STOCKHOLM SCHOOL OF ECONOMICS
Department of Economics
5350 Master's Thesis in Economics
Academic Year 2020-2021

Biased While Betting on Bernie?

A cross-sectional and panel data analysis of prediction error in U.S. election prediction markets from 2015 to 2020

Daniel Evans

December 2020

Abstract

A consensus seems to have emerged that political prediction markets can lose predictive power when certain efficiency criteria are not met. With a cross-sectional dataset of 570 prediction markets about U.S. elections and a panel dataset with 6,465 days of trading from PredictIt.org, I use OLS and correlated random effects models to test whether systematic prediction error is measurable under conditions of questionable market efficiency. In particular, I investigate whether candidates who share traders' ideology and elections that see high levels of voter enthusiasm are associated with higher prediction error due to wishful thinking bias. I also explore whether female and ethnic minority candidates are associated with increased prediction error due to misperceptions about their electability. Finally, I test hypotheses about how prediction error evolves over time and with changes in Google search volume. I do not find strong evidence that any of ideology, enthusiasm, gender, and ethnicity are associated with increased prediction error in these markets. The hypothesis that predictions made further away from election day see more prediction error over a short timespan is strongly supported, but I find no evidence that the duration of trading over a longer timespan or changes in Google search volume matter.

Keywords: prediction markets, political science, behavioral economics, U.S. elections, bias

Supervisor: Magnus Johannesson
Date submitted: December 7, 2020
Date examined: December 16, 2020
Discussant: Niclas Hvalgren
Examiner: Elena Paltseva

Acknowledgements

This thesis would not be complete without expressing thanks to the many people who have offered me their guidance, wisdom, and support through the years, although I can never hope to properly repay the debt of gratitude that they are owed. I would like to thank my supervisor, Magnus Johannesson, for his valuable insight and continuous engagement with the project. Special thanks are owed to my parents, John and Patricia Evans, for their unconditional love and encouragement, and for fostering the senses of determination and intellectual curiosity that continue to motivate my studies to this day. I would also like to thank my sister Gabrielle Evans, her husband, and their young daughters, for their love and for the important reminder that some things in life are more important than work. Finally, although all of my friends are deserving of recognition for the role they play in sustaining my happiness and well-being, I would especially like to thank my classmates Gediminas Goda and Stasia Rudak for their friendship and for the continuous moral support and feedback on my thesis they have offered me throughout this process.

Contents

1. Introduction

Before the United States presidential election on November 3rd, 2020, it was common to hear expressions of uncertainty, anxiety, and suspense about the potential outcome from interested observers around the world. The implications of this election for global peace and prosperity cannot be understated. World leaders of allied and rival states alike were eagerly waiting to know whether U.S. foreign policy would continue to be guided by the untraditional approach of the current occupant of the White House or by a return to the multilateral approach of previous administrations. Large corporations and non-governmental organizations were anticipating that the election results would be a signal about likely changes to regulatory, trade, and fiscal policy in the coming years. With stakes like these associated with U.S. elections, it is unsurprising that many tools have been developed to assist in predicting their outcomes. Among others, these tools include simple opinion polls, advanced forecasting models that synthesize economic, demographic, and polling data into probabilistic forecasts, and prediction markets, the latter of which is the subject of this thesis.

To study these markets, I obtained a cross-sectional dataset of predictions made for 570 different U.S. elections and a panel dataset of 6,465 days of trading from the betting website PredictIt.org. Recent findings have suggested that PredictIt's markets might not meet the criteria to produce efficient market outcomes, which calls into question their informativeness but also creates a novel research opportunity. To give a relevant example of what PredictIt's markets look like, the market "Who will win the 2020 U.S. presidential election?" allowed people to bet money on who they expected the next president to be, with the top two predicted outcomes as of October 27, 2020 being Joe Biden and Donald Trump at about 60% and 40%, respectively (Who will win, 2020). While my dataset does not contain this specific prediction market, as the outcome was not yet known at the time of data collection, it includes similar observations from hundreds of other markets about presidential, senatorial, congressional, statewide, and municipal elections from November 2015 to March 2020. I intend to use this dataset to contribute to the current literature's understanding of the informativeness of prediction markets as a tool for forecasting U.S. election outcomes when market efficiency is in doubt. In particular, I offer evidence from OLS and correlated random effects regressions about whether prediction error in these markets is systematically associated with 1. inherent characteristics of the candidates and elections that correspond to biases about ideology, gender, ethnicity, and voter enthusiasm (judgment bias-based hypotheses), and 2. changes in the availability of information over time, as measured by the duration of trading and fluctuations in Google search volume relating to election-specific keywords (information flow-based hypotheses). The purpose of my identification strategy for the judgment bias-based hypotheses is to isolate the impact of the explanatory variables that is specifically attributable to trader bias, as opposed to other channels or confounding variables that could be associated with inherent characteristics. For each hypothesis, I specify whether I expect the corresponding explanatory variable to have a positive or negative association with prediction error based on existing research from the behavioral economics and political science literatures. In practical

terms, the evidence gathered by this thesis could help in the interpretation and contextualization of forecasts from existing prediction markets: if the results support systematic bias, traders could use the findings to increase their expected profits; otherwise, forecast users could benefit from knowing that the markets are not likely to be biased due to the presence of certain candidate characteristics.

The idea that election forecasting methods need refinement has become a platitude of conventional political wisdom. Although society's collective shock at the outcome of the 2016 U.S. presidential election seems to be the proximate cause of contemporary skepticism about election forecasting, it is not immediately obvious whether the surprise outcome was a failure of the forecasting tools themselves or how they were used and interpreted. For example, the FiveThirtyEight polls-plus forecasting model, which is one of the advanced models mentioned above, gave Hillary Clinton and Donald Trump 71.8% and 28.2% chances of winning, respectively (Silver, 2016). A simple deterministic evaluation of this prediction would conclude that the forecast was wrong since it assigned a higher probability to the outcome that did not occur than the one that did. But the problem with this analysis is that FiveThirtyEight's allocation of probabilities is not implausible on its face: with a sample size of one election, it is not especially surprising that an event with a probability of 28.2% occurred. A more rigorous probabilistic evaluation would require one to determine whether 28.2% was Trump's true chance of victory, given the underlying conditions of the election. Unfortunately, it is impossible to properly verify or falsify this claim, since the election would have to be re-run hundreds of times under the same conditions to see whether Trump actually wins 28.2% of the time. Short of that, one can look at a forecaster's predictions about hundreds of elections to determine whether the prediction method is reliable in general: in Bayesian terms, a forecaster is "well calibrated if, for example, of those events to which he assigns a probability 30 percent, the long-run proportion that actually occurs turns out to be 30 percent" (Dawid, 1982). While being well-calibrated intuitively seems like an useful feature of a forecast, it is no guarantee that the forecast is informative: a forecaster could guarantee perfect calibration in two-candidate races "with no information at all about the forecast problem" by randomly assigning each candidate a 50% chance of winning, since "it can be shown that...pure chance results in half the answers being right and half being wrong" (Appleman, 1960). Indeed, a randomly selected candidate should win 50% of the time. But this "unskilled forecast" provided no new information, despite having achieved calibration. Suppose that another forecaster attempts to make predictions about the same elections based on research and intuition, instead of the random assignment approach. If this forecaster assigns 75% chances of victory, on average, to candidates who go on to win, and their forecasts are also calibrated, then these predictions are much more valuable: the forecaster accurately predicts most outcomes and is still correct about how often their predictions will be wrong. Intuitively, one can assess whether a forecast has achieved some level of informative prediction accuracy by measuring how much better it is than the unskilled forecast approach.

On average, the prediction markets in the dataset assign a 68% probability to winning candidates. The average number of major candidates (defined as candidates who win 5% or more of the final vote) for an election in the dataset is 2.6, although most of them have the median number of 2 candidates.

Therefore, according to a back-of-the-envelope calculation, a forecaster using the random assignment approach would give predictions of $\frac{1}{2.6} = 38.5\%$ (for the average election) or $\frac{1}{2} = 50.0\%$ (for the median election) to winning candidates. Since $68.0\% > 50.0\% > 38.5\%$, the prediction markets appear to be significantly more informative than a random assignment forecaster. However, a perfect forecaster would assign 100% probabilities to winning candidates, so it seems that the markets are subject to quite a bit of prediction error, as well. As suggested above, this thesis will propose several instances in which the amount of prediction error in a given market could be associated with certain underlying variables. If the coefficient of one such variable is found to be statistically significant in a regression, this can be interpreted as evidence of an association (or potentially a causal relationship) between the variable and higher or lower levels of prediction error. This could be helpful in identifying situations where the markets are being over- or under-confident, although it is important to be cautious in interpretation since statistically significant results can be false positives and can sometimes be driven by dubious research practices.

To enhance the credibility of my analysis in the face of these issues, I avoided running any regressions or otherwise analyzing the relationships between my dependent and independent variables until I had made final decisions about the major aspects of my model and communicated them to my supervisor. This included decisions about which variables to include in the model, the level at which to cluster standard errors, what regression methods to use, exclusion criteria for observations, and the hypothesized direction of coefficients. My data was received almost immediately after requesting it, so I could not file a formal pre-registration report, but my practices still limit the perceived or actual opportunity for "p-hacking," i.e. the practice of changing specifications or running multiple statistical tests until statistically significant results are found (Head et al., 2015). Any deviations from the original decisions and their impact on statistical significance are noted within. The most important change to note is that I had to update the panel data specifications after I obtained results from the original models: I added the full set of control variables from the cross-sectional models to the panel data models once I realized I cannot rely on the correlated random effects method to debias the time-invariant independent variables of interest. For the purpose of transparency, the original results are also included in Section 6.

The remainder of the paper is structured as follows. In Section 2, I summarize findings from existing academic research to contextualize the contribution of this thesis and note the sources of information used to motivate my hypotheses. Next, in Section 3 I specify whether I expect each variable to be positively or negatively associated with prediction error based on the literature cited. In Section 4, I explain how each variable was coded and the source of the information, including discussion about the two versions of the dependent variable. In Section 5, I specify the exact models, econometric methods, and identification strategy used. In Section 6, I present my results and discussion about their interpretation. In Section 7, I run robustness checks to test the sensitivity of my results to specification. Finally, in Section 8 I offer final conclusions about each hypothesis, discuss the implications and limitations of my analysis, and suggest potential avenues for future research.

2. Theory and Literature Review

2.1 Background Information about Prediction Markets

Before proceeding further, it will be instructive to offer a sense of the scope and conclusions of existing prediction market research. When discussing PredictIt's markets, the operative example I will use in this section is the "yes" option on the Joe Biden category in the market mentioned in Section 1, corresponding to a prediction that Joe Biden will win the 2020 presidential election. Tziralis and Tatsiopoulos (2007) define prediction markets as "markets that are designed and run for the primary purpose of mining and aggregating information scattered among traders and subsequently using this information in the form of market values in order to make predictions about specific future events." PredictIt's markets certainly meet this definition: traders on the website make predictions in a market by buying or selling shares that correspond to a particular election outcome; this trading activity is then summarized into a price that is interpretable as the predicted probability of that outcome (How to Trade, 2020). Therefore, in buying or selling a Joe Biden share, a trader reveals their beliefs about the probability that he will win the election. In the stock market, traders looking to maximize their returns will only buy shares of assets that they assess to be underpriced, usually based on beliefs about the company's inherent value or on observations about irrational decisions made by other traders. In the same way, a prediction market trader looking to optimize their payoff should buy shares in the Joe Biden category if they think the market is underestimating the likelihood of a Biden win, i.e. if the price is too low. To do this, potential purchasers submit a bid by specifying a maximum price they would be willing to pay for a share. At the same time, current owners of the shares submit an ask by specifying the price at which they would be willing to sell the share. A trade will automatically occur at the ask price if it is lower than or equal to the bid price. The closing price at the end of the day (the number that appears in my dataset) is determined by the price at which the last trade occurred before midnight. The closing price shifts day-to-day across the duration of the market as perceptions about the probability of Biden winning update with the arrival of new information or other changes. Once the outcome is determined and the market is closed, traders are awarded $1 per share of the winning category that they hold. This type of prediction market is known as the continuous double auction format (Christiansen, 2007).

