminars and online seminars), we were able to collect basic information about the presenter. Table A2 shows that presenter characteristics are not systematically different by gender. In the first sample, we proxy for paper quality with publication status—was the paper R&R at a top journal or published at a top journal—and we find that the job market papers presented by men and women were of equal quality.

In both regular and online seminars, male speakers have higher citation counts on average, but this is likely due to seniority (the share of women is higher among more recent PhD cohorts). We do not have information on seniority in the regular seminar sample, but in the online seminar sample, male presenters have three more years of "experience" measured as the number of years since the presenter's first publication.

We show characteristics of the paper presented in Table A3. There are clear differences in fields and in the types of papers being presented, with women presenting a disproportionate share of applied micro, policy relevant papers.

3.3 Human Coding

Three of the five datasets we compiled for this research involve data that was hand-coded by the Seminar Dynamics Collective. Our recruitment of these collaborators occurred in two separate waves. The first wave was recruited to code talks at their home institution—both job market talks and field-specific seminars; the second wave was recruited to code data from the 2019 NBER Summer Institute.

During the first wave, our recruitment was guided by the goal of finding coders at most of the "top thirty" economics departments in the United States. Many collaborators in the first

wave were recruited through an announcement made at a conference on diversity that attracted graduate students from many institutions. Others were recruited by asking for recommendations of potentially interested graduate students from a convenience sampling of faculty at top universities where we did not yet have volunteers. Yet others were recruited through the personal networks of the author team. We ultimately found collaborators at twenty of the "top thirty" departments, as well as a number of other economics departments, resulting in a total of 77 coders who collected data from their home institutions. Among coders in this first wave, 73 percent were female, 73 percent were in an applied micro field, and 36 percent were in the fourth year or higher in their Ph.D. program.

The second wave of recruiting, for the 2019 NBER Summer Institute, was quite different. These coders were recruited largely for their proximity to the conference location in Cambridge, MA, and the pitch we made to recruit them was not about studying gender, but rather the opportunity to attend a high-profile conference while still in graduate school. Moreover, as this is a conference that draws speakers from all over the world, coders had few ties to either presenters or audience members. There were 29 coders involved in this data collection effort (only four of whom had participated in the first wave). Roughly half (53 percent) of NBER coders were female, most (83 percent) specialized in an applied micro field, and about one-third (31 percent) were in their fourth year or higher in their doctoral program.

To collect granular data in real time, we developed an online data collection tool using a common survey data software platform (Qualtrics). Our customized tool presents coders with a series of web-based pages on which they can quickly register their observations by pushing on-screen buttons and also enter open-ended text responses in comment boxes. We intentionally designed the tool to be usable on either a tablet or a laptop so as not to draw attention to the coder during the seminar, thereby reducing the potential for Hawthorne effects. Few coders reported that their activities drew attention or inquiries from other seminar attendees. A detailed description of our tool—including both screenshots and instructions for coders—is available in the appendices of Dupas et al. (2021). Below, we provide a brief description of the key elements.

Before each seminar: Coders entered detailed information about the seminar including the date and time, the title of the paper being presented, the presenter's characteristics (name, gender, and home institution), as well as the seminar characteristics (series name, duration, whether it was a job market talk, the number of men and women in the audience, and any "rules" that governed audience interactions).

During the seminar: Coders collected data in real time about every interaction between the presenter and audience members using time-linked buttons to register the start and end of each interaction. During each interaction, coders pressed buttons registering characteristics (man or woman, professor or student) of the interrupter as well as the type (comment, criticism, suggestion, clarification, or follow-up) and tenor of the interruption. For the job market and regular seminar samples, coders were instructed to record whether the tenor of the interruption was particularly supportive, patronizing, disruptive, demeaning and/or hostile (the default was that that interruption was "neutral"). For the NBER 2019 sample, the NBER asked that we code only positive attributes of the interruption's tenor (if any), and so in that sample coders only recorded whether an interaction was particularly "constructive," "collegial," or "valuable." If coders thought there was nothing notable about the tenor of the interruption, they were instructed to simply leave the field blank.

3.3.1 Potential human biases

It was logistically infeasible to recruit fully "blinded" coders, so our collaborators are a self-selected group of economics students. In particular, women are a minority of academic economists and coders skew more female than the group we are studying, although this is also true of other teams collecting economic data (the U.S. Census Bureau, for instance, was 60 percent female in September 2020). Nonetheless, this robust representation of women among coders and their interest in diversity-related issues (at least for those coding department seminars) might raise concerns that the resulting data are biased toward finding gender differentials. Equally, the fact that coders were pursuing graduate work in economics may create a bias *against* finding gender differentials, as Paredes et al. (2023) find that "exposure to economics causally leads to more gender-biased views."

Dupas et al. (2021) reports a number of analyses that indicate coders did not have particularly progressive gender views, making it unlikely that coder bias would affect the conclusions that can be drawn from our human-coded data. First, results from the Harvard Implicit Assumption Test (IAT) for Gender Career Stereotypes revealed that the great majority of coders were implicitly biased *against* career women, and the average anti-female bias among the coders roughly tracked that of the broader population of IAT-takers. Second, patterns in the subset of seminars where data was collected by male coders is similar to that collected by female coders, as is the subset coded by those with more versus less progressive gender views, and the subset with coders recruited directly vs. through a faculty member at the home institution.

There is also high inter-coder reliability in the number of interruptions for the subset of talks with more than one coder (Table A1). For the more subjective measures, such as the number of interruptions of each type and tenor, which are count variables and highly skewed (with many zeroes), we compute Spearman's rank-order correlation coefficients. We find strongly significant positive correlations across all measures (Table A1). We note that inter-coder correlations on these subjective measures are typically lower when the speaker is female, perhaps suggesting that the audience reaction to these talks may be of a different nature that is harder to classify.[7]

3.4 Machine Coding

Our second approach to data collection relies on computer algorithms to transform audio recordings of presentations into a dataset of presenter-audience interactions. We can do this for two of our five samples. The use of audio-as-data is an emerging frontier in economics research, and our implementation includes several methodological contributions which we describe in the three steps presented in detail below.[8] First, we distinguish the different speakers within an audio recording to count the number of interruptions in a seminar. Second, we analyze each speaker's voice to predict their gender and count the number of interruptions by gender. Third, we label the emotional tone of each voice segment and code the number of interruptions by tone.

3.4.1 Using Diarization to Measure Interruptions

Our raw data consists of hundreds of audio recordings of economics presentations, and our goal is to transform each audio file into a count of the number of interruptions. Thus, our first task is to create a time-stamped roster of who spoke (the presenter or their audience) using a process known as "speaker diarization" for each of our two computer-coded samples.[9]

For the sample of online seminars, we search for changes in the speaker by looking for audio segments that are mixtures of multiple distributions representing different speakers (Park et al., 2022). Following Ravanelli et al. (2021), we then identify the presenter by using voice similarity

[7]Unfortunately we cannot compute inter-coder reliability in the tenor *at the interruption level* (i.e., interruption by interruption) since we cannot match specific interruptions across coders. We instead test the inter-coder reliability on the total number of interruptions of each type.

[8]We make our audio-tone imputation algorithm publicly available (it is downloadable as part of the replication package submitted for this paper). Our hope is that this pre-trained model may prove useful not only for researchers looking to replicate and extend the study of seminar dynamics, but also for other projects, studying political speech, media coverage, workforce or classroom interactions, interview data, and beyond. The interested reader is directed to Appendix D for additional discussion of the machine learning details.

[9]Two different approaches were used due to constraints on the timing and speed with which the data needed to be collected.

to tag each speech segment with a person/voice identifier and inferring the presenter to be the person who speaks the most during a seminar. Other voices are coded as interruptions by audience members. We apply the label of "moderator" to the speaker who introduces the presenter.

For the sample collected at the 2022 NBER Summer Institute, speaker diarization was performed using a software transcription service called "Trint" alongside human annotators. Trint transcribes audio files and tags the different speakers based on pauses and shifts in voice characteristics. Each utterance was then manually verified by two of the authors who also manually coded whether each utterance came from the speaker, discussant, or audience member. In about half of the cases the audio quality of audience comments was sufficiently poor as to be unusable for identifying gender or tone, likely because a microphone was not used. Thus, we use the full sample to analyze the number of interruptions—which requires only that we note whether an interruption occurred— but we use a smaller sample with sufficiently good audio quality for audience speech for more detailed interruption analysis.[10]

3.4.2 Coding Gender

We impute speaker gender using a deep neural network (or DNN), a type of neural network that is commonly used for classification tasks, to predict whether an audio recording includes speech by a man or a woman.[11]

We train this network using a large dataset of short speech recordings labeled with speaker gender called the Mozilla Common Voice Dataset (V4). This dataset is an open-source collection of hundreds of thousands of three to five second audio clips submitted (and validated) by volunteers from around the world. It includes a diverse representation of accents, dialects, and speakers across multiple languages. This diverse dataset improves our model's ability to accurately label speaker gender when we impute gender for presenters and audience members in recordings of economics presentations.

We use a common variety of measures to quantify patterns in audio over time — such as volume, pitch, and timbre — such that each audio clip is then summarized as a 180 variable vector. Through

[10] Appendix D includes more information on the sample preparation.

[11] The architecture of our neural network largely tracks that used in a similar study (Bhattacharya et al., 2022). As such, our model architecture includes a combination of five convolution layers which summarize variation from different, grouped subsets of inputs, and two fully connected layers, where each neuron is connected to every neuron in the previous layer. To explain: Each hidden layer works to highlight combinations of the raw audio measures that are jointly useful in predicting gender and to dampen those that are unrelated to gender. They determine what is and is not relevant based on a set of parameters that are optimized during the training process. Effectively, these steps allow us to focus on the gender-relevant features of the audio measures. The details of the hyperparameters and model features to prevent overfitting, such as dropout layers, are available in the online Appendix D .

an iterative training procedure, the neural network learns to sort audio clips into male and female categories based on variation over the audio vectors. Appendix D provides substantial detail on the training, development, and testing of the gender-classifier network. The model correctly predicts a speaker's gender from their voice for 95.2 percent of recordings in a leave-out subset of the Mozilla data only used for evaluating (not training) the model.

One potential concern with training our model with the Mozilla data is whether an algorithm trained on non-economics-seminar recordings can transfer to imputing gender for speech in the specific context of economics seminars. Robustness checks show that the gender predicted by our algorithm agrees with a name-gender dictionary (used only for the online seminars) as well as with the majority vote from a team of five research assistants (applied to both the online seminars and the NBER 2022 SI) in over 90 percent of all cases.[12]

We then apply the trained algorithm to five-second segments of speech from recordings of online talks and NBER SI 2022 presentations. This produces gender labels for each presenter and speaking audience member in our data. For the seminar presenter, we use the majority gender label from all the five-second clips where they speak during the entire presentation. For audience interruptions, we focus on each utterance—the term we use for an uninterrupted period of speech by one person—and code an audience member as female if the majority of five-second intervals within that utterance are predicted by the gender classifier to come from a woman.

3.4.3 Coding Tone of Voice

We adapt the methods used for gender classification to predict tone of voice – whether a speaker sounds positive, neutral, or negative – from short recordings of speech.

We use two widely used datasets, RAVDESS and CREMA-D, which contain recordings of actors reading the same sentence with different emotional cues.[13] We first aggregate the emotion labels into four tone categories: "Positive" (happy or pleasantly surprised), "Neutral" (neutral or calm), "Negative-Aggressive" (angry or disgusted), and "Negative-Passive" (sad or fearful). We train the model to distinguish between the two negative tones given their different vocal patterns, but combine them into a single "Negative" category in the main analysis. Thus, we have audio data matched with speaker gender and tone and we iteratively train a DNN to predict tone from

[12] More details can be found in Section D.8 of the appendix.

[13] RAVDESS, which includes recordings from 24 actors (evenly split by gender) performing eight emotions (anger, calm, disgust, fear, happy, neutral, sad, and surprised), and CREMA-D, which includes recordings from 91 diverse actors using six emotions (anger, disgust, fear, happy, neutral, and sad).

that audio data. Due to sex-based differences in voice patterns, we estimate two separate tone classifiers—one for male and one for female voices.[14] Again, Appendix D provides details on training and developing the tone prediction DNN.

We then evaluate the model on a subset of data left out of the training procedure. Our tone classifier correctly predicts the tone label as negative (either negative-passive or negative-aggressive) in 72.9 percent of cases for female voices, and 72.1 percent of cases for male voices (the benchmark here would be a 50 percent accuracy rate if the model was assigning tone at random); it correctly predicts the tone label as positive in 49.2 percent of cases for both female and male voices (compared to the 25 percent random benchmark). As with the gender classification, we also had our team of research assistants manually label 80 actor recordings and 80 short clips from our online seminar sample with tone. Among the set of actor recordings, the RAs correctly assigned negative labels for 72.5 percent of cases and positive labels for 65 percent of the cases. For a subset of audio recordings from online seminars, the machine-predicted tone was validated by at least one RA 59 percent of the time. Feedback from the RAs confirmed that classifying the tone of voice was more challenging than classifying gender.

We then apply the trained tone-classifier DNN to five-second segments of speech from recordings of online talks and NBER SI 2022 presentations. This produces tone labels for each speaking audience member in our data. We assign the final tone label to an audio clip by identifying the tone with the highest average probability. To prevent ties, we also require a minimum probability threshold of 50 percent to classify a clip as a non-neutral tone. As mentioned above, after imputing tone, we combine "Negative-Aggressive" and "Negative-Passive" tones into one "Negative" category.

Finally, note that *tone* as coded by the machine is different from *tenor* coded by humans for the in-person seminar samples. The audio model captures only vocal tone (i.e., how something is said), while human coders assess both content and delivery. Unlike the algorithm, coders also observe body language and facial expressions. Despite these differences, both machine- and human-coded measures offer valuable insights into seminar dynamics.

[14] Notice that we are studying the tone of *audience* interruptions, not that of *presenters*, so estimating gender-specific tone classifiers does not create any mechanical link between the measured audience tone and the presenter's gender, which is the focus of our analysis.

4 Analytic Framework

Our analysis explores whether (and how) female presenters are treated differently than similarly situated men. The primary outcomes are the number of interruptions, their type and tone, as well as who interrupted and how.

The core of our analysis involves estimating equations of the following form:

$$Y_{p,r,s,c} = \alpha + \beta_1 Female\ Presenter_p + Controls_{p,r,s,c} + \epsilon_{p,r,s,c} \tag{1}$$

where the subscript p denotes an individual presenter (e.g., Esther Duflo), s denotes a seminar series (e.g., the Stanford Development Economics seminar), r is the research paper they present, and c is the coder (an individual graduate student, or machine-learning algorithm).[15] The dependent variable, $Y_{p,r,s,c}$, is the outcome variable of interest, such as the number of interruptions during a presentation. The variable $Female\ Presenter_p$ is an indicator variable for whether the presenter's gender is female, and so the coefficient of interest, β_1, describes the extent to which women are treated differently than men.[16]

Our selection of presenter-, paper- and audience-level control variables is guided by our research question, which is to compare *similarly situated* men and women. By "similarly situated", we refer to men and women with similar seniority and from similarly-ranked institutions who present papers on similar topics within the same seminar series. As such, we include fixed effects for each *seminar series* (or, for the two NBER samples, the best available proxy), so that our comparisons are limited to, for example, men and women presenting within a specific seminar series (such as the Harvard Labor seminar). These fixed effects effectively also control for seminar field (e.g., labor), location (e.g., Harvard), the rules and norms that apply to a specific seminar series, the prestige of the host institution, and the characteristics of the typical audience at that seminar. For our two NBER samples, our data use agreement required that we only code talks as falling into one of two to three coarse fields (of micro, macro, and other). For these samples, we control for these coarse categories, as well as seminar characteristics, such as duration, format (e.g., regular, discussant, or designated

[15] For human-coded samples, we include a control for coder gender. When a seminar has two coders, we include both observations with a weight of one half.

[16] We experimented with estimating our standard errors in ways that permit various forms of correlation across observations, clustering alternatively by presenter, seminar series, presenter × seminar series, or using Cameron and Miller's multiway clustering, by presenter and seminar series. In each case, the standard error on our coefficient of interest does not change much (typically within 10 percent of the OLS standard error—see Figure B1 for the presenter-level clustering results, and Dupas et al. (2021) and Seré (2023) for other specifications). Thus, for the sake of simplicity, we present the usual OLS robust standard errors.

Q&A time at the end), and rules (e.g., moratoriums on interruptions at the beginning or end of the talk), and we also interact all of these variables. We believe this exhaustive list of attributes is what most effectively controls for each specific NBER program in the absence of program identifiers.

We aim to compare otherwise-comparable presenters, so (when permitted) we also control for *presenter characteristics*. These controls include (when available—this varies across samples): the presenter's rank (senior faculty, junior faculty, or other), their Google Scholar citation count, whether they work at an academic institution, the geographic location of their workplace (the US, Europe, or elsewhere), as well as their department's ranking. Men and women also differ in the topics they research (Dolado et al., 2012; Conde-Ruiz et al., 2022), so the key *paper characteristics* we control for (where possible) are JEL codes, and whether the research presented includes economic theory, data analysis, experimental design, and/or policy analysis.[17]

Some factors vary *within* a seminar series, such as the size and composition of the audience, which often varies from week-to-week. Some of this variation is likely random and hence unlikely to distort our key findings. But some of this variation may reflect an endogenous response to the presenter's identity, the characteristics of the research they're presenting, or other attributes correlated with the presenter's gender. As such, these are outcomes that could reflect mechanisms by which gender matters, and thus we do not include audience size and composition as controls (that is, we seek to avoid what is sometimes called "overcontrol bias.") Yet, these variables are still of interest because they may provide insight into some of the pathways by which men and women are treated differently, and we analyze them as outcomes of interest in separate regressions, following the specification in Equation 1.

We work to make the analysis across the different samples as comparable as possible, but given that each data collection effort included different constraints, some differences remain, and Appendix C outlines the differences in variable construction.

[17] Using the text of the paper presented for job talks and regular seminars and the transcript for online seminars, we use a large language model (LLM) to assign the type labels of theory, data, experimental, and policy. Appendix D details the GPT model used, the prompts, and accuracy metrics for the LLM. These type controls are excluded from column 1 due to insufficient variation to reliably estimate them with JEL controls. To show that type controls do not overturn the results for the job market sample, we pool the job market sample and the regular seminar sample to jointly estimate the coefficients on controls while estimating sample-specific coefficients for presenter gender. We find almost the same results as when we only control for JEL and not type. These results are shown in Table B1.

5 Results: Number of Interruptions

5.1 Total Number of Interruptions

We start by analyzing how often women and men are interrupted in Table 2. Each column shows the results from a different sample, using for each of them the most complete specification that we can run.[18] We present sensitivity to controls in Figure 1.

Table 2: Presenter Gender and Number of Interruptions

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **All samples** (cols. 1–5)
Female presenter	3.768	2.870	0.836	1.940	0.227	1.549
	(1.556)	(1.285)	(0.739)	(0.521)	(0.602)	(0.336)
	{0.016}	{0.027}	{0.259}	{<0.001}	{0.706}	{<0.001}
Controls:						
Seminar series	Yes	Yes	Yes	Yes	Yes	Yes*
Seminar duration	Yes	Yes	Yes	Yes	Yes	Yes*
Presenter's rank	n.a.	n.a.	Yes	Yes	n.a.	Yes*
Presenter's citation count	n.a.	Yes	n.a.	Yes	n.a.	Yes*
Presenter's home institution	Yes	Yes	Yes	Yes	n.a.	Yes*
Paper type	No	Yes	n.a.	Yes	n.a.	Yes*
Paper JEL	Yes	Yes	n.a.	Yes	n.a.	Yes*
Mean (male presenter)	34.39	25.35	14.68	10.93	13.74	15.15
Number of observations	240	292	447	994	467	2,440
Number of unique talks	176	245	442	994	467	2,324

Notes: *when available. Standard errors in parentheses and p-values in brackets. See Table 1 notes for data sources and Appendix C for definitions of controls. "n.a."= not applicable, or not available due to data-use agreements (Cols. 3 and 5). Proxy seminar series FE are constructed for NBER SI samples in accordance with the data-use agreements. Paper type FE control whether the paper includes theory, data, experimental design, and/or policy analysis as determined by an LLM. These type controls are excluded from column 1 due to insufficient variation to reliably estimate them with JEL controls, see footnote 17 and Table B1 for more information. See Figure 1 and Table B2 for analyses of the sensitivity of results to the inclusion of controls for all samples.

