Gender Bias and Property Taxes

By Gordon Burtch[†], & Alejandro Zentner[‡]

Gender bias distorts the economic behavior and outcomes of women and households. We investigate gender biases in property taxes. We analyze records of more than 100,000 property tax appeal hearings and more than 2.7 years of associated audio recordings, considering how panelist and appellant genders associate with hearing outcomes. We show that female appellants are systematically less likely to obtain tax savings from hearings, particularly when facing female ARB panelists. Then, employing a multi-modal large language model (M-LLM), we construct measures of participant behavior and tone from the audio recordings. We observe markedly different behaviors between male and female appellants. Further, in the case of male appellants, we find that these differences in behavior also *depend on the gender of the ARB panelists they face (e.g., male appellants appear to behave systematically more aggressively towards female panelists). In contrast, the behavior of the female appellants varies much less with the gender of the ARB panelists. Regardless, we find that female appellants continue to fare worse when they face female ARB panelists, conditional on the behavior and tone they exhibit in their hearings. Our results are thus consistent with the idea that gender biases are driven, at least in part, by unvoiced beliefs and perceptions on the part of ARB panelists. Our study documents the presence of gender biases in property appraisal appeal hearings and highlights the potential value of generative AI for analyzing large-scale, unstructured administrative data.*

Keywords: Property Tax, Public Finance, Gender Bias, Generative AI, Gender Concordance, Multimodal Large Language Models

Note: Please do not circulate or cite without the authors' permission.

[†] Boston University – gburtch@bu.edu
[‡] UT Dallas – azentner@utdallas.edu
* We thank Emma Wiles, Jetson Leder-Luis, Dokyun Lee, Bin Gu, as well as seminar participants at Southern Methodist University, George Washington University, U Mass-Amherst, and Boston University for helpful comments and suggestions.

I. Introduction

Gender disparities in evaluations, opportunities, and outcomes are well documented across a wide range of economically significant domains, particularly in contexts that involve subjective judgment. A growing body of research demonstrates that, in many cases, gender concordance or discordance—congruence or incongruence of gender between evaluators and those being evaluated—can drive these disparities, influencing outcomes in administrative processes, the labor market, education, healthcare, and criminal justice settings (e.g., Cullen and Perez-Truglia, 2023; Cabral and Dillender, 2024; Takeshita et al., 2020). These effects matter because they suggest that the personal characteristics of evaluators and evaluated individuals may shape critical decisions, often leading to systematic bias.

We examine gender concordance effects in the economically important context of property taxation, specifically in the administrative process related to property appraisal appeals (Nathan, Perez-Truglia and Zentner, 2025). Property taxes represent the second-largest source of tax revenue in the United States, contributing an estimated $547 billion in 2019, more than doubling revenue from corporate income tax (Giaccobasso et al., 2022). Property taxes are particularly important for sub-national governments, which rely on revenue from property taxes to fund essential services such as schools, police, and parks.

In every U.S. county that we are aware of, households can file an appeal over their property appraisal, seeking to reduce their tax burden. Households can appeal their appraised value if they believe it exceeds the market value, if comparable properties have been appraised at lower values leading to an unfair tax burden, or if there are errors in the county's information used to assess the property.

County Appraisal Districts (CADs) in the U.S. rely on data and statistical models to estimate a home's market value. However, these estimates can vary significantly, as evidenced by the wide differences in appraised values assigned to the same home by large companies like Zillow or Redfin. Households leverage this inherent subjectivity in the appraisal process to make the case for an appeal, seeking to reduce their tax burden. In many cases, appeals in our focal state of Texas proceed to a formal, live hearing, wherein a homeowner or their representative presents the case for their appeal in front of a panel of appraisal review board (ARB) members.

We study data on property appraisal appeal hearings obtained from Harris County, Texas (the Houston area) and associated audio recordings. Exploiting exogenous variation in the composition of ARB hearing panels attributable to the fact that panelists are randomly assigned to hearings, we examine how appeal outcomes relate to a hearing's gender composition. We focus on gender concordance between appellants and panelists, documenting several notable findings. We calculate panelist-specific savings adjudication rates and document significant heterogeneity, beyond what could be explained by chance. Further, we calculate estimates of panelist-specific differences in savings adjudication rates by appellant

gender, reflective of anti-female (pro-male) bias, and find that female panelists are systematically less likely to adjudicate in favor of a female appellant as compared to a male appellant. Next, undertaking more formal regression analyses, we find consistent results; we show that female appellants are systematically less likely to be awarded reductions in their home value and that this is particularly true when their ARB panel is comprised of females. The effects we report are substantial; we estimate that female appellants are approximately 4.2 p.p. less likely to obtain any reduction to their home value when faced with an all-female panel as compared to an all-male panel. Further, we estimate that, when facing an all-female (versus all-male) panel, female appellants obtain an approximate 33% smaller reduction to their initial home appraisal, on average. Given the average home appraisal value reduction across all hearings is approximately $20,200, the gender concordance effect we observe translates to a roughly $6,700 difference in final appraised value and an approximate $140 difference in annual property taxes owed (the property tax rate in Harris County is presently 2.13%).

Next, leveraging unstructured audio recordings associated with more than 80,000 of these hearings, i.e., more than 2.7 years of audio, we employ a multi-modal large language model (M-LMM), Gemini 1.5 Flash, to annotate appellant and panelist behaviors and tones in each hearing – a task that would be prohibitively costly employing human analysts. Our goal is to understand why the observed gender differentials might arise. Examining differences in the prevalence of behaviors and tones across hearings by appellant and panel chair gender, we first observe that appellants exhibit much more variation in their tone and behavior than panelists. This can be explained, to some extent, by the fact that panelists receive formal training and are aware that their behavior and statements may be subsequently reviewed in court should the appellant decide to appeal their decision further.

Considering appellant behaviors and tones, we find that female appellants differ, in general, from male appellants in several respects. For example, females are coded as being systematically less likely to 'sound confident,' they are less likely to present formal evidence, and they are less likely to disagree verbally with a panelist.[1] Most important, of course, are behaviors or tones that differ systematically with the gender of ARB panelists, as those behaviors might help explain differential hearing outcomes. And, indeed, we observe that some appellant behaviors and tones *do* differ in ways that depend on the panel chair's gender.

For example, male appellants behave systematically more aggressively toward female panelists. We observe that male appellants are systematically more likely to 'sound hostile', to raise their voice, to sound frustrated, to sound annoyed, to sound defensive, and to interrupt when a panelist is speaking, specifically in front of a female panel chair, as compared to a male panel chair, and as compared to

[1] We consider the possibility that Gemini may be biased in its evaluation of men versus women later, comparing Gemini's annotations for a random sample of audio recordings with those obtained from human workers on Prolific. Notably, we demonstrate substantial agreement between Gemini and humans.

female appellants, broadly. At the same time, we observe that female appellants' are more likely to sound confident in front of a female panel chair, they are more likely to ask a panelist to repeat themselves, and they are more likely to ask a panelist to 'speak up'. However, in each case where female appellants' behavior differs with ARB panel chair gender, that difference is relatively small in magnitude.

Of course, behavioral differences will not necessarily lead to differential hearing outcomes if social norms regarding expected behavior for males and females differ. This leads to our final, and perhaps most notable finding. A unique aspect of our study and dataset is that we have the ability to separate the influence of changes in appellant behavior from the influence of differences in panelist perceptions. By controlling for appellant behavior, we can infer the role of panelist perceptions. That we observe males behaving much more aggressively toward female panels, yet experiencing systematically more positive hearing outcomes already suggests a clear role for differences in panelist perceptions. Examining these effects more explicitly, we find that females remain systematically less likely to obtain home value reductions from female panelists, conditional on the appellant's in-hearing behavior and tone. The magnitude of the conditional effects leads us to conclude that the differential hearing outcomes are largely attributable to unvoiced perceptions or beliefs among ARB panelists.

Our findings have several important implications. The negative, and female-appellant specific gender concordance effect that we document reveals previously unrecognized biases that have the potential to manifest more broadly in fundamental administrative processes across the United States. Property tax appeals directly impact individuals' financial burdens and the amount they contribute toward the provision of public goods and services, e.g., public schools. Fairness in these hearings is essential for maintaining equity and trust in the tax system. Our findings are most likely to extend to other settings where evaluations are undertaken by peers or panels, such as academic peer reviews, corporate board decisions, judicial proceedings, administrative processes (e.g., driver's license issuance), and private decision-making contexts (e.g., loan requests). Our study underscores the need for policymakers to be aware of potential gender biases that may arise in these settings, such that they might consider structural or procedural adjustments, to mitigate their manifestation. Our work also highlights how gender differentials and biases may be evaluated and interrogated by employing large-scale administrative data, including unstructured records of procedural hearings.

II. Related Work

A great deal of literature has considered the impact of demographic concordance on evaluative outcomes. Much of that work has documented the role of gender concordance, showing that shared gender relates positively to evaluation outcomes, from patient evaluations of physicians (Lau et al., 2021; Cabral and

Dillender, 2024), to hiring and compensation (Coffman, Exley and Niederle, 2021; Cullen and Perez-Truglia, 2023), to judicial decisions in the courtroom (Boyd and Nelson, 2017; Knepper, 2018).

The mechanisms behind gender concordance effects are often nuanced and highly context-specific. Demographic concordance can facilitate more positive evaluations due to improved communication and empathy. However, literature dealing with social identity and cultural stereotypes points to a variety of context-specific reasons why gender concordance may benefit an evaluated individual, particularly in terms of gender stereotypes and gender roles. Particularly relevant to our study context, past literature has documented that oversight of household finances is often characterized by a strong gender stereotype, namely the expectation that males may take the lead (Bordalo et al., 2016; Brock and De Haas, 2023). This gender stereotype may cause women to exert less influence in property tax appeal hearings, as women may form negative self-assessments of their competence in the domain (Coffman, 2014). This, in turn, may lead to a positive gender concordance effect if, for example, evaluating women actively seek to overcome the gender stereotype by engaging in activist choice homophily (Greenberg and Mollick, 2017), a behavior that has been observed in the context of entrepreneurial finance, with female funders actively seeking to support female technology entrepreneurs as a means of combating the marginalization of women in tech.

There is also real potential that gender concordance will lead to negative outcomes for women. Research has found that gender dynamics can have a strong influence on negotiation outcomes, with women facing greater expectations and challenges than men (Bowles, Babcock and Lai, 2007; Recalde and Vesterlund, 2023). Literature on social identity suggests that in-group members may judge each other more harshly in settings where they may be concerned about possible accusations of favoritism (Tajfel and Turner, 1979). This idea has been argued to contribute to the 'queen bee' syndrome, wherein women who ascend to leadership positions in male-dominated workplaces may distance themselves from other, more junior women (Ely, 1994).

Of course, beyond the broad literature dealing with gender concordance effects, our paper relates and contributes to the literature studying what motivates households to appeal their property taxes. In this literature, Nathan, Perez-Truglia and Zentner (2025) show that both expected tax savings and filing frictions play significant roles, Giaccobasso et al. (2022) show that households are less likely to appeal if they believe that the government services funded by those taxes will provide more significant personal benefit to them, and Nathan, Perez-Truglia and Zentner (2024) show that households are willing to pay more in taxes if they believe other households are contributing their fair share. Some papers in this literature study property tax appeals heterogeneity by race, gender, and partisanship (Nathan, Perez-Truglia and Zentner, 2025, 2022; Avenancio-León and Howard, 2022; Holz, Novgorodsky and Simon, 2023). While these studies focus on what motivates the

initial decision to appeal, our focus is on biases that affect the process itself, at the ARB hearings stage. Of particular relevance to our study, Holz, Novgorodsky and Simon (2023) describes the results of an experiment that involved sending emails to county appraisers, posing as homeowners, and randomly varying the homeowner's gender. Those authors find that female appraisers are less likely to answer emails from a female homeowner. Although Holz, Novgorodsky and Simon (2023) does not focus on appeal hearings, as we do, their result is initially suggestive of possible negative gender concordance effects in the property tax context. It should be noted, however, that appraisers play a rather marginal role in hearing outcomes in our setting; despite appraisers' presence at the hearings, final decisions are ultimately rendered by the ARB panel.

Finally, our paper contributes to the literature on unstructured data and open-ended survey questions (e.g., Gentzkow and Shapiro, 2010; Gentzkow, Kelly and Taddy, 2019; Stantcheva, 2020, 2023). Large natural language models offer significant potential to expand the use of unstructured data. We advance this area of research by using these generative AI tools to process 2.7 years of audio recordings, a virtually impossible task for humans.

III. Methods

We divide this section into several parts. In the first subsection, we introduce our study context. In the second subsection, we present our data and measures. We then discuss identification and our econometric specification, followed by our approach to processing the unstructured audio data.

A. Study Context

Our study leverages data related to property appraisal appeal hearings conducted in Harris County, Texas. Harris County is the third largest county in the United States, with approximately 4.2 million residents as of 2024. Harris County residents are ethnically diverse[2] and politically moderate,[3] making it an ideal context for our analysis.

In Texas, County appraisal districts (CADs) appraise all homes within their counties and post these appraised values on their websites, sometimes also sending a "Notice of Appraised Value" by mail. These values are typically available by April 15 every year, and households have one month to file an appeal if they disagree with the proposed appraised value.

Protests can be filed for various reasons, including a) market value discrepancies, i.e., homeowners may argue the assessed value is higher than the market value of similar properties recently sold in the county, which often involves comparing the price of their property to the prices of comparable homes from recent

[2]https://www.houstonchronicle.com/projects/2023/houston-population-ethnicity/
[3]https://www.bakerinstitute.org/research/texas-counties

sales, known as "comps," b) unequal appraisal, i.e., the homeowner may argue that their property's proposed taxable value is disproportionately high compared to the taxable value of comparable properties within the same district or neighborhood, potentially signaling unequal appraisal practices leading to unfair taxation, and c) other reasons, such as denied exemptions, errors in the public records (e.g., an incorrect number of bedrooms, bathrooms, or square footage), etc.

Homeowners can file an appeal independently, or hire an agent to appeal on their behalf. After an appeal is filed, CADs may offer a settlement, usually via email. If the CADs do not offer a settlement or homeowners do not accept the proposed settlement, the appeal moves to the ARB, a group of citizens appointed by the local administrative district judge to resolve disputes between property owners and the CADs.

Appraisal District Boards in Texas, such as the Harris County Appraisal District (HCAD) in the Houston area, are political subdivisions of the Texas Comptroller of Public Accounts, and homeowners in every county in Texas can appeal using the same form provided by the state. For this reason, tax appeals operate very similarly across all 241 counties in Texas.