In general, the history of formal academic research into the properties of prediction markets as a tool for forecasting outcomes is relatively sparse, compared to other topics. One of the earliest and most significant contributions to this subfield came with a paper about the forecast accuracy of the Iowa Political Stock Market (IPSM, later renamed the Iowa Electronic Markets or IEM) in the 1988 U.S. presidential election (Forsythe et al., 1992). The IPSM, which also used a continuous double auction format, was set up by academics at the University of Iowa for the purpose of studying whether prediction markets could improve upon the predictive accuracy of polling and other election forecasting methods. Since then, the study of prediction markets as a forecasting tool has expanded

dramatically, with their applications extending to making predictions about outcomes in sports, finance, and world events, and assisting with resource allocation decisions for businesses, among many others (Tziralis & Tatsiopoulos, 2007). As a testament to the continued relevance of this topic, The Journal of Prediction Markets, a journal dedicated solely to the publication of works about this subfield, was founded in 2007 (Williams, 2007).

While this type of formal research started only a few decades ago, the practice of betting money on political outcomes is much older: a historical review found evidence that political betting markets in Italy date back to 1503, and have been in the United States since at least 1796 (Rhode & Strumpf, 2013). Moreover, the theoretical rationale behind prediction markets as a useful tool for forecasting engages with some of the foundational principles of classical economics, namely the efficient markets hypothesis and the rationality assumption. The efficient market hypothesis posits that prices in markets update to accurately reflect all available information, while the rationality assumption refers to the idea that humans are rational economic agents who always seek to maximize their own payoff or utility when making decisions. Although he was not talking about prediction markets, economists like Friedrich Hayek have argued strongly for the idea that markets in general are efficient aggregators of information and that the resulting price reflects the consensus of market participants: the price equilibrium reached by markets is optimal no matter "how little the individual participants" might know and despite the individual participants working for their own self-interest (Hayek, 1945). The efficient market hypothesis also has applications in finance, where some have argued that current stock prices represent the best possible forecast of future stock prices since rational market participants have already synthesized all available information into the price; therefore, trying to "beat the market" by predicting future price fluctuations is pointless, since individuals cannot make a better forecast (Fama, 1970). Following this logic, given enough participants and real economic stakes, the prices in political prediction markets should also reflect the best possible forecast of the outcome that they predict.

2.2 Potential Bias and Inefficiencies in Prediction Markets

These arguments are largely inconsistent with the findings of behavioral economics. Through experimental and other research, behavioral economists have discovered systematic cognitive biases that call into question the idea that market participants are purely rational. These include hyperbolic discounting, which is the observation that people often make suboptimal economic decisions for themselves due to a strong preference for present over future consumption and utility, and loss aversion, which is the strong preference of people to avoid a loss as compared to the potential for making a gain of equal value (Laibson, 1997; Thaler, 1981). With that in mind, it is surprising that Forsythe et. al concluded in their many papers about the IEM that the markets actually do operate efficiently; they admit that this finding seems to "[run] counter to a substantial body of experimental evidence on individual behavior documenting anomalies" (Forsythe et al., 1999). Despite the conclusion of market efficiency, they report substantial evidence from their own (and others') studies and experiments that

"respondents often engage in wishful thinking and respond more often than not that their candidate is likely to win" and that "these biases affect trading behavior on average" (Forsythe et al., 1992). This bias appears to affect traders from all ideological groups. In my example, this would mean that traders who support Joe Biden are more likely to buy shares predicting a Joe Biden victory even when this may be irrational. To explain the apparent paradox of the markets being efficient despite evidence of a systematic bias in trader behavior, the authors consider several potential explanations. One such explanation is the possibility that the traders are representative of the electorate and different ideological groups engage in the same level of wishful thinking, so on average the bias in the market is negligible; they dismiss this explanation, noting that "the data argue against this interpretation" (Forsythe et al., 1992). The relatively nuanced explanation that they settle on is that while the average trader is certainly biased, a group of market participants known as marginal traders end up determining the final price. Marginal traders are thought to be rational market participants who recognize the judgment bias suffered by other traders as an arbitrage opportunity and decide to profit from it. With trader-level data, the authors find evidence of the existence of such participants who invested an above-average amount, did not suffer from wishful-thinking bias, and earned above-average returns; in doing so, they tended to drive the equilibrium price closer to the actual result. This result seems to validate the Hayek hypothesis on some level since prices end up close to the correct level, though the positive outcome seems dependent on there being enough marginal traders who invest sufficient amounts to debias the final price. Markets about obscure elections with little publicly available information may pose a challenge since this could limit the ability of marginal traders to recognize when other traders are being biased. Markets that cap the number of traders or investment amounts might also not have sufficient marginal trader activity to be efficient, even if information is available.

Longer-term research from Berg et. al covering 49 political prediction markets seems to support this intuition. Their results show that the IEM is fairly accurate at predicting the winning candidate's margin of victory, with an average error of 1.49% or 1.58% depending on the measure used, but that accuracy can depend on certain election and market characteristics, noting that "presidential election markets perform better than (typically lower profile) congressional, state and local election markets" and that "markets with more volume near the election" have more accurate predictions (J. Berg et al., 2003). The finding that prediction accuracy can differ depending on election and market characteristics extends to other contexts, with positive results in Taiwan and mostly negative results for several election prediction markets in European countries such as the Netherlands, Austria, Germany, and Sweden (Berlemann & Schmidt, 2001). These researchers suggest that the international differences in market performance could be attributable to the multiparty systems in Europe vs. the two-party system in the U.S., the dynamics of which (multiple candidates) create more uncertainty about the final vote margin. In sum, these findings suggest that election and market features can be drivers of prediction error; this makes it essential to control for such characteristics when trying to measure the effect of a particular variable on prediction error by holding all else equal.

Other key market characteristics than can affect efficiency include fee structures, with prediction market researchers finding that the imposition of "transaction and/or profit fees can lead to mispricing" (J. E. Berg & Rietz, 2019). PredictIt charges a 10% fee on traders' profits and a 5% fee on trader withdrawals (How to Trade, 2020). Furthermore, an agreement with U.S. federal regulators limits each market to a maximum of 5,000 traders and each trader to a maximum of $850 invested in a given contract (Terms & Conditions, 2020). Therefore, some have suggested that PredictIt's markets might suffer from these efficiency issues, with "contracts hosted by PredictIt" being "chronically mispriced" due to PredictIt's fee structures and betting limits "limiting the ability of traders to capture arbitrage profits" (Stershic & Gujral, 2020).

Given the suggestive evidence about the potential inefficiencies in PredictIt's markets, it is surprising that more comprehensive reviews of prediction accuracy in PredictIt's markets (and other markets structured differently from the IEM) are not often conducted. Indeed, the literature is relatively concentrated, with about 43% of all political prediction market research having been written about the efficient-seeming IEM as of 2012 (Tziralis & Tatsiopoulos, 2007). This may contribute to a popular perception that political prediction markets are efficient and more accurate than other prediction methods, despite questions about the generalizability of conclusions from the IEM and mixed results from other market research. But even the IEM studies tend to be small: the longer-term IEM research mentioned above had just 49 election markets. Other election prediction market research (such as the international studies I mention above) has also been limited in scope and scale, with the paper about markets in Germany having just 25 market observations. Therefore, a main differentiating contribution of this thesis is a dramatic expansion in the number of markets studied to 570. This larger sample size will help me determine if observed patterns from other election markets hold when the sample is expanded to a wider range of elections, including many scarcely watched ones that might not typically be featured in the IEM. Additionally, other political prediction market studies tend to measure prediction error as the difference between the vote share in election results vs. the estimated vote share according to the markets; by contrast, PredictIt's markets are mostly structured as a binary or categorical prediction about which candidate will win instead of a predicted vote share. For this reason, I use a different measure of prediction error from most other studies. Other researchers also tend to focus on the question of whether individual trader biases are measurable at the trader level and drive inefficiencies at the market level. Since this question has been studied extensively and trader-level data was not available from PredictIt, I choose a different research focus. My hypotheses are comparatively novel in that they largely focus on whether fundamental features of the candidates and elections themselves are associated with prediction error in the market, though my argument for the existence of these effects is based on hypothesized trader biases: if biases corresponding to these characteristics exist in the average trader, then whether they are measurable at the market level will likely depend on whether marginal traders are aware of others' biases and are not prevented from profiting from them. My hypothesis about the relationship between Google search volume and error on a given day is new, although the use of Google search volume as a proxy for the level of attention

being paid by individuals to a particular topic has some precedence (Jiang, 2016).

2.3 Research Relating to the Hypotheses

Given PredictIt's fee structures and other rules that limit the influence of marginal traders, it seems plausible that systematic prediction error stemming from judgment bias in traders could be measurable in PredictIt's markets despite limited evidence of this from past research. To explore potential sources of bias, I consulted academic research about the interplay between demographics and perceptions of U.S. politics. Unlike IEM traders who tend to come from a selective pool of people, participation in PredictIt's markets is freely open to any visitors of the website. PredictIt's user base tends to be "young, male, and relatively affluent from major U.S. cities like New York and San Francisco" (Perticone, 2018). These characteristics match those of the "Solid Liberals" ideological group as classified by the Pew Research Center, whose members are "among the youngest typology groups", are "financially comfortable", and are more likely to live in an urban area than any other group (C. Doherty et al., 2017). Pew notes that these types of voters tend to "overwhelmingly express liberal attitudes on virtually every issue" and to be politically engaged in support of progressive candidates and causes. One famous example of a candidate these voters tend to support is Bernie Sanders, whose base has been described as comprising "younger" and "very liberal" Democrats (Bronner & Bacon Jr., 2020). Another example is Andrew Yang, the champion of progressive policies such as a universal basic income, whose base tends to skew young and male (Skelley, 2019). The particular demographic makeup of traders raises the prospect that prices could be affected by wishful thinking bias in markets for elections that features these types of candidates.

Further research reveals other potential biases in the markets. In particular, Americans tend to believe that women and ethnic minorities are less electable than men and white Americans because they perceive other voters to be prejudiced in their voting choices (Mercier et al., 2020). This is despite the fact that "no evidence of any direct, consistent, or substantial impact" of gender on electoral outcomes has been found and that "minority Democrats and minority Republicans [perform] as well as their white co-partisans" (Dolan, 2013; Juenke & Shah, 2016). Others authors note that these misperceptions are also measurable among people who are highly politically engaged, including local party chairs (D. Doherty et al., 2019). In other words, experience in and knowledge about politics does not necessarily lead to more accurate beliefs about electability. These findings suggest that markets for elections with female and ethnic minority candidates could be subject to systematic mispricing due to the prevalence of misperceptions about their electability. But a caveat worth mentioning is that although female candidates tend to win general elections at the same rate as male candidates, researchers such as Lawless and Pearson (2008) have found that "primary elections are not gender neutral" and that this "hamper[s] women's entrance into public office," often meaning that only better qualified women in terms of "electoral experience and fundraising success" will run for higher office. This raises the possibility that female candidates are different from male candidates, on average, in ways that could affect electoral success other than gender; if traders are aware of

these dynamics in elections involving female (and possibly minority) candidates, the hypothesized relationship might not exist.