We report β_1, the coefficient on *Female Presenter*, which measures the average number of extra interruptions that a woman receives when presenting her research, relative to a similarly-situated man. Thus, the coefficient in the first column of Table 2 shows that women on the job market received, on average, 3.8 extra interruptions per presentation than men (p-value 0.016). This corresponds to a 11 percent increase compared to the mean of 34.4 interruptions experienced by

[18] The differences in control variables across columns reflect constraints specific to each sample. For instance, job market candidates are too early in their careers for citations to be meaningful. For the NBER sample in 2019, coders recorded the *rank* of each presenter's institution, but under our agreement with the NBER, they were not permitted to record the presenters' names or the names of their institutions, so we could not subsequently count their citations. The agreement with the NBER in 2022 was even more strict, and we were not allowed to record any details about the paper or the author. Appendix C includes full definitions of the variables.

male presenters. The second column shows that in the sample of regular (in-person) departmental seminars, the gender gap is 2.9 extra interruptions (+11 percent from a base of 25, p-value 0.027).

Next, we turn to conference presentations in the third and fifth columns, which report findings from two separate years of the NBER Summer Institute. The data in the third column was collected by human coders, while the data in the fifth was computer coded. We first note that interruptions are fewer in (typically shorter) conference presentations, with a mean number of interruptions of 14.7 and 13.7, respectively, in columns 3 and 5 of Table 2. The gender gap is not statistically different from zero in either of the two NBER samples, but for 2019 we cannot rule out an effect size (in percentage terms) as large as that in the seminar samples. The null result for the 2022 NBER sample is somewhat tighter. It may be due to the heightened awareness that interactions between audience and speakers were being studied in 2022. Given that the average presentation time was 7 minutes longer in NBER SI 2022 compared to NBER SI 2019 (Table 1), the number of interruptions in NBER SI 2022 is particularly low overall. Our final sample is both our largest, and most heterogeneous, as it includes all online presentations without discussants (around 85 percent are online seminars and 15 percent appear to be online conference presentations). It is also quite different from the earlier samples, as online presentations tend to involve much less back-and-forth, with only 11 interruptions on average when a man presents (column 4 of Table 2). The interruption gender gap in this sample is 1.9 (+18 percent, p-value<0.001).

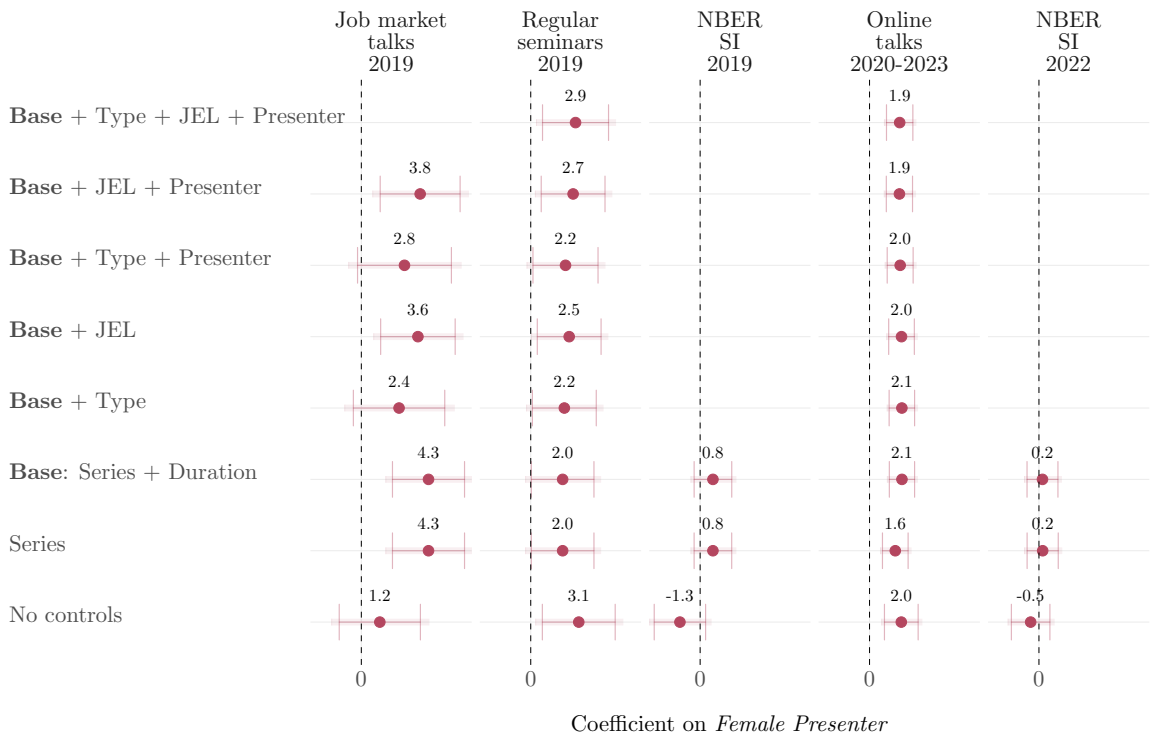
Figure 1 shows how the results change (or not) as we peel away each layer of controls. It shows that the results are robust provided the seminar series is controlled for.[19] This highlights the critical importance of controlling for the audience/field.[20] In all analyses that follow, we use the specification with the most complete possible controls for each sample (i.e., for each sample, the first specification with a coefficient estimate in Figure 1).

Our estimates of the interruption gender gap—measured as the count of extra interruptions made to women—appear to vary quite a bit across samples, but upon closer inspection, they vary in ways that one might expect. Job talks and regular seminars are relatively long (around 80 minutes, on average) and interactive, while conference presentations are typically brief (40 minutes), and highly structured. Moreover, in-person seminars tend to yield more free-flowing dialogue than online seminars and conferences, which tend to be more stilted. Each of these patterns is evident

[19] Table B2 shows the results in table form for the specification with no controls, and the specification with comparable controls across all 5 samples.

[20] For the two NBER samples, our data agreements precluded that we collect information on presenter details, as well as information on the specific *program* a talk was in (the equivalent of the seminar series). For these samples we create a proxy program/series fixed effect using broad fields, talk format, and duration.

Figure 1: Gender Gap in Interruptions: sensitivity to controls



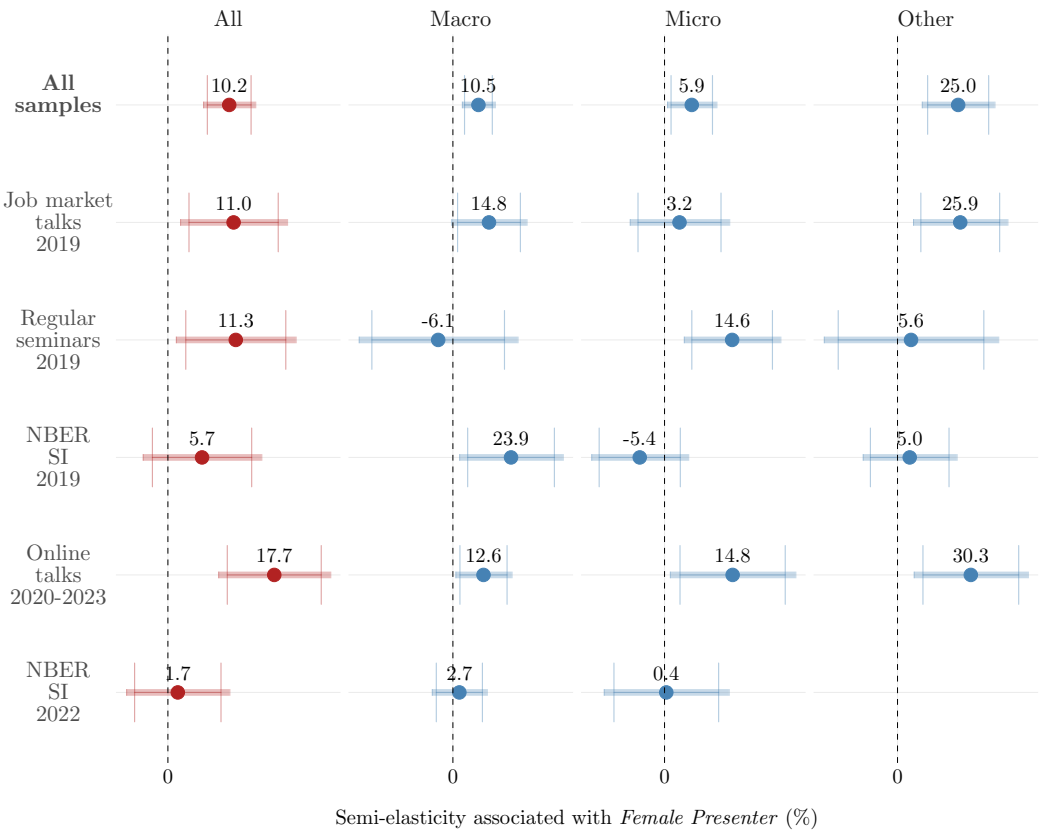
Notes: The "no controls" specification includes a control for coder gender in the human-coded samples (cols. 1-3). "Series" refers to (a) the university hosting the job talks (col. 1), (b) the seminar series (col. 2), (c) a proxy for the NBER program using talk format and duration (cols. 3 and 5), and (d) a proxy for the seminar series using information on both the YouTube channel from which the recording was sourced and language processing of the video title and description (col. 4). "Type controls" refers to the four binary variables indicating whether the paper makes (i) a theoretical point, (ii) an empirical contribution, (iii) uses an experimental design, and (iv) addresses a policy question, as inferred by LLM from the paper (cols. 1 and 2) or video transcript (col. 4).

in the average number of interruptions per talk (shown at the top of the lowest bottom panel of Table 2), which vary tremendously across samples. There are also substantial differences in the sample sizes and statistical precision of each estimate.[21]

To summarize the results, Figure 2 presents these coefficients as semi-elasticities (evaluated at the sample mean), and show the corresponding 90 and 95 percent confidence intervals. Pooling all samples together yields an estimated gender interruption gap of 10.2 percent (with a standard error of 2.1).

[21] In the sample of online talks, a subset of presentations (157 out of 994) may possibly be conference presentations (based on the name of the online series). The number of interruptions in these presentations is substantially lower than in regular online seminars, and the estimated effect of having a female presenter is smaller and not statistically significant in that subsample (estimate = 0.09, standard error = 0.91).

Figure 2: Gender Gap in Interruptions: Semi-elasticities



Notes: Percentage change from coefficients on *Female presenter* from Table 2 and Table B3. As per the data use agreement, all talks in the NBER 2022 sample were labeled as either "Macro" or "Micro". The independent variable is a binary variable indicating the presenter is female, and the dependent variable is the total number of interruptions. The coefficient is divided by the mean for male presenters. Whiskers and the ends of the error bars indicate 90 percent and 95 percent confidence intervals, respectively. For legibility, the scale varies across vertical panels.

5.2 Differences by Field

To explore whether there are important differences by field, we break each sample into three broad categories—microeconomics, macroeconomics, and others. Our categories follow those of Chari and Goldsmith-Pinkham (2017) and are assigned based on the paper (job market sample and online talks sample), or on the seminar series/program name (regular seminars sample and NBER samples—see Section C.5.2 in the online appendix for details.) The regression specification largely follows Equation 1, but instead of having one coefficient of interest, we now have three: β_1^{macro}, β_1^{micro}, and β_1^{others}, which are the coefficients on *female presenter* × *macro seminar*, *female presenter* × *micro seminar*, and *female presenter* × *other seminar field*, respectively (recall we have seminar series fixed effects soaking up the gender-neutral effects of seminar fields).

The results in percentage terms (semi-elasticities) are presented in blue in Figure 2, and the regression results in levels are shown in Table B3. While the results in the pooled estimation, or in the largest sample (online talks), suggest the phenomenon of more interruptions for female speakers affects all fields, the results in the smaller samples are noisier and mixed. In job market talks, the pattern is most pronounced in non-micro fields, while we see the opposite in regular seminars, though we acknowledge that the regular seminar sample includes only 41 macro talks, only 8 of which by female speakers, so drawing clear conclusions for macro is difficult with that sample alone.

6 Who Interrupts, How, for What and When?

This section provides additional analyses aimed at understanding the nature of engagement in economics seminars, and how it differs by presenter gender. The analysis here is guided by two questions which we think are potentially useful to think through the value of interruptions. First, where does the additional engagement for female presenters come from? We consider audience size, as well as the gender of those who interrupt. Second, do interruptions directed at female presenters differ qualitatively? Here, we consider the tenor, tone, type and timing of interruptions.

6.1 More Engagement: Attendance and Characteristics of Interrupters

Table 3 studies how presenter gender influences the quantity and composition of seminar attendees. The measures we have for attendance and composition vary somewhat across our samples. For the human-coded samples, we counted the number of attendees by gender and flagged whether the seminar had unusually high attendance. For the online talk sample, we are unable to have a measure of true attendance because we do not have information on virtual viewers, but we count the number of distinct audience speakers to measure participation as an admittedly imperfect proxy for attendance. We have no information for attendance at NBER SI 2022.

We find that female presenters have larger seminar audiences (especially for job market talks). This greater attendance at female talks is driven by a mix of male and female audience members. The increase in female audience members is typically larger (in percentage terms) than the increase in male audience members, and is particularly large for job market talks by female presenters (35 percent more women vs. 5 percent more men, Table 3).

To what extent is the greater attendance at women's talks explaining the greater number of

interruptions women receive? We first note that the finding that women are interrupted more often holds—in roughly similar size and significance—in analyses that control for audience size (Table A4). Since audience size is endogenous, controlling for audience size may introduce a bias. As such, the conditional results should be interpreted with caution. Nevertheless, it suggests that audience size is unlikely to explain the results on interruptions.

Table 3: Attendance and Participation

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022
Coef. on female presenter:					
Total Attendance	5.052	2.034	0.387	0.324	n.a.
	(1.570)	(1.092)	(2.736)	(0.110)	
	{0.002}	{0.064}	{0.888}	{0.003}	
Attendance by women	3.435	1.180	0.128	0.138	
	(0.839)	(0.611)	(1.081)	(0.074)	
	{<0.001}	{0.055}	{0.905}	{0.061}	
Attendance by men	1.617	0.850	0.186	0.185	
	(1.335)	(0.740)	(2.139)	(0.100)	
	{0.227}	{0.252}	{0.931}	{0.065}	
Higher than usual attendance	0.051	0.001	-0.026		
	(0.070)	(0.066)	(0.040)		
	{0.471}	{0.984}	{0.508}		
Mean for male presenter:					
Total Attendance	39.16	25.60	62.58	3.72	
Attendance by women	9.53	8.64	17.22	0.98	
Attendance by men	29.63	16.96	45.41	2.74	
Higher than usual attendance	0.24	0.22	0.21		
Number of observations	233	290	446	994	
Number of unique talks	171	243	441	994	

Notes: Each cell corresponds to a different regression, in which the attendance variable on the left is regressed on the presenter being female, for the sample in that column. Specifications identical to those shown in Table 2. Rather than attendance, the online seminars sample (4) uses the number of different speakers in the audience. The NBER SI 2022 dataset does not include attendance as per the data-use agreement.

We turn to analyzing the source of the extra interruptions that women receive in Table 4. Similar to our analysis of attendance by audience gender, we compare interruptions by audience gender. The dependent variable is either the number of interruptions by men, the number of interruptions by women, or the share of interruptions by women.

Considering first interruptions by men, the positive coefficients for four of the samples imply that female presenters receive more interruptions from men than male presenters (+1.2 in the pooled specification, p-value<0.001, corresponding to a 9 percent increase). In the job market

sample, the extra interruptions by men (+3.1 in column 1 of Table 4) outnumber the extra male audience members (+1.6 in column 1 of Table 3). The same is true for the online talks sample (+1.6 interruptions from males vs. 0.2 distinct additional male voices in the audio recording). This further suggests that a larger audience may not be the only explanation for the greater number of interruptions received by women presenters.

Table 4: Presenter Gender and Number/Share of Interruptions, by Interrupter Gender

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **All samples** (cols. 1–5)
Panel A: Number of interruptions by men						
Female presenter	3.118	1.030	0.746	1.556	-0.500	1.185
	(1.516)	(1.213)	(0.678)	(0.451)	(0.896)	(0.332)
	{0.041}	{0.397}	{0.272}	{0.001}	{0.577}	{<0.001}
Mean (male presenter)	31.04	20.63	12.17	8.65	10.69	13.02
Number of observations	240	291	447	994	257	2,229
Number of unique talks	176	244	442	994	257	2,113
Panel B: Number of interruptions by women						
Female presenter	1.001	1.933	0.126	0.384	1.028	0.609
	(0.537)	(0.670)	(0.228)	(0.404)	(0.578)	(0.228)
	{0.064}	{0.004}	{0.581}	{0.342}	{0.077}	{0.008}
Mean (male presenter)	3.19	5.00	2.46	2.29	4.31	3.02
Number of observations	240	291	447	994	257	2,229
Number of unique talks	176	244	442	994	257	2,113
Panel C: Share of interruptions by women						
Female presenter	0.032	0.039	0.016	0.005	0.053	0.019
	(0.015)	(0.023)	(0.015)	(0.022)	(0.030)	(0.012)
	{0.042}	{0.099}	{0.308}	{0.818}	{0.085}	{0.102}
Mean (male presenter)	0.09	0.20	0.19	0.22	0.30	0.21
Number of observations	240	291	447	945	257	2,180
Number of unique talks	176	244	442	945	257	2,064
Effect panel A = panel B (p-value)	0.189	0.515	0.387	0.053	0.153	0.152

Notes: Standard errors in parentheses and p-values in brackets. Specifications identical to those shown in Table 2. The number of observations in Col. 5 is reduced from the sample in Table 2 to those seminars with sufficiently good audio quality for audience members where we can estimate the gender of the interrupter (and later the tone). The coefficient estimate in Table 2 for the subsample shown in Table 4 Col. 5 is 0.571, with a standard error of 0.371. The mean number of interruptions is 15.01.

Turning to interruptions by women, the positive coefficients in all five samples indicate that women are responsible for some of the gender-interruption gap (+0.6 interruptions by women, p-value=0.008, corresponding to a 20 percent increase). There are two samples where additional interruptions by women even outnumber additional interruptions by men: regular seminars, and NBER SI 2022. On average, women account for only 21 percent of interruptions during presentations by men, but when a woman is presenting, the share of interruptions by women increases by

1.9 percentage points (close to a 20 percent increase, p-value 0.102) in the pooled sample. This increase is observed in four of the five samples, and significant at the 10 percent level in three (Panel C of Table 4).

The dual finding that women presenters (i) have additional audience members, particularly women, and (ii) elicit more interruptions from female audience members, could suggest an important role model effect—in which case the dynamic implications of some of the differential treatment could be positive for women's representation in the profession.[22]

After seeing where the additional interruptions come from, we next explore whether they are qualitatively different.

6.2 Tenor, Tone, Type and Timing of Interruptions

While a more engaged audience can be beneficial in principle, not all interruptions are beneficial. Audience members interrupt in various ways and for various reasons. In this section, we categorize interruptions in three key ways. First, we measure the tenor and tone of interruptions. Second, we categorize interruptions by type (questions, suggestions, criticisms, etc.). Third, we study interruptions where the audience talks over the presenter.

6.2.1 Tenor and Tone of Interruptions

We consider two attributes of interruptions, tenor and tone. First, we consider context-dependent "tenor" (defined by Merriam-Webster as "the drift of something spoken or written"): in our human-coded samples, an interruption is labeled as neutral (the default), negative or positive based on a combined assessment of what is said, how it is said, and the context of the entire presentation. Second, we consider context-independent tone: in our computer-coded samples, an interruption is predicted to be negative or positive purely on how the audience member sounds (i.e., tone of voice). Both of these measures are of interest. The first captures important nuances that ultimately reveal the impact of an interruption's intent. The second provides a one-dimensional but consistent metric for tone across seminars that is mechanically objective. Evidence that they measure different things can be found in the sample means: in the human-coded samples, the share of interruptions coded as negative in tenor is lower than 5 percent, while in the computer-coded samples, the share of interruptions coding as having a negative tone of voice is around 50 percent.

[22] Dupas et al. (2021) document that female presenters are more likely to receive interruptions from *graduate students*, especially female students.