In Texas, as in many other states, property tax protests may be filed by or on behalf of any resident. Any given protest, should it proceed to a formal hearing, is adjudicated by a panel of ARB members drawn (randomly) from a pool of several hundred. To serve as a member of the ARB in Harris County (the focus of this study), an individual must apply and undergo an interview with a local district judge or their representative. Applicants for the position must meet several requirements. Applicants must have been a county resident for at least two years, and must not hold any conflicts of interest. An applicant must also be willing and able to attend ARB hearings on any given weekday or Saturday and must be willing and able to attend hearings conducted online or in person, at the discretion of the appellant.[4] Importantly, appellants are not made aware of the identity of their ARB panel in advance of the hearing. Hence, appellants have no ability to self-select, e.g., deciding whether a husband or wife might attend, based on their panel's composition.

Our structured data pertain to all property tax protests that proceeded to a formal hearing with the Harris County ARB, from 2013 through mid-2023, where the appellant appearing at the hearing was an individual homeowner.[5] The initial sample includes a total of 114,515 protest hearings. Additionally, we consider audio recordings associated with 80,197 of these hearings.

[4]Official requirements are detailed on the Harris County ARB panelist application form.

[5]Generally, the attendee that appears at a given hearing may be a property owner or a representative of a professional agency that files property tax protests on behalf of their clients. Our study focuses on individuals who represent their own property because protest agents are professionals. As a result, professional agents are likely to experience very different outcomes. Protest agents frequently represent large volumes of clients in front of the ARB; hence, they have more experience with the protest process, and they are also more likely to be familiar with ARB panelists through the course of their work.

B. Data and Measures

Structured Data. — Each protest hearing record includes information on the protested property value assessment, as well as the revised assessment (if any) that emerged as an outcome of the hearing, i.e., the adjusted property value. We consider the difference between these two measures. Specifically, we construct a binary indicator of $I(Saved)$ and a continuous, non-negative measure of the percentage saved, i.e., $PctSaved$, where the latter reflects the percentage decline from the original assessed value. The former value thus reflects savings that appellants receive at the extensive margin, while the latter reflects savings at both the extensive and intensive margin. In approximately 3,400 cases, the final assessed property value exceeds the original property value assessment. This is a rare occurrence, which likely reflects that pieces of information came to light during the hearing that was detrimental to the appellant's case, leading the ARB to realize that the initial assessment was too low. We omit these hearings from our analyses.

Each protest hearing record also includes the names of the ARB panelists who participated in the hearing, the appraiser, and the individuals who attended the formal hearing as the appellant(s). We parse the first and last names of each party from the hearing record. We then impute the gender of each based on their parsed first name.[6] In turn, we use this gender information to construct an indicator of appellant gender, as well as a count-based measure of the number of $FemalePanelists$ that were present.

Each protest record includes a hearing case description field, populated with categorical values conveying details of the hearing, including, most notably, the medium via which the hearing took place. Beginning in 2020, due to the COVID-19 pandemic, Harris County began allowing protesters to elect that their hearing takes place online, via WebEx or Zoom. This practice was continued with the decline of the pandemic and remains in place today. A value of VRTL indicates that a hearing was conducted online, via either WebEx or Zoom. This information is used to construct a binary indicator of the presence of this code in the hearing record, $Online$.

We also consider several variables as controls. We construct dummies reflecting

[6]This imputation is based on Social Security Administration (SSA) data (Tzioumis, 2018; Kaplan, 2021). It should be noted that multiple parties may attend the hearing representing the property in question. In such cases, we gender code the first name that is listed. We focus our analysis on the first appellant listed because, unlike the other names of ARB panelists and the appraiser, which are standardized and consistently reported – suggesting these values are populated from an IT system – all appellant names are reported together in a single field, and the formatting is highly inconsistent – presumably reflecting manual data entry. We elaborate on the implied measurement error later and the steps we take to address that error leveraging gender detected based on voices in our audio recordings. Gender inferences are made based on empirical probabilities derived from the SSA data. For example, if a particular first name is held by 100 people in the United States SSA records, and 98 of those people were identified as female, then we would infer a 98% probability that the individual in our sample is female. Our approach employs a majority threshold in assigning gender. However, our results are robust to using stricter thresholds, e.g., 80%, 90%, 95%, and indeed they grow stronger.

the tax year in which a protest was filed. Further, we construct a time-series measure of the cumulative number of property tax protest hearings that appear in the data before the tax year of a given protest, *PriorProtests*, as well as an indicator of whether the most recent protest, should it exist, resulted in savings for the appellant, *RecentSavings*. Finally, we construct two dummy indicators reflecting the basis or rationale of the protest, whether the appellant was arguing a market value discrepancy, and whether they were arguing an unequal appraisal value. All other arguments are taken as the reference category.

Lastly, it is important to note that, beginning in 2022, homeowners began to receive the option to present their case in front of a one-member or three-member panel. Accordingly, a subset of the hearing records document only one ARB panelist. We will consider estimations that model outcomes are a function of only the gender of the ARB panel chair. For hearings conducted before a three-member panel, we will also consider estimations modeling outcomes as a function of the count of female (vs. male) panelists.

Table 1 reports descriptive statistics for the variables based on our structured dataset of 114,515 hearings.[7] We see that the mean assessed home value in this sample is approximately $407,000, that approximately 20% of hearings have been conducted online, and that approximately 65% of hearings result in some tax savings for the appellant, translating to an average reduction of roughly $23,000 to the assessed home value. Further, concerning gender, we see that roughly 34% of hearing appellants are female, that approximately 55% of ARB panels are chaired by a female, and that the three-member panels include 1.55 females on average, suggesting strong gender balance in the ARB panelist pool.

Before we proceed further, we can assess the ARB's claim that panelists are, in fact, randomly assigned to hearings. We can evaluate this via a balance check, assessing covariate imbalance in hearing characteristics between ARB panels chaired by a male and female panelist. We assess covariate imbalance employing standardized mean differences (SMD), i.e., the difference in means in units of the pooled standard deviation (Rosenbaum and Rubin, 1985).[8] The results of the balance test are reported in Table 2. All variables exhibit an SMD well below 0.10, a common threshold for practically meaningful imbalance in the literature (Austin, 2009).

Audio Data. — In addition to our structured sample, we obtained audio recordings associated with thousands of hearings. This data is valuable because it allows

[7]Note that this sample reflects the subset of all structured hearing recordings, namely the subset for which we successfully i) employ regular expressions to parse a first name and ii) predict gender for all hearing attendees based on those first names, including all ARB panelists, appraisers, and the first appellant listed. Should we fail to parse a name that can be matched to SSA data, we ultimately omit the hearing recording. As a result, we omitted approximately 59,000 hearing records from the original sample where the gender prediction failed.

[8]Unlike t-tests, the standardized mean difference is not influenced by sample size. The standardized mean difference also enables a comparison of the relative balance across variables measured in different units.

Table 1—Descriptive Statistics: Structured Data

Variable	Mean	SD
Female Appellant	0.343	0.475
Female Panel Chair	0.552	0.497
Count Female Panelists	1.574	0.949
Any Saved	0.651	0.477
Assessed Home-value Decrease	22,946.383	63,540.795
Online	0.204	0.403
Home Value	407,010.893	537,843.339
Tax Year	2018.120	3.200
Protests Prior	0.621	1.122
Recent Savings	0.241	0.428

Table 2—Standardized Mean Differences in Hearing Characteristics By Panel Chair Gender

	Male		Female		
Variable	**Mean**	**SD**	**Mean**	**SD**	**SMD**
Appellant Female	0.346	(0.476)	0.341	(0.474)	-0.0114
Log(Home Value)	12.5	(0.851)	12.5	(0.857)	0.0795
Prior Protests	0.604	(1.11)	0.634	(1.13)	0.0274
Recent Savings	0.237	(0.425)	0.245	(0.430)	0.0195
Online	0.196	(0.397)	0.211	(0.408)	0.0394
Tax Year	2018	(3.24)	2018	(3.17)	0.0726

us to construct measures of what took place during hearings, with the potential to shed light on the sources of any bias we might document in hearing outcomes, based on our structured dataset. Additionally, we can leverage the recordings to explore any possible role of measurement error around appellant gender, considering alternative measures based on an appellant's voice. Given the ARB indicates a typical hearing should last about 15 minutes, we omit recordings that are excessively long or short in duration (more than 45 minutes or less than 3 minutes). We then match the resulting recordings to structured hearing records in our sample, achieving a match for 80,197 hearings, our working sample. The total audio duration in this sample is approximately 2.7 years (more than 23,650 hours). Listening to each recording and coding its characteristics manually would be infeasible. Accordingly, we process the audio recordings employing Google's Gemini 1.5 Flash, a multi-modal large language model (LLM) capable of processing lengthy audio input. For each audio recording, we ask Gemini to "listen" to the file and respond to several questions about the behavior and tone of the 'appellant who speaks the most' and the 'panelist who speaks the most', replying with a consistently (JSON) formatted response.

We use the annotations from Gemini to i) construct an alternative, complementary indicator of appellant gender, ii) explore possible differences in how appellants act depending on their gender and the gender of ARB panelists, iii)

examine how different behaviors and characteristics relate to hearing outcomes, depending on appellant and ARB panelist gender, and iv) whether hearing outcomes differ by appellant and ARB panelist gender, conditional on the in-hearing behavior and tone of appellants. It should be noted that we coded features that capture both behavior and tone of the 'appellant who speaks the most' and of the 'panelist who speaks the most'. This is important because the appellant who speaks most might not be the first appellant listed in the structured hearing record. Further, the panelist who speaks the most might not be the chair. It should also be noted that we did not attempt to code the gender of all three ARB panelists in the case of three-member panels, because the number of people in the room would be rather numerous, and the task would require a combination of speaker identification and gender prediction, a rather complex task, one for which we would lack confidence in Gemini Flash's ability.

Table 3 reports descriptive statistics for the variables derived from our audio dataset. We provide more detail about the specific behaviors and tones we code from the audio files in the Appendix. A particularly notable pattern we observe is very little variation among the tonal and behavioral measures exhibited by the ARB panelist. Panelists are almost always coded as sounding confident, they never sound angry or frustrated, etc. By contrast, appellants exhibit wide variation in behavior and tone. This difference is likely driven by the fact that ARB panelists follow a well-defined script and procedure during the hearings. Moreover, ARB panelists receive formal training on how to behave and what statements they should or should not make during hearings. This is particularly important because ARB panelists are generally aware that the outcomes of protest hearings may appealed in court.

C. Econometric Specification

Our initial specification, detailed in Equation (1), is intended to examine how homeowner and panelist gender relate to hearing outcomes, in terms of property tax savings. In the equation, hearings are indexed by i, and tax years are indexed by t. For measures of $HomeValueSavings$, we focus our initial estimations on a binary indicator of whether the homeowner achieves any reductions in their property taxes via a reduction in the assessed home value at the end of the hearing, $AnySaved^{g}_{i,t}$). We model that binary indicator as a function of the ARB panel chair's gender, $FemalePanelChair_{i,t}$, tax-year fixed effects, home value bucket fixed effects, where buckets reflect ranges of $100,000, appraiser fixed effects, a measure reflecting the count of prior property tax protest hearings that appear in our sample for the same property up to that point, i.e., from 2013 forward, a binary indicator of whether the hearing is conducted online or offline, and a binary indicator of whether the most recent protest hearing, if it exists, resulted in some appraisal reduction (tax savings) for the property owner. Our key parameter of interest in this analysis is β_1, which captures any differences in hearing outcomes depending on the gender of the panel chair, conditional on case characteristics.

Table 3—Descriptive Statistics: Audio Data

Variable	Mean	SD
Appellant Sounds Confident	0.896	0.305
Appellant Disagrees with Panelist	0.759	0.428
Appellant Presents to Evidence	0.589	0.492
Appellant Sounds Hostile	0.111	0.314
Appellant Raises Voice	0.146	0.353
Appellant Asks Questions	0.718	0.450
Appellant Sounds Assertive	0.932	0.252
Appellant Sounds Frustrated	0.809	0.393
Appellant Sounds Annoyed	0.769	0.422
Appellant Sounds Defensive	0.244	0.429
Appellant Interrupts Panelist	0.335	0.472
Appellant Asks Panelist to Speak Louder	0.065	0.246
Appellant Sounds Nervous	0.023	0.149
Appellant Asks Panelist to Repeat	0.348	0.476
Panelist Sounds Frustrated	0.005	0.071
Panelist Sounds Annoyed	0.004	0.067
Panelist Sounds Angry	0.000	0.015
Panelist Sounds Nervous	0.000	0.007
Panelist Sounds Confident	0.999	0.024
Panelist Sounds Defensive	0.006	0.078
Panelist Sounds Hostile	0.000	0.015
Panelist Raises Voice	0.000	0.016
Panelist Asks Appellant to Speak Louder	0.019	0.136
Panelist Asks Appellant to Repeat	0.133	0.339
Panelist Interrupts Appellant	0.104	0.305
Panelist Disagrees with Appellant	0.327	0.469

(1)

$$HomeValueSavings_{i,t} =$$

$$\beta_1 \cdot FemalePanelChair_{i,t} + \beta_2 \cdot PriorProtests_{i,t} + \beta_3 \cdot Online_{i,t}$$

$$+ \beta_4 \cdot RecentSavings_{i,t} + \beta_5 \cdot MarketEquity_{i,t} + \beta_6 \cdot UnequalAppraisal_{i,t}$$

$$+ \sum_{v=2}^{V} \eta_v \cdot HomeValueBucket_{i,t}^{v} + \sum_{\tau=2014}^{2023} \lambda_\tau \cdot TaxYear_{i,t}^{\tau} + \sum_{a=2}^{A} \nu_a \cdot Appraiser_{i,t}^{a} + \epsilon_{i,t}$$

We also conduct estimations that replace our binary indicator of a female panel chair gender with a series of three dummies reflecting the count of female panelists (taking 0 as a reference), limiting the estimation to only those hearings that involved three-member ARB panels. Further, we report a series of estimations incorporating account (property) fixed effects to ensure our results are not a function of unobserved imbalances in the features of the properties involved.

We next replace our binary outcome, *AnySaved*, with a non-negative measure of the dollar amount saved, *AmountSaved*. Given the right-skewed nature of this measure and that the scale of the amount saved will vary with the baseline valuation, we take the logarithm of the amount saved, and employ it as our dependent variable in an OLS regression. Our coefficients in these regressions thus reflect approximate percentage effects. When performing all estimations, we cluster our standard errors in two dimensions, namely by property account number and by appraiser.

Subsequently, we turn to the audio recordings. We begin by reporting a visual, descriptive analysis of the measures obtained from the M-LLM, to explore differences in coded appellant behaviors and tones by gender, both of the appellant and the ARB panelist. We then examine how each behavior or tone associates with hearing outcomes, conditional on the appellants' and ARB panelist genders. Finally, we incorporate the behavioral and tonal measures into our baseline regression specification, to understand whether any effects of ARB panelist gender persist conditional on appellant behavior.