As alluded to throughout this section, one of the driving factors behind the (potential) accuracy of prediction markets is their ability to synthesize publicly available information into better predictions than individuals would be able to make. Unsurprisingly, research cited above has found that, in general, elections with more information available about them (e.g. presidential) tend to have better predictions. Consistent with this, since there is less information available when trading at earlier points in time, one would expect traders' predictions to improve measurably as election day gets closer. The evidence about this is mixed: in trying to measure the incidence of a particular type of bias, Restocchi et al. (2018) find that predictions become less biased closer to election day when measured in the short run, but this pattern is not consistent in the long run. Other authors have not been able to establish any link between prediction accuracy and the duration of the market over longer timespans despite evidence that the quality of (for example) polling data improves as election day gets closer (Berlemann & Schmidt, 2001). This is suggestive that the relationship between market duration and prediction error could be nonlinear or dependent on other factors. Complicating matters further, the release of new information as election day approaches is not likely to be uniformly distributed: political actors often time the release of new information to manipulate the political consequences. This emphasizes the importance of capturing the dynamic flow of information when trying to estimate its impact on prediction error. But regardless of timing, one would still expect that the revelation of new information should allow for better forecasts. The counterargument is that traders may not interpret new information rationally: indeed, behavioral economics and psychology researchers have found evidence that people's interpretation of information is affected by recency bias, i.e., the "tendency to emphasize the importance of recent experience in estimating future events" even when older information might be more useful (Phillips-Wren et al., 2019). Furthermore, it could be that "trading occurs on the arrival of new information" because "market participants have heterogeneous priors, and hence differ in their interpretations of public information" (Rothschild & Sethi, 2013). While these findings raise the possibility of traders failing to benefit from new information, I would still argue that the arrival of information should improve traders' predictions, due to previously cited research about markets with more information having better forecasts.

3. Hypotheses

Based on the literature review conducted, I propose the following hypotheses. I indicate here whether the variables are expected to be associated with more or less prediction error, the expected sign of the coefficient, and whether the hypothesis is considered to be judgment bias-based or information flow-based. This information is summarized in Table 1.

Table 1: Hypotheses Summary

Hypothesis	Association	Sign	Type
Progressive (H1)	More error	+	Judgment bias
Female (H2)	More error	+	Judgment bias
Minority (H3)	More error	+	Judgment bias
(Log) Enthusiasm (H4)	More error	+	Judgment bias
(Log) Duration of Market (H5a)	More error	+	Information flow
Days Until Election (H5b)	More error	+	Information flow
Google Search Volume (H6)	Less error	-	Information flow

Hypothesis 1. Predictions made about elections with at least one candidate whose ideology is identified as progressive will have more prediction error. As mentioned in the literature review, the demographics of PredictIt traders are not representative of the U.S. electorate, with a strong skew toward urban, male, and young voters. Research from C. Doherty et al. (2017), Skelley (2019), and Bronner and Bacon Jr. (2020) on the political preferences of this demographic group indicates that they tend to be supportive of progressive candidates, such as Bernie Sanders and Andrew Yang. Combined with research from Forsythe et al. (1992) about wishful thinking bias in prediction market traders of all ideologies, I hypothesize that prediction error in PredictIt's markets could be systematically and positively associated with the presence of progressive candidates.

Hypothesis 2. Predictions made about elections with at least one candidate who is female will have more prediction error. Political science research cited in the literature review from Mercier et al. (2020) and Dolan (2013) indicates that American voters, including ones who are highly politically engaged (as traders might be), tend to hold misperceptions about the electability of female candidates (D. Doherty et al., 2019). Based on this, I hypothesize that prediction error in PredictIt's markets could be systematically and positively associated with the presence of female candidates.

Hypothesis 3. Predictions made about elections with at least one candidate who is an ethnic minority will have more prediction error. Political science research cited in the literature review from Mercier et al. (2020) and Juenke and Shah (2016) indicates that American voters, including ones who are highly politically engaged (as traders might be), tend to hold misperceptions about the electability of ethnic minority candidates (D. Doherty et al., 2019). Based on this, I hypothesize that prediction error in PredictIt's markets could be systematically and positively associated with the presence of

ethnic minority candidates.

Hypothesis 4. Predictions made about elections where voters are enthusiastic will have more prediction error. This is another potential instance of wishful thinking bias as observed by Forsythe et al. (1992). Higher levels of enthusiasm among the electorate means that traders could be more likely to make irrational bets. Additionally, enthusiasm could drive supporters of candidates to sign up for the betting website and make predictions in favor of their preferred candidate. Potential reasons why traders could be enthusiastic about an election are varied. One possibility is that traders could be enthusiastic about a particular candidate. If enthusiasm is driven by the presence of a progressive, female, or minority candidate, then it could be a bad control for the other three independent variables; I address this possibility with a robustness check in Section 7. Otherwise, traders might be enthusiastic about an election because it is framed as uniquely decisive or impactful. An example of the latter could be the closely watched 2017 special election in Georgia's 6^{th} district, which attracted an unusual amount of attention because it was expected to be competitive and was framed as a referendum on President Trump's job performance during his first year in office (Barrow, 2017).

Hypothesis 5a. (Cross-sectional model). Predictions made in markets that have longer durations (i.e. more days of trading) will have more prediction error. For markets in the cross-sectional model, I take the average prediction error of a market over its duration as the dependent variable observation. Longer durations imply that predictions are being made further away from election day with less information available to traders. Since the average level of information available over the life of the market is lower in markets with longer durations, predictions in those markets should see more error. I base this hypothesis on the classical understanding of how traders in markets use information to improve forecasts such as from Hayek (1945) and Fama (1970), but more recent research from Restocchi et al. (2018) and Berlemann and Schmidt (2001) give mixed results that suggest that the importance of duration of trading could depend on whether the market is operated in the short run or the long run. As seen in Section 4, I run both a cross-sectional and a panel data model. This hypothesis is specific to the cross-sectional model where markets are allowed to differ in their durations of trading, and is not tested in the panel data setting.

Hypothesis 5b. (Panel data model). Predictions made further away from election day will have more prediction error. In the panel model, the amount of prediction error measured on a particular day in a given market (i.e. a market-day pair) is the dependent variable observation. Traders have less access to information further away from election day, so I expect to see more error in those observations. I once again base this hypothesis on the classical understanding of how traders in markets use information, but in this instance the findings from Restocchi et al. (2018) about the relationship between the number of days and prediction error in the short term are more supportive, since the panels are restricted to 15 days. For the panel dataset only (i.e. this does not affect the cross-sectional dataset), I apply an exclusion criterion to drop observations such that each market has 15 days of trading to ensure that the panels are balanced and comparable between markets. This

hypothesis is specific to the panel data model since it describes a relationship over time that cannot be tested in the cross-sectional setting.

Hypothesis 6. Predictions made on days with more Google search volume about an election will have less prediction error. I track search volume relating to keywords that uniquely identify each election. High levels of search volume relating to an election on a given day should be a proxy for the revelation of information, e.g., a scandal, which should decrease prediction error as more information is available to the traders. Ideally, search volume will capture the nonlinear flow of information approaching election day referred to in the literature review, as the revelation of information can come on any day. Crucially, it seems plausible that the interests of searchers of U.S. election-related terms on a given day will overlap with the interests of U.S. election market traders. Indeed, search volume as a proxy for the level of attention paid to a certain topic has been shown to be associated with trader behavior and price fluctuations in the stock market, as in Jiang (2016). The argument for the sign of this hypothesis is also based on the classical understanding of how traders in markets use information, but the findings from Phillips-Wren et al. (2019) and Rothschild and Sethi (2013) about recency bias and the heterogeneous interpretation of information suggest that results for this hypothesis could lead to different conclusions if traders misinterpret new information. This hypothesis is specific to the panel data model since it describes a relationship over time that cannot be tested in the cross-sectional setting.

4. Dataset and Variables

4.1 Dataset and Exclusion Criteria

As mentioned above, the main dataset containing the predicted probabilities of outcomes was obtained from PredictIt.org. The dependent variable was calculated entirely from this original dataset, but most of the independent variables were obtained and coded manually from external sources.[1] Each variable description in the below subsections contains more information about where it was sourced.

The original dataset had market data about 623 elections, 53 of which were removed due to meeting certain exclusion criteria that were determined before any regressions or analyses were run. The final dataset contains 570 unique election and market observations. In particular, I excluded:

1. Markets that represent duplicate observations for elections that are already in the dataset (e.g. differently phrased questions about the same election), since they have identical independent variable information. When given the choice, I retained the market that was already phrased as a binary question to simplify data processing; there were no cases of duplicate observations with more than one binary market. 6 observations were dropped due to this criterion.

2. Markets whose outcomes were determined by a cause other than election results, e.g. deaths, retirements, or court cases. 17 observations were dropped due to this criterion.

3. Markets that are not specific to a district, state, or city, e.g., the Democrats Abroad primary and the United States-wide general election, since proper demographic information on these observations does not exist or it is a composite of all other observations. 2 observations were dropped due to this criterion.

4. All observations relating to local elections smaller than mayoral elections, such as city council and town supervisor, due to lack of independent variable data for some. 4 observations were dropped due to this criterion.

5. Markets about referendums, since they do not have candidates to test demographic hypotheses about. 24 observations were dropped due to this criterion.

   I note that the cross-sectional dataset is unaffected by the sixth criterion. For the purpose of obtaining a balanced panel (i.e. one where every panel has an equal number of market-day observations), I exclude observations from the panel dataset if they meet the following criterion:

6. Market-day observations outside of the 15 days preceding election day. I do this in order to ensure the consistency of comparison between panels. I exclude market-day observations that come on or after election day because I am only interested in prediction error in markets before

[1] After I completed the coding of the externally sourced variables, I took a random sample of 50 observations and re-checked their dependent and independent variable values to test their accuracy. This process might not detect every potential coding error in the dataset, but hopefully should detect systematic mistakes that could induce significant measurement error.

any official election results are reported. Markets are typically kept open for trading through election day and often days later, especially if the results are unclear. Since many elections did not have 15 full days of trading before election day, I was left with 15 days of trading in 431 election markets after applying this exclusion criterion. Therefore, the final number of market-day observations in the panel dataset is 6,465.

4.2 Predicted Probability and Dependent Variables

Predicted Probability Variables: p_m and $p_{m,t}$

This variable is denoted as p_m in the cross-sectional model and $p_{m,t}$ in the panel data model and ranges from 0 to 1. It is defined as the probability that traders in a given market m attribute to the winning category, t days before the election (if there is a t subscript). I note that this variable is not the dependent variable and it is not used in any regressions. However, it is used to construct the dependent variable. How this variable is coded depends on whether the underlying market is *Binary-by-nature* or *Binary-by-adaptation*. Markets in the former category are already structured as binary choices due to category design. An example is the market "*Which party will win the U.S. Senate race in Maryland in 2016?*" for which traders could only bet "Democratic" or "Republican." The price of the winning category ("Democratic" because it is known that a Democrat won this race, in retrospect) as a percentage of the total price of all categories is interpreted as the traders' predicted probability of the event happening. In the cross-sectional model, p_m is the average of the probabilities implied by the closing prices across the entire duration of the market. The average probability of "Democratic" over the lifetime of the market was 97%, so $p_m = 0.97$ in the cross-sectional specification. In the panel data version, $p_{m,t}$ is the probability implied by the closing price on a given day in the market. On 2016-11-07, the probability of "Democratic" was 94% vs. 6% for Republican, so this observation is coded as $p_{m,t} = 0.94$ in the panel data specification.