Table 5 presents our analysis of the tenor and tone of interruptions, disaggregated by presenter gender. The specifications in this table match those provided in Table 2 and follow Equation 1, but the dependent variable is the number of interruptions rated as having a positive (negative) tenor/tone. In the human-coded samples, interruptions classified as negative included labels such as "patronizing," "disruptive," "demeaning," or "hostile," while positive interruptions are coded as "supportive," (samples 1 and 2) or "constructive," "collegial," or "valuable" (sample 4). Again, these measures are based on a coder's context-dependent assessment of interruptions from audience members. For computer-coded data, negative tones are interruptions that sound "negative-aggressive" (capturing emotions like "anger" or "disgust") and "negative-passive" (capturing "sadness" or "fearfulness"), while positive interruptions are those categorized as "happy" or "positively surprised."

Table 5: Presenter Gender and Tenor/Tone of Interruptions

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **Human-coded** (Cols. 1-3)	(7) **Computer-coded** (Cols. 4-5)
Panel A: Negative interruptions							
Female presenter	0.515	0.369		0.919	0.640	0.461	0.861
	(0.278)	(0.308)		(0.372)	(0.708)	(0.195)	(0.329)
	{0.065}	{0.233}		{0.014}	{0.367}	{0.019}	{0.009}
Mean (male presenter)	1.68	0.61		6.03	7.08	1.05	6.24
Number of observations	240	291		994	257	531	1,251
Number of unique talks	176	244		994	257	420	1,251
Panel B: Positive interruptions							
Female presenter	-0.030	-0.102	0.389	0.925	-0.270	0.104	0.675
	(0.374)	(0.615)	(0.416)	(0.404)	(0.400)	(0.280)	(0.334)
	{0.937}	{0.868}	{0.351}	{0.022}	{0.501}	{0.711}	{0.043}
Mean (male presenter)	1.26	3.56	2.70	3.81	3.22	2.64	3.69
Number of observations	240	292	447	994	257	979	1,251
Number of unique talks	176	245	442	994	257	863	1,251

Notes: Standard errors in parentheses and p-values in brackets. Specifications identical to those shown in Table 2. Figure A1 shows sensitivity of the results to the inclusion of controls. Col. 3: Negative interruptions were not labeled as such for the NBER 2019 sample as per the data collection agreement. Col. 5: Sample size for NBER 2022 reduced from the sample in Table 2 to those seminars with sufficiently good audio quality for audience members for tone analysis.
Human-coded negative interruptions include those labeled as patronizing, disruptive, demeaning and/or hostile, and positive interruptions were labeled as supportive, constructive, collegial, or valuable. Machine-coded negative interruptions were predicted to be either negative-aggressive (angry or disgust) or negative-passive (sad or fearful) and positive interruptions were predicted to be happy or positively surprised based on training data on tones from voice actors.

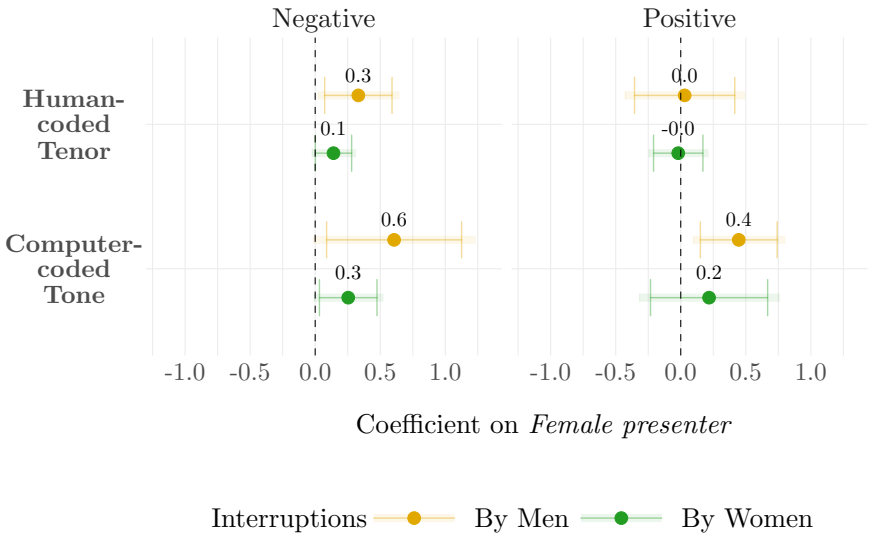
Results indicate that female presenters experience a higher number of interruptions coded as negative compared to their male counterparts. The pooled estimates suggest that female speakers

receive on average 0.5 additional interruptions rated by fellow economists as negative in tenor (p-value=0.024, a 44 percent increase), and 0.9 additional interruptions rated by the machine as negative in tone (p-value=0.009, a 14 percent increase) (columns 6 and 7 of Table 5). The point estimates are systematically positive across the four samples for which we could record negative tenor or tone, but the sample-specific results are noisy.

The results on positive interruptions are more heterogeneous. We find no gender gap in interruptions with a positive tenor (column 6 of Table 5). There is however an increase in interruptions with positive tone in the machine-coded samples (column 7), though this entirely driven by the large online talks sample. In that sample, the magnitude of the effect is similar for positive and negative interruptions, in contrast with all other samples.

Figure 3 shows the tone results disaggregated by the gender of the interrupter (the sample-specific regressions are shown in Table A5). The extra negative interruptions, be they negative in tenor or tone, come from both men and women in the audience. There is however some heterogeneity across samples (Table A5). In both the job market and online samples, the main source of additional negative interruptions received by female presenters come from men. In contrast, in the NBER SI 2022 sample, the additional negative-in-tone interruptions all come from women. In regular seminars, the additional negative-in-tenor interruptions come from both groups.

Figure 3: Interruptions by Tenor/Tone and Gender of Interlocutor



Notes: The independent variable is a binary variable indicating the presenter is female, and the dependent variables are the number of interruptions by men (orange) or by women (green). The left panel contains the estimates for negative interruptions and the right panel the estimates for positive interruptions. Whiskers and the ends of the error bars indicate 90 percent and 95 percent confidence intervals, respectively. Table A5 shows the results in table form, including means of the dependent variables and sample-specific estimates.

To what extent is the increase in interruptions that are negative in tenor and tone driven by the fact that women receive *more* interruptions? To answer this, we look at the *share* of interruptions that are rated as negative or positive (Table A6). For the two human-coded samples, the share of interruptions coded as negative is low (only 5 percent in job talks and 2 percent in regular seminars) because coders were asked to rate the tenor only if they felt that it was non-neutral. In those human-coded seminars, we see increases in the share of negative interruptions of 1.5 and 2 percentage points, respectively, with p-values of 0.086 and 0.177. The increase in the share of interruptions rated as negative in tenor is significant at the 5 percent level (+1.8 percentage points, corresponding to a 53 percent increase) when we pool across the human-coded samples. For the computer-coded samples, we do not see a difference in the share of negatively-toned interruption between female and male presenters. Together, these results suggest that the increase in negative-in-tenor interruptions is not just the mechanical consequence of the gap in the number of interruptions, while the increase in negative-in-tone interruptions in online talks likely is.

6.2.2 Types of Interruptions

During presentations, interruptions from audience members can take various forms, including questions, suggestions, critiques, and comments. These types of interruptions influence seminar dynamics in different ways. For example, interruptions that are questions may help the audience follow or clarify the presenter's ideas, while non-question interruptions, such as comments or criticisms, can potentially disrupt the flow of the presentation by introducing new topics or undermine the presenter if not presented constructively. Suggestions could be either helpful if presented in a positive way, or harmful if raised from a deficiency standpoint. Clarification questions are often asked too early ("the answer is in two slides!") and could be avoided, and follow-ups could at times be postponed to *after* the presentation is over. In this section, we analyze how the type of interruption varies between female and male presenters across four of our five samples. As in the analysis of tenor and tone, we look at effects both in levels and in shares.

For the human-coded samples, volunteers were asked to label interruptions as comments, clarifications, suggestions, criticisms, or follow-ups when applicable, leaving them unlabeled if none of these categories applied. For the online seminar sample, transcripts of the presentations were processed to identify interruptions by detecting question marks.[23] Interruptions where an inter-

[23]Recognizing potential transcription errors, these transcripts were automatically processed through an automated language processing tool for punctuation standardization, ensuring consistent placement of question marks. This processing was strictly limited to punctuation correction; for example, transforming a mispunctuated statement like

rogation mark was detected were classified as interrogative, and those without were classified as declarative.

Table 6: Presenter Gender and Types of Interruptions

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **Human-coded** (cols. 1-3)
Coef. on female presenter:						
Type of interruption:						
Non-labeled	2.350	0.555	1.203		n.a	1.102
	(1.312)	(0.999)	(0.506)			(0.475)
	{0.075}	{0.579}	{0.018}			{0.021}
Suggestions	0.751	0.673	0.426			0.540
	(0.400)	(0.382)	(0.283)			(0.195)
	{0.062}	{0.080}	{0.132}			{0.006}
Comments	0.046	0.131	-0.214			-0.108
	(0.638)	(0.599)	(0.355)			(0.277)
	{0.942}	{0.828}	{0.546}			{0.696}
Criticism	-0.417	0.158	-0.010			-0.065
	(0.449)	(0.385)	(0.235)			(0.183)
	{0.354}	{0.682}	{0.967}			{0.720}
Clarifications	1.646	1.119	-0.603			0.494
	(1.079)	(0.745)	(0.539)			(0.416)
	{0.129}	{0.135}	{0.264}			{0.235}
Follow-ups	-0.199	0.390	-0.065			0.049
	(0.509)	(0.359)	(0.327)			(0.220)
	{0.696}	{0.278}	{0.842}			{0.822}
Interrogative				1.182		
				(0.276)		
				{<0.001}		
Declarative				0.760		
				(0.383)		
				{0.048}		
Mean for male presenter:						
Non-labeled	12.38	10.01	3.41			6.72
Suggestions	2.32	2.40	2.65			2.53
Comments	5.09	4.20	3.33			4.07
Criticism	3.21	1.74	2.12			2.18
Clarifications	13.62	8.64	5.41			8.40
Follow-ups	3.01	1.82	1.85			2.15
Interrogative				4.78		
Declarative				6.16		
Number of observations	240	292	447	994		978
Number of unique talks	176	245	442	994		863

Notes: Each cell corresponds to a separate regression. Standard errors in parentheses. Specifications identical to those shown in Table 2. Information on the type of interruption was not recorded for the NBER 2022 sample as per the data use agreement.

The results, shown in Table 6, indicate that the additional interruptions that women receive

"what are the implications of this policy on inflation" into the correctly punctuated "What are the implications of this policy on inflation?".

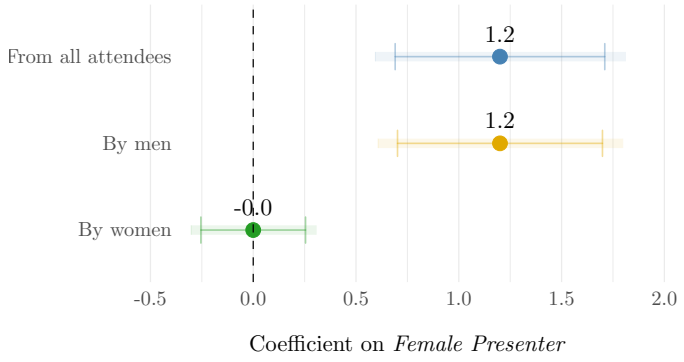
include a mix of actual questions and non-questions. In the online sample, female presenters receive 1.2 additional interruptions coded as questions and 0.8 additional interruptions coded as declarative. Consistent with this, in the human-coded samples, where we can see the type of interruptions with more granularity, women receive on average 1.1 more generic (non-labeled) questions and 0.5 more suggestions per talk.

Here again, the question is whether this increase in suggestions is entirely driven by the greater level of interruptions overall, or whether this type of interruption is more common for female speakers. The analysis on the *share* of interruptions of each type shows no significant gender difference for any of the outcomes considered, which means that women just get more of everything in equal proportion (Table A7).

6.2.3 Timing of Interruptions: Talking Over the Presenter (Online Talks Only)

One final way that we assess how male and female speakers may be treated differently is whether the audience chooses to speak over them. We can explore this in the sample of online presentations. In particular, recall that the speaker diarization takes narrow clips of speech and asks whether they are best modeled as a univariate distribution (reflecting a single speaker), or as a multivariate distribution (reflecting the influence of multiple voices). We use the latter to create a count of the number of interruptions that involve an audience member talking over the presenter. We find that just over 50 percent of interruptions in online seminars are mid-sentence.

Figure 4: Mid-Sentence Interruptions: Talking Over the Presenter



Notes: Data comes from the sample of online talks (Col. 4 of Table 1). "Talking over the presenter" is determined by computer algorithm. 54 percent of interruptions are mid-sentence in an average online seminar, and this does not vary based on the gender of the audience member interrupting. The figure shows coefficient estimates of three separate regressions following the specifications in Table 2 (Col. 4). The independent variable is a binary variable indicating the presenter is female, and the dependent variable is the total number of interruptions that talk over the presenter (last row), the number of such interruptions by men (row 2), and by women (row 1). Whiskers and the ends of the error bars indicate 90 percent and 95 percent confidence intervals, respectively.

We present the coefficient of interest in Figure 4, which shows that female speakers are talked over by the audience an extra 1.2 times per seminar (with a standard error of 0.3), relative to similarly situated men. All of these additional mid-sentence interruptions are by male audience members. Recall from Table 4 that in this sample the gender differential in the number of interruptions by male audience members is 1.6. This means that 75 percent of the extra interruptions by males for female presenters involve them being interrupted mid-sentence. As a point of comparison, in an average seminar, "only" 54 percent of all interruptions by males are mid-sentence.

6.3 Interruptions Over Time

After our data collection effort started, numerous economics departments and conferences, including NBER, set new ground rules for how seminars should operate. These ground rules range from simple actions like no interruptions in the first ten minutes and raising one's hand to be called on by the presenter, to having a moderator who guides seminar interactions and maintains a professional environment. Due to data limitations, we are not in a position to study whether such rules mitigate the effect of presenter gender.

However, within the large sample of online seminars, which span the period 2020 to 2023, we are able to test for a potential change over time. We find that the gender gap in the number of interruptions does not appear to have reduced in the second half of the study period (Table A8). The number of additional interruptions rated as negative has not declined either, but positive interruptions may have increased. A potential positive trend would point towards a reduction in the incidence of women being "talked over", but the estimates are noisy. Of course, it is possible that the trends *online* do not represent trends for in-person seminars.

7 Concluding remarks

Analyzing data from around 2,000 economics talks between 2019 and 2023, we show that female presenters are interrupted approximately 11 percent more times than their male counterparts, controlling for field, format, duration, and seminar series. The evidence points to two main factors explaining this. First, women have larger audiences, and more women in particular. As a result, they receive more interruptions from women. But this is not the sole explanation, since over half of the additional interruptions come from men. The second factor appears to be a shift in audience behavior in response to the presenter's gender. This shift leads a given-size audience to interrupt

female speakers more often.

Our data also reveal that interruptions in economics seminars, across speaker types, are: (i) very common, (ii) often commentary rather than questions, (iii) rated as negative (in terms of tone of voice) half the time, (iv) openly negative in *tenor* (e.g., hostile) in a few instances, and (v) often cutting off the presenter mid-sentence. These patterns suggest that the increased engagement women presenters receive in seminars may not be entirely benign. Indeed, we find that women receive a higher number of interruptions rated by fellow economists as negative in tenor, and disproportionately so. They also receive more interruptions with a negative tone of voice. What's more, the many interruptions, even when well intended, *de facto* reduce the time that presenters have to make a case for themselves and their research. As a result, the difference in the sheer number of interruptions between women and men presenters could have a disparate impact on women economists.

Yet, we also show that female presenters have more women attendees. Interestingly, women are also among those who interrupt female presenters more frequently. This pattern suggests a possible role model dynamic at work—similar to how women often take on additional mentoring responsibilities for junior female colleagues. The presence of women presenters may create an environment where female participants feel more comfortable engaging in academic discourse during economics seminars.

We employ a combination of objective and subjective methods to identify, quantify and rate interruptions. As a methodological note, studies on interruptions often report inconsistent results due to varying methodologies and unrepresentative samples (Baker, 2015; Tannen, 1993). Our analysis is based on a large dataset representative of multiple economics sub-fields and settings, some coded by humans but others containing dialogues coded by machine learning algorithms that are unaffected by observer bias. To our knowledge, this is one of the first studies in the social sciences to apply algorithmic audio processing, providing a comprehensive examination of audience participation and academic interactions.

Our findings likely do not reflect an explicit desire by seminar attendees to treat men and women differently but rather point to implicit biases. Such biases may extend beyond seminar interactions, potentially influencing career advancement for women economists. Increasing awareness of our own biases and adopting practices to mitigate them can provide the foundation to support a more welcoming environment for academic discourse.[24]

[24]The AEA Task Force on Best Practices for Professional Conduct in Economics recommends "setting and enforcing

The psychology and sociology literatures provide useful insights on ways to reduce implicit bias (see Correll (2017) for an overview). In particular, research suggests that we can mitigate the impact of implicit bias by "slowing ourselves down" (Eberhardt, 2020) and going systematically through a list of pre-defined criteria (Correll, 2017). For instance, audience members might ask themselves: *How important is the answer to this question at this moment? Could I find the answer if I looked through the paper? What is the likelihood that the answer will be provided later in the presentation?*

Another insight from this literature is to adhere to professional norms enforced by a moderator. In the last couple of years, following our early results, a number of leading economics departments and conferences (including the NBER) have surveyed their members, discussed potential remedies, and set new ground rules for how they want their seminars to operate. These ground rules range from simple actions like no questions in the first ten minutes and raising one's hand to be called on by the presenter, to having a moderator who guides seminar interactions and maintains a professional environment. Such interventions seem particularly important to ensure accurate assessment of job market candidates and/or appropriate modeling of behavior for training the next generation of economists.

rules of responsible behavior by attendees at conference and seminar presentations" (Bayer et al., 2021). The brochure can be downloaded here: https://www.aeaweb.org/resources/best-practices/brochure.

References

AEA (2021). Committee on the status of women in the economics profession (cswep) annual survey of u.s. economics departments, united states, 1994-2020.

Alexopoulos, M., Han, X., Kryvtsov, O., and Zhang, X. (2024). More than words: Fed chairs' communication during congressional testimonies. *Journal of Monetary Economics*, 142:103515.

Allgood, S., Badgett, L., Bayer, A., Bertrand, M., Black, S. E., Bloom, N., and Cook, L. D. (2019). Aea professional climate survey: Final report. *American Economic Association*.

Bai, Z. and Zhang, X.-L. (2021). Speaker recognition based on deep learning: An overview. *Neural Networks*, 140:65–99.

Baker, M. (2015). Over half of psychology studies fail reproducibility test. *Nature*, 27:1–3.

Bayer, A., Kalemli-Özcan, S., Pande, R., Rouse, C. E., Smith, A. A., Suárez Serrato, J. C., and Wilcox, D. W. (2021). Best practices for economists: Building a more diverse, inclusive, and productive profession. *AEA Papers and Proceedings*, 111:824–59.

Bhattacharya, S., Borah, S., Mishra, B. K., and Mondal, A. (2022). Emotion detection from multilingual audio using deep analysis. *Multimedia Tools and Applications*, 81(28):41309–41338.

Bisbee, J., Fraccaroli, N., and Kern, A. (2025). Yellin' at yellen: Hostile sexism in the federal reserve congressional hearings. *The Journal of Politics*.

Blair-Loy, M., Rogers, L. E., Glaser, D., Wong, Y. L. A., Abraham, D., and Cosman, P. C. (2017). Gender in Engineering Departments: Are There Gender Differences in Interruptions of Academic Job Talks? *Social Sciences*, 6(1):1–19.

Boustan, L. and Langan, A. (2019). Variation in women's success across phd programs in economics. *Journal of Economic Perspectives*, 33(1):23–42.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2022). Gender differences in peer recognition by economists. *Econometrica*, 90(5):1937–1971.

Card, D., DellaVigna, S., Funk, P., and Iriberri, N. (2023). Gender gaps at the academies. *Proceedings of the National Academy of Sciences*, 120(4):e2212421120.

Carter, A. J., Croft, A., Lukas, D., and Sandstrom, G. M. (2018). Women's visibility in academic seminars: Women ask fewer questions than men. *PLOS ONE*, 13(9):e0202743.

Chari, A. and Goldsmith-Pinkham, P. (2017). Gender representation in economics across topics and time: Evidence from the nber summer institute. Working Paper 23953, National Bureau of Economic Research.