When conducting these final regressions leveraging audio data, we incorporate the appellant gender values that Gemini infers based on the hearing audio. This is important because, as noted above, when dealing with structured hearing records, we gender code only the first name listed among all appellant attendees, and it is possible the coded individual is not the individual who led the conversation. Further, our parsing may be subject to error, given the name format varies across records. And, the gender predictions based on Social Security Administration data may also be occasionally incorrect. Accordingly, by incorporating Gemini's predictions about appellant gender based on voice, we obtain an independent indicator of appellant gender,[9] one for which the measurement error is presumably independent of that in our name- and regular-expression-based gender predictions. Most importantly, we can consider estimates from a sub-sample of records among which our name-based and audio-based gender predictions agree, i.e., a sample likely characterized by relatively little measurement error.

Our goals with this second set of regressions, incorporating the data from the audio recordings, are three-fold. First, we hope to ensure that our baseline findings are not a product of endogeneity and bias due to measurement error in appellants' gender predictions obtained via name-parsing using regular expressions. Second, we hope to assess how homeowner behaviors and tone translate to hearing outcomes depending on panel and homeowner gender. Third, and most importantly, we hope to determine whether any systematic differences in hearing

[9]We are somewhat less concerned with gender predictions associated with the ARB panelists because i) formatting of ARB panelist names is consistent throughout the data, suggesting that the values are drawn from an IT system, and ii) the confidence associated with SSA-based gender predictions for ARB panelists is generally high. Further, as mentioned above, we are less comfortable relying on Gemini's gender predictions for ARB panelists, the chair or otherwise, because there are typically three panelists, hence accurate prediction requires both accurate speaker disambiguation/identification followed by a gender inference, implying much more measurement error.

outcomes that depend on panel and appellant gender persist after we condition on measures of homeowner behavior and tone during the hearing. Were we to observe the latter, it would suggest that the sources of the gender concordance effects are primarily manifesting via unvoiced factors that are unrelated to appellant behavior (e.g., differences in perceptions or beliefs on the part of ARB panelists).

IV. Descriptive Evidence

Assuming the assignment of ARB panelists is a true experiment, we can perform a simple OLS regression, relating the *AnySaved* dummy to a binary indicator of whether the ARB panel chair is female, absent any controls, and the result will bear a plausibly causal interpretation. Doing so, we obtain the results in Table IV, where we observe evidence of a negative gender concordance effect for female appellants, yet no evidence for any such effect among male appellants.

Table 4—Effect of Female Panel Chair on Probability of Any Reduction in Appraisal Value, by Appellant Gender

Appellant Gender	**Female**	**Male**
Intercept	0.653*** (0.006)	0.656*** (0.006)
Female Panel Chair	-0.016*** (0.006)	-0.004 (0.006)
Num.Obs.	39,309	75,206
R2 Adj	2.435e-4	2.46e-6
RMSE	0.4787	0.4757

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser.

Having established a baseline estimate, we next consider various descriptive analyses that speak to the presence of panelist-specific bias, and whether any such bias is gender dependent. We begin by plotting estimates of the average rate at which each ARB panelist grants home value reductions to appellants, with 95% confidence intervals. When calculating these values, we limit our attention to ARB panelists who participated in at least 50 hearings during our observation period, to avoid noisy estimates. Further, when calculating these values, we first demean the indicator of *AnySavings* with respect to the tax year, to account for the fact that the rate at which savings were awarded differed systematically over time due to rapid increases in average home value. A cap is imposed on the percentage that property appraisal values can increase from one year to the next. Because home values rose rapidly at some point, property appraisals began to systematically undervalue homes, leading to a decline in the rate at which hearings would result in additional tax savings. Were we to ignore tax year when constructing these estimates, ARB panelists who served in different years would

exhibit marked variation in their rate of awarding savings due to these market-level shifts.

We depict the resulting estimates in Figure 1, ordering panelists along the horizontal axis from least to most likely to adjudicate savings. The average probability that savings are adjudicated spans a range of approximately 40 percentage points, from the most 'critical' to the most 'favorable' ARB panelist.

Even if ARB panelists exhibited no systematic bias in their decisions, a normal distribution of savings adjudication rates across panelists might nonetheless be expected due to random variation in case characteristics. Accordingly, to explore whether the observed distribution truly reflects panelist-specific biases in decision-making (i.e., whether the distribution truly differs from what we might expect due to chance), we perform a permutation test (Good, 2013). We implement our test via a repeated shuffling of the hearing outcome variable (i.e., *AnySaved*) over hearings to recover artificial, simulated ARB panelist-specific estimates of savings adjudication rates under the null hypothesis that adjudication rates are, in fact, homogeneous across panelists. That is, by breaking the relationship between the outcome and assigned ARB panelist identities, we effectively simulate the null distribution of our panelist-specific test statistics absent any panelist-specific bias. We can then compare the observed distribution of test statistics against that null distribution to arrive at an empirical p-value. This approach effectively retains the structure of panelist case assignments while enabling us to test the null hypothesis that panelists have no systematic effect on outcomes.

We summarize the observed degree of panelist-specific bias using the variance of resulting savings adjudication rates, across panelists. It is important to note that our approach provides a conservative test of systematic bias. Variance in the outcome attributable to panelist bias (panelist-specific treatment effects, in a sense) is unaffected by the shuffling procedure; that is, the shuffling procedure does not eliminate the bias, it simply redistributes it across hearings uniformly. Because the variation in the outcome due to bias remains, the distribution we observe under the null reflects an inflated variance relative to what would occur in the true absence of panelist bias. As a result, our permutation test is relatively conservative or robust. In any case, we expect greater variation in savings adjudication rates in the true data, compared to the null distribution, to the extent that panelist-specific bias is at play.

In Figure 2, we plot the true set of estimates in red, atop the simulated estimates from 1,000 randomly shuffled samples shown in gray. The simulated distributions show what would result from the objective cases – i.e., the distribution of savings if ARB panelists did not exhibit systematic bias. It is readily apparent that the true distribution has much fatter tails than any simulated distribution, such that the variance in panelists' rates of savings adjudication is wider than can be explained by chance alone.

Delving deeper into this analysis, we can contrast hearing outcomes per panelist among hearings split by appellant gender, to gain a sense of the extent to

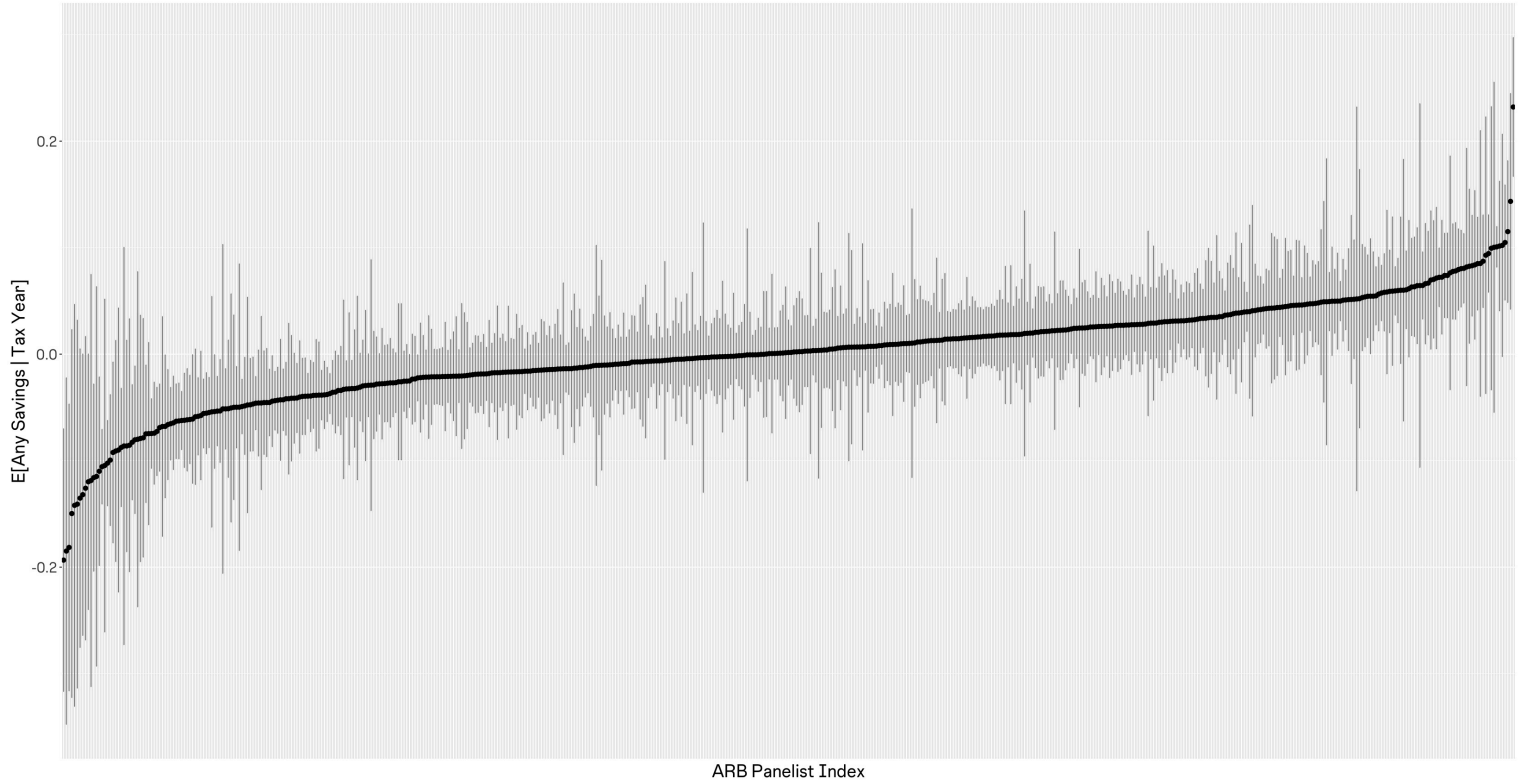


Figure 1. Average rate at which tax savings are awarded per ARB panelist, de-meaned by tax year.

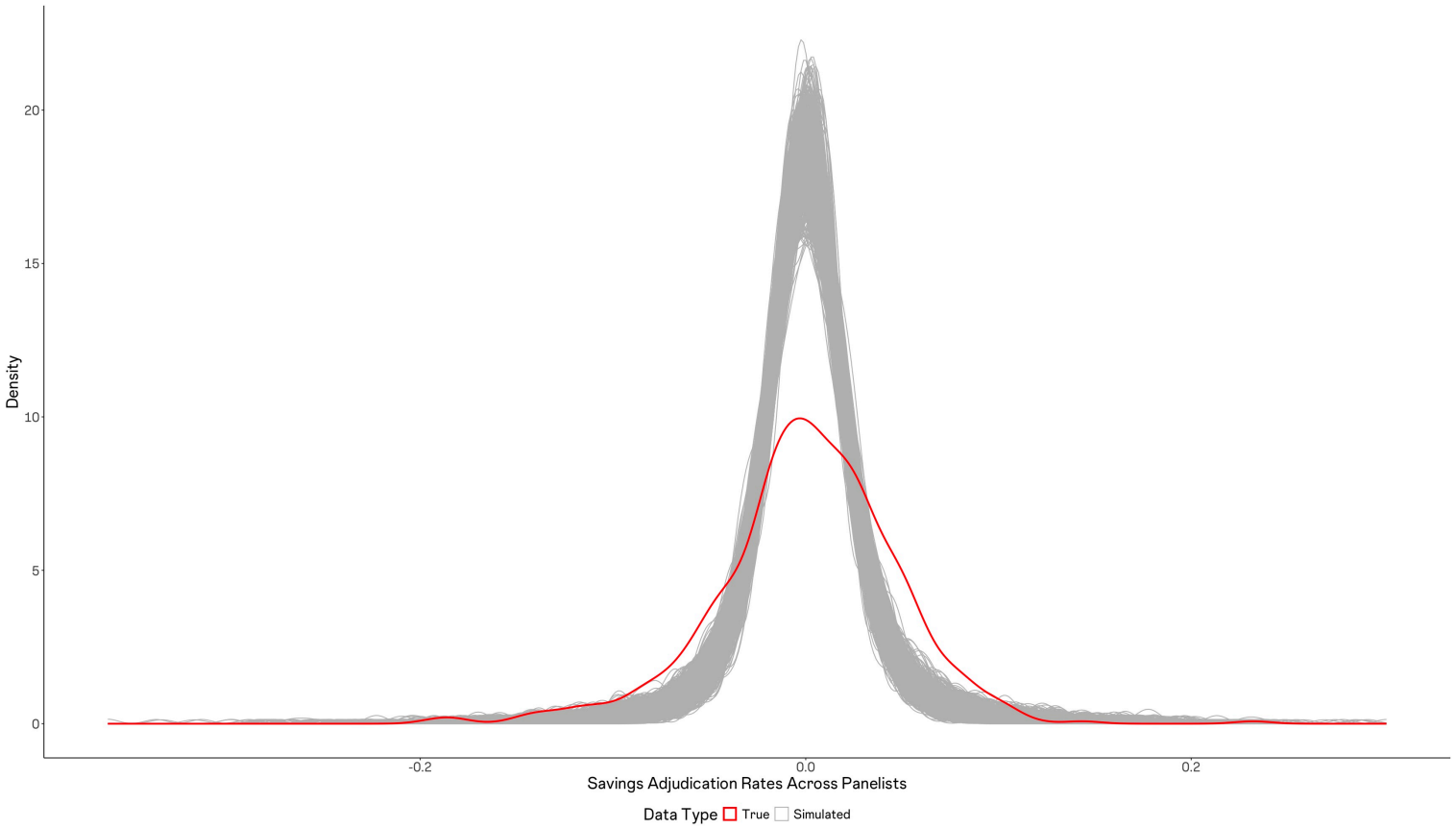


FIGURE 2. PERMUTATION TEST EVALUATING PANELIST BIAS; VARIANCE OF TRUE ESTIMATES YIELDS AN EMPIRICAL P-VALUE OF 0.

which panelists adjudicated decisions might depend on homeowners' gender. We identified all ARB panelists who participated in at least 50 hearings and had also participated in at least 25 hearings involving homeowners of each gender. We calculate the rate at which each panelist awarded a home value reduction, i.e., the mean of our *AnySaved* measure across hearings in which they participated. Finally, we calculate the difference between the resulting values across homeowner genders, i.e., E[AnySaved | Female] - E[AnySaved | Male]. Finally, grouping ARB panelists based on their gender and ordering them from anti-female bias (left) to pro-female bias (right), we plot the resulting value curves in Figure 3.

Inspecting the figure, two facts are immediately apparent. First, a) 55 percent of male panelists exhibit anti-female decision-making, as compared to 62 percent of female panelists, and b) the anti-female tendencies of the most anti-female panelists are much stronger than the anti-male bias of the most anti-male panelists. This latter observation is based on the fact that the most anti-female panelist, on the extreme left of the horizontal axis, adjudicates savings for female appellants with 20% lower probability than male appellants and, by contrast, the most anti-male panelist, on the extreme right, adjudicates savings for male appellants with approximately 15% lower probability than female appellants.