On the other hand, markets that are *Binary-by-adaptation* are not structured as binary choices, but I adapt them to be binary for the purposes of this thesis. An example is the market "*Who will win the 2016 Massachusetts Democratic primary?*", for which traders could bet on the categories *Hillary Clinton*, *Bernie Sanders*, and *Martin O'Malley*. In these cases, I redefine the market to be a binary question about whether traders think that the winning category will occur. In retrospect, it is known that Hillary Clinton won, so I restructure the market to be "*Will Hillary Clinton win the 2016 Massachusetts Democratic primary?*" The price of the winning category as a percentage of the total price is interpreted as the probability that the event occurs ("yes"), while the prices of the losing categories are summed into a combined price and probability ("no"). The average probability over the lifetime of the market was 62% for Hillary Clinton, 38% for Bernie Sanders, and approximately 0% for Martin O'Malley, so $p_m = 0.62$ in the cross-sectional specification. This is interpreted as 62% chance of "yes" and 38% chance of "no." On 2016-02-29, the probabilities attributed to Hillary Clinton, Bernie Sanders, and Martin O'Malley were 81%, 18%, and 1%, respectively. The probability variable for this market-day observation would be coded as $p_{m,t} = 0.81$.

To give a sense of the distribution of this variable, I computed summary statistics for p_m and $p_{m,t}$ and presented them in Table 2.

Table 2: Independent Variable Summary Statistics

	Mean	SD	25^{th} Pct	Median	75^{th} Pct	Observations
p_m	0.68	0.22	0.52	0.71	0.87	570
$p_{m,t}$	0.74	0.24	0.60	0.84	0.94	6,465

As mentioned before, the average probability that traders assign to winning candidates is 68% in the cross-sectional dataset, as compared to 74% in the panel dataset. A t-test for a difference between these means strongly rejects the null hypothesis that there is no difference at the 1% significance level, with $p = 0.00$. Evidently, the application of the sixth exclusion criterion has changed the composition of observations of the dependent variable (which is constructed from the probabilities) between the cross-sectional and panel datasets. Traders make somewhat more accurate predictions in the panel dataset perhaps because the average prediction in that dataset is made closer to election day or because the elections excluded from the panel dataset are fundamentally different in some way. In Section 7, I run a robustness check to determine whether the omission of these election market observations introduces selection bias to the results of my cross-sectional regressions. Unlike the mean, the standard deviations for each dataset look fairly similar and both indicate a fairly wide range of probabilities in the data. The median of 0.71 in the cross-sectional dataset is fairly consistent with its mean of 0.68, but it indicates a slight skew in the data. The larger disparity between the mean (0.74) and median (0.84) in the panel dataset indicates that observations in the panel dataset are strongly concentrated in the right tail of the distribution, but with a significant leftward skew dragging the mean downwards. This is also likely to be an effect of the panel transformation: in the last 15 days of a market, traders seem to become extremely confident and consistently give high predicted probabilities about the subset of uncompetitive elections, since there is little uncertainty left. This is further supported by the higher 25^{th} and 75^{th} percentiles of probabilities in the panel dataset than the cross-sectional dataset. Probabilities relating to markets about competitive elections likely continue to fluctuate within the lower ranges of the data as uncertainty is still high over the last 15 days of a market. These results are suggestive that the variance of prediction error could be dependent on the duration of trading, the number of days until the election, and the competitiveness of elections, all which are independent variables. This issue will be accounted for through the use of heteroskedasticity-robust clustered standard errors.

With the probability variables, I calculate the amount of prediction error to use as the dependent variable. More positive values indicate more error. I coded both a binary version (denoted by the b superscript) and a continuous version of the dependent variable (denoted by the c superscript).

Binary Dependent Variables: y_m^b and $y_{m,t}^b$

In this specification of the dependent variable, I indicate that traders in a given market m, t days

until the election (if applicable), correctly predicted the outcome of the election, if their predicted probability for the winning category was greater than 50%. This is coded as a 0, i.e. there is no prediction error. Otherwise, I code 1 if they attributed a probability of exactly 50% or lower for the winning category, indicating prediction error. I formalize the coding rule as follows:

$$y_m^b = \begin{cases} 0, & p_m > 0.5 \\ 1, & p_m \leq 0.5 \end{cases}$$

$$y_{m,t}^b = \begin{cases} 0, & p_{m,t} > 0.5 \\ 1, & p_{m,t} \leq 0.5 \end{cases}$$

A major advantage of this specification of the dependent variable is its intuitiveness and simplicity: the traders' prediction was either correct or incorrect. However, the main issue is that the value is essentially only interesting for competitive elections. Traders will almost never fail to predict elections with obvious outcomes. Therefore, the amount of variation in the dependent variable could be minimal depending on the proportion of competitive elections in a finite dataset, which would make it hard to detect whether the effects of independent variables exist.

Continuous Dependent Variables: y_m^c and $y_{m,t}^c$

In this specification of the dependent variable, I do not indicate whether traders made a correct prediction. Instead, I calculate the distance of the predicted probability of the winning category (p_m or $p_{m,t}$) from the deterministic probability, which is 1 because it is known that the winning category happened. Essentially, this dependent variable measures how underconfident traders were in their assessed probability of the correct outcome. For example, if $p_m = 0.95$, this would give a value of $y_m^c = 0.05$, indicating that the traders were very confident in predicting the correct outcome. If $p_m = 0.10$, this would give a value of $y_m^c = 0.90$, indicating the traders were very confident in their incorrect prediction. Values of y_m^c closer to 0 are indicative of less prediction error since they signify higher predicted probabilities for correct outcomes. I formalize the coding rule as follows:

$$y_m^c = 1 - p_m$$

$$y_{m,t}^c = 1 - p_{m,t}$$

A major advantage of the continuous specification is that it contains more variation than the binary one. Since it does not rely on the sharp cutoff at $p_m > 0.50$, this makes it possible to utilize observations from all elections instead of just competitive ones near the discontinuity, since traders' varying levels of confidence about the outcomes of uncompetitive elections could be driven by independent variables. One disadvantage is that a continuous scale of prediction error is a more abstract concept to interpret than the binary result of a prediction being correct or incorrect. Another is the fact that the specification effectively punishes traders for predicting that elections will be

competitive, i.e. $y_m^c \approx 0.50$, when in fact that might be a good prediction if the outcome is truly a tossup. I mitigate this issue with the introduction of the k_m controls vector, as I discuss in the control variables subsection.

In coding this variable, the choice was made to use an absolute error score instead of the Brier score. The Brier score for an observation would simply be the error squared, i.e. $y_m^c = (1 - p_m)^2$; this method of calculating error is a common way of measuring the prediction error of forecasts (Brier, 1950). Each approach has advantages and disadvantages, but I opted for the absolute score instead of the Brier score as the main measure since the squaring function provides unequal weights to observations at different points along the spectrum from 0 to 1. In practical terms, under a Brier scoring rule, observations with high error scores are weighted more than observations with low error scores, which are disproportionately attenuated to 0. A high absolute error score of 0.99 barely changes to become 0.9801 while a low error score of 0.15 almost disappears at 0.0225 under the Brier approach. To determine whether my results are sensitive to choice of scoring rule, I perform a robustness check on my results in Section 7 where I re-run the continuous dependent variable specifications with the Brier score instead.

4.3 Independent Variables of Interest

Major Candidate is a Progressive (H1): r_m

r_m takes a value of 1 if a major candidate ($> 5\%$ of the vote) is considered progressive, and 0 otherwise. It is time-invariant in the panel data model, so it does not have a time subscript. Based on Hypothesis 1, which predicts positive association with prediction error, I expect $\beta_1 > 0$.

This data was sourced from lists of endorsed candidates from 2015 through 2020 available on the websites of groups that are dedicated to promoting progressive candidates. All of the progressive candidates identified were Democrats, but the majority of Democrats in the dataset were not endorsed by these groups, which indicates that the groups make a distinction between party affiliation and progressive ideology. The groups are the Progressive Democrats of America, Progressive Change Campaign Committee, Indivisible, Justice Democrats, Sunrise Movement, and Our Revolution. The choice of groups was sourced mainly from exploratory research about progressive endorsements from Rakich and Conroy (2020), but Sunrise Movement and Progressive Democrats of America are new additions that I made. 167 markets with progressive candidates were identified, with 76 of these being in clusters, i.e. repetitions of Bernie Sanders in the 2016/2020 Democratic primary clusters. I include further discussion about my approach to clustering standard errors at the conclusion of Section 5. This leaves 91 independent observations out of 570 that are coded as progressive in addition to the 76 observations in clusters.

Major Candidate is Female (H2): f_m

f_m takes a value of 1 if a major candidate ($> 5\%$ of the vote) is female, and 0 otherwise. It is time-invariant in the panel data model, so it does not have a time subscript. Based on Hypothesis 2, which predicts positive association with prediction error, I expect $\beta_2 > 0$.

This data was sourced from ballotpedia.org, an online U.S. election encyclopedia, or other public news sources. I determined whether a major candidate was female by reading publicly available information about the major candidates. 344 markets out of 570 were identified as having female candidates, with 272 of them existing as independent observations outside of the main clusters.

Major Candidate is an Ethnic Minority (H3): e_m

e_m takes a value of 1 if a major candidate ($> 5\%$ of the vote) is an ethnic minority, and 0 otherwise. It is time-invariant in the panel data model, so it does not have a time subscript. Based on Hypothesis 3, which predicts positive association with prediction error, I expect $\beta_3 > 0$.

This data was sourced from ballotpedia.org or other public news sources. For each election, I used publicly available reporting or information about the candidates to determine whether they identify as an ethnic minority. In each observation coded as 1, I was able to find information from reliable sources indicating that the candidate is an ethnic minority. For example, in the "Which party will win the 2018 Maryland gubernatorial race?" market, I coded the ethnic minority variable as 1 due to publicly available reporting about one of the candidates: "If he wins, Jealous would become Maryland's first African-American governor" (Foran, 2018). Additionally, to ensure the completeness of my listing, I checked lists or membership groups of politicians who are ethnic minorities in order to ensure that I had correctly coded the candidates in my dataset. For example, I checked the membership list of the Congressional Hispanic Conference, a group for Hispanic Republicans, and ensured that all of its members who show up in my dataset are coded as ethnic minorities. I define ethnic minorities in the United States as African-American, Hispanic-American, Asian-American, Middle-Eastern American, Indian-American, Native-American, or any category of ethnic self-identification other than just white. 186 markets out of 570 were identified as having ethnic minority candidates, with 140 of them existing as independent observations outside of the main clusters.

Enthusiasm (H4): $\log \frac{s_m}{n_m}$

s_m is the total amount of fundraising that comes from individuals (i.e. excluding corporations, political groups, and candidate donations to themselves) who gave to candidates in an election, divided by the number of voters n_m, as a proxy for enthusiasm and engagement. The fundraising variable s_m by itself is strongly driven by population, so I normalize it by the numbers of voters n_m in order to find elections that have an unusual dollar amount of donations per voter. This score, which I designed for the purpose of this thesis, works as expected: the market for the 2017 special election in Georgia's 6^{th} district that I mentioned before comes in at the 99^{th} percentile of enthusiasm scores. This variable is time-invariant, so it does not have a time subscript. Based on Hypothesis 4, which predicts positive association with prediction error, I expect $\beta_4 > 0$.

This is publicly available data from the FEC, state government websites, and the campaign finance watchdog followthemoney.org. This variable should not be skewed by big-dollar donations from a few donors since donations from individuals directly to candidates have a maximum donation limit – e.g., \$2,800 at the federal level (FEC, 2020). Therefore, fundraising from individuals should be a good measure of enthusiasm since it is mostly driven by the number of people who donate, not by big

donations.

It is log-transformed because there should be a decreasing impact as enthusiasm increases to extremely high levels. The data is highly skewed, with right-tail outliers like small congressional elections that attract an unusual amount of money and attention relative to the number of voters.