Chen, J., Liu, Q., and Kim, M. (2022). Gender gap in tenure and promotion: Evidence from the economics ph.d. class of 2008. *Southern Economic Journal*, 88(4):1277–1312.

Conde-Ruiz, J. I., Ganuza, J.-J., García, M., and Puch, L. A. (2022). Gender distribution across topics in the top five economics journals: a machine learning approach. *SERIEs*, 13(1):269–308.

Correll, S. J. (2017). Sws 2016 feminist lecture: Reducing gender biases in modern workplaces: A small wins approach to organizational change. *Gender & Society*, 31(6):725–750.

Costello, A. M., Fedorova, E., Jin, Z., and Mihalcea, R. (2022). Editing a woman's voice. *arXiv Working Paper*.

Davenport, J. R., Fouesneau, M., Grand, E., Hagen, A., Poppenhaeger, K., and Watkins, L. L. (2014). Studying gender in conference talks–data from the 223rd meeting of the american astronomical society. *arXiv preprint arXiv:1403.3091*.

Dolado, J. J., Felgueroso, F., and Almunia, M. (2012). Are men and women-economists evenly distributed across research fields? some new empirical evidence. *SERIEs*, 3:367–393.

Doleac, J. L., Hengel, E., and Pancotti, E. (2021). Diversity in economics seminars: Who gives invited talks? *AEA Papers and Proceedings*, 111:55–59.

Donald, S. G. and Hamermesh, D. S. (2006). What is discrimination? gender in the american economic association, 1935-2004. *American Economic Review*, 96(4):1283–1292.

Dupas, P., Modestino, A. S., Niederle, M., Wolfers, J., and Collective, T. S. D. (2021). Gender and the dynamics of economics seminars. Working Paper 28494, National Bureau of Economic Research.

Eberhardt, J. L. (2020). *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do.* Penguin.

Ginther, D. K. and Kahn, S. (2004). Women in economics: Moving up or falling off the academic career ladder? *Journal of Economic Perspectives*, 18(3):193–214.

Ginther, D. K. and Kahn, S. (2021). Women in academic economics: Have we made progress? *AEA Papers and Proceedings*, 111:138–42.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The voice of monetary policy. *American Economic Review*, 113(2):548–84.

Handlan, A. and Sheng, H. (2023). Gender and tone in recorded economics presentations: Audio analysis with machine learning. *Unpublished*.

Hengel, E. (2022). Publishing while female: Are women held to higher standards? evidence from peer review. *The Economic Journal*, 132(648):2951–2991.

Hinsley, A., Sutherland, W., and Johnston, A. (2017). Men ask more questions than women at a scientific conference. *PLOS ONE*, 12:e0185534.

Jacobi, T. and Schweers, D. (2017). Justice, interrupted: The effect of gender, ideology, and seniority at supreme court oral arguments. *Va. L. Rev.*, 103:1379.

James, A., Kashyap, M., Victoria Chua, Y. H., Maszczyk, T., Núñez, A. M., Bull, R., and Dauwels, J. (2018). Inferring the Climate in Classrooms from Audio and Video Recordings: A Machine Learning Approach. In *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, pages 983–988. ISSN: 2470-6698.

Jarvis, S. N., Ebersole, C. R., Nguyen, C. Q., Zhu, M., and Kray, L. J. (2022). Stepping up to the mic: Gender gaps in participation in live question-and-answer sessions at academic conferences. *Psychological Science*, 33(11):1882–1893.

Kendrick, K. H. and Torreira, F. (2015). The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):255–289.

Knox, D. and Lucas, C. (2021). A dynamic model of speech for the social sciences. *American Political Science Review*, 115(2):649–666.

Koffi, M. (2021). Gendered citations at top economic journals. *AEA Papers and Proceedings*, 111:60–64.

Koffi, M. (2025). Innovative ideas and gender (in) equality. *American Economic Review (Forthcoming)*.

Liem, C. C., Langer, M., Demetriou, A., Hiemstra, A. M., Sukma Wicaksana, A., Born, M. P., and König, C. J. (2018). Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening. In *Explainable and interpretable models in computer vision and machine learning*, pages 197–253.

Lundberg, S. and Stearns, J. (2019). Women in economics: Stalled progress. *Journal of Economic Perspectives*, 33(1):3–22.

McCallum, A. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Miller, M. G. and Sutherland, J. L. (2022). The effect of gender, party and seniority on interruptions at congressional hearings. *American Political Science Review*, page 1–19.

Naim, I., Tanveer, M. I., Gildea, D., and Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., and Gatica-Perez, D. (2014). Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 16(4):1018–1031.

Paredes, V., Paserman, M. D., and Pino, F. J. (2023). Does economics make you sexist? *Review of Economics and Statistics*, pages 1–47.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). Speechbrain: A general-purpose speech toolkit. arXiv preprint arXiv:2106.04624.

Salem, V., McDonagh, J., Avis, E., Eng, P. C., Smith, S., and Murphy, K. G. (2021). Scientific medical conferences can be easily modified to improve female inclusion: a prospective study. *The Lancet Diabetes & Endocrinology*, 9(9):556–559.

Sarsons, H. (2017). Recognition for group work: Gender differences in academia. *American Economic Review: P& P*, 107(5):141–45.

Sarsons, H., Gërxhani, K., Reuben, E., and Schram, A. (2021). Gender differences in recognition for group work. *Journal of Political economy*, 129(1):101–147.

Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., Manohar, V., Dehak, N., Povey, D., Watanabe, S., et al. (2018). Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge. In *Interspeech*, pages 2808–2812.

Seré, M. (2023). Don't stop me now: Gender attitudes in academic seminars through machine learning. *Unpublished*.

Simonsohn, U. (2021). Madam Speaker: Are female presenters treated worse in econ seminars? *Data Colada blog, April 30*.

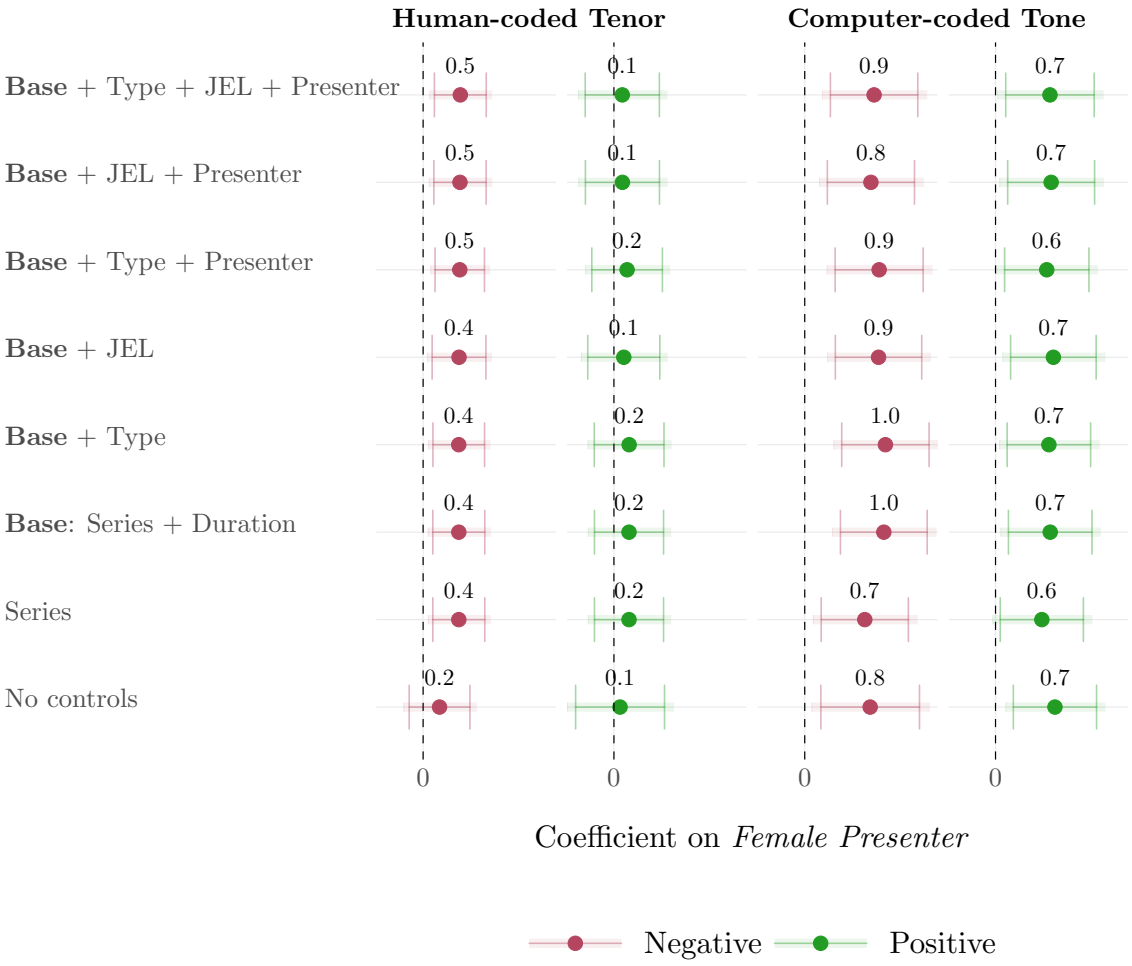
Tannen, D. (1993). *Gender and Conversational Interaction*. Oxford University Press.

Teodorescu, M., Ordabayeva, N., Kokkodis, M., Unnam, A., and Aggarwal, V. (2022). Determining systematic differences in human graders for machine learning-based automated hiring. *Brookings Working Paper Series*.

Wu, A. H. (2020). Gender Bias among Professionals: An Identity-Based Interpretation. *The Review of Economics and Statistics*, 102(5):867–880.

A Appendix Figures and Tables

Figure A1: Gender Gap in *Tenor* and *Tone* of Interruptions: Sensitivity to Controls



Notes: Human-coded samples include the job market talks, regular seminars and NBER SI 2019. Coders coded whether the tenor of the interruption was non-neutral, and if so, noted if it was "supportive", "collegial", or "valuable" (coded here as positive), or "demeaning", "patronizing", "disruptive" or "hostile" (coded here as negative). Computer-coded samples include the NBER SI 2022 and online talks. For computer-coded data, negative tones are interruptions that sound "negative-aggressive" (capturing emotions like "anger" or "disgust") and "negative-passive" (capturing "sadness" or "fearfulness"), while positive interruptions are those categorized as "happy" or "positively surprised." See details in Appendix C2.

Table A1: Inter-Coder Reliability: Correlation Coefficients

	All	Female presenter	Male presenter
Panel A: General			
Total	0.92	0.83	0.94
	{<0.001}	{<0.001}	{<0.001}
Number of observations	84	32	52
Panel B: Type			
Number of interruptions labeled as:			
Non-labeled	0.32	0.33	0.30
	{0.006}	{0.082}	{0.047}
Suggestion	0.47	0.38	0.53
	{<0.001}	{0.040}	{<0.001}
Comment	0.37	0.16	0.48
	{0.001}	{0.411}	{0.001}
Criticism	0.47	0.35	0.53
	{<0.001}	{0.069}	{<0.001}
Clarification	0.54	0.54	0.53
	{<0.001}	{0.003}	{<0.001}
Follow-up	0.16	-0.01	0.31
	{0.183}	{0.954}	{0.040}
Number of observations	74	29	45
Panel C: Tenor			
Number of interruptions labeled as:			
Positive	0.32	0.41	0.23
	{0.013}	{0.030}	{0.193}
Negative	0.50	0.48	0.50
	{<0.001}	{0.010}	{0.007}
Neutral	0.90	0.74	0.95
	{<0.001}	{<0.001}	{<0.001}
Number of observations	64	30	34

Notes: Table shows Spearman's rank-order correlation coefficients. Panel A includes all talks coded by two coders. Panel B excludes talks coded by two "super-type-coders" who labeled all interruptions with a type. Panel C exclude talks coded by four "super-tenor-coders" who assigned a non-neutral tenor to all interruptions or rated all non- neutral interruptions as positive or negative. Results on tenor (type) are unaffected when all talks coded by super-tenor (super-type) coders are excluded.

Table A2: Presenter Characteristics, by Gender

	Female	**Male**	**All**	**T-test (Female = Male)**
	Mean	Mean	Mean	P-value
Panel A: Job market talks (2019)				
PhD Institution				
Top 1–6 Economics Departments	0.55	0.51	0.53	0.743
Top 7–20 Economics Departments	0.16	0.33	0.26	0.104
Oth. U.S. Academic Institutions	0.06	0.08	0.07	0.780
Oth. Non-U.S. Academic Institutions	0.23	0.08	0.14	0.070
Placement				
Tenure Track	0.87	0.82	0.84	0.525
Tenure Track Top10	0.19	0.20	0.20	0.910
Tenure Track Top20	0.32	0.37	0.35	0.687
Post-doc	0.03	0.04	0.04	0.847
Non-Academic Job	0.10	0.06	0.07	0.562
Paper				
Published	0.13	0.10	0.11	0.714
Published Top 5	0.06	0.06	0.06	0.953
R&R Top 5	0.16	0.12	0.14	0.628
Published or R&R Top 5	0.23	0.18	0.20	0.651
Observations	31	49	80	
Panel B: Regular seminars (2019)				
Has Google Scholar Page	0.90	0.95	0.93	0.203
Total citations (x 10,000)	0.62	0.82	0.76	0.297
Top 1–6 Economics Departments	0.19	0.18	0.18	0.919
Top 7–20 Economics Departments	0.20	0.27	0.25	0.287
Oth. U.S. Academic Institutions	0.40	0.33	0.35	0.338
Oth. Non-U.S. Academic Institutions	0.17	0.14	0.15	0.545
Non-academic Institutions	0.04	0.08	0.07	0.311
Observations	70	150	220	
Panel C: Online Talks (2020-2023)				
Total citations (x 10,000)	0.48	1.24	1.00	0.000
Years of experience	8.97	11.70	10.82	0.000
Top 1–6 Economics Departments	0.11	0.13	0.12	0.459
Top 7–20 Economics Departments	0.17	0.18	0.17	0.722
Oth. U.S. Academic Institutions	0.37	0.28	0.31	0.006
Oth. Non-U.S Academic Institutions	0.28	0.34	0.32	0.082
Non-academic Institutions	0.08	0.08	0.08	0.758
Observations	290	611	901	

Notes: In each panel, an observation is a unique presenter. E.g., if a given job market candidate presented her paper in many places and shows up six times in our dataset (presenting in six different settings), we consider this as just one observation in this summary statistics table. We coded placement and paper outcomes in September 2022 by looking at speakers' websites and online CVs. See Appendix D for details on the definition of presenter's home institution and citations, respectively.

Table A3: Paper Characteristics, by Gender

	Female	**Male**	**All**	**T-test (Female = Male)**
	Mean	Mean	Mean	P-value
Panel A: Job market talks (2019)				
Field				
Micro	0.59	0.41	0.50	0.016
Macro	0.23	0.24	0.24	0.854
Other	0.17	0.34	0.26	0.010
Type				
Empirical	1.00	0.83	0.91	0.000
Experimental	0.10	0.11	0.11	0.891
Policy	0.86	0.52	0.69	0.000
Theory	0.26	0.68	0.47	0.000
Observations	86	90	176	
Panel B: Regular seminars (2019)				
Field				
Micro	0.85	0.71	0.75	0.019
Macro	0.10	0.19	0.16	0.080
Other	0.05	0.10	0.09	0.190
Type				
Empirical	0.90	0.84	0.86	0.189
Experimental	0.25	0.25	0.25	0.996
Policy	0.73	0.61	0.65	0.089
Theory	0.37	0.48	0.44	0.133
Observations	78	167	245	
Panel C: Online talks (2020-2023)				
Field				
Micro	0.29	0.30	0.30	0.875
Macro	0.46	0.41	0.42	0.132
Other	0.25	0.30	0.28	0.135
Type				
Empirical	0.72	0.64	0.66	0.007
Experimental	0.15	0.09	0.11	0.013
Policy	0.55	0.50	0.52	0.177
Theory	0.38	0.49	0.45	0.001
Observations	318	676	994	

Notes: Characteristics at the paper level and unit of observation is unique talks. See Appendix C for details on the definition of a paper's field and type, respectively.

Table A4: Number of Interruptions: Controlling for Attendance

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) **All Samples** (cols. 1–3)
Female presenter	3.335	2.743	0.835	1.837
	(1.720)	(1.315)	(0.742)	(0.615)
	{0.054}	{0.038}	{0.261}	{0.003}

Notes: This table reproduces Table 2 for the samples for which attendance information is available, and adds a control for attendance.

Table A5: Interruptions by Tenor/Tone and Gender of Interlocutor

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **Human-coded** (cols. 1-3)	(7) **Computer-coded** (cols. 4-5)
Panel A: Negative interruptions by men							
Female presenter	0.437	0.181		0.762	0.016	0.332	0.607
	(0.259)	(0.205)		(0.357)	(0.669)	(0.158)	(0.315)
	{0.093}	{0.377}		{0.033}	{0.981}	{0.036}	{0.055}
Mean (male presenter)	1.57	0.49		5.19	5.98	0.87	5.35
Panel B: Negative interruptions by women							
Female presenter	0.078	0.211		0.156	0.624	0.141	0.254
	(0.078)	(0.165)		(0.147)	(0.326)	(0.085)	(0.135)
	{0.320}	{0.203}		{0.288}	{0.057}	{0.098}	{0.059}
Mean (male presenter)	0.11	0.11		0.84	1.10	0.11	0.89
Panel C: Positive interruptions by men							
Female presenter	-0.112	-0.025	0.229	0.653	-0.339	0.030	0.446
	(0.344)	(0.527)	(0.345)	(0.214)	(0.264)	(0.234)	(0.179)
	{0.744}	{0.963}	{0.508}	{0.002}	{0.201}	{0.899}	{0.013}
Mean (male presenter)	1.06	2.89	2.48	2.00	1.92	2.37	1.99
Panel D: Positive interruptions by women							
Female presenter	0.073	-0.079	-0.012	0.257	0.069	-0.019	0.218
	(0.066)	(0.247)	(0.184)	(0.336)	(0.274)	(0.115)	(0.274)
	{0.275}	{0.749}	{0.950}	{0.444}	{0.802}	{0.867}	{0.426}
Mean (male presenter)	0.20	0.65	0.82	1.10	1.29	0.67	1.14
Number of observations	240	291	447	994	257	978	1,251
Number of unique talks	176	244	442	994	257	862	1,251

Notes: Standard errors in parentheses and p-values in brackets. See Table 1 notes for data sources and Table 2 for the list of controls.

Table A6: Tenor/Tone of Interruptions (Shares)

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020–2023	(5) NBER SI 2022	(6) **Human-coded** (Cols. 1–3)	(7) **Computer-coded** (Cols. 4–5)
Panel A: Negative interruptions							
Female presenter	0.015	0.020		-0.007	0.021	0.018	-0.001
	(0.009)	(0.015)		(0.022)	(0.032)	(0.008)	(0.019)
	{0.086}	{0.177}		{0.754}	{0.516}	{0.023}	{0.957}
Mean (male presenter)	0.046	0.024		0.542	0.468	0.034	0.527
Number of observations	240	291		945	257	531	1,202
Number of unique talks	176	244		945	257	420	1,202
Panel B: Positive interruptions							
Female presenter	0.011	-0.025	0.029	0.019	-0.024	0.011	0.009
	(0.008)	(0.028)	(0.037)	(0.022)	(0.021)	(0.020)	(0.018)
	{0.149}	{0.387}	{0.425}	{0.393}	{0.256}	{0.586}	{0.598}
Mean (male presenter)	0.030	0.182	0.238	0.342	0.204	0.179	0.313
Number of observations	240	292	447	945	257	979	1,202
Number of unique talks	176	245	442	945	257	863	1,202

Notes: Standard errors in parentheses and p-values in brackets. Specifications identical to those shown in Table 5 but where the dependent variable is the share of tone interruptions rather than count. Col. 3: Negative interruptions were not labeled as such for the NBER 2019 sample as per the data collection agreement. Col. 5: Sample size for NBER 2022 reduced from the sample in Table 2 to those seminars with sufficiently good audio quality for audience members for tone analysis.
Human-coded negative interruptions include those labeled as patronizing, disruptive, demeaning and/or hostile, and positive interruptions were labeled as supportive, constructive, collegial, or valuable. Machine-coded negative interruptions were predicted to be either negative-aggressive (angry or disgust) or negative-passive (sad or fearful) and positive interruptions were predicted to be happy or positively surprised based on training data on tones from voice actors.