V. Main Regression Results

We now turn to our regressions, wherein we can account for changes over time via tax year dummies, and we can also explore within property variation in hearing outcomes across hearings, depending on panel composition. We begin with an estimation of the relationship between ARB board chair gender and hearing outcome, splitting the sample by appellant (homeowner) gender, considering our binary indicator of whether any reduction resulted in assessed home value, i.e., *AnySaved*, employing Ordinary Least Squares (OLS) regression, i.e., a Linear Probability Model (LPM). The results are presented in Table 5.

The upper panel presents regression results from models that include fixed effects by tax year, home value buckets, and appraiser, whereas the lower panel replaces home value bucket fixed effects by account fixed effects.[10] We observe consistent results across the board in terms of gender effects. The probability that female homeowners obtain reductions in their home value assessment, and thus their property taxes, is systematically lower (higher) when they face a female (male) ARB chair. Considering the magnitude, the difference in probability is approximately 1.4 pp. Considering three-member panel hearings only, and replacing the binary indicator of ARB chair gender with dummies reflecting the count of female panelists (the reference value is 0, implying an all-male panel), we see the same pattern. Female homeowners' probability of obtaining any tax

[10]We were unable to obtain a response from the Harris County Appraisal District, but speaking with representatives of the Dallas County Appraisal District, they indicated that the property account number remains tied to a particular home even if it changes owners. The DCAD representative indicated they believed the same was true of all counties in Texas.

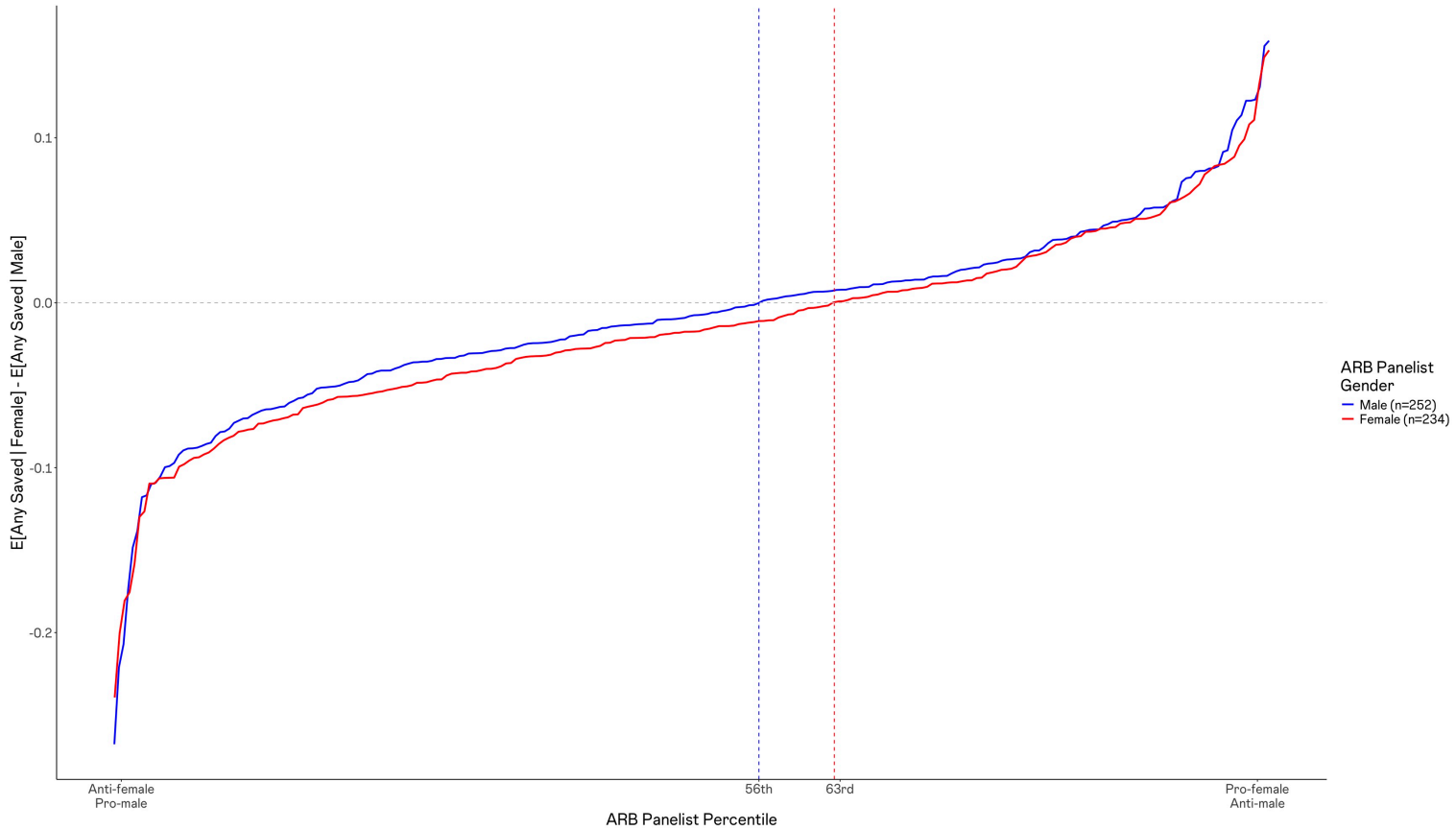


Figure 3. Female vs. Male Differential per ARB Panelist

Table 5—Effect of ARB and Appellant Gender on Any Reduction in Home Value

Appellant Gender	**Female**	**Male**	**Female**	**Male**
Tax Year, Home Value Bucket & Appraiser FEs				
Female Panel Chair	-0.014*** (0.005)	-0.004 (0.004)		
1 Female Panelist			-0.025*** (0.009)	0.002 (0.007)
2 Female Panelists			-0.031*** (0.009)	0.004 (0.008)
3 Female Panelists			-0.042*** (0.010)	-0.007 (0.009)
Online	-0.040*** (0.011)	-0.031*** (0.008)	-0.030*** (0.011)	-0.021*** (0.009)
Market Equity	0.115*** (0.014)	0.125*** (0.009)	0.116*** (0.014)	0.127*** (0.009)
Unequal Appraisal	0.036*** (0.006)	0.035*** (0.005)	0.037*** (0.006)	0.036*** (0.005)
Count Prior Protests	-0.057*** (0.003)	-0.049*** (0.003)	-0.059*** (0.004)	-0.051*** (0.003)
Recent Savings	0.089*** (0.008)	0.060*** (0.006)	0.087*** (0.008)	0.061*** (0.006)
Num.Obs.	39,309	75,206	35,453	67,732
R2	0.070	0.070	0.063	0.059
R2 Adj.	0.061	0.065	0.053	0.054
RMSE	0.46	0.46	0.46	0.45
Tax Year, Account & Appraiser FEs				
Female Panel Chair	-0.020** (0.010)	-0.005 (0.006)		
1 Female Panelist			-0.044*** (0.017)	0.015 (0.011)
2 Female Panelists			-0.036*** (0.017)	0.018 (0.011)
3 Female Panelists			-0.048*** (0.020)	0.010 (0.013)
Online	-0.007 (0.022)	0.014 (0.015)	0.006 (0.024)	0.024 (0.016)
Market Equity	0.066*** (0.022)	0.061*** (0.016)	0.062*** (0.023)	0.061*** (0.016)
Unequal Appraisal	0.020 (0.013)	0.045*** (0.008)	0.014 (0.014)	0.049*** (0.008)
Count Prior Protests	-0.037*** (0.009)	-0.020*** (0.005)	-0.037*** (0.009)	-0.019*** (0.006)
Recent Savings	-0.239*** (0.010)	-0.268*** (0.006)	-0.243*** (0.010)	-0.268*** (0.007)
Num.Obs.	39,309	75,206	35,453	67,732
R2	0.862	0.829	0.863	0.829
R2 Adj.	0.270	0.285	0.260	0.266
RMSE	0.18	0.20	0.18	0.19
Mean(Any Saved)	0.644	0.654	0.661	0.674

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB member.

savings is monotonically decreasing in the number of female panelists they face, such that the difference in the probability of tax savings between panels comprised entirely of males versus entirely of females is approximately 4.2 pp. Next, we consider the bottom panel of estimates, where we incorporate home account fixed effects, enabling us to focus strictly on variation in the panel gender composition across multiple hearings that relate to the same home. We again observe consistent results, though statistical significance declines, perhaps unsurprisingly. Importantly, by contrast, when we consider gender concordance effects for male homeowners, we observe no statistically significant effects, in any of the estimations. Indeed, all coefficients associated with our chair and panelist gender dummies indicate precisely estimated null results.

Beyond the gender effects, our other coefficients reveal interesting patterns. First, online hearings are systematically less likely to result in tax savings, though these effects disappear following our inclusion of account fixed effects, which suggests that those results may be driven by self-selection. For example, perhaps homeowners are more likely to request a hearing online when they anticipate their case is weaker and less likely to result in savings.

Second, we observe that, when appellants have previously protested and obtained any savings in their most recent hearing, this generally associates positively with current hearing outcomes when we do not include an account fixed effect. Once we introduce the account fixed effect, the coefficient turns negative. These results suggest that protest hearings involving certain owners, appellants, or property characteristics are systematically more likely to achieve reductions to an initial appraisal value. However, once we account for those stable, cross-sectional differences, recent success in a protest hearing reduces the likelihood of subsequent favorable outcomes. That is, a fresh win appears to reduce the likelihood of obtaining another reduction. This dynamic can likely be explained by the fact that recent savings achieved in a prior protest will generally imply that some error has been identified in the prior appraisal process, and that error has been rectified. Once errors have been addressed, the chance of success in subsequent protests will naturally decline, as arguments for any reduction will be reduced. Third, we observe a consistent negative effect of prior protest hearing counts in all regressions. Given our estimations condition on whether a property has obtained savings on the most recent prior appeal, this effect could indicate that appraisal review board panels, observing a lengthy history of prior protests within the property's file, may view the appellant's case with greater skepticism.

That said, it is important to also recognize that several administrative restrictions limit how property taxes can be adjusted. These restrictions influence the incentives for households to appeal an appraisal, as well as the potential savings households may obtain from an appeal. Further, these restrictions make it difficult to evaluate the temporal evolution of appeals and savings arising from them (Giaccobasso et al., 2022). Most notably, property tax increases are capped at 10% annually. Households already under the cap have less incentive to appeal

since their efforts would not immediately reduce their property taxes.

We next explore our measure of the dollar amount saved. As noted earlier, these estimations employ the logarithm of the amount saved as the outcome. The coefficient estimates are thus interpretable as approximate percentage effects. Considering Table 6, we observe the same general pattern of effects as in Table 5. Whereas female homeowners' case outcomes differ systematically depending on the gender of the ARB panelists, outcomes for male homeowners do not. Considering our estimations in column 1, which employs the same right-hand side model as in column 1 of Table 5, we see that female homeowners save approximately 12.2% less (more) when faced with a female (male) ARB panel chair.[11] Further, we see that the amount saved declines with the addition of more female panelists, such that the difference in home value reduction obtained by female homeowners when facing an all-female versus all-male panel is approximately 32.8%.[12] Incorporating account fixed effects, we observe consistent findings in the case of ARB chair gender, though statistical significance declines slightly.

VI. Audio Analysis

We next leverage a multi-modal large language model (LLM), specifically Gemini 1.5 Flash, to process all audio recordings associated with nearly all owner-attended ARB hearings in our original sample. Ultimately, we work with a subset of 80,197 audio recordings, namely those we can reliably match to hearings from our original sample of structured hearing records. Some structured hearing records in our sample do have associated recordings. Further, in some cases, we have multiple hearings for the same property in the same tax year (e.g., if separate hearings were held to adjudicate different protest reasons), and we lack any means of correctly discerning which recording associates with which hearing. Employing Gemini, we construct measures that quantify appellant behavior during a hearing. Using this data, we hope to understand i) any systematic differences that female homeowners may exhibit depending on the gender composition of the panel they face, and ii) any systematic differences in the relationship between homeowners' behavior and hearing outcomes, depending on ARB panelist gender.

As noted earlier, we consider a measure of hearing length (in seconds), extracted from the metadata of each audio recording. Further, we consider Gemini's assessment of the fraction of time an appellant spent talking during a hearing, binary indicators of how the appellant sounds, e.g., confident, frustrated, nervous, defensive, assertive, etc., and binary indicators of appellant actions during the hearing, e.g., referencing evidence, raising their voice, interrupting a panelist, disagreeing with a panelist, saying thank you, asking questions, asking a panelist to speak louder, or asking a panelist to repeat themselves.

[11] exp(-0.13) - 1 = -0.122
[12] exp(-0.397) - 1 = -0.328

Table 6—Effect of ARB and Appellant Gender on Log(Dollar Reduction in Home Value)

Appellant Gender	**Female**	**Male**	**Female**	**Male**
Tax Year, Home Value Bucket & Appraiser FEs				
Female Panel Chair	-0.131*** (0.053)	-0.062 (0.045)		
Count Female Panelists: 1			-0.233*** (0.092)	0.004 (0.073)
Count Female Panelists: 2			-0.304*** (0.087)	0.009 (0.078)
Count Female Panelists: 3			-0.397*** (0.102)	-0.103 (0.087)
Online	-0.436*** (0.101)	-0.447*** (0.080)	-0.363*** (0.110)	-0.359*** (0.085)
Market Equity	1.353*** (0.139)	1.549*** (0.094)	1.361*** (0.138)	1.561*** (0.094)
Unequal Appraisal	0.203*** (0.055)	0.215*** (0.052)	0.206*** (0.056)	0.219*** (0.052)
Count Prior Protests	-0.553*** (0.033)	-0.491*** (0.027)	-0.568*** (0.034)	-0.512*** (0.027)
Recent Savings	0.918*** (0.075)	0.632*** (0.056)	0.894*** (0.079)	0.640*** (0.058)
Num.Obs.	39,309	75,206	35,453	67,732
R2	0.082	0.084	0.077	0.075
R2 Adj.	0.073	0.079	0.067	0.070
RMSE	4.50	4.53	4.46	4.49
Tax Year, Account & Appraiser FEs				
Female Panel Chair	-0.211*** (0.098)	-0.104 (0.064)		
Count Female Panelists: 1			-0.400*** (0.164)	0.066 (0.115)
Count Female Panelists: 2			-0.334*** (0.168)	0.094 (0.108)
Count Female Panelists: 3			-0.476*** (0.194)	0.006 (0.128)
Online	-0.071 (0.214)	0.055 (0.149)	0.086 (0.229)	0.154 (0.167)
Market Equity	0.753*** (0.219)	0.864*** (0.153)	0.698*** (0.232)	0.850*** (0.159)
Unequal Appraisal	0.158 (0.131)	0.430*** (0.079)	0.080 (0.137)	0.455*** (0.084)
Count Prior Protests	-0.421*** (0.089)	-0.252*** (0.053)	-0.406*** (0.093)	-0.246*** (0.060)
Recent Savings	-2.320*** (0.100)	-2.662*** (0.065)	-2.390*** (0.104)	-2.670*** (0.068)
Num.Obs.	39,309	75,206	35,453	67,732
R2	0.862	0.826	0.863	0.826
R2 Adj.	0.271	0.270	0.261	0.253
RMSE	1.74	1.98	1.72	1.95
Mean(Amount Saved)	$20,206.00	$24,379.00	$20,678.00	$24,904.00

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB member. Log transformation is applied as Ln(y+1) in the case of $0 savings.