Duration of Trading and Days Until Election (H5a, H5b): $\log d_m$ and $d_{m,t}$

For the cross-sectional OLS model, $\log d_m$ is equal to the log of the total number of days available in a market m, i.e. its duration of trading. It is time-invariant in the cross-sectional models, so it does not have a time subscript. For the panel data model, $d_{m,t}$ is equal to t, the number of days left until the election, for a given market-day observation. Since it is time-variant in the panel models, it has a time subscript. Based on Hypotheses 5a and 5b, which predict positive association with prediction error, I expect $\beta_{5a} > 0$ and $\beta_{5b} > 0$.

This cross-sectional variable is log-transformed because the distribution of the data is skewed. Additionally, the change in prediction error that happens when the number of days is high vs. when the number of days is low should be different, i.e., there should be smaller marginal increases in effect size as the number of days goes up. Findings from Restocchi et al. (2018) as cited in the literature review are supportive of the potential nonlinear relationship between the duration of trading and prediction error. The panel data specification of days is not log-transformed since each panel has a uniform distribution of days from 0 to 15 and there are unlikely to be significant differences in the marginal effects of days on prediction error over such a short timeframe.

Google Search Volume (H6): $g_{m,t}$

$g_{m,t}$ measures the total amount of Google search volume relating to market m, t days before the election. Since it is time-variant, it has a time subscript. Based on Hypothesis 6, which predicts negative association with prediction error, I expect $\beta_6 < 0$.

Data for this continuous variable is sourced from publicly available Google search volume data. Changes in Google search volume should reflect underlying changes in the amount of information available about an election. An alternative proxy for this variable could be the number of mentions of an election in the media on a given day. For an example of how this variable is coded, for the market "Which party will win the 2016 Vermont gubernatorial race?" I track search volume for the term "Vermont gubernatorial election 2016." The keywords used will be consistent between markets. Google automatically normalizes search volume over the time period requested: the market-day observations with the least search volume within the panel will have a value of 0 and those with the most will have 100. This makes the variable useless for comparisons of search volume between markets, so it is not included in the cross-sectional version of the model.

4.4 Control Variables

Availability of Polling: a_m

a_m indicates whether there is publicly available and easily accessible polling relating to this elec-

tion. This variable is time-invariant, so it does not have a time subscript. I sourced this data from the realclearpolitics.com website, which hosts a popular poll aggregator. If an election has its own page of polls listed on the aggregator, then it is coded as a 1. To see what this web page looked like during different election cycles, I used the Wayback Machine available on archive.org to see which elections had their polls featured. My motivation for the inclusion of this variable is the research from J. Berg et al. (2003) and Berlemann and Schmidt (2001) documenting the importance of election characteristics in determining prediction error. Elections for which there is easily accessible polling information are likely to be fundamentally different from other elections because they tend to be more closely watched (i.e. competitive senatorial races instead of congressional elections). This variable is likely to be correlated with independent variables of interest, such as enthusiasm (for closely watched races), and accessible polling should affect the incidence of prediction error since it represents more information available to traders.

Trade Volume: $\log v_m$

$\log v_m$ is equal to the log of the total sum of trade volume that occurred relating to the winning category of a given market over its duration. This variable is log-transformed because the data is highly skewed and I expect a declining impact of volume as it increases. My motivation for the inclusion of this variable is the research from J. Berg et al. (2003) that finds a relationship between trade volume and prediction accuracy. A high volume of trade in a market likely indicates an election with lots of publicly available information, which should decrease the amount of prediction error. In addition, the incidence of progressive, female, or ethnic minority candidates could be higher in more closely watched elections. This variable should control for the effect of elections being closely watched, but it is likely to be highly collinear with similar variables that capture the nature of the election. I also include the log of volume of the market over its duration in the panel data model to control for differences between markets, but I exclude a time-variant version of volume because it is very likely to capture the same effect as and be a bad control for Google search volume on a given day.

Competitiveness Vector: $k_m = k_m^{toss} + k_m^{lean} + k_m^{likely} + k_m^{safe}$

For each election observation, I code a dummy variable indicating its expected level of competitiveness. A value of 1 for k_m^{toss} would indicate that the election is considered to be a tossup. k_m^{lean}, k_m^{likely}, and k_m^{safe} are indicators that one candidate is favored to win in the election to varying degrees of likelihood. My motivation for the inclusion of these variables is the research from J. Berg et al. (2003) and Berlemann and Schmidt (2001) documenting the significance of election characteristics in determining prediction error; one such important election characteristic is inherent competitiveness. Prediction error values are often highly dependent on the inherent competitiveness of elections, which is not always adequately controlled for based on the control variables included. I note that previous prediction market research, such as that done about the IEM, does not include these controls because they are unnecessary: in those markets, prediction error is measured as the difference between traders' predicted vote margin and actual vote margin. For them, there is no reason why competitive

elections should see more prediction error than noncompetitive elections since traders are not judged on whether they correctly pick the winning candidate. By contrast, traders in PredictIt's binary prediction markets will see more prediction error when they wrongly predict a candidate will win, even if the race is competitive and difficult to predict. Therefore, I eliminate bias by introducing this variable since the inherent competitiveness of elections is likely to be correlated with explanatory variables of interest (e.g. progressive candidates could affect the competitiveness of elections, perhaps due to the median voter theorem as described by Holcombe (2006)).

There are many possible ways to measure or create a proxy for this variable. One might be to assume that competitiveness is primarily driven by structural features of candidates, elections, and electorates that tend to be relatively constant over time, e.g. that Florida is competitive because its demographics make it a swing state, and it will remain so over short timespans. There seems to be some merit to this: in my dataset, a simple regression of the margin of victory in an election on the margin of victory from the most recent iteration of the same election (e.g. predicting a congressman's 2016 margin of victory from the 2014 margin of victory) yields evidence of a strongly positive and statistically significant relationship ($p = 0.00$). But this simple relationship ignores the fact that levels of competitiveness do change from cycle to cycle, as seen by the big swing toward Republicans in Wisconsin from 2012 to 2016. Therefore, a variable capturing competitiveness that is individualized to a specific election and cycle would more precisely capture the control variable of interest. Luckily, there are forecasters such as FiveThirtyEight who create such predictions for most election observations in my dataset.

According to FiveThirtyEight's methodology, in a two-candidate race, a tossup prediction indicates that each candidate has between a 40% and 60% chance of winning, according to historical outcome data. Lean refers to between 60% and 75%, likely between 75% and 95%, and safe 95% or above (Silver, 2020a). These ratings were designed to correspond to qualitative ratings from the Cook Political Report, a professional political forecaster. For each election observation, I gather data about priors from one of the three following sources:

1. **FiveThirtyEight:** I consider this to be the highest quality prior available. FiveThirtyEight's quantitative models makes a rigorous assessment of polling, demographic, and historical data in order to develop a tossup, lean, likely, or safe prior for each election.

2. **Cook Political Report:** the Cook Political Report's experts specialize in assigning qualitative predictions to elections, which correspond to the same categories as above.

3. **Manually coded priors:** in some cases, priors were not available from the above two sources, so I assigned priors based on available information, such as polling. To assist in this, FiveThirtyEight provides historical data about average margins of victory for races in each category of competitiveness. Historically, the margin of victory for tossup races was 0 points; for lean races, it was 7 points; for likely, it was 12 points; and for safe, it was 34 points (Silver, 2020b). Therefore, I looked at the most recent poll conducted before the election to develop a prior: if

the poll suggests a margin of victory close to 0 points, I assigned tossup; if it is closer to 7, 12, or 34, I assigned lean, likely, and safe, respectively. In rare cases, in the absence of polls or other forecasts, I looked at qualitative reports describing people's expectations about the race to develop a prior. As a check for the reasonableness and comparability of this method to the previous two, I compared the final margins of victory between the FiveThirtyEight/Cook priors and the coded ones. They appear to be approximately comparable: the average tossup margin of victory was 4.4% vs. 5.3% for external and coded priors, respectively. For lean, they were both exactly 7.5%. For likely, they were 10% vs. 15% for external and coded priors, which are both quite close to the historical figure of 12%. For safe, they were 27% vs. 47% for external and coded, respectively. This last difference is large but it is a reflection of the fact that the safe category has essentially no upper bound. Clearly, candidates winning by either 27% or 47% are running in safe elections. Furthermore, elections in the coded set tended to be less competitive ones, such as primaries for popular incumbents, who often win by 50% or more, which explains the higher margins of victory.

Given the above, I would argue that these priors are the best measure of competitiveness available for inclusion in the model as they are a reliable estimate for the expected margin of victory (and therefore competitiveness) of a given election.[2] But due to the existence of different possible approaches for measuring competitiveness and the potential for measurement error in the manually coded priors, I present results from robustness checks omitting the competitiveness variables and using the past margin of victory approach in Section 7.4 to test for the sensitivity of results to this particular coding. I note that I decided to run this robustness check after running my original specifications.

Other Controls Vector: x_m

Finally, I introduce a set of additional control variables to capture other potential covariation between the independent and dependent variables. As with others, my motivation for the inclusion of these variables is the research from J. Berg et al. (2003) and Berlemann and Schmidt (2001) documenting the importance of election characteristics in determining prediction error. These characteristics include variables describing the nature and environment of the election, i.e. whether the election is a senatorial, congressional, presidential, state/local election, and whether it is a general election or a primary. Similarly, I include dummies for the year the election is held in in order to capture time period-specific shocks that might affect all election markets in an election cycle. These variables are all dummy variables sourced directly from the Ballotpedia.org page about each of the elections and coded as a 1 if they describe the election and a 0 if they do not. I also include a variable for the number of major candidates due to specific results from Berlemann and Schmidt (2001) indicating that the greater number of candidates in European elections (as compared to American ones) could be a determinant of prediction error. A candidate is classified as major if they earned more than 5%

[2] A regression of current-year election margins of victory on these priors reveals their strong predictive power: the likely, lean, and tossup dummies are associated with highly statistically significant ($p = 0.00$ for all) and progressively smaller margins of victory, as compared to the omitted safe dummy. The R-squared statistic of this regression is 51%, indicating that they do a better job of explaining margin of victory than past margin of victory data, which had an R-squared of 17%.

of the final vote total in the election. This information is also sourced from Ballotpedia.

Furthermore, I include control variables capturing the demographic characteristics of the states, districts, and cities where elections are held, again as motivated by the above researchers. These include median household income and the percentage of the district that is 65 or older, that is female, that is white, that is African-American, that is Hispanic-American, that has a high school degree, and that has a bachelor's degree. The demographic data for each district, state, or city is obtained from Census.gov. I made the choice to include terms capturing demographic differences instead of location dummies because of the lack of observations for many states and districts and the fact that location-based prediction error is driven largely by underlying characteristics. For example, I could include a Wisconsin dummy because of the large polling and prediction errors observed in this state during the time period in question, but researchers have argued that this error was driven by the overrepresentation of college graduates in polls (Kennedy et al., 2018). Therefore, by controlling for education characteristics instead of location, I also account for the effect of education levels on prediction error in states and districts that might only have one observation and therefore cannot have a location dummy.

Each of these variables captures some aspect of the nature or location of the election that could potentially impact the independent and dependent variables. For example, it could be that whether an election has an ethnic minority candidate is correlated with the ethnic diversity of the district. And the ethnic diversity of the district could affect prediction accuracy if certain ethnic groups are under-represented in available polling. Another example is year: for example, my year dummies help to account for the polling error common to elections held in 2016 that caused pollsters (and likely traders) to underestimate winning probabilities for Republican candidates. The 2016 dummy is likely also correlated with independent variables, such as enthusiasm, since people are more engaged in politics during presidential election years.

4.5 Descriptive and Summary Statistics

In Table 3, I present a matrix showing the number of observations in the dataset that correspond to each possible combination of election type and year. This is to give the reader an overview of the timespan and nature of the elections subject to analysis.