Table A7: Types of Interruption (Shares)

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **Human-coded** (Cols. 1-3)
Coef. on female presenter:						
Type of interruption:						
Non-labeled	0.034	0.011	0.044		n.a	0.026
	(0.033)	(0.027)	(0.032)			(0.020)
	{0.304}	{0.688}	{0.168}			{0.180}
Suggestions	0.014	0.011	0.022			0.017
	(0.010)	(0.014)	(0.021)			(0.012)
	{0.175}	{0.445}	{0.307}			{0.162}
Comments	-0.004	-0.025	-0.030			-0.023
	(0.017)	(0.018)	(0.025)			(0.015)
	{0.808}	{0.161}	{0.228}			{0.111}
Criticism	-0.022	0.001	-0.006			-0.008
	(0.012)	(0.013)	(0.016)			(0.009)
	{0.073}	{0.912}	{0.712}			{0.407}
Clarifications	-0.010	-0.011	-0.035			-0.014
	(0.025)	(0.025)	(0.025)			(0.016)
	{0.696}	{0.650}	{0.163}			{0.362}
Follow-ups	-0.018	0.003	-0.009			-0.007
	(0.014)	(0.013)	(0.016)			(0.009)
	{0.197}	{0.847}	{0.584}			{0.486}
Interrogative				0.014		
				(0.016)		
				{0.365}		
Declarative				-0.014		
				(0.016)		
				{0.371}		
Mean for male presenter:						
Non-labeled	0.35	0.37	0.27			0.31
Suggestions	0.07	0.10	0.21			0.15
Comments	0.15	0.18	0.24			0.21
Criticism	0.10	0.07	0.13			0.11
Clarifications	0.41	0.36	0.32			0.36
Follow-ups	0.09	0.08	0.10			0.09
Interrogative				0.41		
Declarative				0.59		
Number of observations	240	292	447	945		979
Number of unique talks	176	245	442	945		863

Notes: Standard errors in parentheses and p-values in brackets. Specifications identical to those shown in Table 6 but where the dependent variable is the share of interruptions of each type rather than count. Information on the type of interruption was not recorded for the NBER 2022 sample as per the data use agreement.

Table A8: Presenter Gender and Interruptions, Over Time

	(1) Total interruptions	(2) Negative interruptions	(3) Positive interruptions	(4) Mid-sentence interruptions
Female presenter, 2020-2021	1.754	0.886	0.776	1.342
	(0.486)	(0.417)	(0.308)	(0.334)
	{<0.001}	{0.034}	{0.012}	{<0.001}
Female presenter, 2022-2023	2.696	1.053	1.532	0.624
	(1.413)	(0.650)	(1.381)	(0.618)
	{0.057}	{0.106}	{0.267}	{0.313}
Controls:				
Seminar series	Yes	Yes	Yes	Yes
Seminar duration	Yes	Yes	Yes	Yes
Presenter's rank	Yes	Yes	Yes	Yes
Presenter's citation count	Yes	Yes	Yes	Yes
Presenter's home institution	Yes	Yes	Yes	Yes
Paper topic	Yes	Yes	Yes	Yes
Mean (male presenter)	10.93	6.03	3.81	5.79
Number of observations	994	994	994	994

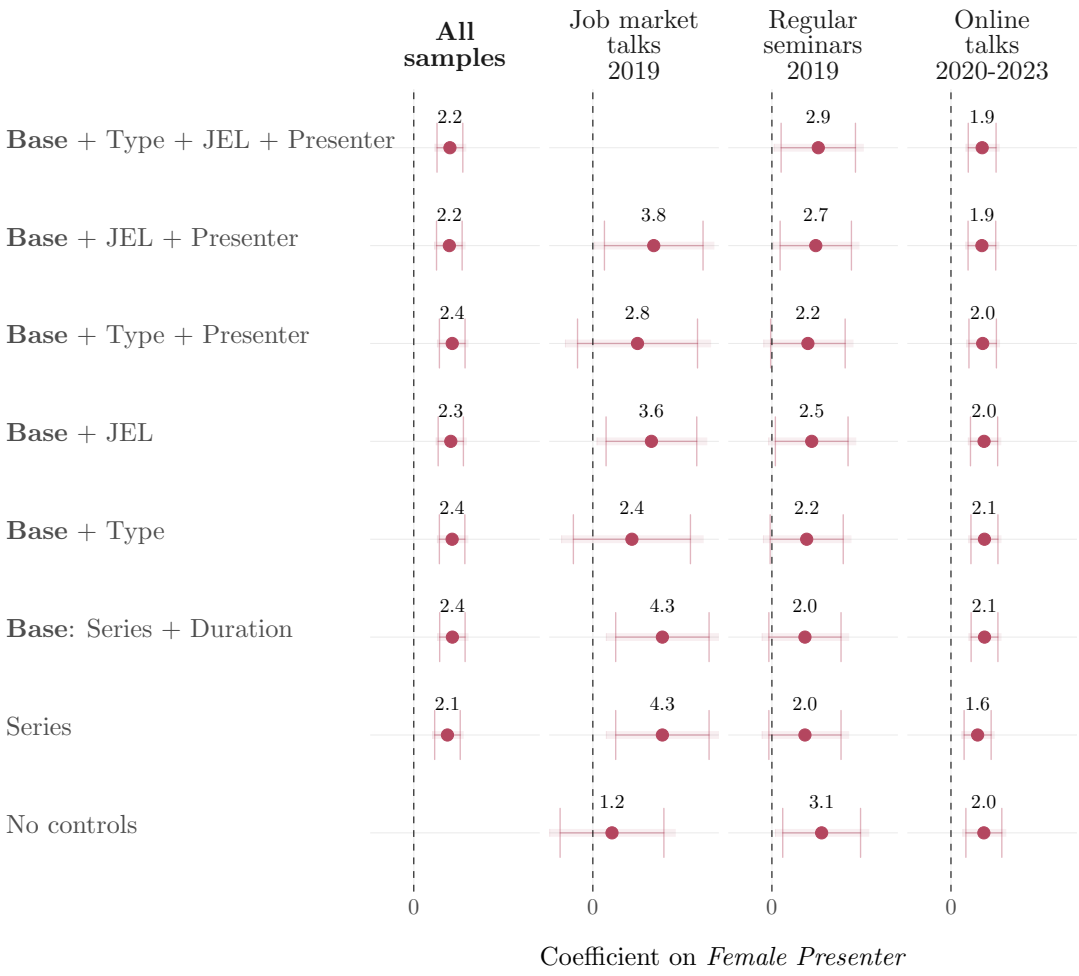
Notes: Sample restricted to Online Seminars (col. 4 of Table 1). There are 210 talks in the period 2022-2023, 61 by female presenters. Interruptions can be positive, negative, or neutral. Standard errors in parentheses and p-values in brackets.

Online Appendix

Table of Contents

B Additional Figures and Tables

Figure B1: Gender Gap in Interruptions: Sensitivity to Controls with Standard Errors Clustered at the Presenter Level



Notes: This reproduces Figure 1, with the additional of clustering at the presenter level, for the three samples for which we have individual identifiers and can cluster the standard errors at the presenter level.

Table B1: Main Outcomes for Job Market Talks: Reliability of Paper-Type Controls

	(1) Job market talks JEL only	(2) Job market talks Type only	(3) Job market talks JEL + Type	(4) Job market talks Joint	(5) Regular seminars JEL+Type	(6) Regular seminars Joint	(7) Pooled sample JEL+Type
Panel A: Total Attendance							
Female presenter	5.052	3.624	3.362	6.128	2.034	2.519	4.169
	(1.570)	(1.843)	(1.836)	(1.586)	(1.092)	(1.193)	(0.997)
	{0.002}	{0.051}	{0.069}	{<0.001}	{0.064}	{0.035}	{<0.001}
Panel B: Total Interruptions							
Female presenter	3.768	2.763	3.458	3.445	2.870	2.788	3.095
	(1.556)	(1.828)	(2.073)	(1.484)	(1.285)	(1.252)	(0.973)
	{0.016}	{0.132}	{0.097}	{0.021}	{0.027}	{0.027}	{0.002}
Panel C: Negative interruptions							
Female presenter	0.515	0.365	0.220	0.544	0.369	0.389	0.461
	(0.278)	(0.325)	(0.315)	(0.284)	(0.308)	(0.281)	(0.195)
	{0.065}	{0.262}	{0.486}	{0.056}	{0.233}	{0.167}	{0.019}
Panel D: Positive interruptions							
Female presenter	-0.030	-0.205	-0.285	-0.433	-0.102	-0.166	-0.291
	(0.374)	(0.419)	(0.490)	(0.416)	(0.615)	(0.563)	(0.368)
	{0.937}	{0.626}	{0.562}	{0.298}	{0.868}	{0.768}	{0.429}
Panel E: Types of Interruptions							
Coef. on female presenter:							
Type of interruption:							
Non-labeled	2.350	1.550	1.681	2.073	0.555	0.452	1.211
	(1.312)	(1.592)	(1.599)	(1.314)	(0.999)	(0.990)	(0.800)
	{0.075}	{0.331}	{0.294}	{0.115}	{0.579}	{0.648}	{0.131}
Suggestions	0.751	0.493	0.709	0.629	0.673	0.612	0.620
	(0.400)	(0.451)	(0.449)	(0.394)	(0.382)	(0.357)	(0.262)
	{0.062}	{0.276}	{0.116}	{0.111}	{0.080}	{0.087}	{0.018}
Comments	0.046	-0.340	-0.029	-0.489	0.131	0.244	-0.099
	(0.638)	(0.769)	(0.849)	(0.622)	(0.599)	(0.570)	(0.430)
	{0.942}	{0.658}	{0.972}	{0.432}	{0.828}	{0.669}	{0.817}
Criticism	-0.417	-0.373	-0.101	-0.481	0.158	0.128	-0.157
	(0.449)	(0.523)	(0.517)	(0.454)	(0.385)	(0.372)	(0.288)
	{0.354}	{0.476}	{0.845}	{0.290}	{0.682}	{0.731}	{0.586}
Clarifications	1.646	2.104	1.965	1.965	1.119	1.094	1.502
	(1.079)	(1.138)	(1.330)	(1.038)	(0.745)	(0.738)	(0.621)
	{0.129}	{0.066}	{0.141}	{0.059}	{0.135}	{0.139}	{0.016}
Follow-ups	-0.199	-0.200	-0.204	-0.131	0.390	0.392	0.147
	(0.509)	(0.503)	(0.539)	(0.472)	(0.359)	(0.333)	(0.284)
	{0.696}	{0.691}	{0.705}	{0.782}	{0.278}	{0.240}	{0.604}

Notes: Col. (1) is the preferred specification for the job market sample from Table 2. Col. (2) includes estimation for the job market sample with controls for JEL and Type controls. Col. (4) is the preferred specification for the regular seminar sample with all controls, including JEL and Type controls, from Table 2. "Joint" cols. (3) and (5), are estimated together. The control variables (including JEL and Type) have coefficients that are estimated jointly, and separate coefficients are estimated for the job market sample and for the regular seminar sample. Col. (6) estimates one female presenter coefficient for the pooled job market and regular seminar samples and uses both JEL and type controls.

There are only 39 unique job market papers by female presenters, limiting our ability to estimate on all coefficients reliably. To show that "type" controls do not overturn the results for the job market sample, we pool the job market sample and the regular seminar sample to jointly estimate the coefficients on controls while estimating sample-specific coefficients for presenter gender. We find almost the same results (column 4) as when we only control for JEL and not type (column 1). In contrast, the JEL+Type specification estimated on the job market sample alone tends to deviate from the results in the jointly estimated specification (column 3 vs. column 4). For this reason, we favor the JEL-only specification for the job market sample. For the regular seminar sample, the strong similarity in results between columns 5 and 6 indicates that the JEL+Type specification is appropriate.

Table B2: Number of Interruptions: Sensitivity to Controls

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020-2023	(5) NBER SI 2022	(6) **All Samples** (cols 1–5)
Panel A: No controls						
Female presenter	1.184	3.077	-1.295	2.034	-0.532	0.970
	(1.581)	(1.418)	(1.001)	(0.656)	(0.751)	(0.420)
	{0.455}	{0.031}	{0.196}	{0.002}	{0.479}	{0.021}
Controls:	None	None	None	None	None	None
Panel B: Most comparable controls						
Female presenter	4.303	2.044	0.820	2.076	0.227	1.660
	(1.403)	(1.219)	(0.732)	(0.492)	(0.602)	(0.326)
	{0.002}	{0.095}	{0.263}	{<0.001}	{0.706}	{<0.001}
Controls:						
Seminar series	Yes	Yes	Yes	Yes	Yes	Yes
Seminar duration	Yes	Yes	Yes	Yes	Yes	Yes
Mean (male presenter)	34.39	25.35	14.68	10.93	13.74	15.15
Number of observations	240	292	447	994	467	2,440
Number of unique talks	176	245	442	994	467	2,324

Notes: This table replicates Table 2 with different control specifications. See Figure 1 for full sensitivity analysis.

Table B3: Presenter Gender and Number of Interruptions, by Broad Field

	(1) Job market talks 2019	(2) Regular seminars 2019	(3) NBER SI 2019	(4) Online talks 2020–2023	(5) NBER SI 2022	(6) **All Samples** (cols. 1–5)
Female presenter						
x Macro	5.308	-1.777	4.128	1.477	0.413	1.645
	(2.809)	(4.862)	(1.872)	(0.691)	(0.896)	(0.544)
	{0.060}	{0.715}	{0.028}	{0.033}	{0.645}	{0.003}
x Micro	1.220	3.774	-0.750	1.478	0.045	0.929
	(2.056)	(1.368)	(0.745)	(0.694)	(0.808)	(0.430)
	{0.554}	{0.006}	{0.315}	{0.034}	{0.956}	{0.031}
x Other	7.662	0.838	0.440	3.258		3.182
	(2.926)	(2.745)	(0.861)	(1.294)		(0.974)
	{0.010}	{0.760}	{0.610}	{0.012}		{0.001}
Mean for male presenter:						
Macro	35.75	29.31	17.26	11.73	15.53	15.66
Micro	37.59	25.76	13.95	10.02	11.72	15.75
Other	29.60	15.03	8.73	10.77	n.a	12.72
Number of observations:						
Macro	51	41	215	419	240	924
Micro	124	224	173	294	227	1,084
Other	65	27	59	281		432

Notes: Standard errors in parentheses and p-values in brackets. Specifications identical to those shown in Table 2. As per the data use agreement, all talks in the NBER 2022 sample were labeled as either "Macro" or "Micro". Col. 5: For the smaller sample used in **??** and Table 5, the coefficient for female x micro is 0.349, with a se of 1.163, and the coefficient for female x macro is 0.759, with a se of 1.357. The control mean for Macro seminars is 16.22, and 13.27 for Micro.

C Variable Descriptions

This section includes detailed descriptions of variables used in this paper. Additional details on human coders (including tests for coder biases) can be found in Dupas et al. (2021). Additional details on machine coding can be found in Appendix D.

C.1 Number of Interruptions

Number of interruptions is the number of times that people other than the presenter spoke during the presentation.

Job market and departmental seminars: The number of interruptions was recorded by human coders using the Qualtrics tool developed by Dupas et al. (2021). Each time the presenter was interrupted, the coder registered it in the software.

NBER SI 2019: Same as for the job market and departmental seminars.

Online seminars: The number of interruptions is based on the number of times non-presenters spoke in the presentation recordings. We identify when there are changes between speakers on this large sample using the automatic, diarization procedure from Seré (2023). Then the dominant speaker is labeled as the presenter and the remaining speakers are labeled as audience. We drop extremely short interruptions and start counting interruptions once the presenter starts giving the presentation (after the introduction). We additionally identify moderators based on the non-presenter speaking in the beginning and drop seminars that are dominated by the moderator asking questions (representing reading questions from an online chat). The full analysis is in Seré (2023) and is qualitatively consistent with the results in this paper.

NBER SI 2022: The number of interruptions is based on the number of times non-presenter and non-discussants spoke in presentation recordings. We create time-stamps for when each person is speaking using Trint. Manually, we label if that is the presenter, discussant, or an audience member talking. We drop interruptions that are extremely short. We also do not include discussions in the room of the presentation after the formal talk is concluded (sometimes the recording continued after the presentation ended). Additionally, we drop introductions from the moderator at the beginning of the talk to introduce the speaker.

C.2 Tenor and Tone of Interruptions

Job market and departmental seminars: Tenor of interruptions for this sample was assessed by the human coders using the Qualtrics tool. Coders could leave an interruption unlabeled (in fact, this was the default option), or could be labeled as patronizing, disruptive, demeaning, hostile, or supportive. We consider the first four labels as indicative of "Negative" tenor, and the latter one as "Positive" tenor. The labels were assigned according to the coder's perception of the interruption.

The specific written instructions to the coders were: "You should think of this as assessing the interaction towards the speaker. We aren't asking you to code the intention of the person making the comment, nor how it was taken by the speaker. But rather, it's your assessment of the tenor of the comment in a scientific setting. *You may leave this blank for many interjections, only filling this up if you think it is warranted.* You can click as many buttons as you deem appropriate, so, an interaction can be, e.g. just supportive, just patronizing or both. Or nothing. Use your judgment."

NBER SI 2019: Tenor of interruptions for this sample was assessed by the human coders using the Qualtrics tool. The default was that the question was neutral (no label). Positive interruptions were interruptions labeled as particularly valuable, constructive or collegial according to the coder's perception of the interruption. Due to the data-use agreement with NBER, we were not able to count negative-tenor interruptions. Coders received the same instructions as for the two other human-coded samples mentioned above.

Online seminars: Tone of interruptions is imputed using a deep neural network to predict tone from audio recordings. Negative and positive tone-of-voice here represent "how it sounds" to the machine compared to the tones of voice actors used to train the algorithm. Tone from the machine does not include "what it is said," body language, or context from earlier in the seminar. Negative interruptions are those the machine predicted to be either negative-aggressive (angry or disgusted) or negative-passive (sad or fearful). Positive interruptions are those the machine predicted as positive (happy or positively surprised). We train separate tone-prediction models for male and female speakers due to average differences in voice pitch across gender. We require the deep neural network to assign over a 50 percent probability of a positive or negative tone, otherwise the interruption is labeled as neutral. A more detailed description of training and robustness is in Appendix D.

NBER SI 2022: Same as for the online seminars.

C.3 Types of Interruptions

Job market and departmental seminars: Type of interruptions for this sample was assessed by the human coders using the Qualtrics tool. Coders had the option to record if an interruption had any of the following (non-exclusive) characteristics: clarification, criticism, suggestion, comment, or follow-up. A clarification is asking the presenter to clarify a previous point or a something shown on the slides. A criticism is a direct critique or challenge to the presenter's work. A suggestion is suggesting that the presenter conducts additional analyses or modifies their analysis in a specific way. A comment is a broader category capturing a non-question interruption that does not neatly fall under criticism or suggestion. A follow-up is an interruption that picks up on a previous interruption by the same or another audience member.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: There are two classification schemes for interruption types in this sample: first, whether an interruption is a question (interrogative) or not (declarative). Second, whether an interruption involves the audience talking over the presenter or not.

Interruptions are categorized as either interrogative or declarative depending on whether the transcript of the interruption includes a question mark (?). We feed the presentation transcript through standard text-cleaning software to ensure appropriate punctuation is included to match the words used. Seré (2023) includes two alternative specifications of this: one uses pattern-matching approach that scans the text of each interruption for specific phrases and structures commonly associated with questions, such as "do you think," "what is," or "how can." The other uses a text-prediction algorithm to take text as input and predict where it was a question or not using a BERT classifier. All methods return similar results.

Interruptions are also categorized by whether the audience talking over the presenter or not. Using the audio recording and diarization algorithm, we are able to see if audio wavelengths are best modeled as a univariate distribution (reflecting a single speaker), or as a multivariate distribution (reflecting the influence of multiple voices). Therefore, audience member speech (an interruption) and is best modeled with a multivariate distribution is labeled as an interruption where the audience member talking over the presenter.

NBER SI 2022: We were not allowed to save transcripts of the seminars due to our data-use agreement with NBER for this sample. Therefore we were not able to generate data on interruption types.

C.4 Speaker Characteristics

C.4.1 Speaker's Gender

Job market and departmental seminars: Speaker gender for the presenter, discussant, or audience member for this sample was recorded by the human coders using the Qualtrics tool. This was based on the coder's observation from the seminar and name of the presenter, discussant, or audience member.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: Presenter, discussant, and audience gender is inferred from an audio processing algorithm that compares the similarity of the speaker's voice to those of the voices of male and female voice actors in a variety of datasets. Appendix D includes more detail on the algorithm structure and evaluation.