A. *The Quality of LLMs' Audio Coding*

Before considering the results, it is worth considering the time and cost savings from relying on a multimodal LLM for this annotation exercise. The average recording is 15 minutes long. The total length of audio is approximately 2.8 years in length. Had we hired a human annotator for the task, paying minimum wage, i.e., $7.25/hr, the total cost would have amounted to approximately $177,000. In contrast, our total cost of employing Gemini 1.5 Flash for this exercise was approximately $450. The time to complete this annotation task would also have required years of effort (perhaps weeks or months, in practice, were we to have employed many coders in parallel). By contrast, processing recordings sequentially via the Gemini API (i.e., without parallel processing) required approximately 14 days. With parallel processing, the task could have been completed in a matter of days.

Before we proceed, it is helpful to obtain some sense of Gemini's accuracy. We have a couple of Gemini-coded values that we can compare against 'ground truth' labels of a sort, derived from the structured hearing data. In particular, we examine Gemini's inferences about whether the appellant seems to be the property owner versus an agent hired on the owner's behalf, where the reality is that all appellants in our data are property owners. Additionally, we examine Gemini's inferences about whether a hearing sounds like it took place online or in person, comparing the result with the actual medium recorded in the hearing record. The results of these comparisons appear in Table 7 and 8.

Considering Gemini's ability to infer appellants' online attendance at their hearing, it is important to recognize that this inference depends on explicit cues in the hearing audio, e.g., appellant references to their screen sharing and audio configuration. To the extent an appellant attending online has no difficulty with their technology setup, cues of online attendance will be absent. Conversely, for hearings that truly took place offline, there should never be cues of online attendance. Our result here is entirely consistent with that idea. Among hearings that took place offline, it makes sense that no such cues would be present. Accordingly, as we would hope, Gemini rarely infers that an offline hearing occurred online. Gemini mistakenly labels offline hearings as online in just 1.2% of instances (827 out of 69,312 hearings). By contrast, among online hearings, Gemini fails to detect a verbal cue indicating online attendance by the appellant in approximately 39.8% of instances (4,334 out of 10,885 hearings).

Table 7—Cross-tab of Actual vs. Inferred Online vs. Offline Hearing Attendance

	Gemini Offline	Gemini Online
True Offline	68,485	827
True Online	4,334	6,551

Considering Gemini's inferences about whether the appellant is the property owner, again, this will depend on explicit reference being made to the appellant's ownership of the property, or some other indicative statements, serving as cues that the appellant lives at the property. Such cues should be relatively common, as ARB panels often ask the attendees whether they own the property. Further, appellants will often reference living at the location, the price at which they purchased the property, and so on. It is thus reassuring to see that Gemini correctly identifies the appellant as the property owner in approximately 92.8% of our hearings sample (as noted earlier, our sample is exclusively comprised of owner-attended hearings).

Table 8—Gemini-inferred Property Owner vs. Hired Agent

Gemini's Inference	Count
Agent	5,758
Owner	74,439

B. Descriptive Results

We now turn to Gemini's responses about the appellants' and panelists' behaviors and tone. In Figure 4 we graphically depict the incidence rate for each behavior or tone for appellants and panelists, reporting split sample means and 95% confidence intervals for each combination of attendee gender and ARB chair gender. The values we depict are re-centered around the population mean incidence for the respective behavior or tone, to facilitate easier comparisons of cross-gender differences. For example, we observe that, in the case of male appellants facing male ARB chairs, appellants are systematically more likely to sound confident (they are labeled as such 1.75% more often than the average incidence across our entire sample). In contrast, female appellants are systematically more likely to ask questions of the panel when facing a female ARB chair, approximately 1.6% more often than the overall average.

Several notable patterns are immediately apparent from the plot. First, as we observed in our descriptive statistics for the labels, the coded behavior and tones of ARB panelists exhibit much less variance than those of appellants. Across most measures of behavior and tone, Gemini codes panelists quite homogeneously across hearings. The only exceptions relate to our measures of whether panelists ask the appellant to 'Speak Up', ask them to 'Repeat' themselves, 'Interrupt' them, or disagree with them. Of greatest note is that female panelists are systematically less likely to verbally disagree with a female appellant, which contradicts what we might expect given that female appellants tend to fare worse with female ARB panels.

Considering the appellants' behavior, we observe immediate and stark differences between males and females. For example, male appellants are less likely to be coded as sounding nervous and more likely to be coded as sounding confident. Males are also more likely than females to verbally disagree with the ARB panel, they are also more likely to present formal evidence, and they are more likely to raise their voice or sound hostile. Interestingly, however, some of these more aggressive behaviors are mainly targeted toward female ARB panelists. Considering how appellants' behavior *differs* depending on ARB chair gender, we see that male appellants are more likely to interrupt female panelists, more likely to raise their voice toward female panelists, more likely to sound hostile toward female panelists, and more likely to sound defensive, compared to their behavior when facing a male ARB panelist.

By contrast, female appellants exhibit relatively few differences in their behavior between male and female ARB panelists. Even for those behaviors and tones where female appellants' behavior does differ based on ARB chair gender, the differences are relatively small in magnitude, e.g., no more than a 0.5 pp difference in the probability a behavior or tone is exhibited. Further, the differences are statistically significant in only four cases: i) asking a panelist to repeat themselves, ii) asking a panelist to 'speak up', and iii) sounding confident. Notably, however, in these cases, the behavior or tone in question is more pronounced when a female appellant faces a female rather than a male ARB chair. As such, we do observe evidence of some small behavioral and tonal differences on the part of female appellants depending on the ARB panel chair's gender that *may* explain the less favorable outcomes of female appellants, to some extent, though the relatively small degree of variation in behavior suggests that this explanation is unlikely to explain differential hearing outcomes entirely.

Before we proceed further, it is also helpful to consider the contrast between our regular-expression, Social Security Administration-based predictions of appellant genders and the predictions that Gemini makes based on audible voice in the recordings. As noted earlier, we would anticipate that measurement error will arise with each approach, but that the errors between each measure should be independent of one another. Whereas our regular-expression approach focused on the first name appearing in the formal attendee field, our request of Gemini is to predict the gender of the appellant who 'speaks the most' during the hearing. The order of names entered in the structured data is not necessarily reflective of which appellant led the case presentation at the hearing. That said, for Gemini to succeed at this task, it is not only required to discern gender based on voice, it must also correctly identify which speaker is the appellant. Moreover, even the ultimate task of inferring gender from the vocal pitch, etc., for the identified speaker can be challenging. The audio recordings vary significantly in their quality, and for some voices, even with clear audio, it may be difficult to determine a speaker's gender with any degree of confidence (see Schanke, Burtch and Ray, 2024, for a lengthy discussion of gender inference around audible vocal character-

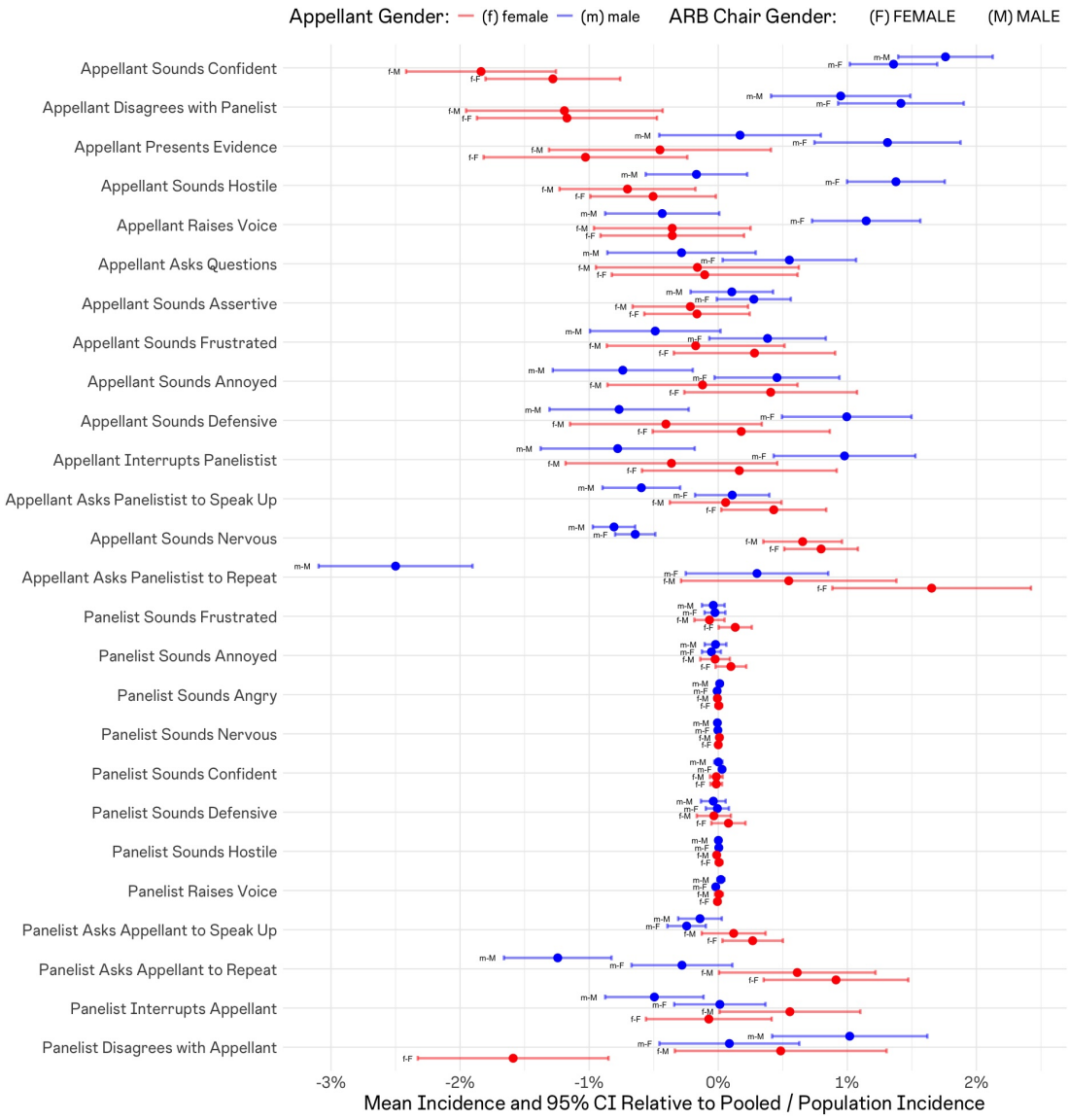


Figure 4. Appellant and Panelist Behavior by Own and ARB Panel Chair Gender. Panelist Behavior Refers to the Panelist Who 'Speaks the Most' During the Hearing, and Appellant Behavior Refers to the Appellant Who 'Speaks the Most' During the Hearing. We explain the definition of each tone and behavior in the Appendix.

istics). We present the cross-tabulation of Gemini and regular-expression-based gender inferences for appellants in Table 9, where we see that the two approaches agree with one another in 77.8% of cases. This high level of agreement suggests that there is more signal than noise in each approach we take, yet, as expected, there also appears to be substantial error with each.

Table 9—Cross-tabulation of Gender Inference: Gemini (Audio) vs. Regular Expression (SSA)

	Regular Expression: Female	Regular Expression: Male
Gemini: Female	17,077	7,145
Gemini: Male	10,628	45,347

Given the above, before we proceed with our analysis of the impact of appellant behaviors on hearing outcomes, we will first explore the sensitivity of our earlier findings by relying on either measure, separately, or their intersection. We first repeat the baseline regressions on our audio-recorded sample, splitting the sample based on appellant gender as determined by Gemini from audio. The results, presented in Table 10, are generally consistent with our earlier findings. The estimates are slightly attenuated relative to our earlier estimates, but the general pattern of effects remains consistent. Further, in Table 11, we again repeat the estimations, this time employing only those hearings where Gemini's prediction about appellant gender agrees with our regular-expression-based predictions. Again, we find consistent results. Repeating the same exercise with our *AmountSaved* outcome, we obtain the results in Tables 12 and 13. Again, we observe consistent results in each case. Broadly, the stability of our estimates is reassuring, as it suggests that measurement error in the gender predictions, whatever its nature, does not drive or materially influence our findings.

Now, perhaps due to social norms, whatever behavior and tones may be exhibited by appellants, they may also be perceived differentially by ARB panelists, potentially in ways that also depend on ARB panelists' genders. That is, female ARB panelists may perceive female appellants' behavior differently than male ARB panelists do, conditional on the behavior exhibited. We consider this possibility next, in two ways. First, we attempt to understand whether and to what degree appellants' behavior differentially correlates with hearing outcomes, depending on their gender and that of the ARB panel. Second, we repeat our analysis of the impact of gender (dis)concordance on hearing outcomes, this time conditioning on appellants' behavior within their hearing.