Table 3: Election Type and Year Matrix

	Observations by Election Type					
	Municipal	Statewide	Congressional	Senatorial	Presidential	Total
2015	1 [0%]	3 [1%]	0 [0%]	0 [0%]	0 [0%]	4 [1%]
2016	0 [0%]	10 [2%]	29 [5%]	23 [4%]	146 [25%]	208 [36%]
2017	16 [3%]	7 [1%]	8 [1%]	3 [1%]	0 [0%]	34 [6%]
2018	1 [0%]	65 [11%]	176 [31%]	45 [8%]	0 [0%]	287 [50%]
2019	3 [1%]	5 [1%]	3 [1%]	0 [0]	0 [0]	11 [3%]
2020	0 [0%]	0 [0%]	0 [0%]	0 [0%]	26 [4%]	26 [4%]
Total	21 [4%]	90 [16%]	216 [38%]	71 [13%]	172 [29%]	570 [100%]

observations in each category as a percentage of the total dataset, in brackets

The groupings of observations into election type and year follow predictable patterns. Observations about presidential elections, which comprise 29% of the dataset, can only be found in 2016 and 2020, since presidential elections in the U.S. only happen every four years. The vast majority of these observations are in 2016 because the 2020 general election markets were not yet resolved at the time of data collection. Only results from presidential primaries are included in 2020. 50% of all observations in the dataset are from 2018 because there was a strong focus on the hundreds of races that would determine whether Democrats capture control of the U.S. House of Representatives and protect their vulnerable Senate seats; the congressional and senatorial observations from this year make up 31% and 8% of the dataset, respectively. The other key races of interest in that year were statewide (largely gubernatorial) elections across dozens of states. 2015, 2017, and 2019 have very few observations because they are non-federal election years, but some states and cities hold their statewide and municipal elections in off-cycle years. Furthermore, special elections for congressional and senatorial seats can be held in off-cycle years when the incumbent vacates the seat unexpectedly due to resignation or death. A t-test for the difference in the means of prediction error between general presidential elections and other types of elections shows that they are associated with less prediction error, confirming the finding from J. Berg et al. (2003).

In Table 4, I present summary statistics about the independent variables that describe election

and market characteristics in the dataset. For the numerical independent variables, these statistics are intended to give a sense of the central tendency, range, and skewness of each one. For the binary independent variables, the mean indicates what percent of the dataset is described by the variable. Summary statistics for the panel data days and Google search volume variables are not included because their values are normalized between market panels.

Table 4: Independent Variable Summary Statistics

	Mean	SD	25^{th} Pct	Median	75^{th} Pct	Min	Max
Progressive	0.29	0.45	0	0	1	0	1
Female	0.60	0.49	0	1	1	0	1
Minority	0.33	0.47	0	0	1	0	1
Enthusiasm	14.62	16.29	5.66	10.06	16.43	0.654	176.9
Duration of Trading	143.4	154.6	16	90	200	1	648
Availability of Polling	0.19	0.39	0	0	0	0	1
Volume	105,686	222,211	3,430	14,340	98,308	2	1,999,407
Number of Candidates	2.61	1.10	2	2	3	1	8
General	0.54	0.499	0	1	1	0	1
Primary	0.46	0.499	0	0	1	0	1
Tossup	0.16	0.37	0	0	0	0	1
Lean	0.19	0.39	0	0	0	0	1
Likely	0.27	0.44	0	0	1	0	1
Safe	0.38	0.49	0	0	1	0	1

Interestingly, the majority of elections (60%) in the dataset feature at least one major female candidate. This is much higher than the percentages for progressive (29%) and ethnic minority (33%) candidates. The summary statistics for the enthusiasm, duration of trading, and volume variables provide evidence for the assertion that they have rightward skews: their means of 14.62, 143.4, and 105,686 are all greater than their medians of 10.06, 90, and 14,340, respectively. Their standard deviations are also greater than their means, indicating a concentration of data in the higher ranges of the distribution. These results support the logarithmic forms of each variable. Unsurprisingly, the median number of candidates in an election is 2 because the U.S. has a two-party system and the majority (54%) of elections in the dataset are general elections between a Republican and a Democrat. The 46% of observations in the dataset that represent primary elections tend to have a higher number of candidates. The values range from 1 in uncompetitive elections with only one major candidate to 8 as seen in one of California's multiparty jungle primaries. These outliers represent an extremely small percentage of the dataset and the mean, SD, and median are not consistent with a significant rightward skew, so a logarithmic form would not be appropriate for this variable. Polling is easily available on RealClearPolitics' website for approximately 19% of elections. Finally, 16% of elections in the dataset are coded as tossups, while 19%, 27%, and 38% are coded to the progressively less competitive lean, likely, and safe categories.

5. Identification Strategy and Methods

5.1 Identification Strategy

A key identifying assumption required to obtain unbiased estimators in the cross-sectional ordinary least squares (OLS) setting is that all variables that confound the relationship between the independent variables of interest and the dependent variable have been controlled for. Any such variables that are not included in the regression will cause omitted variable bias (OVB) in the coefficients of my independent variables. In reality, it is not possible to include all such variables in a regression due to the risk of having an overspecified model and constraints regarding the availability of data. Therefore, researchers need to exercise judgment in deciding which control variables to include based on knowledge about the underlying subject and considerations about which potential sources of OVB are most likely to threaten the identification of coefficient estimates. My primary justification for the inclusion of specific variables in the regression is in Section 4, while in this section I offer a general explanation of my reasoning for the inclusion of variables from four summarized categories. Given the inherent complexity of U.S. elections and prediction market forecasting, it will be instructive to group potential sources of OVB and independent variables into the following groups:

1. **Candidate Controls.** Characteristics of the major candidates who run in the elections. From this group, I include variables capturing whether a candidate is progressive, female, or an ethnic minority. In my estimation, these are the most salient demographic cleavages in U.S. society that could affect traders' decisions when making predictions about candidates. An example of an excluded control is candidate religious affiliation since this characteristic has mostly fallen out of focus in contemporary U.S. politics, with Americans being "now nearly universally willing to vote for a Jewish or Catholic presidential candidate" (McCarthy, 2019). This research indicates that Americans are much less willing to vote for Atheist or Muslim candidates, but these candidates are exceedingly rare, with only three Muslims and "no known atheists" in the U.S. Congress as of 2019 (Smith, 2019). In contrast, ideology, gender, and ethnicity continue to attract attention as salient cleavages in U.S. politics and they remain powerful determinants of voter perceptions, as seen in the literature review.

2. **Election Controls.** Characteristics of the elections themselves. I include variables capturing enthusiasm, election year, election type, the number of candidates, the availability of polling, and inherent competitiveness. Each of these variables captures an important feature of the election that is likely to be correlated with the independent variables and prediction error. An example of an excluded control is absolute fundraising amounts. Since political science literature "has not able to conclusively establish a causal connection" between spending and voting behavior, I assess that fundraising and spending are unlikely to be correlated with prediction error after other election characteristics are controlled for (Dawood, 2014).

3. **Electorate Controls.** Characteristics of the districts or states where elections are held. I list all such characteristics in Section 4. Generally, they were chosen based on an expectation that they will be correlated with the nomination of progressives, women, and ethnic minorities and candidates and the history of demography-based polling errors (and therefore prediction error) in U.S. elections. An example of an excluded control is the percentage of the district that is Asian-American since this group is an extremely small proportion of the American electorate and has not historically been associated with systematic polling errors.

4. **Market Controls.** Characteristics of the markets themselves. I include total trade volume and the duration of trading. As argued earlier, these are likely to be correlated with the independent variables and prediction error. An example of excluded controls are the demographic characteristics of traders. I requested but not could obtain this data from PredictIt in time to write the thesis. In any case, the demographic traits of traders could be bad controls if, for example, the presence of a progressive candidate attracts progressive-leaning traders. This effect could be a channel through which the progressive variable impacts prediction error through wishful thinking bias, so its omission is appropriate.

OVB is not the only threat to identification in cross-sectional regressions. Including control variables that are proxies or contain similar information to independent variables of interest can attenuate the magnitude and significance of their coefficients. For example, including a control for the ideological scores of candidates would likely take away some of the explanatory power of the progressive variable. Similarly, including bad controls (i.e. mediator variables) can also bias coefficients by capturing some of the effect of independent variables. If a control could be the outcome of an independent variable, then it is likely to be a bad control; a potential example of this is if voter enthusiasm is an outcome of a candidate being a progressive, female, or minority. This could especially be a problem for the progressive variable, since its correlation coefficient with enthusiasm is 0.27, as compared to 0.09 and 0.02 for female and minority, respectively. By including enthusiasm in the regression, I potentially introduce bias if enthusiasm is a channel through which the progressive variable impacts prediction error. The decision about whether to include this variable in the same regression as the other independent variables comes down to a tradeoff between omitted variable bias (if enthusiasm from other sources is correlated with the independent variables) and bias created by controlling for a mediator variable. In Section 7, I test the robustness of the results from my main specifications to models where enthusiasm is excluded as an independent variable; I decided to run this test after I ran my analyses for the first time.

I address arguments that other variables could be bad controls for the independent variables of interest below. Election year, election type, the number of candidates, and the availability of polling are immutable characteristics or are essentially determined by the level and type (e.g. presidential vs. senatorial) of election, not by independent variables of interest. The competitiveness control variables and trade volume could certainly be an outcome of the progressive, female, minority, and enthusiasm variables. But since I am trying to measure the effects of these independent variables on prediction

error caused by *bias*, it makes sense to control for the channels of actual competitiveness and trade volume. If progressive candidates tend to cause more competitive elections, then the channel of increased prediction error is competitiveness, not bias. Similarly, if progressives attract more trade volume and that improves prediction accuracy due to the greater pool of information among traders, then this is also a channel other than bias. Therefore, I improve measurement of the coefficient of interest by controlling for these unrelated channels. All of the district characteristic controls are inherent features and not caused by the independent variables. Finally, the duration of trading in the market is determined by the level and type of election, not independent variables of interest.

In the panel data setting, more robust estimation techniques exist to help with debiasing coefficients. Unfortunately, these techniques are only effective for time-variant independent variables of interest, not time-invariant ones like progressive, female, and ethnic minority. Since I plan to test Hypotheses 5b. and 6, I will also proceed with running panel data models. Typically, when working with panel data, the choice is between a fixed effects (FE) specification and a random effects (RE) specification (Wooldridge, 2012). Fixed effects models work by calculating the means of the dependent and independent variables over the life of the panel and subtracting them from the model, thereby creating demeaned versions of each variable. Since time-invariant factors are fixed over the duration of the panel, their value each period is equal to their mean and their demeaned value is 0. They are therefore dropped from the model when using fixed effects. This makes fixed effects models quite effective at debiasing estimators since they account for all unobserved time-invariant factors specific to each panel unit that could confound the relationships between the dependent and independent variables, not only the ones I controlled for.

Since unobserved time-invariant heterogeneity is not a problem in this setting, the key remaining assumption for the identification of time-variant independent variables in fixed effects models is that all time-variant sources of heterogeneity are included in the model. This is a fairly plausible assumption in this scenario since the panels are run over 15 days and there should be minimal variation in underlying factors that affect prediction error; characteristics of the candidates, elections, districts, and markets should essentially be fixed over such a short time period. Exceptions are likely to include the reduction in uncertainty that comes as Election Day gets closer and changes in the availability of information, such as from a candidate scandal. I capture these effects with the days variable and the Google search volume variable. I note that it is common in panel data regressions to use two-way fixed effects models that account for unobserved time period-specific heterogeneity as well as unit-specific heterogeneity. An example would be if an economic recession in some time period affected all elections in my dataset; in this case, including time period-specific fixed effects would drop out this source of heterogeneity that affected all units in a given time period. While two-way fixed effects models tend to be more effective at debiasing coefficients than simple unit-specific fixed effects models, they would not be appropriate in this setting. An unusual feature of my dataset is that, while the panels are balanced at 15 days each, they are run over different time periods. Some election panels are run in the 15 days leading up to election day in 2016, while others are run in the 15 days before election

day 2018, and many other dates. Therefore, there are not common shocks like recessions that affect all observations that would necessitate the use of a two-way fixed effects model. Election date is a time-constant characteristic of an election that varies between panels, so time period-specific shocks would be accounted for with the regular unit-specific fixed effects estimator.