NBER SI 2022: Same as for the online seminars.

C.4.2 Presenter's Home Institution

Fixed effects for the presenter's home institution use the institution's type, rank, and/or country.

Job market and departmental seminars: Presenter home institution for this sample was recorded by the human coders using the Qualtrics tool. We have too many singletons to include institution-specific fixed effects for each presenter's home departments so we instead group home institutions into coarser clusters by type and rank, yielding 5 institution-rank fixed effects for each of: Top 1-6 economics departments; Top 7-20 economics departments; Other U.S. academic institutions; Other institutional academic institutions; Non-academic institutions. University ranks for top 10 and 20 institutions is from the US News and World Report 2017 Rankings.

NBER SI 2019: Presenter home institution for this sample was observed by the human coders. Coders then referenced an excel sheet that listed the University ranks for top 10 and 20 institutions is from the US News and World Report 2017 Rankings. Coders then recorded whether the presenter's home institution was a top 20 institution, a non-top 20 institution, or a non-academic institution using the Qualtrics tool.

Online seminars: The online seminars sample includes 7 fixed effects: Top 1-6 economics departments; Top 7-20 economics departments; Other U.S. academic institutions; Other EU academic institutions; Other UK academic institutions; Non-US/EU/UK institutions; and Central Banks, IMF, World Bank, etc. The additional fixed effects, compared to the human-collected sample, are due to the increased heterogeneity of this sample, which encompasses a broader set of presentations and conferences from a more diverse international context, not limited to the U.S.

NBER SI 2022: Information on individual presenters could not be collected for this sample, in agreement with NBER.

C.4.3 Presenter's Citation Count

We collect information on citations using Google Scholar profiles when we were allowed to collect the presenter's identity. For the NBER SI 2019 and NBER SI 2022 samples, our data-use agreements with NBER prevented us collecting the name of presenters and therefore we were not allowed to collect citations. When a new video was detected, it was downloaded and processed, and the presenter's name—extracted from the video—was used to scrape their Google Scholar profile. This process was conducted in multiple waves between 2022 and 2023, so citation counts reflect the moment the video was downloaded. As a result, presenters who appeared more than once may have different citation counts across observations, reflecting values close to the time of each seminar.

Job market and departmental seminars: Citation counts were obtained from the Google Scholar profile of the presenter for the 2019 departmental seminars. For the job market talks, it was not meaningful to collect citation data, as presenters are typically completing their PhDs when entering the job market and few have any citations.

Online seminars: Citation counts were obtained from the Google Scholar profile of the presenter.

C.4.4 Presenter's Rank and Seniority

Job market and departmental seminars: We do not collect seniority measures for presenters in job talks or departmental seminars. For job talks it was not meaningful to collect seniority data, as presenters are typically completing their PhDs and entering the job market.

NBER SI 2019: Presenter's seniority (categorical) for this sample was observed by the human coders and recorded using the Qualtrics tool for three broad tiers per our data use agreement with the NBER. The three tiers were senior (tenured) faculty member, junior (untenured) faculty member, and other (student / post-doc / non-academic).

Online seminars: Presenter's seniority (continuous) was computed as the difference between the year in which the presenter published their first paper (obtained from Google Scholar) and the year in which the seminar took place (2020 to 2023).

NBER SI 2022: Our data-use agreement with NBER for this sample prevented us collecting the name of presenters and therefore we were not able to look up presenters' seniority.

C.5 Research Characteristics

C.5.1 Paper/Presentation Topic (JEL)

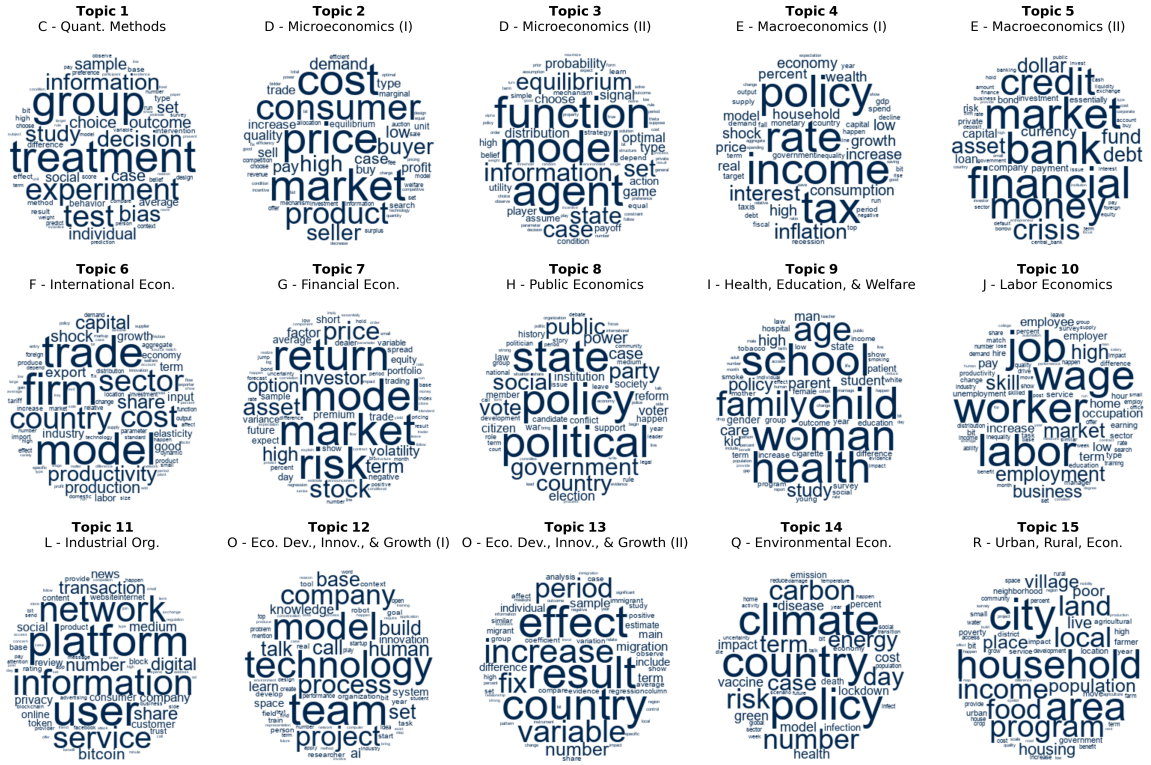
Job market and departmental seminars: JEL codes were typically drawn from the title page of the paper, and if multiple codes were listed we chose the most frequent, and in the event of a tie, chose the most relevant. If JEL codes were not listed, we chose the most relevant based on the content of the paper. We then constructed 7 fixed effects corresponding to single-digit JEL codes: C (math and quant methods), D (micro), E+F+G (macro, international, and financial), H+I (public, health, education and welfare), J (labor), L (industrial organization), Other codes.

NBER SI 2019: For this sample we did not collect paper topics since we were not allowed to record identifying information about the presenter, their paper, or the program.

Online seminars: Presentation topics are identified using a topic modeling approach applied to the audio transcripts of the seminars and are matched to JEL codes. For that, a pre-processed transcripts were analyzed as the corpus. A "topic" in this context is defined as a cluster of words that frequently appear together. The analysis was conducted using Mallet, a toolkit for topic modeling that includes sampling-based implementations of Latent Dirichlet Allocation (McCallum, 2002). The number of topics was selected using the coherence-score of models with varying number of topics, and the one with 15 topics achieved the highest coherence value before a plateau in the score occurs. The word clouds in Figure C2 display the most probable words for each topic. We match the word clouds to the closest JEL code using a natural language processing tool. Then we group the 15 topics into five groups of JEL codes: macro, micro, finance, econometrics, and general.

NBER SI 2022: For this sample we did not collect paper topics, and due to our data-use agreement we were unable to match papers to recordings after our initial data processing had completed (which included deleting recordings).

Figure C2: Paper Topics in Online Seminars



C.5.2 Research Field

For our analysis in grouping seminar series by field, we create the broader research field categories of micro and macro. For some samples we also categorize into theory/econometrics and others.

Job market and departmental seminars: For presentations in regular departmental seminars, fields are based on seminar series. For example, all papers presented in the Harvard macroeconomics seminar are classified as macro. We grouped seminars into three broad categories: micro, macro, and theory/econometrics. For the job market talks, we coded field ex-post based on the paper's research topic (as described above).

NBER SI 2019: Human coders observed which program the paper was presented in and then used an excel spreadsheet listing the research field based on the manual labels of programs from Chari and Goldsmith-Pinkham (2017). We categorized fields into the broad categories of macro, micro, and finance, as per our data use agreement with NBER for this sample.

Online seminars: The research field for the seminar series is an aggregation from the paper topics into three categories: macro, micro, and others. This aggregation process was assisted by a natural language processing tool. Specifications run with series' field labeled by a research assistant return similar results as those from the topic model.

NBER SI 2022: Research field comes from manual labels of programs from Chari and Goldsmith-Pinkham (2017). Due to our data-use agreement for this sample, we categorized programs as either macro or micro and were unable to have a third category for finance or more granular fields. Handlan and Sheng (2023) include a table which shows the program assignment to different fields.

C.5.3 Paper/Presentation Types

We code four additional characteristics for paper/presentation topics using an LLM: (1) Theory: Is the primary purpose of this paper developing economic theory? (2) Data: Is the primary purpose of this paper analyzing empirical data? (3) Experimental: Does this paper involve a field or lab experiment? (4) Policy: Is this paper substantially engaged with analyzing economic policies? Details on this process are in the appendix Section D.10.

Job market and departmental seminars: We use the entire paper text for the paper corresponding to the presentation for this sample with the LLM to code the the four type categories (theory/-data/experimental/policy).

NBER SI 2019: By agreement with the NBER for this sample, individual presentation titles (and corresponding papers) were not allowed to be coded.

Online seminars: We use the transcripts from the online seminar presentations with the LLM to code the the four type categories (theory/data/experimental/policy).

NBER SI 2022: As with the NBER SI 2019 sample, we were not allowed to code the presentation title due to our agreement.

C.6 Seminar Duration

All measures of seminar duration are measured in minutes. There are some decisions we made in selecting our samples that affect the average duration measures. For the NBER SI samples, we drop very short presentations, including the lightning rounds and egg timers which are under 15 minutes. For the NBER SI 2022 sample, we also drop presentations where the recording was interrupted due to technical issues. For the online seminar sample, we exclude seminars shorter that 20 minutes and longer than 120 minutes.

Job market and departmental seminars: Seminar duration was recorded by the human coder using the Qualtrics tool. It is measured as the actual number of minutes between the seminar's start time and end time based on when coders clicked the start and end buttons, rather than the scheduled time.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: Seminar duration was measured from the presentation recordings. The start of the presentation starts when the main presenter begins speaking.

NBER SI 2022: Seminar duration was measured from the presentation recordings. The start of the presentation starts when the main presenter begins speaking and ends once the presentation, discussion, and formal questions conclude. We manually verified when the presentation ends and people break for small conversations, and do not include measures for after the end of the presentation.

C.7 Seminar series fixed effects

We use seminar series fixed effects to control for differences in seminar culture across series. For the NBER SI samples, we were unable to record the exact program a presentation took place in due to our data-use agreement. As a second best, we construct a proxy version of fixed effects using data permitted in our agreements.

Job market and departmental seminars: Seminar series was recorded by human coders using the Qualtrics tool. We subsequently cleaned the seminar names to remove typos and fix differences in capitalization Recognizing that job market talks bring a different audience and norms, we code all job market talks within a department as one series (eg "Harvard job market seminars") while other seminars are coded the usual way (eg "Harvard Labor Seminar").

NBER SI 2019: By agreement with the NBER for this sample, individual program meetings (which correspond to conference tracks) were not allowed to be coded. We use a combination of the conference track field (e.g., micro, macro, and finance) (following Chari and Goldsmith-Pinkham (2017)) as well as the format-style to create a "proxy" seminar series fixed effect. Format style was recorded by the coders and included attributes such as whether the session was joint between multiple programs and whether it included a discussant, a Q&A session at the end, and a moratorium on questions at the beginning. Using both field and format, we then constructed a set of saturated "Field $\times$ Format" fixed effects which yields 15 categories (broad field $\times$ format combinations).

Online seminars: Seminar series were identified directly with the YouTube channel from which the recordings were sourced. However, in instances where this was not sufficient, natural language processing (NLP) techniques were applied to both the title and description of each recording within the same YouTube channel. For example, applying these techniques to the titles and descriptions of seminars from the CEPR channel, originally yielding 168 seminars, enabled their subsequent categorization into 12 distinct seminar series.

NBER SI 2022: Similar to above, individual program meetings were not coded due to a data-use agreement with NBER. We use a combination of the field and format-style to create a "proxy" seminar series fixed effect. Each program was coded into two broad fields: micro and macro (see

Handlan and Sheng (2023)). We also document the presentation format as whether there was a discussant and how long the discussant presented. While we do not directly observe whether a presentation has a Q&A session at the end and a moratorium on questions at the beginning, we use the distributional attributes of audience interruptions to proxy for Q&A and moratorium. We label seminars with more than 90 percent of the interruptions occurring in the last 25 percent of the total duration as having a Q&A session, and seminars with no interruptions in the first ten minutes as having a moratorium. We then constructed a set of saturated "Field × Format" fixed effects that is comparable to what we have for the NBER SI 2019 sample.

C.8 Human-Coder Characteristics

Job market and departmental seminars: There were 77 coders who collected our seminar samples, of whom 73 percent were female, 73 percent were in an applied micro field, and 36 percent were in the fourth year or higher in their Ph.D. program.

NBER SI 2019: There were 29 coders involved in this data collection effort (four of whom had participated in the earlier department seminar study). Roughly half (53 percent) of the NBER coders were female, most (83 percent) specialized in an applied micro field, and 31 percent were in their fourth year or higher in their doctoral program.

C.9 Attendance and Participation

Job market and departmental seminars: Attendance was recorded by human coders using the QUaltrics tool. Coders counted the number of people in attendance as well as the number of female and male attendees in the audience. Finally, they also record whether attendance was higher than normal.

NBER SI 2019: Same as for job market and departmental seminars.

Online seminars: Participation is measured using the number of unique audience members who speak in the presentation. From the recording, we do not know attendance but can proxy this by the diversity of speakers from the audience. Audience members are tracked throughout a presentation by algorithmically matching different interruptions together based on the similarity of speech patterns.

NBER SI 2022: We were unable to measure attendance in-person from the recordings of presentations in this sample because the camera only points at the presenter. We also did not track audience members throughout the presentation.

D Details of Machine Learning Approach

In this appendix section, we provide more context, details, and validations for the machine learning approaches of this paper.

Economic research applying machine learning approaches to analyzing audio data is both relatively new and expanding into new sub-fields. One set of studies analyzes the voices of macroeconomic policymakers, focusing on the tone of Federal Reserve officials in speeches (Alexopoulos et al., 2024; Bisbee et al., 2025; Gorodnichenko et al., 2023). Other studies focus on the workplace, and researchers have investigated how audio analysis of job interviews can aid recruiters in screening candidates (Liem et al., 2018; Naim et al., 2015; Nguyen et al., 2014; Teodorescu et al., 2022). In educational settings, audio recordings of teacher and student behavior have been used to assess classroom climate and learning James et al. (2018). And some have analyzed vocal cues in Supreme Court arguments (Knox and Lucas, 2021). Given that advances in computing power paired with machine learning methods have made audio-as-data accessible to the research community, one of the contributions of our paper is to provide some guidelines for economists in the appropriate use of these methods.

D.1 Data Sources

There are three data sources used for training and developing the model from the computer science literature: Mozilla Common Voice, RAVDESS, and CREMA-D.

Mozilla Common Voice (V4) is a dataset with short audio clips (3-5 seconds) of people reading standardized sentences with corresponding labels for whether the speaker is female, male, or other. We restrict our sample to clips that are agreed to be female or male speakers by multiple crowd-sourced ratings. Our starting sample is 507,184 audio clips that we downsample to get balanced data. There are many versions of the Common Voice dataset that grows over time. We use this data to train the gender classification algorithm.

RAVDESS is a database of audio clips (3-5 seconds) of 12 female and 12 male actors reading standardized sentences multiple times using different directed emotions (angry, disgust, fear, sad, neutral, calm, surprised, and happy) and intensity (normal or strong intensity). There are 1440 clips total. Speaker gender is self-identified and provided by speakers themselves. We group together emotions into four tones: angry and disgust make negative-aggressive, fear and sad make negative-passive, neutral and calm make neutral, and surprised and happy make positive. We use this data to train the tone-of-voice classification algorithm with CREMA-D.

CREMA-D is an audio database with short audio clips (3-5 seconds) with 91 diverse actors reading standardized sentences using different directed emotions (angry, disgust, fear, sad, neutral, and happy). There are 7,442 clips total, which we downsample to balance across tones. Gender is self-identified by the speaker, and tone is validated by crowd-sourced ratings. All speakers are speaking English in the recordings. We group together emotions into four tones: angry and disgust make negative-aggressive, fear and sad make negative-passive, neutral is neutral, and happy is

positive. From this dataset, we leverage audio clips and corresponding labels for speaker gender and tone-of-voice. We use this data to train the tone-of-voice classification algorithm with RAVDESS.

D.2 Data Processing

Each audio clip needs to be processed and converted from sound waves into numerical vectors that we can input into our deep learning model. Audio signals are complex—they comprise many frequencies, with each frequency registering at a different intensity, and these intensities change over time. As a result, we follow the literature on voice recognition (including Gorodnichenko et al. (2023), Bai and Zhang (2021), Sell et al. (2018)) and this section describes that process.

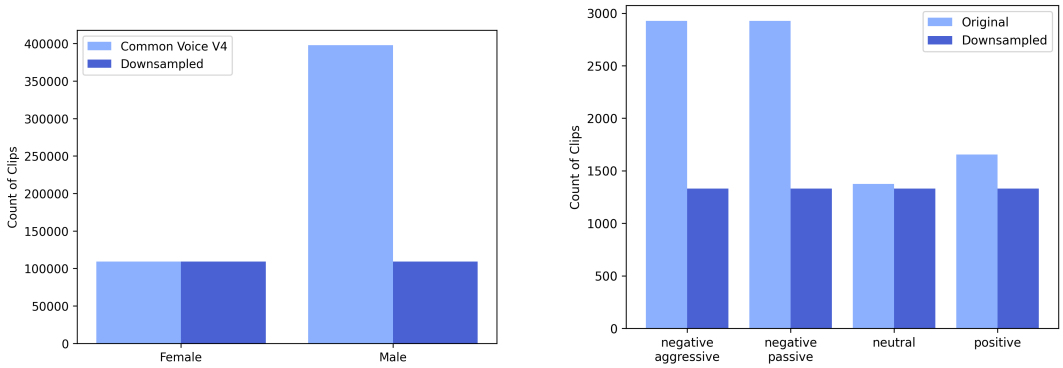
For each short audio clip (about 5 seconds in length), we use a python package called Librosa to extract three types of audio measures: Mel spectrogram, MFCC, and chroma scale values. We have a separate number to measure the intensity on different frequencies (pitches or notes) in the Mel, MFCC, and Chroma. These three types of scale capture slightly different qualities: (1) "Mel Spectro-gram Frequencies" capture the volume (in decibels) of each pitch within a clip; (2) "Chromatic Scale Frequencies" capture the musical scale (but not octave) of each clip; and (3) "Mel Frequency Cepstral Coefficients" capture the timbre of each clip—the elements of sound that make trumpet sound different from a human singing the same musical note. In total, our audio vector is 180 dimensions. The richness of this summary is necessary because gendered differences in speech patterns reflect a combination of frequencies and intensities over intervals of time—and relies on stable acoustic features such as pitch and timber—rather than a single feature.

On a technical note, we set the sampling rate to 16 kHz for all our data. This rate is appropriate for speech data, as it captures the relevant frequencies for human voice while minimizing unnecessary data. The CREMA-D dataset, which is used in our study, is saved at 16 kHz, so we adopt this as the standard rate for consistency. We also use mono-channel sampling.