For the first analysis, we conduct four subsample regressions, splitting on combinations of appellant gender and ARB chair gender. When doing so, we focus on the *AnySaved* outcome, for the sake of simplicity. We also focus strictly on cases where Gemini's gender predictions align with our regular-expression-based gender predictions to minimize potential measurement error. We work with ap-

Table 10—Effect of ARB Gender on Home Value Reduction (DV = AnySaved) by (Voice-based, Gemini-detected) Appellant Gender

Appellant Gender	**Female**	**Male**	**Female**	**Male**
Tax Year, Home Value Bucket & Appraiser FEs				
Female Panel Chair	-0.011* (0.006)	-0.003 (0.005)		
Female Panelists (1)			-0.016 (0.010)	-0.002 (0.008)
Female Panelists (2)			-0.020* (0.011)	0.003 (0.007)
Female Panelists (3)			-0.038*** (0.012)	-0.006 (0.008)
Online	-0.042*** (0.012)	-0.014 (0.010)	-0.030** (0.014)	-0.005 (0.010)
Market Equity	0.131*** (0.016)	0.113*** (0.009)	0.133*** (0.016)	0.113*** (0.009)
Unequal Appraisal	0.035*** (0.007)	0.038*** (0.005)	0.037*** (0.007)	0.038*** (0.005)
Count Prior Protests	-0.056*** (0.004)	-0.054*** (0.003)	-0.060*** (0.005)	-0.057*** (0.003)
Saved Last Protest	0.076*** (0.009)	0.074*** (0.006)	0.076*** (0.009)	0.077*** (0.006)
Num.Obs.	24,222	55,975	22,030	52,674
R2	0.07	0.055	0.064	0.051
R2 Adj.	0.056	0.048	0.049	0.044
RMSE	0.46	0.45	0.46	0.45
Tax Year, Account and Appraiser FEs				
Female Panel Chair	-0.028* (0.015)	0.004 (0.008)		
Female Panelists (1)			-0.027 (0.025)	0.009 (0.015)
Female Panelists (2)			-0.040 (0.025)	0.021 (0.014)
Female Panelists (3)			-0.024 (0.030)	0.013 (0.016)
Online	0.090** (0.036)	0.018 (0.020)	0.058 (0.042)	0.030 (0.023)
Market Equity	0.096*** (0.034)	0.042* (0.022)	0.102*** (0.035)	0.044** (0.022)
Unequal Appraisal	0.052** (0.021)	0.040*** (0.010)	0.055** (0.022)	0.040*** (0.010)
Count Prior Protests	-0.016 (0.013)	-0.019*** (0.007)	-0.012 (0.014)	-0.021*** (0.007)
Saved Last Protest	-0.240*** (0.016)	-0.268*** (0.009)	-0.240*** (0.017)	-0.270*** (0.009)
Num.Obs.	24,222	55,975	22,030	52,674
R2	0.91	0.863	0.91	0.862
R2 Adj.	0.247	0.266	0.241	0.257
RMSE	0.14	0.17	0.14	0.17
Mean(Any Saved)	0.647	0.677	0.664	0.687

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB member.

Table 11—Effect of ARB on Home Value Reduction (DV = AnySaved) by (Gemini-detected = Regex-based) Appellant Gender

Appellant Gender	**Female**	**Male**	**Female**	**Male**
		Tax Year, Home Value Bucket & Appraiser FEs		
Female Panel Chair	-0.016** (0.008)	0.000 (0.005)		
Female Panelists (1)			-0.027** (0.012)	0.003 (0.009)
Female Panelists (2)			-0.038*** (0.013)	0.010 (0.008)
Female Panelists (3)			-0.042*** (0.013)	0.002 (0.009)
Online	-0.053*** (0.015)	-0.010 (0.011)	-0.048*** (0.017)	-0.001 (0.012)
Market Equity	0.127*** (0.019)	0.116*** (0.010)	0.129*** (0.019)	0.117*** (0.010)
Unequal Appraisal	0.035*** (0.008)	0.039*** (0.005)	0.037*** (0.008)	0.039*** (0.005)
Count Prior Protests	-0.057*** (0.005)	-0.052*** (0.003)	-0.059*** (0.006)	-0.055*** (0.003)
Saved Last Protest	0.077*** (0.011)	0.068*** (0.006)	0.074*** (0.011)	0.070*** (0.006)
Num.Obs.	17,077	45,347	15,577	42,497
R2	0.079	0.058	0.075	0.054
R2 Adj.	0.060	0.050	0.054	0.045
RMSE	0.46	0.45	0.45	0.45
		Tax Year, Account and Appraiser FEs		
Female Panel Chair	-0.039** (0.019)	0.008 (0.009)		
Female Panelists (1)			-0.043 (0.032)	0.018 (0.016)
Female Panelists (2)			-0.063* (0.032)	0.029* (0.015)
Female Panelists (3)			-0.052 (0.038)	0.026 (0.018)
Online	0.102** (0.049)	0.023 (0.024)	0.104* (0.055)	0.028 (0.027)
Market Equity	0.019 (0.047)	0.043* (0.024)	0.022 (0.048)	0.046* (0.024)
Unequal Appraisal	0.025 (0.027)	0.045*** (0.011)	0.029 (0.028)	0.046*** (0.011)
Count Prior Protests	-0.039** (0.016)	-0.015* (0.008)	-0.028 (0.018)	-0.013 (0.008)
Saved Last Protest	-0.233*** (0.020)	-0.269*** (0.010)	-0.240*** (0.022)	-0.272*** (0.010)
Num.Obs.	17,077	45,347	15,577	42,497
R2	0.923	0.867	0.924	0.866
R2 Adj.	0.226	0.274	0.224	0.264
RMSE	0.13	0.17	0.13	0.17
Mean(Any Saved)	0.650	0.679	0.665	0.689

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB member.

Table 12—Effect of ARB Gender on Home Value Reduction (DV = Log(AmtSaved+1) by (Voice-based, Gemini-detected) Appellant Gender

Appellant Gender	**Female**	**Male**	**Female**	**Male**
Tax Year, Home Value Bucket & Appraiser FEs				
Female Panel Chair	-0.140** (0.063)	-0.037 (0.046)		
Female Panelists (1)			-0.155 (0.099)	-0.021 (0.078)
Female Panelists (2)			-0.233** (0.105)	0.022 (0.077)
Female Panelists (3)			-0.399*** (0.121)	-0.075 (0.081)
Online	-0.438*** (0.117)	-0.263*** (0.097)	-0.325** (0.139)	-0.177* (0.104)
Market Equity	1.484*** (0.158)	1.394*** (0.091)	1.496*** (0.158)	1.398*** (0.092)
Unequal Appraisal	0.221*** (0.065)	0.216*** (0.047)	0.232*** (0.067)	0.214*** (0.047)
Count Prior Protests	-0.558*** (0.041)	-0.544*** (0.030)	-0.595*** (0.045)	-0.573*** (0.030)
Saved Last Protest	0.770*** (0.082)	0.777*** (0.059)	0.758*** (0.086)	0.799*** (0.058)
Num.Obs.	24,222	55,975	22,030	52,674
R2	0.083	0.067	0.078	0.064
R2 Adj.	0.069	0.060	0.063	0.057
RMSE	4.47	4.48	4.43	4.45
Tax Year, Account and Appraiser FEs				
Female Panel Chair	-0.326** (0.151)	-0.028 (0.082)		
Female Panelists (1)			-0.235 (0.238)	0.035 (0.146)
Female Panelists (2)			-0.395 (0.245)	0.110 (0.139)
Female Panelists (3)			-0.279 (0.284)	0.021 (0.165)
Online	0.844** (0.355)	0.155 (0.200)	0.586 (0.410)	0.211 (0.234)
Market Equity	1.100*** (0.331)	0.637*** (0.215)	1.155*** (0.339)	0.665*** (0.219)
Unequal Appraisal	0.487** (0.206)	0.367*** (0.102)	0.495** (0.217)	0.356*** (0.101)
Count Prior Protests	-0.201 (0.127)	-0.236*** (0.069)	-0.156 (0.138)	-0.236*** (0.077)
Saved Last Protest	-2.354*** (0.153)	-2.627*** (0.085)	-2.363*** (0.164)	-2.662*** (0.088)
Num.Obs.	24,222	55,975	22,030	52,674
R2	0.909	0.861	0.909	0.860
R2 Adj.	0.244	0.254	0.235	0.245
RMSE	1.41	1.73	1.39	1.72
Mean(Amount Saved)	$19,336	$23,373	$19,876	$23,648

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB member.

TABLE 13—EFFECT OF ARB GENDER ON HOME VALUE REDUCTION *(DV = LOG(AMTSAVED+1)* BY (GEMINI-DETECTED = REGEX-BASED) APPELLANT GENDER

Appellant Gender	**Female**	**Male**	**Female**	**Male**
		Tax Year, Home Value Bucket & Appraiser FEs		
Female Panel Chair	-0.181** (0.078)	-0.016 (0.052)		
Female Panelists (1)			-0.263** (0.110)	0.017 (0.085)
Female Panelists (2)			-0.388*** (0.121)	0.075 (0.085)
Female Panelists (3)			-0.415*** (0.128)	-0.008 (0.093)
Online	-0.542*** (0.141)	-0.238** (0.109)	-0.490*** (0.169)	-0.159 (0.118)
Market Equity	1.439*** (0.185)	1.459*** (0.099)	1.453*** (0.185)	1.462*** (0.099)
Unequal Appraisal	0.202** (0.080)	0.227*** (0.056)	0.214*** (0.081)	0.223*** (0.056)
Count Prior Protests	-0.566*** (0.051)	-0.531*** (0.033)	-0.591*** (0.055)	-0.557*** (0.034)
Saved Last Protest	0.806*** (0.099)	0.715*** (0.067)	0.767*** (0.103)	0.741*** (0.065)
Num.Obs.	17,077	45,347	15,577	42,497
R2	0.092	0.071	0.089	0.068
R2 Adj.	0.073	0.063	0.068	0.059
RMSE	4.44	4.48	4.41	4.45
		Tax Year, Account and Appraiser FEs		
Female Panel Chair	-0.387** (0.191)	0.013 (0.093)		
Female Panelists (1)			-0.332 (0.306)	0.099 (0.161)
Female Panelists (2)			-0.572* (0.306)	0.163 (0.153)
Female Panelists (3)			-0.484 (0.354)	0.137 (0.183)
Online	0.924* (0.475)	0.209 (0.237)	1.031* (0.535)	0.188 (0.273)
Market Equity	0.261 (0.461)	0.668*** (0.237)	0.287 (0.476)	0.698*** (0.239)
Unequal Appraisal	0.191 (0.271)	0.398*** (0.115)	0.205 (0.278)	0.399*** (0.115)
Count Prior Protests	-0.407*** (0.153)	-0.190** (0.078)	-0.314* (0.168)	-0.166* (0.085)
Saved Last Protest	-2.353*** (0.194)	-2.657*** (0.099)	-2.447*** (0.209)	-2.700*** (0.103)
Num.Obs.	17,077	45,347	15,577	42,497
R2	0.923	0.865	0.924	0.864
R2 Adj.	0.230	0.262	0.226	0.252
RMSE	1.29	1.71	1.27	1.70
Mean(Amount Saved)	$19,788	$24,291	$20,291	$24,540

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors clustered by account and appraiser; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB member.

proximately three-quarters of the audio-recorded sample (roughly 62,000 records across the four sample splits). We report the result of these four estimations below, in Table 14. We observe, in most cases, that appellants' behaviors and tone have little to no impact on hearing outcomes, with a handful of exceptions. Note that we report these regressions in Table 14 to explore correlations; that is, the results do not necessarily represent causal effects. For example, we see that presenting formal evidence has a significant, positive relationship with home value reductions, and this effect is relatively homogeneous across all gender-pair sub-samples, increasing the probability of a value reduction by approximately 6-7 pp. Further, we observe that 'sounding confident' has a positive impact on hearing outcomes specifically when ARB panels are chaired by men. We also observe that 'sounding assertive' has a significant positive impact on hearing outcomes, specifically for male appellants. Importantly, however, the only effect we see that would suggest an explanation for why females fare worse (better) in front of female (male) ARB chairs is that female appellants obtain a significant improvement in their outcomes when they 'sound defensive' in front of a female ARB chair, yet not when they face a male ARB chair. A particularly salient takeaway from these estimates is that, despite our observation that males are systematically more aggressive in their behavior in front of female ARB chairs (raising their voice, interrupting, disagreeing, sounding hostile), this has no material consequence for their hearing outcomes. This suggests that the behavior is tolerated and perhaps even expected.

We now incorporate controls for our Gemini-coded features into our regression specification, focusing on the same outcome measures. As before, we split our sample by appellant gender, incorporating the same set of controls, e.g., measures of whether the hearing is conducted online, the count of prior protest hearings for the property account during our larger sample period, an indicator of whether the most recent hearing, should it exist, resulted in savings, and fixed effects for the home value bucket (or account), the appraiser, and the tax year. Here, we focus on the sample where Gemini's gender predictions agree with the regular-expression-based predictions, the goal being to minimize any role of measurement error. The results of these regressions are reported in Tables 15 and 16.

In all-female homeowner models, we again observe a negative and significant gender concordance effect. The point estimates we observe are roughly equivalent to those observed in our baseline models, reported in Tables 5 and 6, which suggests that appellant behaviors and tones do not mediate the gender-concordance effect.

Finally, we repeat the estimation accounting for high-dimensional embedded representations of the broader, unstructured content of hearing audio recordings, rather than conditioning upon specific Gemini-coded features. We achieve this employing a double machine learning framework (Chernozhukov et al., 2018), an approach that has been taken in other recent literature to account for unstructured data in a relatively non-parametric manner (Manzoor et al., 2024). The

Table 14—Effect of Appellant Behavior by Gender Composition of Appellant and Panel Chair on Home Value Reduction *(DV = AnySaved*

Appellant-Chair Gender	**Female-Female**	**Male-Female**	**Female-Male**	**Male-Male**
App Raises Voice	0.009 (0.031)	-0.007 (0.017)	0.016 (0.029)	0.021 (0.018)
App Disagrees	-0.014 (0.013)	-0.030*** (0.008)	-0.018 (0.016)	-0.021* (0.011)
App Interrupts	-0.009 (0.012)	0.012 (0.008)	-0.006 (0.015)	0.004 (0.009)
App Refers to Evidence	0.062*** (0.010)	0.063*** (0.007)	0.065*** (0.012)	0.069*** (0.008)
App Sounds Confident	0.012 (0.017)	0.009 (0.012)	0.041*** (0.019)	0.026*** (0.012)
App Sounds Frustrated	0.007 (0.018)	-0.014 (0.009)	0.027 (0.019)	-0.012 (0.011)
App Sounds Defensive	0.023* (0.013)	-0.010 (0.009)	-0.003 (0.013)	-0.010 (0.010)
App Sounds Assertive	0.013 (0.024)	0.023* (0.013)	-0.001 (0.028)	0.049*** (0.016)
App Sounds Nervous	0.001 (0.028)	0.015 (0.024)	0.033 (0.026)	0.065*** (0.030)
App Sounds Angry	0.022 (0.043)	0.015 (0.026)	0.017 (0.038)	-0.031 (0.028)
App Sounds Accusatory	-0.048 (0.041)	-0.027 (0.021)	-0.027 (0.041)	0.000 (0.028)
App Asks Questions	-0.011 (0.012)	0.010 (0.008)	0.003 (0.013)	0.008 (0.008)
App Asks Repeat	-0.004 (0.012)	0.003 (0.007)	-0.005 (0.013)	-0.011 (0.009)
App Asks Speak Louder	-0.003 (0.023)	0.025* (0.013)	-0.017 (0.024)	-0.007 (0.017)
Num. Obs.	9,278	24,752	7,799	20,595
R2	0.089	0.063	0.090	0.060
R2 Adj.	0.056	0.049	0.049	0.042
RMSE	0.46	0.45	0.45	0.45
Mean(Any Saved)	0.642	0.678	0.659	0.680

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by account and appraiser; Gender compositions correspond to appellant gender (where Gemini coding agreed with regular-expression-based prediction) followed by panel chair gender (e.g., Female-Female indicates both appellant and panel chair are female).