But for my purposes, the main disadvantage of FE is that my first four explanatory variables would be dropped from the model. They are time-invariant and therefore perfectly collinear with fixed effects, so it is impossible to include and estimate them. The typical alternative specification in this scenario is the random effects model, which would allow me to estimate the regression including these variables. For random effects models, the assumption about including all time-variant sources of heterogeneity is still required but remains plausible for reasons discussed above. More critically, the additional RE assumption is that the independent variables of interest are uncorrelated with unobserved heterogeneity in all time periods. That is to say that differences in the mean and variance of Google search volume within and between panels must not be driven by unobserved variables. This is a similar assumption to the one from cross-sectional regressions, so the possibility of OVB still threatens identification. For this reason, the ability of fixed effects estimators to drop all unobserved sources of time-invariant heterogeneity makes them a more attractive choice to estimate the coefficients of time-variant variables than random effects estimators.

Instead of either the traditional FE or RE estimator, I use the correlated random effects (CRE) estimator to obviate my concerns about each of them. The CRE approach is a mix between the fixed effects and random effects methods that allows me to drop out all sources of time-invariant unobserved heterogeneity that affect my time-variant variables, as in the FE model, while still keeping and estimating the time-invariant independent variables of interest, as in the RE model (Wooldridge, 2012). Instead of assuming that the time-variant variables of interest are uncorrelated with unobserved heterogeneity as required by the latter model, the CRE model directly models the correlation that could exist between them. This is done by generating an average value for each time-variant independent variable over the life of the panel and including it as a control. This control debiases the time-variant coefficients by partialing out their mean value: that is, the control demeans the variable, just like in a fixed effects analysis. This procedure should produce identical estimates to the FE approach for the time-variant variables.

The fixed effects analysis being performed in the CRE model does not work for the time-invariant independent variables, though, since their demeaned value is 0. Since it is impossible to perform the FE analysis for these constant variables, it is still necessary to include the full set of controls from the cross-sectional regression in order to argue that their coefficients are unbiased. The coefficient estimates for the time-invariant variables will be at least somewhat different from the cross-sectional estimates since the composition of observations has changed. The effects of this change on my power to detect effects is ambiguous, since the panel analysis drops 139 markets but gains thousands of new observations that are highly correlated within their clusters. Since the dropping of observations could lead to selection bias if the omission of observations is correlated with the independent and

dependent variables, in Section 7 I run a robustness check to test whether the omission changes the conclusions about the independent variables of interest in my cross-sectional regressions. This also helps to make a cleaner comparison between the cross-sectional and panel data models. I note that I made the decision to run this robustness test after initially running the regressions.

I note two additional considerations regarding my choice of panel data models. Typically, researchers who choose random effects over fixed effects will point to the results of a Hausman test to justify their choice; the null hypothesis of this test is that the error term is uncorrelated with independent variables (i.e. the key RE assumption is appropriate) as evidenced by the lack of a statistically significant difference between FE and RE coefficients (Wooldridge, 2012). As I mentioned before, the FE and CRE approaches should produce identical coefficients; for each model, I run a Hausman test as a formality in Section 6 and demonstrate that this is the case. The second consideration is that an assumption of fixed effects analyses (and CRE analyses by extension) is that the treatment effect of the independent variable is homogenous between units; fixed effects estimates are biased in the presence of heterogeneous treatment effects because this leads to the misweighting of observations in the calculation of fixed effects coefficients (Gibbons et al., 2018). This would be violated if the treatment effect of days on prediction error differed significantly depending on whether it is, for example, the 2016 Democratic primary in New Hampshire or the 2018 senate race in Texas being treated. In the panel data segments of Section 6, I will test for the possibility of bias from heterogeneous treatment effects using the test developed by Gibbons et al. (2018). The test compares the coefficients generated from a regular fixed effects analysis to coefficients from a heterogeneous treatment effects-robust fixed effects estimator to see if there is a significant difference. I note that I decided to run the Hausman and the heterogeneous treatment effects tests after initially running regressions.

Now that the identification strategy has been adequately defined, I will specify my models and methods in greater detail. The coefficient outputs for each model will be subject to a two-tailed test for statistical significance at the 5% level in order to assess the strength of evidence obtained about the hypothesis. That is, if the p-value of a coefficient is less than 0.05, then I consider that sufficient evidence to reject the null hypothesis that $\beta_j = 0$.[3]

[3]Since I am running 22 significance tests at the 5% level, it would be unsurprising if one of them appeared to be significant by random chance under the assumption that the variable outcomes are uncorrelated. When multiple tests are being run, it is sometimes appropriate to use the Bonferroni correction to account for the chance of a false positive. However, the Bonferroni correction is inappropriate in instances where the significance of one test is correlated with the significance of another test; this is clearly the case in my regressions since many variables repeat between the models. Since observing significance for a variable in one model raises the chance of observing significance for that variable in another model, I will not use the Bonferroni correction: it would be excessively conservative and increase the chance of observing a false negative.

5.2 Models

Cross-sectional Linear Probability Model Specification

$$y_m^b = \overbrace{\beta_1 r_m + \beta_2 f_m + \beta_3 e_m + \beta_4 \log \frac{s_m}{n_m} + \beta_{5a} \log d_m}^{\text{Independent Variables of Interest}} + \overbrace{a_m + \log v_m + k_m}^{\text{Important Controls}} + x_m + \epsilon_m$$

In this iteration of the model, I use OLS to run a Linear Probability Model with clustered standard errors. I regress the binary dependent variable on the predictors. Discussion about the clustering of standard errors can be found at the conclusion of this section. In regressions with a binary dependent variable interpreted as a probability, the choice is between running a linear probability model (LPM) and a probit/logit model, each of which has advantages and disadvantages (Wooldridge, 2012). The LPM is popular because the coefficients it produces come with an intuitive interpretation. The main disadvantage is that the LPM can produce predicted dependent variable values greater than 1 or less than 0, which are impossible as probabilities. While probit/logit models do not have the same disadvantages, their coefficients are difficult to properly interpret. However, the LPM's disadvantage is less problematic in settings where the main research focus is on the average predicted effect of an independent variable on probabilities, like in this thesis, as opposed to when the focus is on the appropriateness of the predicted probabilities themselves (Chatla & Shmueli, 2013). For this reason, I proceed with the LPM.[4]

Cross-sectional Continuous Specification

$$y_m^c = \overbrace{\beta_1 r_m + \beta_2 f_m + \beta_3 e_m + \beta_4 \log \frac{s_m}{n_m} + \beta_{5a} \log d_m}^{\text{Independent Variables of Interest}} + \overbrace{a_m + \log v_m + k_m}^{\text{Important Controls}} + x_m + \epsilon_m$$

In this iteration of the model, I use OLS with clustered standard errors. I regress the continuous dependent variable on the predictors. Since this variable is no longer interpretable as a probability, there is no choice to make between the LPM and the probit/logit models. While impossible predicted outcomes outside of the 0 to 1 range could still occur, the justification used above remains valid.

Panel Data Linear Probability Model Specification

$$y_{m,t}^b = \overbrace{\beta_1 r_m + \beta_2 f_m + \beta_3 e_m + \beta_4 \log \frac{s_m}{n_m} + \beta_{5b} d_{m,t} + \beta_6 g_{m,t}}^{\text{Independent Variables of Interest}} + \overbrace{a_m + \log v_m + k_m}^{\text{Important Controls}} + x_m + \epsilon_{m,t}$$

In this iteration of the model, I run a panel data version of the LPM with clustered standard errors. I regress the binary dependent variable on the predictors. As discussed above, I employ the correlated random effects estimator.

[4] After running the regressions, I also confirmed that the incidence of predicted values outside of the unit interval is quite rare in all four main specifications.

Panel Data Continuous Specification

$$y^c_{m,t} = \overbrace{\beta_1 r_m + \beta_2 f_m + \beta_3 e_m + \beta_4 \log \frac{s_m}{n_m} + \beta_{5b} d_{m,t} + \beta_6 g_{m,t}}^{\text{Independent Variables of Interest}} + \overbrace{a_m + \log v_m + k_m}^{\text{Important Controls}} + x_m + \epsilon_{m,t}$$

In this iteration of the model, I run a panel data version of my model with clustered standard errors. I regress the continuous dependent variable on the predictors. I once again opt to use the CRE estimator.

5.3 Standard Errors

The final component of my model to discuss is my approach to standard errors. Using heteroskedasticity robust standard errors is often insufficient to account for the relationships in the error terms that exist between observations in the presence of clustered observations. But the traditional guidance about how to cluster standard errors often leads to flawed or ambiguous conclusions. In their seminal paper on clustered standard errors, Abadie et al. (2017) argue that the traditional approach of deciding how to cluster standard errors on the basis of whether there are expected to be correlations between the error terms of observations is incorrect: "Typically the stated motivation is that unobserved components of outcomes for units within clusters are correlated... We take the view that clustering is ... either a sampling design or an experimental design issue." For this thesis, there was no sampling performed at a clustered level, so the experimental design issue is what is relevant: "clustering can also be an experimental design issue, when clusters of units, rather than units, are assigned to a treatment." (Abadie et al., 2017) I argue that the assignment of the progressive, female, and ethnic minority treatments were made to clusters in certain cases. For example, the assignment of female was made at the level of the 2016 Democratic presidential primaries since the fact that Hillary Clinton participated in the Iowa caucus in the cluster also implies that she will participate in other ones in the cluster, like the New Hampshire primary. In other words, there is no variation in the independent variables within this cluster because there cannot be: the assignment mechanism applied the treatment to all of these observations simultaneously. Similar arguments can be made for the 2016 Republican presidential primaries, the 2016 general election, and the 2020 Democratic presidential primaries.

Clearly, the proposed clusters are extremely different in size. In the cross-sectional dataset, the 2016 Democratic primary has 49 observations, the 2016 Republican primary has 46 observations, the 2016 general election has 51 observations, and the 2020 Democratic primary has 26 observations. Then the other hundreds of single-observation "clusters" are essentially considered independent in the cross-sectional versions of the model. I discuss the applicability of clustering to the panel data model in the final paragraph. The presence of single-observation clusters make it impossible to calculate some model-level statistics, e.g. F stats, due to the lack of variation within the clusters, but singletons alone do not threaten inference. However, the variability in size violates one of the key assumptions

of using the cluster robust variance estimator: "CRVE is consistent under three key assumptions:... A3. Each cluster contains an equal number of observations." (Mackinnon & Webb, 2016). As an alternative suggestion for valid inference despite the violation of this assumption, the authors propose using the wild cluster bootstrap procedure: "Section 4 presents Monte Carlo evidence using simulated datasets with a continuous test regressor and either equal cluster sizes or ones proportional to state populations. We show that inference based on CRVE t statistics can perform poorly in the latter case... In contrast, the wild cluster bootstrap procedure always performs extremely well." For the cross-sectional models, this can be implemented easily in Stata with the *boottest* command. However, the *boottest* command does not currently support the *re* function of the *xtreg* command that I use to run the correlated random effects models. In these cases, I instead use the *bootstrap* command in order to specify the correct standard errors generation approach.