D.3 Splitting Data into Training, Validation, and Testing Samples

To train our machine learning model to predict gender or tone labels from measures of audio, we need to first set up training, validation, and testing subsets of our data. The testing set is a leave-out set that is not used in developing the model in any way and serves as our out-of-sample accuracy benchmark. The remainder of our sample is split into different training and validation sets. The training subset is used to estimate the parameters of the network via an iterative process known as gradient descent which adjusts model parameters to minimize the distance between predicted gender attributions and the known gender of the speaker. The validation subset determines when to stop updating parameters to prevent overfitting. We employ an ensemble learning approach called "bagging" which trains multiple classification models based on different subsets of the data and then averages their predictions to assign a final label. We train five different models each with unique subsets of the data and stop training when no further improvement is observed on a validation subset. How we estimate multiple versions of our algorithm is explained below, but first we discuss creating the different training, validation, and testing sets in more detail.

Figure D1: Gender Balance Through Downsampling



(a) Common Voice Dataset

(b) RAVDESS and CREMA-D Datasets

To balance categories in our dataset, we apply down sampling. For the Common Voice data there are many more male observations than the 109,290 female observations. Therefore, we randomly select 109,290 male clips to create balance. This decreases the overall sample size by about half, but 218,580 is still sufficiently large for our analysis. For the combined RAVDESS and CREMA-D dataset, we have imbalance across the tones. Here we down sample the CREMA-D data and keep all the RAVDESS data such that each tone-of-voice category has 1330 observations (also balanced by gender) before data augmentation.

We then remove a random 10 percent of observations from the sample to make our testing or leave-out sample that we use to evaluate the out-of-sample accuracy of our models. For the Common Voice dataset, we require there to be an equal number of male and female clips in the testing sample. For the RAVDESS and CREMA-D datasets, we require the testing sample to have equal representation across the tones and to not split actors between training and testing sets. In these later datasets the same speaker supplies multiple clips. By restricting speakers to not span the training and testing sets, we are preventing artificially high accuracy coming from the algorithm learning how one speaker talks (called "leakage").[25]

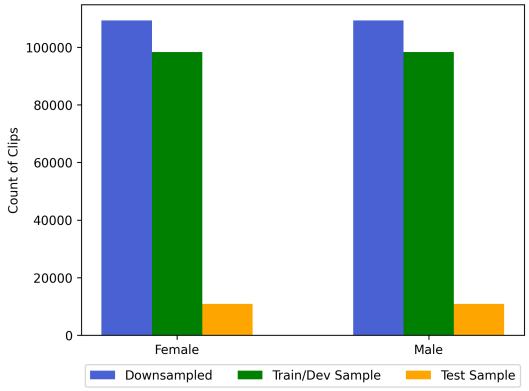
The remaining 90 percent of our data is then used for training our model. We use an ensemble learning technique called "bagging" where our final prediction is actually the majority prediction across multiple, separately trained models. In our case, we train the same model architecture five different times using different subsamples data for training and validation. In short, training data is used to improve the fit of the model while validation data tells us when to stop iteratively updating parameters. First we will explain how the data is divided and then we will explain the training and validation producer more below.

This data is split into five different subsets (called folds) that are representative (balanced labels). As before, we do not allow speakers to be present in multiple folds. We then create one

[25]Data leakage can make an algorithm seem more accurate and generalizable than it truly is. With data leakage, the researcher may conclude their algorithm performs very well out of sample, when in fact, the observations in the testing set are from the same actor and tone as that in the training set. The validation and testing sets should not be used directly to shift model weights.

Figure D2: Training, Validation, and Testing Samples for Gender and Tone

(a) Gender: Common Voice Dataset



(b) Tone: RAVDESS and CREMA-D Datasets

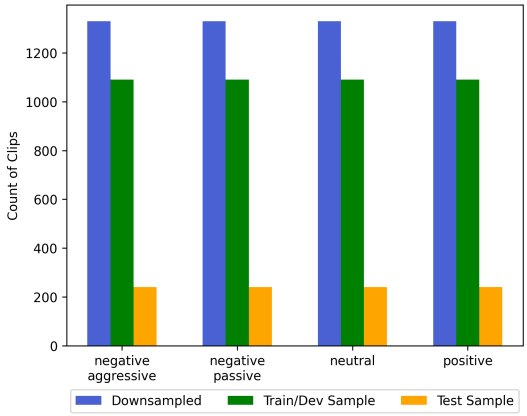
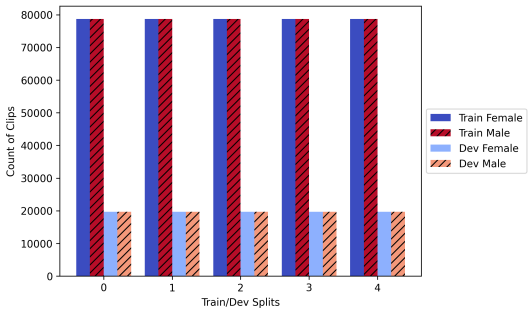
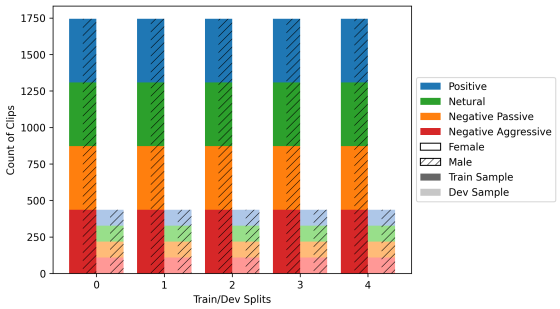


Figure D3: Five-Fold Cross-Validation Splits for Gender and Tone

(a) Gender: Common Voice Dataset



(b) Tone: RAVDESS and CREMA-D Datasets



"split" of the data by having one of the five folds be the validation/development sample and the remaining four folds act as the training sample. Below are graphs showing that in each of the five ways of splitting the data (having one fold be a validation/development set and the rest allocated to training) still has a balance in labels across the samples. We then train a separate model for each of these subsets, which we will later aggregate in our ensemble learning approach.

D.4 Data Augmentation

The RAVDESS and CREMA-D datasets alone provide a small sample to train a tone-classifier. Accordingly, we apply a technique called data augmentation, where we minimally perturb the audio data but assign it the same label as the original clip. This provides us with synthetic data observations we can use for training our models. We do not apply data augmentation for the gender classifier because we already have a large dataset with the Common Voice dataset.

This technique has its roots in image classification. For example, if you are trying to train a model to identify pictures of cats, you can rotate a photo of a cat by 90 degrees, and it should still be labeled as a cat. While this is obvious to humans, the machine benefits by seeing different

angles and rotations. For each rotation, color shift, or stretching of an image, the researcher can produce a "new" observation that helps the model learn how to classify images of cats. We apply the audio analysis version of augmentation where we slightly speed up, slow down, pitch up, pitch down, and add white noise to an audio clip. The underlying component of the audio clip remains the same, but the numerical measure of that audio clip varies slightly. This helps the algorithm by observing more data and becoming more robust to overfitting.

Specifically, we perform time-stretching where the speed of the audio is adjusted to 0.81 times (slow down) and 1.07 times (speed up) the original rate. Similarly, we apply pitch-shifting by raising the pitch by +1 semitone (pitch up) and lowering it by -3.5 semitones (pitch down). Gaussian white noise is also added to simulate real-world environmental noise. Moreover, we trim the first and last second of each audio file to remove non-speech sections while preserving the core speech content. These augmentations increase our sample size for training and development by *six* times. Each augmented clip is assigned to the same training and development set that the original clip belonged to. We ensure that the leave-out testing sample is not subjected to any data augmentation to maintain the integrity of testing performance.

D.5 Model Architecture

For this model we adapt the architecture from Bhattacharya et al. (2022). The authors are also working with tone-of-voice classification and are able to achieve a very accurate classification model. We adapt the model in four clear ways: first we use different audio feature inputs (totaling 180 features instead of their 193). Second, we are classifying audio into four tone categories rather than six emotion categories. Third, we use audio of only English speakers and their project focuses on comparing tones across languages by working with English, German, and Italian recordings. Fourth, we require strict sample-splitting procedures to prevent data leakage. We apply the same architecture for our tone model for gender classification and simply change the number of classes from four to two. Table D1 details the hyperparameters we used for training the audio-classification models.

Table D1: Hyperparamters

Parameter	Value
Loss Criterion	Cross Entropy Loss
Optimizer	Adam
Learning Rate	1e-4
Decay Rate	1e-6
Momentum	0.1
Random Seed	2024
Training Epoch	200
Patience	20
Batch	64

D.6 Model Training

For both gender and tone classification tasks, we trained separate sets of models using five-fold cross-validation to ensure robust performance and reduce overfitting. Each model's training process was carried out with a consistent set of hyperparameters and the same overall architecture, allowing us to maintain comparability across models and leverage ensemble predictions for greater accuracy.

Data Loading and Preprocessing Pipelines: Each dataset was preprocessed through a standardized pipeline to ensure consistency across samples. The data preparation routine includes balancing the data by downsampling categories and filtering. For instance, subsets were created to train models on gender-specific samples or to control for emotion intensity. Additionally, for tone classification, we maintained balance across the categories by applying downsampling and using multiple datasets in tandem to ensure representation across tones and genders.

Feature Standardization and Scaling: To ensure uniform input distributions and facilitate model convergence, each audio feature vector was normalized. This step adjusts all features to have a mean of zero and a standard deviation of one, which is essential for models with deep architectures. Normalization minimizes the risk of gradient vanishing or exploding during training, especially for neural networks that are sensitive to unscaled inputs.

Augmentation-Specific Filters and Conditions: For tone classification, data augmentation was selectively applied to increase robustness without compromising the dataset's original structure. As mentioned before, augmentations included time-stretching (speed adjustments to 0.81 and 1.07 times the original rate), pitch-shifting (+1 and -3.5 semitones), and adding Gaussian white noise to simulate real-world environmental conditions. Each augmented clip retained its original label and was assigned to the same fold as its non-augmented counterpart to prevent cross-fold leakage. Additionally, we filter samples to include only original, non-augmented clips in some models, allowing control over augmentation intensity per task.

Specific Subset Selection for Training Samples: Subsets were created where clips were restricted to actors with ID numbers below a set threshold, allowing control over sample composition. Similarly, we filtered out specific emotion types and intensity levels, such as strong or surprised tones, depending on the experimental conditions. This approach helped maintain a high level of control over sample characteristics, balancing variation in the data without introducing artificial biases across model versions.

Cross-Validation and Ensemble Prediction: We applied five-fold cross-validation to ensure balanced representation across labels and mitigate overfitting. Each fold served as a validation set once, while the remaining four folds were used for training. To prevent data leakage, no individual speaker appeared in both training and validation sets for any fold. For each task, an ensemble approach was used: we trained five models with different cross-validation splits, and the final prediction for each sample was determined by a majority vote across these models, enhancing prediction reliability and minimizing potential biases.

Batch Training and Hyperparameters: Each model was trained in batches of 64 samples using a learning rate of 1×10^{-4}, weight decay of 1×10^{-6} , and a momentum of 0.9. Training

continued for up to 200 epochs, with early stopping applied after 20 epochs of non-improvement in validation loss. This setup, consistent across both gender and tone classification, provided a balance between model flexibility and convergence stability.

Library and Environment Specifications: The training and analysis were implemented in Python using torch for deep learning, pandas for data manipulation, and sklearn for scaling and performance metrics.

Evaluation Metrics and Performance Logging: During training, we monitored validation performance using metrics such as accuracy, precision, recall, and F1 score, providing a comprehensive assessment of model robustness. These metrics helped optimize model selection and confirmed performance consistency across folds. All evaluation results were logged per epoch to track model behavior over time and identify optimal stopping points.

By including these steps and conditions, we ensure that the model training approach is fully documented and reproducible. Each step reflects a commitment to rigorously maintaining dataset balance, preventing data leakage, and achieving model robustness suitable for both gender and tone classification tasks.

D.7 Model Performance

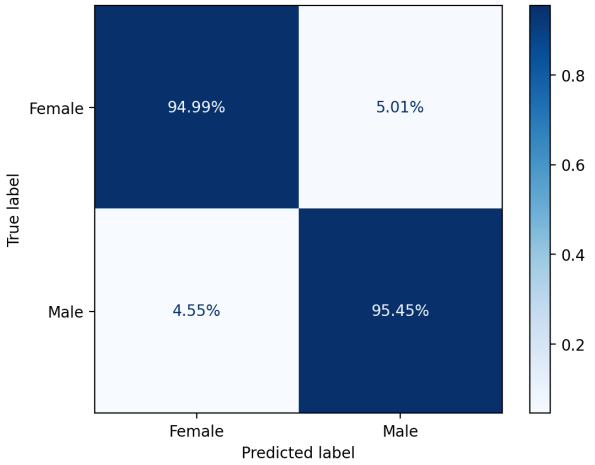
In this section we report the out-of-sample evaluation of our trained machine learning models. Recall, we set aside a sample of voice actor recordings that are not used in estimating parameters or determining when to stop training the model. This set is the leave-out sample. Below, we describe our evaluations of the gender-classifier and the tone-classifier models. The next section includes additional validation exercises with alternative methods, including direct human validation.

The gender classification model is trained to sort audio recordings into male or female speakers. The training data is constructed by volunteers who identify as male or female, described earlier in the appendix. We find a very high accuracy of 95 percent, indicating a reliable model for labeling male and female speakers. Figure D4 is the confusion matrix for this algorithm on the leave-out sample. It shows that 94.99 percent of female recordings are correctly labeled as being female speakers, and 95.45 percent of male recordings are predicted to be male.

The tone classification model is trained to categorize audio into four distinct tones: Negative Aggressive (anger and disgust), Negative Passive (sadness and fear), Neutral (neutral and calm), and Positive (happiness and surprise). Three versions of the ensemble model were trained, each using a different subset of the data: one trained exclusively on female speakers, one on male speakers, and a combined model using both female and male data. Performance was evaluated on a hold-out sample to measure out-of-sample accuracy. The model trained on female-only data achieved an accuracy of 62.5 percent, while the model trained on male-only data achieved 53.3 percent. The combined model, which included data from both genders, achieved an accuracy of 56.3 percent, indicating some performance variation based on the gender-specific data used for training. The gender-specific models are used in the main specifications of the paper.

Now let us discuss the accuracy metrics by tone groups we use in our analysis: negative, positive,

Figure D4: Confusion Matrix for Gender Classification (Leave-out Sample)



Notes: The true label is the label provided by the voice actor, and the predicted label comes from the trained algorithm applied to this leave-out sample that was not used in training. Values shows accuracies within rows (e.g. to top row includes the percent of female clips predicted as female or male).

and neutral. In the paper we group the negative tone categories together so Negative-Aggressive and Negative-Passive are jointly called Negative. Of the Negative observations in the testing sample, the female model correctly predicts Negative 72.9 percent of the time and the male model does so 72.1 percent. Because we add the two negative categories, the random benchmark for Negative would be 50 percent. For Positive labels the random benchmark is still 25 percent. Both the male and female models correctly identify Positive audio clips 49.2 percent of the time. Again, these accuracy benchmarks are much better than random and perform similarly to our human validation set as described in the next section.

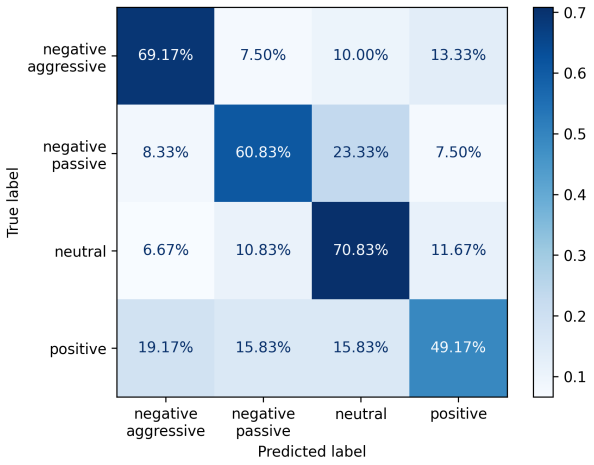
Figure D5 below displays the confusion matrices for the tone classification model trained on separate gender-based datasets. The first figure shows the performance of the model trained exclusively on female data, while the second figure corresponds to the model trained on male data. Each confusion matrix provides insights into the model's classification accuracy across four tone categories: Negative Aggressive, Negative Passive, Neutral, and Positive.

In both figures, the diagonal entries represent the proportion of correctly classified samples. The off-diagonal entries capture misclassifications, where a tone category was incorrectly predicted as another. For instance, the model trained on male data some confusion between Negative Aggressive and Positive tones (mostly driven by actors talking very loudly).

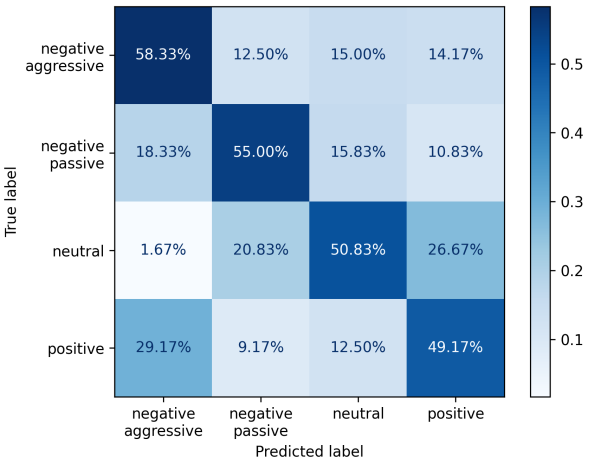
These confusion matrices highlight that while both gender-specific tone models achieve reasonable performance, there are variations in classification accuracy across tone categories, likely due to inherent differences in the tone expression between genders. However, for negative and positive categories we use in the paper, we find similar accuracies across speaker gender. The matrices provide a useful diagnostic for identifying which tones are more challenging to classify accurately.

Figure D5: Confusion Matrices for Tone Classification (Leave-out Sample)

(a) Female-Tone Model



(b) Male-Tone Model



Notes: The true label is the intended tone from the voice actor, and the predicted label comes from the trained algorithm. Each matrix shows the model trained with data from a one gender, and tested on the leave-out sample for that same gender. The leave-out samples were not used in training the models. Values shows accuracies within rows (e.g. to top row includes the percent of negative-aggressive clips predicted as the four different tones).

D.8 Additional Validation with Alternative Methods

In addition to evaluating the machine learning models with their accuracy on leave-out samples, we have performed substantial robustness checks with human coders and alternative coding specifications on both the data from voice actors and the economics seminar data. Overall, we find that the algorithm does as well as human coders in classifying gender and does reasonably well in classifying tone.

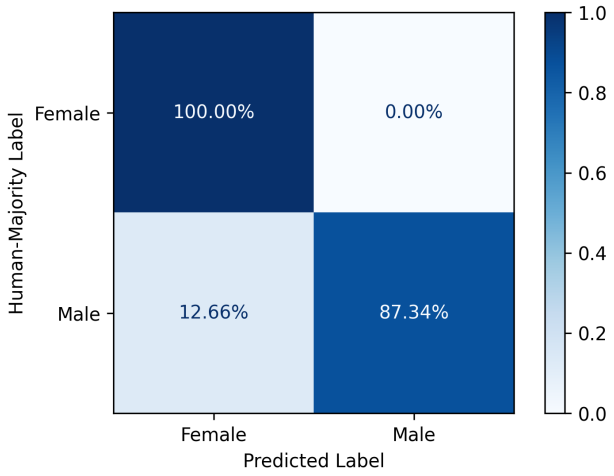
D.8.1 Human Validation

We performed three different validation exercises for the gender and tone classification with a team of human coders where they listened to samples of online economic seminars and labeled voice actor recordings. For each exercise, we have between four to six humans providing labels for audio data that we use to benchmark our algorithm. A common metric we report is the "human-majority label" which requires that a majority of the human coders agree on either a gender or tone label. Additionally, the range of accuracies of individual human coders provides an informative benchmark for the difficulty of our classification tasks, and we can compare our algorithm metrics to that range.

In the first task, we compare human-coded gender labels with the computer-coded gender labels for 140 recordings from the online seminar sample. Figure D6 summarizes the results. For this task we had five coders. The recording samples were each 5 seconds long and included a mix of male and female speakers. For all but one recording, there was a clear majority label. Our gender-classification neural network perfectly predicts 100 percent of the recordings that human-coded as

female. For speakers labeled as male, our algorithm also predicts male 87.3 percent of the time. The extremely high accuracy relative to the actor data presented in Figure D4, along with additional validations in the next section using names, suggests that the gender classification algorithm is highly accurate.