Table 15—Effect of ARB on Home Value Reduction *(DV = AnySaved)* by (Gemini-detected = Regex-based) Appellant Gender

Appellant Gender	**Female**	**Male**	**Female**	**Male**
Female Panel Chair	-0.015* (0.008)	0.000 (0.005)		
Female Panelists (1)			-0.026** (0.012)	0.002 (0.008)
Female Panelists (2)			-0.036*** (0.013)	0.010 (0.008)
Female Panelists (3)			-0.039*** (0.013)	0.003 (0.009)
Market Equity	0.126*** (0.019)	0.118*** (0.010)	0.128*** (0.019)	0.118*** (0.010)
Unequal Appraisal	0.034*** (0.008)	0.038*** (0.005)	0.036*** (0.008)	0.038*** (0.005)
Online	-0.046*** (0.015)	-0.008 (0.011)	-0.041** (0.017)	0.000 (0.012)
Count Prior Protests	-0.059*** (0.005)	-0.053*** (0.003)	-0.061*** (0.006)	-0.056*** (0.003)
Saved Last Protest	0.073*** (0.011)	0.066*** (0.007)	0.070*** (0.011)	0.069*** (0.006)
Gemini-coded Audio Features				
App Raises Voice	0.004 (0.020)	0.007 (0.013)	0.013 (0.021)	0.008 (0.013)
App Disagrees	-0.015 (0.010)	-0.027*** (0.007)	-0.013 (0.011)	-0.023*** (0.007)
App Interrupts	-0.009 (0.009)	0.008 (0.006)	-0.008 (0.010)	0.008 (0.007)
App Refers to Evidence	0.067*** (0.008)	0.069*** (0.006)	0.062*** (0.008)	0.064*** (0.006)
App Sounds Confident	0.027** (0.012)	0.016** (0.008)	0.020 (0.013)	0.011 (0.008)
App Sounds Frustrated	0.022* (0.013)	-0.009 (0.007)	0.024* (0.013)	-0.006 (0.007)
App Sounds Defensive	0.009 (0.009)	-0.008 (0.007)	0.005 (0.009)	-0.012* (0.007)
App Sounds Assertive	-0.001 (0.017)	0.036*** (0.009)	-0.003 (0.018)	0.036*** (0.010)
App Sounds Nervous	0.006 (0.019)	0.027 (0.019)	-0.005 (0.019)	0.029 (0.019)
App Sounds Angry	0.025 (0.029)	-0.002 (0.019)	0.022 (0.029)	-0.006 (0.019)
App Sounds Accusatory	-0.034 (0.030)	-0.018 (0.017)	-0.038 (0.030)	-0.014 (0.017)
App Asks Questions	-0.004 (0.009)	0.008 (0.006)	0.006 (0.009)	0.011* (0.006)
App Asks Repeat	-0.004 (0.009)	-0.006 (0.006)	-0.004 (0.009)	-0.005 (0.006)
App Asks Speak Louder	-0.004 (0.016)	0.012 (0.011)	-0.007 (0.017)	0.011 (0.011)
Num. Obs.	17,077	45,347	15,577	42,497
R2	0.085	0.065	0.080	0.059
R2 Adj.	0.065	0.056	0.058	0.050
RMSE	0.46	0.45	0.45	0.45
Mean(Any Saved)	0.650	0.679	0.665	0.689

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by account and appraiser; Panelist genders are genders derived from names on hearing records; Appellant genders reflect genders derived from names on hearing records and inferred by Google Gemini on hearing audio; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB panelist; Estimations incorporate home value bucket FEs.

Table 16—Effect of ARB on Home Value Reduction (*DV = Log(AmtSaved+1)* by (Gemini-detected = Regex-based) Appellant Gender

Appellant Gender	**Female**	**Male**	**Female**	**Male**
Female Panel Chair	-0.168** (0.076)	-0.018 (0.051)		
Female Panelists (1)			-0.250** (0.110)	0.013 (0.083)
Female Panelists (2)			-0.368*** (0.120)	0.075 (0.084)
Female Panelists (3)			-0.388*** (0.127)	-0.005 (0.092)
Market Equity	1.427*** (0.183)	1.482*** (0.101)	1.448*** (0.183)	1.486*** (0.101)
Unequal Appraisal	0.195** (0.080)	0.224*** (0.056)	0.213*** (0.080)	0.222*** (0.056)
Online	-0.443*** (0.140)	-0.202* (0.109)	-0.396** (0.166)	-0.131 (0.117)
Count Prior Protests	-0.587*** (0.052)	-0.543*** (0.033)	-0.609*** (0.056)	-0.567*** (0.034)
Saved Last Protest	0.752*** (0.100)	0.693*** (0.067)	0.711*** (0.103)	0.715*** (0.065)
Gemini-coded Audio Features				
App Raises Voice	0.062 (0.196)	0.069 (0.129)	0.149 (0.203)	0.067 (0.129)
App Disagrees	-0.165* (0.099)	-0.318*** (0.068)	-0.142 (0.107)	-0.288*** (0.074)
App Interrupts	-0.130 (0.093)	0.057 (0.062)	-0.124 (0.095)	0.052 (0.064)
App Refers to Evidence	0.856*** (0.074)	0.856*** (0.054)	0.813*** (0.078)	0.812*** (0.053)
App Sounds Confident	0.252** (0.116)	0.195** (0.077)	0.183 (0.126)	0.146* (0.078)
App Sounds Frustrated	0.329*** (0.123)	0.025 (0.070)	0.347*** (0.126)	0.056 (0.073)
App Sounds Defensive	0.108 (0.088)	-0.067 (0.066)	0.073 (0.091)	-0.110 (0.067)
App Sounds Assertive	0.019 (0.167)	0.368*** (0.091)	0.011 (0.175)	0.369*** (0.097)
App Sounds Nervous	0.143 (0.190)	0.259 (0.184)	0.011 (0.190)	0.256 (0.184)
App Sounds Angry	0.215 (0.286)	0.063 (0.190)	0.194 (0.286)	0.045 (0.193)
App Sounds Accusatory	-0.344 (0.297)	-0.270 (0.171)	-0.387 (0.292)	-0.239 (0.173)
App Asks Questions	-0.011 (0.086)	0.155** (0.062)	0.076 (0.091)	0.188*** (0.064)
App Asks Repeat	0.007 (0.087)	-0.026 (0.055)	0.006 (0.092)	-0.011 (0.056)
App Asks Speak Louder	-0.068 (0.160)	0.084 (0.102)	-0.102 (0.168)	0.070 (0.107)
Num. Obs.	17,077	45,347	15,577	42,497
R2	0.102	0.080	0.097	0.077
R2 Adj.	0.082	0.072	0.076	0.068
RMSE	4.42	4.45	4.39	4.43
Mean(Amount Saved)	$19,788	$24,291	$20,291	$24,540

Notes: *** p<0.01, ** p<0.05, * p<0.1; Robust standard errors in parentheses clustered by account and appraiser; Panelist genders are genders derived from names on hearing records; Appellant genders reflect genders derived from names on hearing records and inferred by Google Gemini on hearing audio; Observation counts decline in columns 3 and 4 due to the omission of hearings involving only one ARB panelist; Estimations incorporate home value bucket FEs.

double machine learning framework is useful here because it simultaneously enables us to accommodate many covariates, i.e., our embedding dimensions and their likely non-linear interactive relationships with appellant gender, panelist gender, and hearing outcomes. Our implementation models the relationships between our nuisance parameters (structured covariates and text-embedding dimensions), the outcome, and the ARB panelist gender via a pair of random forest models. The text-embedding for each hearing audio recording was obtained by first passing the recording through Open AI's Whisper Medium speech-to-text model. Because Whisper model inference is relatively slow (compared to inference with Gemini Flash, at least), we conduct this exercise focusing strictly on hearing audio records from 2022. The textual transcription for each hearing recording was then passed through Google's text-embedding-004 model, yielding a high-dimensional vector representation of the spoken content. Finally, each embedding dimension is treated as a distinct covariate in our estimation, which we implement employing the *DoubleML* package in *R*, specifically its implementation of a partial linear specification. We focus on the linear probability model that considers the *Any Saved* outcome. The results of these estimations are presented in Table 17, where we again observe consistent results.

Table 17—Double Machine Learning Estimation of the Effect of ARB Chair Gender on Any Home Value Reduction, by Appellant Gender, Conditional on Hearing Transcript Embedding

Appellant Gender	**Female**	**Male**
Female Panel Chair	-0.027* (0.016)	-0.002 (0.012)
Num.Obs.	3,731	7,266
Mean(Any Saved)	0.510	0.518

Notes: *** p<0.01, ** p<0.05, * p<0.1; Appellant and panelist genders are derived from names on hearing records; Estimations account for home value bucket FEs, the same set of controls and hearing transcript embeddings; Estimation performed on hearings conducted for tax year 2022.

That the gender concordance effects persist for female homeowners across both approaches, conditioning on hearing activity, suggests that, although appellants' (or panelists') behavior may contribute to differential hearing outcomes, other unmeasured (unvoiced) factors continue to play a significant role.

VII. Discussion

Our study provides new insights into the role of gender concordance in property tax protest hearings, revealing significant biases that disproportionately disadvantage female appellants. We find that female homeowners are systematically less likely to secure reductions in their assessed property values, particularly when they appear before female-dominated appraisal review board (ARB) panels.

These gender concordance effects are substantial, with female appellants facing a notably lower probability of success when confronted with female panelists than male panelists.

We further explore the behavioral dynamics underlying these disparities by leveraging unstructured audio data processed through a multi-modal large language model (Gemini 1.5 Flash). Although we find that female homeowners exhibit differences in behavior and tone depending on the gender composition of the ARB panel, these differences are generally modest. Further, controlling for appellant behavior or hearing activity more generally, we continue to find that female appellants are systematically less likely to achieve favorable outcomes when faced by female (versus male) panelists, suggesting that the biases we document are likely rooted in unobserved or implicit factors influencing ARB decision-making.

Our findings contribute to the growing literature on gender biases in evaluation processes by highlighting a previously undocumented source of bias in the context of property taxation. Our results imply that gender biases in such fundamental administrative processes can exacerbate existing inequalities, with tangible financial consequences for affected individuals. This is particularly concerning given the role of property taxes in funding essential public services, meaning that gender-based disparities in tax burdens can have broad societal implications.

Moreover, our study underscores the importance of considering gender dynamics in seemingly neutral administrative settings and calls for policymakers to reassess the structure and procedures of ARB hearings. Structural adjustments, such as diversifying panel compositions or introducing bias-mitigation training, may help reduce the influence of implicit biases.

Finally, our use of large-scale administrative data combined with unstructured audio analysis demonstrates the potential for advanced AI tools to enhance our understanding of complex social dynamics. The ability to systematically analyze behavioral and tonal cues at scale opens new avenues for research into decision-making processes, especially in contexts where bias may be subtle or difficult to detect using traditional methods.

Our work is, of course, subject to several limitations. First, we focus on a single geography, Harris, County, Texas. As such, generalizability is a question. Houston may have a unique sociocultural climate, and the local market may be subject to idiosyncrasies in terms of norms and laws, as well as its property appraisal appeals process. That said we have observed broad evidence of gender-concordant bias across the entire pool of ARB panelists (see Figure 1), suggesting this phenomenon may be more general, cutting across political and social groups. We can evaluate the generalizability of our results using the SANS framework (selection, attrition, naturalness, and scalability) developed in (List, 2020). Regarding selection, our sample is representative of hearings in Harris County, the target population. However, we note potential selection bias, as females may be less likely to participate in hearings due to anticipated biases against them. Notably, in our setting, participants do not know the gender of the panel members before

the hearing. Given the similarity of hearings across Texas's 241 counties, our sample may also be representative of hearings statewide. Considering attrition, experiences, and outcomes in prior hearings could influence individuals' willingness to participate in future hearings, potentially affecting the composition of future samples. Regarding naturalness, both homeowners and panelists operate within a natural margin, engaging in real decisions rather than an artificial task designed for research purposes. Regarding scalability, there are no clear reasons to anticipate a voltage drop if the analysis is scaled to other counties in Texas. However, once our findings on biases in this context become public, behavior may change. For instance, the Texas state comptroller, who oversees hearings in all counties in the state, may share our findings with county ARBs, and appellants might adjust their behavior in future hearings. In sum, we view our findings as a WAVE1 insight in the nomenclature of (List, 2020). Replications are necessary to determine whether our results generalize to other administrative processes, and future work can and should explore these dynamics by leveraging administrative data from other areas of the United States.

Second, our analysis relies on the accuracy of gender imputations based on first names, which, although generally reliable, may introduce some degree of misclassification. This potential misclassification could slightly influence our estimates of gender effects, though we expect such biases to be minimal given the robustness of our results. Notably, employing gender codings from Gemini's audio annotations rather than appellant names yields consistent results, suggesting that any measurement error here does not explain our findings. Nonetheless, future research could improve upon this by using direct measures of gender where available.

Third, while we leverage a state-of-the-art multi-modal large language model (Gemini 1.5 Flash) to analyze audio recordings, the model's inferences about tone and behavior are not without error. Although we validate some of these inferences against known ground truths, there remains the possibility that certain nuances of human interaction—especially those influenced by socio-cultural context—are missed or misinterpreted by the model. This limitation suggests that human coding, or a hybrid approach combining machine learning with human validation may be a valuable avenue for future research.

Fourth, while we control for various structured and unstructured factors, there may still be unobserved variables influencing both the behavior of appellants and the decisions of ARB panelists. Our analysis of behavior and tone during hearings focuses primarily on what can be quantified, which may overlook subtler aspects of interpersonal dynamics that could influence outcomes.

Overall, these limitations highlight the need for further research. Despite these limitations, our findings provide compelling evidence of gender (dis)concordance effects in property tax protest procedures, with significant implications for policy and practice. Further, our work highlights the significant potential of multimodal LLMs for processing information from large-scale unstructured administrative data, to better understand procedural processes and outcomes.

VIII. Conclusion

Our study sheds light on a previously under-explored area of gender bias within the context of property tax protest hearings. We document significant gender-concordance effects, where female homeowners face systematic disadvantages in achieving favorable outcomes, particularly when their cases are adjudicated by female ARB panelists. These findings suggest that implicit biases, rather than observable behaviors, play a crucial role in driving these disparities. The implications of our study extend beyond the specific context of property taxation, highlighting the broader risks of gender bias in administrative processes that involve subjective evaluations.

Our results underscore the importance of considering gender dynamics in the design and oversight of evaluative processes, particularly those that impact individuals' financial well-being. As policymakers and practitioners work to enhance fairness and equity in tax systems and other administrative domains, our findings point to the need for interventions that can mitigate gender-based biases. Future research should continue to investigate these dynamics across different regions and contexts, leveraging diverse methodologies to better understand and address the underlying causes of such biases. By doing so, we can move closer to ensuring that all individuals, regardless of gender, are treated equitably in critical administrative proceedings.

Beyond these contributions, our results could also affect future behavior by appellants and panelists, reducing biases, in which case they would represent a public service. Indeed, when individuals know they are being recorded, their behavior often changes due to the observer effect, where heightened scrutiny influences actions. Teachers, for instance, tend to modify their teaching practices during evaluations, focusing more on structured lessons and student engagement to meet professional standards (Kane et al., 2013). Similarly, judges may exhibit greater formality in court proceedings to align with public and appellate expectations, especially when media coverage is present (Lim, Snyder and Strömberg, 2015). Legislators similarly adjust their rhetoric during televised sessions, emphasizing performative or emotional appeals to resonate with constituents (Osnabrugge, Hobolt and Rodon, 2021). This shift toward transparency and visibility creates both opportunities for accountability and challenges in maintaining authentic behavior. Research consistently shows that such surveillance amplifies the emphasis on appearance and messaging, often at the expense of substance (Ash and MacLeod, 2017; Boussalis et al., 2021). Finally, our paper illustrates how advancements in AI can enable the analysis of recordings and data previously deemed unsuitable for examination, uncovering insights akin to how DNA revolutionized cold case investigations and helped to solve decades-old mysteries.

References

Ash, Elliot, and W. Bentley MacLeod. 2017. "Transparency and Emotionality in Political Speech." *American Economic Journal: Microeconomics*, 9(1): 1–30.

Austin, Peter C. 2009. "Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples." *Statistics in medicine*, 28(25): 3083–3107.

Avenancio-León, Carlos F, and Troup Howard. 2022. "The assessment gap: Racial inequalities in property taxation." *The Quarterly Journal of Economics*, 137(3): 1383–1434.

Bordalo, Pedro, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2016. "Stereotypes." *The Quarterly Journal of Economics*, 131(4): 1753–1794.

Boussalis, Constantine, Travis G. Coan, Mirya R. Holman, and Stefan MÜller. 2021. "Gender, Candidate Emotional Expression, and Voter Reactions During Televised Debates." *American Political Science Review*, 115(4): 1242–1257.

Bowles, Hannah Riley, Linda Babcock, and Lei Lai. 2007. "Social incentives for gender differences in the propensity to initiate negotiations: Sometimes it does hurt to ask." *Organizational Behavior and human decision Processes*, 103(1): 84–103.

Boyd, Christina L, and Michael J Nelson. 2017. "The effects of trial judge gender and public opinion on criminal sentencing decisions." *Vand. L. Rev.*, 70: 1819.

Brock, J Michelle, and Ralph De Haas. 2023. "Discriminatory lending: Evidence from bankers in the lab." *American Economic Journal: Applied Economics*, 15(2): 31–68.

Cabral, Marika, and Marcus Dillender. 2024. "Gender Differences in Medical Evaluations: Evidence from Randomly Assigned Doctors." *American Economic Review*, 114(2): 462–499.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters."

Coffman, Katherine Baldiga. 2014. "Evidence on self-stereotyping and the contribution of ideas." *The Quarterly Journal of Economics*, 129(4): 1625–1660.

Coffman, Katherine, Christine Exley, and Muriel Niederle. 2021. "The role of beliefs in driving gender discrimination." *Management Science*, 67(6): 3551–3569.

Cullen, Zoë, and Ricardo Perez-Truglia. 2023. "The old boys' club: Schmoozing and the gender gap." *American Economic Review*, 113(7): 1703–1740.

Ely, Robin J. 1994. "The effects of organizational demographics and social identity on relationships among professional women." *Administrative science quarterly*, 203–238.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica*, 78(1): 35–71.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature*, 57(3): 535,Äì74.

Giaccobasso, Matias, Brad C Nathan, Ricardo Perez-Truglia, and Alejandro Zentner. 2022. "Where do my tax dollars go? tax morale effects of perceived government spending." National Bureau of Economic Research.

Good, Phillip. 2013. *Permutation tests: a practical guide to resampling methods for testing hypotheses.* Springer Science & Business Media.

Greenberg, Jason, and Ethan Mollick. 2017. "Activist choice homophily and the crowdfunding of female founders." *Administrative Science Quarterly*, 62(2): 341–374.

Holz, Justin, David Novgorodsky, and Andrew Simon. 2023. "Racial Inequality in Property Tax Appeals: Evidence from Field Experiments with Homeowners and Assessors." *Working Paper*.

Kane, Thomas J., Douglas O. Staiger, Daniel F. McCaffrey, and Thomas Miller. 2013. "Have we identified effective teachers? Validating measures of effective teaching using random assignment." *MET Project Research Paper*.

Kaplan, J. 2021. "Predictrace: Predict the race and gender of a given name using census and social security administration data." *R package version*, 2(0).

Knepper, Matthew. 2018. "When the shadow is the substance: Judge gender and the outcomes of workplace sex discrimination cases." *Journal of Labor Economics*, 36(3): 623–664.

Lau, Emily S, Sharonne N Hayes, Annabelle Santos Volgman, Kathryn Lindley, Carl J Pepine, Malissa J Wood, and American College of Cardiology Cardiovascular Disease in Women Section. 2021. "Does

patient-physician gender concordance influence patient perceptions or outcomes?" *Journal of the American College of Cardiology*, 77(8): 1135–1138.

Lim, Claire S. H., Jr. Snyder, James M., and David Strömberg. 2015. "The Judge, the Politician, and the Press: Newspaper Coverage and Criminal Sentencing across Electoral Systems." *American Economic Journal: Applied Economics*, 7(4): 103–35.

List, John A. 2020. "Non est Disputandum de Generalizability? A Glimpse into The External Validity Trial." *NBER Working Paper No. 27535*.

Manzoor, Emaad, George H Chen, Dokyun Lee, and Michael D Smith. 2024. "Influence via ethos: On the persuasive power of reputation in deliberation online." *Management Science*, 70(3): 1613–1634.

Nathan, Brad C, Ricardo Perez-Truglia, and Alejandro Zentner. 2025. "My Taxes are Too Darn High: Why Do Households Protest their Taxes?" *American Economic Journal: Economic Policy. Forthcoming*.

Nathan, Brad, Ricardo Perez-Truglia, and Alejandro Zentner. 2022. "Is the Partisan Divide Real? Polarization in Preferences for Redistribution." *AEA Papers and Proceedings*, 112: 156–62.

Nathan, Brad, Ricardo Perez-Truglia, and Alejandro Zentner. 2024. "Paying Your Fair Share: Perceived Fairness and Tax Compliance." *NBER. Working Paper 32588*.

Osnabrugge, Moritz, Sara B. Hobolt, and Toni Rodon. 2021. "Playing to the Gallery: Emotive Rhetoric in Parliaments." *American Political Science Review*, 115(3): 885–899.

Recalde, Maria P, and Lise Vesterlund. 2023. "Gender differences in negotiation: can interventions reduce the gap?" *Annual Review of Economics*, 15(1): 633–657.

Rosenbaum, Paul R, and Donald B Rubin. 1985. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician*, 39(1): 33–38.

Schanke, Scott, Gordon Burtch, and Gautam Ray. 2024. "Digital Lyrebirds: Experimental Evidence That Voice-Based Deep Fakes Influence Trust." *Management Science*.

Stantcheva, Stefanie. 2020. "Understanding Economic Policies: What do People Know and How Can they Learn?" *Harvard University Working Paper*.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics*, 15(Volume 15, 2023): 205–234.

Tajfel, H., and J. C. Turner. 1979. "An integrative theory of intergroup conflict *In:* 'The Social Psychology of Intergroup Relations'." , ed. Mary Jo Hatch and Majken Schultz, 56–65. Monterey, CA: Brooks/Cole.

Takeshita, Junko, Shiyu Wang, Alison W Loren, Nandita Mitra, Justine Shults, Daniel B Shin, and Deirdre L Sawinski. 2020. "Association of racial/ethnic and gender concordance between patients and physicians with patient experience ratings." *JAMA network open*, 3(11): e2024583–e2024583.

Tzioumis, Konstantinos. 2018. "Demographic aspects of first names." *Scientific data*, 5(1): 1–9.

APPENDIX A: Audio Annotation with Google Gemini 1.5 Flash

For each audio recording, we obtain annotations using Google's 'genai' API (Application Programming Interface), using Python. All API requests employed 'gemini-1.5-flash-001' and requested a JSON formatted response. Audio annotations for hearings from 2022 were obtained in late August of 2024. Additional hearing records were obtained for 2013 through 2021 in early October of 2024 and were coded in the latter half of that month. For every recording, a zero-shot prompt was employed, using the default temperature setting (a value of 1.0). The exact prompt for the coding exercise was as follows.

```
The following audio file is a recording of an appraisal review board hearing
in Houston. Listen carefully and answer the following questions. Be sure to
use exactly one of the provided response options for each question, when
indicated, and never respond with 'null'. Please analyze the following audio
file and respond with the answers to each question in the specified
JSON format. Ensure that the JSON keys correspond exactly to the questions
asked. The JSON structure should look like this:

{
  "appellant_attending_online": "ANSWER YES or NO",
  "board_member_who_speaks_most_gender": "ANSWER MALE or FEMALE",
  "count_people_present_for_property_owner": "ANSWER WITH A NUMBER, e.g., 2",
  "appellant_asks_board_member_or_appraiser_to_repeat": "ANSWER YES or NO",
  "appellant_asks_board_member_or_appraiser_to_speak_louder":
"ANSWER YES or NO",
  "appellant_ever_sounds_accusatory_or_hostile": "ANSWER YES or NO",
  "appellant_sounds_defensive": "ANSWER YES or NO",
  "appellant_sounds_assertive": "ANSWER YES or NO",
  "appellant_asks_board_member_or_appraiser_any_questions":
  "ANSWER YES or NO",
  "appellant_presents_formal_evidence": "ANSWER YES or NO",
  "appellant_type_of_evidence_if_any":
"ANSWER WITH FREE FORM TEXT OR NOT APPLICABLE",
  "appellant_property_owner_or_hired_agent":
"ANSWER WITH OWNER or AGENT",
  "appellant_who_speaks_most_gender": "ANSWER WITH MALE or FEMALE",
  "appraiser_gender": "ANSWER WITH MALE or FEMALE",
  "appellant_time_speaking_during_hearing":
  "ANSWER WITH A PERCENTAGE VALUE, e.g., 50%",
  "appellant_ever_raises_voice": "ANSWER WITH YES or NO",
  "appellant_sounds_frustrated": "ANSWER WITH YES or NO",
  "appellant_sounds_annoyed": "ANSWER WITH YES or NO",
```

```
  "appellant_sounds_nervous": "ANSWER WITH YES or NO",
  "appellant_sounds_confident": "ANSWER WITH YES or NO",
  "appellant_sounds_angry": "ANSWER WITH YES or NO",
  "appellant_ever_disagrees_with_board_member_or_appraiser":
"ANSWER WITH YES or NO",
  "appellant_ever_interrupts_board_member_or_appraiser":
"ANSWER WITH YES or NO",
  "board_member_or_appraiser_ask_appellant_any_questions":
"ANSWER WITH YES or NO",
  "board_member_or_appraiser_ever_asks_appellant_to_repeat":
  "ANSWER WITH YES or NO",
  "board_member_or_appraiser_ever_ask_appellant_to_speak_louder":
  "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_accusatory_or_hostile":
  "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_defensive": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_confident": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_nervous": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_ever_raise_voice": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_frustrated": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_annoyed": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_sound_angry": "ANSWER WITH YES or NO",
  "board_member_or_appraiser_ever_disagree_with_appellant":
  "ANSWER WITH YES or NO",
  "board_member_or_appraiser_ever_interrupt_appellant":
  "ANSWER WITH YES or NO"
}
```

The resulting JSON responses were then parsed into numeric measures. Although the Gemini API allows users to configure requests to produce JSON formatted responses, the formatting of responses is not guaranteed with Flash (it *is* guaranteed with Gemini 1.5 Pro, but inference with Gemini Pro is much slower, and an equivalent length request/response costs approximately 20x as much to obtain as from Gemini Flash. Some audio recordings, when annotated, therefore, resulted in malformed responses and were excluded from our sample.

Appendix: Explanation of Tones and Behaviors Coded in Audio Recordings

The following descriptions define the tones and behaviors coded from audio recordings of property tax protest hearings in Harris County, Texas. These features represent key elements of appellant communication and interaction during hearings. The coding was performed using Gemini 1.5 Flash and reflects distinct and interpretable aspects of speech and behavior that are generally comprehensive in the context of such hearings.

- **Confidence**: Confidence is characterized by a steady, clear, and self-assured vocal delivery, often marked by a lack of hesitation and a consistent tone. Projecting confidence can instill a sense of credibility and competence in listeners.
- **Assertiveness**: Assertiveness involves a direct and purposeful communication style that is neither passive nor aggressive. It reflects attempts to control the flow of conversation and discourse while respecting others' perspectives.
- **Hostility**: Hostility is identified through a sharp, cutting tone, often paired with elevated volume, sarcasm, or aggression. It reflects underlying anger or irritation and may alienate listeners.
- **Frustration**: Frustration is conveyed through strained or uneven vocal tones, frequent sighs, or a quicker-than-usual pace, signaling dissatisfaction or impatience with the proceedings.
- **Annoyance**: Annoyance is a less intense and more transient form of frustration, often marked by clipped or curt vocal delivery, with signs of irritation but not outright aggression.
- **Defensiveness**: Defensiveness is indicated by a protective or guarded tone, characterized by justifications or explanations in response to perceived criticism.
- **Nervousness**: Nervousness is detected through hesitations, voice tremors, pitch variability, and speech errors, reflecting discomfort or lack of confidence.
- **Disagrees**: This behavior captures instances where one party verbally expresses disagreement with another party's statements, either explicitly or implicitly. Disagreement is important as it can shape the dynamics of argumentative discourse.
- **Presents Evidence**: This behavior indicates instances where the appellant references, describes, or provides documentation, photos, spreadsheets, etc., in support of their argument. Presenting evidence is a critical behavior in structured hearings, as it demonstrates preparation and engagement.

- **Raises Voice**: This behavior captures instances of increased vocal volume, which may signal emphasis, frustration, or attempts to assert control in the interaction.
- **Asks Questions**: This includes any inquiries from one party, e.g., the appellant, to another, e.g., the panel, reflecting an attempt to seek clarification or challenge the hearing process. Question-asking is a critical component of participatory dialogue.
- **Interrupts**: This behavior reflects overlaps in speech where one party interjects before another has finished speaking. Interruptions may indicate urgency, frustration, or a desire to assert control.
- **Asks to Speak Louder**: This behavior reflects difficulties in hearing or understanding another party, prompting requests for increased vocal volume.
- **Asks to Repeat**: This behavior reflects instances where one party requests another to restate their points, e.g., due to inaudibility or confusion.

These coded tones and behaviors were selected for their relevance to hearing dynamics and their grounding in existing communication and behavioral research. They aim to comprehensively capture the range of appellant interaction styles, contributing to the analysis of procedural outcomes and dynamics.