Figure D6: Confusion Matrix for Gender Classification (Human Validation)



Notes: The online seminar subsample for human validation includes 140 short recordings of a mix of presenters and audience members. The human-majority label is the label agreed on by at least three of the five human coders. The human-majority labeled 60 clips as female and 79 clips as male. One clip had no majority label, and the algorithm predicted that to be male. Values show the agreement percentages within rows (e.g. the top row shows the share of clips predicted to be female and male among the clips labeled as female by human coders).

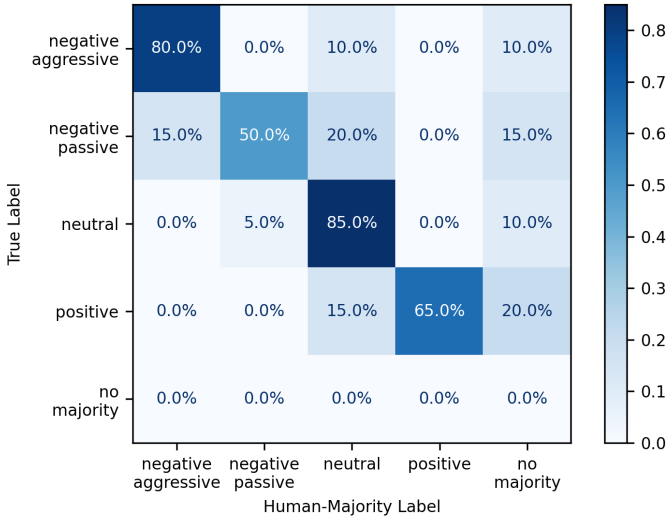
For the second and third tasks, we asked our human-coders to classify the tone (negative-aggressive, negative-passive, neutral, and positive) for a sample of voice actor recordings and a sample of the online seminar recordings. When training the human-coders how to label "like the machine does," we presented them with a total of 12 training examples from voice actors across the four tone groups and two genders. Importantly, we gave them the following instructions: "Classify tone based on speech patterns and sounds, not on the words that are being said. Again, we are classifying based on what someone sounds like, not what they are saying."

Classifying tone is a more difficult task than classifying gender, due to both the increase in categories and the nuances of tone itself. Accordingly, the first exercise is to evaluate how well the human coders do in labeling the voice actor data. Figure D7 shows the confusion matrix comparing the human-majority tone label to the actor's intended tone for a sample of 80 recordings. The accuracies in the different categories range from 50 percent to 85 percent, and the overall accuracy is 70 percent. We can see the fifth column represents cases when the human coders dramatically disagreed and highlights the challenges of classifying tone. For individual coders, the range of their individual total accuracy scores was between 58.8 percent and 75 percent accuracy. When splitting the sample into female and male actors, we find similar accuracy ranges.

The human coders were able to predict the actors' tone with an accuracy that was on par with

the machine-predictions on the leave-out sample, albeit slightly higher (reported in Figure D5). There is only about a 10 percentage point difference between the machine and the human coders (57.9 percent compared to 70 percent), where the machine's accuracy out of sample is on par with the lowest accuracies from the individual coders (58.8 percent). Our interpretation is that the machine is labeling tone about as well as some humans.

Figure D7: Confusion Matrices for Tone Classification (Human Benchmark)



Notes: This sample includes 80 recordings of voice actors balanced by gender and tone. The true label is the intended tone from the voice actor. The human-majority label is the label agreed on by at least three of the four human coders. There were many observations where there was no majority. Values shows accuracies within rows (e.g. to top row includes the percent of negative-aggressive clips predicted as the four different tones).

We take this a step further and ask our human coders to also listen to a set of 160 recordings from online seminars. Of those, 80 observations are short, five-second clips like the inputs to the algorithm and 80 observations are longer audio recordings (at the utterance level where we choose the tone based on average tone probabilities over the five-second intervals of the utterance). As an internal check, a random number of clips did not include speech to ensure the coders correctly labeled them as non speech (in the end there were 20 and all were correctly identified).

We look to the remaining 140 recordings to compare between human labels and the machine predictions. At least one human coder agrees with the machine prediction for 55.5 percent of this sample (52.5 percent for the 51 five-second clips, and 58.8 percent for the 59 longer utterance recordings). Human-majority labels only exist for 78.6 percent of the 140 recordings of economists (meaning there is substantial heterogeneity in opinions from the human coders on how to label tones). The machine algorithm agrees with the human-majority 30 percent of the time on the 110 clips where a human-majority exists.

Given the human-coders reported challenges in analyzing tone like the machine (based on their accuracy with the actor data and comments from the coders), it is difficult to attribute the disagreement to something with the algorithm, the small sample, or how humans understand tone.

We interpret this as a disagreement between the machine and the human coders on the intensity thresholds between tones. The human coders are much more likely to code "neutral" tones, especially on the economic seminar sample. The human-majority labels 44 percent of recordings as "neutral" tone, comparatively the algorithm labels "neutral" to 27 percent. This pattern also shows up in Figure D7, but is less pronounced.

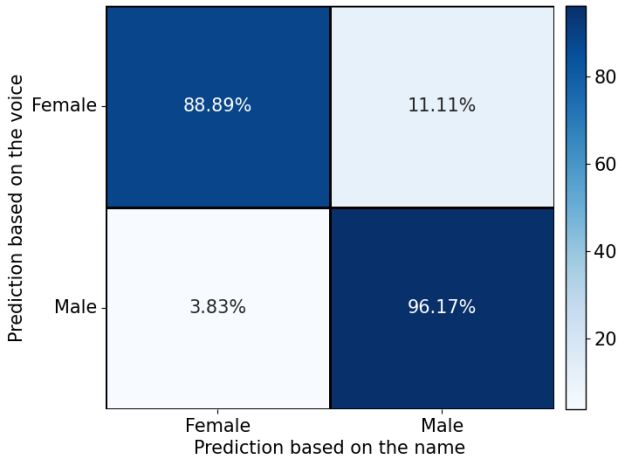
Finally, we note that this additional human validation of tone was not possible for the NBER SI 2022 sample because we were not able to save and share presentation recordings due to our data-use agreement with NBER. Nevertheless, we see the validation out-of-sample on the online seminar data encouraging about the accuracies of the models compared to how humans hear gender and tone.

D.8.2 Gender Classification Using Names

Besides using audio data to classify speakers' gender, another popular method is to use the speakers' names. We use a large dictionary, called Genderize, that maps names to genders. This mapping is probability-weighted to account for names that are used for both men and women. We use this on the names of speakers in the online seminar sample directly. For the NBER SI 2022 sample, we were not allowed to record presenter names. Accordingly, we use the NBER SI 2022 program to create bounds for the number of female presenters using female authorship.

The confusion matrix in Figure D8 compares the name-based gender prediction to the audio-based prediction that is the main method of the paper. There is a high degree of agreement between the approaches. Because is a risk of error with the name approach - for example, the dictionary associates "Ariel" as a common female name but it can also be a male name - we do not expect a perfect match even for a perfect algorithm. However, the large amount of agreement across our robustness approaches gives us preponderance of the evidence that the algorithm does well in capturing female and male speaking patterns.

Figure D8: Gender Prediction with Voice vs. Name (Online Seminars)



Notes: This is based on the online seminar sample where we have data on presenter name. There is large agreement between the two approaches. Values show the agreement percentages within rows (e.g. the top row shows the percent of presenters predicted to be female who have names that are labeled as female or male).

We also perform a name-gender classification verification for the NBER SI 2022 sample, but we are limited to a looser validation using the publicly available NBER SI programs (which include participants' names). We are able to do a direct comparison with discussants (where we find 30.9 percent of discussants are classified as female by name, and 31 percent when classified by voice). For presenters, we have to do an indirect comparison using author composition .

We create two bounds to benchmark the number of female presenters we impute for the NBER SI 2022. The upper bound is the share of papers with at least one female author, representing the extreme case where the female coauthor for every paper with at least one female author was presenting. The lower bound is the share of papers with only female authors, representing the other extreme, where female coauthors don't present when there are other male coauthors.

Using author names from the NBER SI 2022 programs, we apply the name-gender classifier to measure the upper and lower bounds for female presenter share. 9.6 percent of papers have only female authors and therefore must have a female presenter. 51.5 percent of papers have at least one female author and therefore *could* have a female presenter. This is a fairly wide interval. Nevertheless, our finding that 33.8 percent of presenters at NBER SI 2022 were female is safely in the middle of that interval and therefore seems reasonable.

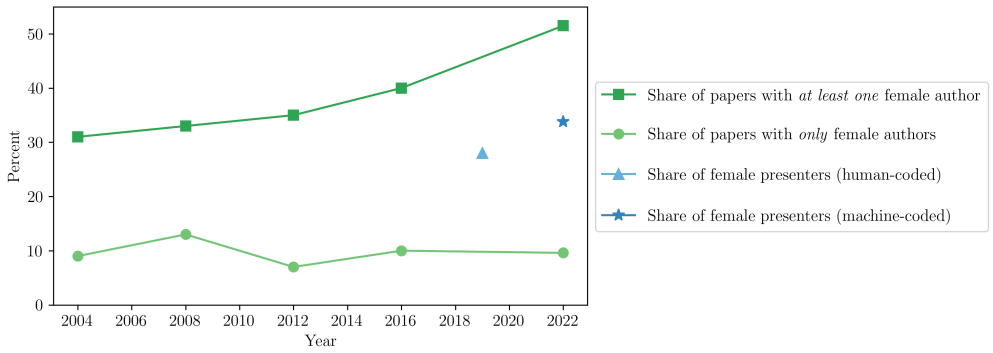
For additional context, we plot additional years of information for the NBER SI author and presenter female shares in Figure D9. We plot the share of female presenters in the NBER SI 2022 sample compared to the NBER SI 2019 sample. For the upper and lower bounds using female authorship, we plot data points prior to 2022 directly from tables in Chari and Goldsmith-Pinkham (2017) and we extend those measures to 2022.

Ultimately, this plot shows that the gender-classification we have for NBER SI 2022 produces a reasonable share of female presenters. Both the numbers from 2022 and the human-coded share

from 2019 are around 30 percent. Additionally, it is within the upper and lower bounds we estimate from female authorship.

Finally, this figure also shows a continued increase in the share of papers at NBER SI that have a female author. This new result is an extension to one table of Chari and Goldsmith-Pinkham (2017) and shows encouraging progress towards the profession's goal of inclusivity at top conferences, like the NBER SI.

Figure D9: NBER Summer Institute Shares of Female Authors and Presenters



Notes: The share of papers with at least one female author and with only female authors (green squares and circles) is from Chari and Goldsmith-Pinkham (2017) for 2004, 2008, 2012, and 2016 and is based on the NBER SI programs. We extend this to NBER SI 2022 using the online posted programs and the name-gender classifier. The share of female presenters for NBER SI 2019 (blue triangle) comes from the human-coders in the audience. The share of female presenters for NBER SI 2022 (blue star) comes from the gender-prediction algorithm and presenters' voices.

D.9 Additional Details on Computer-Coded Samples

In this section we provide additional detail for the selection of seminars in the online seminar and NBER SI 2022 samples.

D.9.1 Online Seminars

To be included in the sample, each seminar had to be part of a seminar series organized or sponsored by an economics department based in either the United States or Europe. Seminars sponsored by highly regarded research institutions such as the American Economic Association (AEA) and the Centre for Economic Policy Research (CEPR) were also included.[26] Once a set of YouTube channels hosting these seminars was identified, all available seminars were downloaded. This process was carried out periodically between early 2022 and early 2023, ensuring that the complete archive of each channel was collected, resulting in a total of 2,398 online seminars. Several filtering steps were subsequently applied to obtain the final dataset.

Although all seminars took place during the COVID-19 pandemic, a random manual inspection—particularly of the final seminars in each series—was conducted to ensure that no hybrid

[26]For example, the American Economic Association provides an extensive list of online seminars at https://www.aeaweb.org/resources/online-seminars.

seminars were included in the sample.

To focus on standard academic presentations in economics, seminars shorter than 20 minutes or longer than 110 minutes were excluded. The presenter of each seminar was defined as the individual with the longest total speaking time. If a second speaker had more than 15 minutes of speaking time, the seminar was excluded to ensure that only events with a clearly defined presenter were retained. The moderator was identified using a three-step heuristic based on typical seminar structure. First, if the first speaker was not the presenter, that speaker was designated as the moderator. Second, if no moderator was identified using this rule and the seminar had only two participants, the non-presenter was assigned as the moderator. These two rules allowed identification of a moderator in 90 percent of the seminars. For the remaining cases, the last non-presenter speaker was designated as the moderator, under the assumption that moderators often both introduce and conclude the seminar. A seminar was classified as moderator-dominated if the moderator accounted for at least 51 percent of the interruptions; such seminars were excluded from the analysis as they are unlikely to reflect the typical interaction dynamics of economics seminars.

Since the analysis includes presenter-related variables such as citation counts, only seminars where the presenter's Google Scholar profile could be reliably identified were retained. As explained in the main text, this was done using a combination of NLP and web scraping methods, followed by manual verification to ensure the matched profile corresponded to the actual presenter, avoiding potential name confusions. In addition, two seminar series—originally included through the automated procedure—were discarded as they mostly featured computer science presentations. While including them did not alter the main results, they were removed to ensure disciplinary consistency. The final dataset consists of 994 seminars. This includes 30 seminars where the speakers were identified as belonging to other fields (mostly computer science or biostatistics) but were retained because they presented in economics seminars. As these speakers tend to exhibit different characteristics (e.g., citation counts five times higher than the sample average), a control variable was added to indicate their presence.

D.9.2 NBER SI 2022

To focus on standard presentations in economics, we drop the short presentations (the egg timers and lightning rounds) and panels. We also do not include measures from programs where participants opt-ed out of the study. Finally, we also do not include analysis of audience speech for samples with poor audio quality. This produces a sample of 257 presentations.

The presentations at NBER SI 2022 were in hybrid format and required audience members and the presenter to speak into microphones. Therefore, we run a pass over the sample to select talks where the audio quality of audience members is sufficiently good in terms of timing and volume. We validate the timing manually and we identify a volume threshold, under which we say the audience audio is insufficient. We use the root mean square of the audio waveform as the volume threshold, because it directly reflects how well the audio was captured by the microphone, and it is mathematically related to many other standard sound measures, such as peak-to-peak amplitude

and intensity.

D.10 Assignment of Research Types with LLMs

Standard JEL codes describe the substantive field within which each research paper lies. In addition, we seek to code a second dimension, coding the research style or method. This matters because, for instance, a public finance developing theory about the asymmetric information between a taxpayer and the government may elicit a different audience with different engagement than a public finance paper analyzing empirical results from a field experiment. As such, we create the following four variables for each presentation.

1. Theory: Is the primary purpose of this paper presented developing economic theory?
2. Empirical: Is the primary purpose of this paper analyzing empirical data?
3. Experiment: Is this paper focused on a field or lab experiment?
4. Policy analysis: Is the primary purpose of this paper evaluating a specific program or policy, studying a specific law or regulation change, or simulating policy responses.

These groups are not mutually exclusive, and it is not unusual for a paper to engage in several of these dimensions at once.

We use a large language model to code these variables. For the seminar samples, we feed it the underlying research papers, and for the online seminars, we feed it a transcript of the entire seminar. (We cannot create these variables for the NBER samples where we were not permitted to code identifying information about the paper being presented.)

D.10.1 LLM Details

Here we provide the details of the LLM approach, including model, parameters, and prompts. We use OpenAI's GPT-4o model, and feed the data through the Assistants API. The Assistants API provides a straightforward form of "Retrieval Augmented Generation" in which the LLM has access not only to the instructions (below), but also the uploaded file when making evaluations. Each paper is uploaded and evaluated independently ("zero shot" learning), so that the evaluation of no one paper affects the interpretation of another. As is standard, we use zero temperature to minimize the variation in GPT responses. And because we use the API, none of the data is reincorporated into GPT's training. Specifically, for each paper or transcript, we:

1. Upload the transcript or the paper (using a readable rather than scanned pdf) to Open AI using its File API.
2. Set the instructions (listed below) for this project-specific Assistant, and give the Assistant access to that specific paper (or transcript).

3. Ask the Assistant to "Please read the attached economics research paper carefully, and assess its characteristics using the rubric I've given you." This prompt also includes a request for structured outputs so that the output is machine-readable.

4. Extract the variable values and validation explanations

The instructions that guide the LLM Assistant are as follows:

```
Context: You are a professor of economics at a leading economics department, and you
    study the economics research process. As part of your research, you read recent
    papers closely, and categorize them according to a well defined rubric. You are
    careful and thorough as you read, and thoughtful about how you categorize each new
    research paper, considering them within the context of the sort of research papers
    you would read in top-tier economics journals, like the AER, QJE, JPE, ReStud, or
    Econometrica.

Task: Please read the attached research paper carefully, paying attention not only to
    what it analyzes, but how the analysis is conducted. I will guide you through a
    series of questions, asking you to describe in a structured way the attributes of
    that research paper. Be prepared to explain your reason for each answer.

Instructions: First, collect the following data about each paper:
1. Title
2. Author(s)
3. Abstract or whole paper: Was this file the entire research paper, or just an abstract?
4. Number of pages
5. JEL codes: The JEL codes should be listed on the first page, but if you can't find any
    , create your own based on your reading of the paper (be sure to put an asterisk next
     to any you create).
6. Keywords: The keywords should be listed on the first page, but if you can't find any,
    create your own based on your reading of the paper (be sure to put an asterisk next
    to any you create).

Second, and most importantly, please answer the following questions. In each case,
    provide a simple yes or no answer, and an indication of your confidence in your
    assessment (on a scale of 1= "very unsure", 2= "quite unsure", 3- "somewhat unsure" 4
     "neutral" 5="somewhat confident" 6="quite confident" 7="very confident"). Provide a
    few relevant bullet points explaining the reasons for your assessment, as well a few
    relevant bullet points providing the most important reasons why a colleague might
    disagree with or doubt your assessment. These questions are not mutually exclusive:
1. Is the primary purpose of this paper developing economic theory?
2. Is the primary purpose of this paper analyzing empirical data?
3. Is this paper focused on a field or lab experiment? (This category does not include
    natural experiments.)
4. Is the primary purpose of this paper evaluating a specific program or policy, studying
     a specific law or regulation change, or simulating policy responses in ways that
```

```
someone might cite as evidence in support or opposition to a policy. (Please look for
 more than peripheral references to policy implications.)
```

Our evaluation of online seminars uses the transcript of the presenter's speech, excluding interruptions, rather than the underlying research paper as input.. As such, we tweak these instructions, so that we tell the LLM that it is not reading "the attached research paper", but rather instruct the LLM that it is being provided with "the transcripts of research seminars" and asking about "this paper" we asked about"the paper being presented."

D.10.2 LLM Evaluation

To evaluate the accuracy of our labels, we created a human-labeled validation set. We randomly selected 75 papers for the validation set: 20 from the job market sample, 15 from the regular seminar sample, and 40 from the online seminar sample. Each paper is independently labeled by two of the authors. The pair resolved any disagreements and submitted consensus labels. A separate author developed the LLM prompt and was blinded to these labels. Table D2 summarizes the LLM accuracy for each label. Overall, the pooled evaluation shows very high accuracy for the LLM, ranging from 82.7 percent to 97.3 percent agreement. The precision for policy-labels implies the LLM was more likely to label a paper/presentation as policy compared to the human-labels – likely driven by papers discussing policy implications.

Table D2: LLM Accuracy

Sample	Metric	Theory	Empirical	Experimental	Policy	Sample Size
Job Talk	Accuracy	0.750	0.900	0.950	0.800	20
	Precision	1.000	0.944	0.667	0.692	
	Recall	0.667	0.944	1.000	1.000	
Regular Seminar	Accuracy	0.800	0.933	1.000	0.867	15
	Precision	1.000	1.000	1.000	0.778	
	Recall	0.700	0.923	1.000	1.000	
Online Seminars	Accuracy	0.875	0.925	0.975	0.775	40
	Precision	0.917	0.960	0.833	0.533	
	Recall	0.880	0.923	1.000	0.800	
Pooled	Accuracy	0.827	0.920	0.973	0.800	75
	Precision	0.951	0.964	0.846	0.649	
	Recall	0.780	0.930	1.000	0.923	

Notes: Table shows accuracy of the LLM predictions for a paper's type. Ground-truth labels were manually created by the authors. Accuracy is defined as the number correctly predicted, divided by the total. Precision is the number of true positives, divided by the number of true and false positives. Recall is the number of true positives, divided by the number of true positives and false negatives.