Finally, I note that the large clusters based on the treatment assignment and bootstrapped standard errors approach will remain in the panel data model. Additionally, each election not contained in a bigger cluster will become its own cluster of 15 market-day observations, which accounts for autocorrelation between observations in a panel.

6. Results and Interpretation

This section presents regression results for each iteration of the model. The output featured includes estimated coefficients and p-values for the independent variables of interest. For the cross-sectional models in Tables 5 and 6, I first run the regression without any control variables and include the results in column (1). To get a sense for how the different groups of control variables identified in Section 5 affect the coefficients, in each subsequent column I add a group of controls. Column (2) features regression results including the Election Controls group. Column (3) adds the Electorate Controls group. Finally, column (4) adds the Market Controls group and represents the full model. For the panel data models in Tables 7 and 8, I first run the regression as a simple fixed effects regression in the (FE) column to prove the equivalence of its estimators with the correlated random effects (CRE) estimators; this regression necessarily excludes all the fixed independent variables. Next, I run the CRE model with all of the independent variables of interest and without controls and include the results in (CRE1). The next three columns follow the same pattern as the cross-sectional models, with columns (CRE2), (CRE3), and (CRE4) adding the Election Controls, the Electorate Controls, and the Market Controls, respectively, to arrive at the full model in the lattermost column. I note that the results in columns (1) - (3) in Tables 5 and 6 and (FE) - (CRE3) columns in Tables 7 and 8 are only included to facilitate discussion: any statistically significant results from these columns will not be interpreted and are not considered to be evidence about the hypotheses, since they are incomplete models. I note that I decided to include these additional columns of results after initially running my regressions.

For each independent variable of interest, I include discussion about the sign, magnitude, and statistical significance of the coefficient estimate obtained. If coefficients are measured to be statistically significant at the 5% or the 1% level, they are superscripted with one or two asterisks, respectively. I do not report standard errors because they are based on the assumption that the underlying distribution is approximately normally or t-distributed. Bootstrapping imposes no assumption of normality, so the *boottest* command in Stata does not report standard errors. In addition, the *boottest* 95% confidence intervals are not symmetrically distributed around the point estimate, so it would be impossible to impute standard errors from the confidence interval.

6.1 Cross-sectional Binary Model Regression Output

Table 5: Cross-sectional Binary DV Specification Results

	Incremental Models			Full Model
	(1)	(2)	(3)	(4)
Progressive (H1)	0.053	−0.024	−0.022	−0.023
	(0.77)	(0.53)	(0.57)	(0.55)
Female (H2)	−0.053	−0.056	−0.056	−0.056
	(0.15)	(0.11)	(0.11)	(0.11)
Minority (H3)	−0.012	0.025	0.039	0.039
	(0.82)	(0.48)	(0.31)	(0.32)
(Log) Enthusiasm (H4)	0.035	−0.001	−0.001	−0.002
	(0.10)	(0.95)	(0.95)	(0.92)
(Log) Duration of Trading (H5a)	0.002	−0.008	−0.005	−0.007
	(0.92)	(0.51)	(0.65)	(0.62)
Election Controls?	No	Yes	Yes	Yes
Electorate Controls?	No	No	Yes	Yes
Market Controls?	No	No	No	Yes
Observations	570	570	570	570

p-values in parentheses

$^{*}\, p < 0.05, \quad ^{**}\, p < 0.01,$

It is readily apparent from the results in Table 5 that this regression does not offer evidence to support any of the hypotheses. A p-value lower than 0.05 would be required to conclude that I have obtained some evidence in favor of the hypothesis, given my statistical significance threshold of 5%. As indicated by the p-values on the independent variables of interest in column (4), which range from 0.11 to 0.92, I am unable to reject the null hypothesis that $\beta_j = 0$ for any of them. As these coefficients are essentially indistinguishable from 0, it is not informative to include discussion about their magnitude and signs. There were no changes in significance caused by the inclusion or omission of control groups. The magnitudes of some variables appeared to change by quite a lot, but discussion about these changes, which recur to some extent in the other specifications, will be more instructive in the subsequent sections.

The failure to measure statistically significant coefficients on the independent variables of interest could be for a number of reasons. One is the possibility that the true effect of each independent variable of interest is in fact $\beta_j = 0$, in which case the results are appropriate. Another possibility is that there was insufficient power to detect an effect that may actually exist because of too few observations or insufficient variation in the dependent variable: in this specification, only 131 out of 570 observations are coded as incorrect predictions, which is a small comparison group to pair with the 439 correct prediction observations. A third possibility is that the coefficients are being biased towards zero by the omission of unobserved variables or the inclusion of certain highly collinear control variables, as I discussed in Section 5. In the following specification, I attempt to address the challenge posed by the lack of dependent variable variation.

6.2 Cross-sectional Continuous Model Regression Output

Table 6: Cross-sectional Continuous DV Specification Results

	Incremental Models			Full Model
	(1)	(2)	(3)	(4)
Progressive (H1)	0.041	0.018	0.023	0.021
	(0.58)	(0.32)	(0.20)	(0.23)
Female (H2)	−0.043	−0.047*	−0.046*	−0.045*
	(0.09)	(0.03)	(0.02)	(0.02)
Minority (H3)	0.003	0.019	0.046*	0.046*
	(0.93)	(0.31)	(0.04)	(0.04)
(Log) Enthusiasm (H4)	0.040*	0.011	0.013	0.009
	(0.01)	(0.30)	(0.26)	(0.42)
(Log) Duration of Trading (H5a)	0.007	0.001	0.003	−0.003
	(0.46)	(0.91)	(0.60)	(0.74)
Election Controls?	No	Yes	Yes	Yes
Electorate Controls?	No	No	Yes	Yes
Market Controls?	No	No	No	Yes
Observations	570	570	570	570

p-values in parentheses

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

The p-values corresponding to every coefficient in column (4) of Table 6 got closer to 0 once more variation in the dependent variable was introduced, with the exception of the duration of trading variable. In particular, the p-values for the coefficients on the female and minority independent variables dropped below the significance threshold of 0.05, which allows me to reject the null hypotheses that $\beta_2 = 0$ and $\beta_3 = 0$ in favor of the alternatives that $\beta_2 \neq 0$ and $\beta_3 \neq 0$. However, in the case of the female variable, the significance of this coefficient does not support Hypothesis 2 and is not consistent with prior research about perceptions of electability as cited in my literature review: the sign on the coefficient is negative, indicating that election observations having female candidates are associated with less prediction error. One explanation could be the findings from Lawless and Pearson (2008) about female candidates who run for office being different from male candidates in ways that could affect actual or predicted success. Another possibility is that the general observation of Americans' misconceptions about female candidates does not extend to the unrepresentative subgroup of traders who use PredictIt; to the contrary, the presence of female candidates could garner interest and motivate traders to become more informed about the election they are betting on, which results in better predictions, perhaps due to normative beliefs about the representation of women in politics. Another noteworthy observation about this variable is its remarkably consistent magnitude (ranging from -0.043 to -0.047) as it progresses through the incremental models: it seems like its magnitude is not dependent on the inclusion or exclusion of any particular control variables, although the coefficient did not attain significance until the second specification. But one should still exercise additional skepticism about this result: since the sign of the coefficient goes against

the prior developed from existing literature, this increases the likelihood that the result is a false positive, i.e., that the statistically significant effect observed is due to random chance. Keeping that in mind, the interpretation of the coefficient is that the presence of a female candidate in an election decreases the expected amount of prediction error by 0.045, ceteris paribus. The sign of the minority variable is consistent with Hypothesis 3 and with literature about Americans' misperceptions about electability, as it indicates that the presence of an ethnic minority candidate increases the expected amount of prediction error by 0.046, ceteris paribus. As such, this is suggestive evidence in favor of Hypothesis 3. In contrast with the female variable, the magnitude and significance of this variable varies quite a lot between the incremental models, which suggests this result is sensitive to the choice of control variables. Hypotheses 1, 4, and 5a are still unsupported due to the lack of statistically significant coefficients. The lack of evidence for wishful thinking bias from the coefficients for H1 and H4 is consistent with the conclusion from Forsythe et al. (1999) that market-level outcomes are not affected by wishful thinking bias despite bias in individual traders. Finally, the lack of significance for the duration of trading variable is unsurprising given the mixed or null evidence obtained by other researchers about the relationship between duration and prediction error, in particular over longer timespans. I now proceed with estimating the panel data models.

6.3 Panel Data Binary Model Regression Output

Table 7: Panel Data Binary DV Specification Results

	Incremental Models				Full Model
	(FE)	(CRE1)	(CRE2)	(CRE3)	(CRE4)
Progressive (H1)		0.113	0.075	0.080	0.079
		(0.06)	(0.10)	(0.08)	(0.09)
Female (H2)		−0.041	−0.046	−0.046	−0.045
		(0.23)	(0.18)	(0.20)	(0.22)
Minority (H3)		−0.021	0.031	0.051	0.053
		(0.61)	(0.39)	(0.21)	(0.22)
(Log) Enthusiasm (H4)		0.04*	0.002	−0.002	−0.005
		(0.04)	(0.94)	(0.93)	(0.82)
Days Until Election (H5b)	0.005**	0.005**	0.005**	0.005**	0.005**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Google Search Volume (H6)	0.000	0.000	0.000	0.000	0.000
	(0.35)	(0.27)	(0.16)	(0.16)	(0.18)
Time-Invariant Predictors?	No	Yes	Yes	Yes	Yes
Election Controls?	No	No	Yes	Yes	Yes
Electorate Controls?	No	No	No	Yes	Yes
Market Controls?	No	No	No	No	Yes
Observations	6,465	6,465	6,465	6,465	6,465

p-values in parentheses

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

For the purpose of facilitating a better comparison between the different iterations of the model,

I first run the panel data specification using the binary dependent variable as seen in column (CRE4) of Table 7. Compared directly with the first model, the conclusions from this specification are largely similar. The progressive, female, minority, and enthusiasm variables are statistically insignificant, so I have obtained no evidence for Hypotheses 1 - 4. The magnitudes of the female (-0.045) and minority (0.053) are quite similar to their magnitudes in previous specifications even while statistically indistinguishable from 0.

Furthermore, the panel specification allows me to test the time-varying Hypotheses 5b and 6. The p-value of 0.00 and sign of β_{5b} allow me to reject the null hypothesis that $\beta_{5b} = 0$ and offers fairly strong support in favor of Hypothesis 5b. Consistent with the hypothesis, one further day of distance from the election is associated with a predicted increase of 0.50% in the likelihood of prediction error, ceteris paribus, for a maximum of 15 total days. Therefore, the model expects that a prediction 15 days from election day is 7.0% more likely to be incorrect than one made 1 day before election day. This is consistent with the result from Restocchi et al. (2018) that, in the short run, prediction error does measurably decrease as election day gets closer. The p-value on the coefficient on the Google volume variable is insufficient to offer evidence in favor of Hypothesis 6. The results from H5b and H6 did not measurably change in magnitude or significance between any specification in Table 7; this is unsurprising because the FE and CRE estimators should produce essentially identical results for these variables.[5]

As mentioned in Section 1, my choice of specification for the panel data models changed from the original plan after I ran the regressions. The results for the original specification correspond to the (CRE1) column in Table 7. As can be seen, in that specification the enthusiasm variable was measured to be significant at the 5% level. This effect disappears with the inclusion of control variables in subsequent models and in the final model. The days variable was significant in the original specification and remains so in the full model, so no change is noted regarding that variable.

6.4 Panel Data Continuous Model Regression Output

As happened in the previous instance of switching from the binary to the continuous dependent variable, the p-values of most coefficients in column (CRE4) of Table 8 got closer to 0. This increase was only sufficient to offer evidence for H1: with its p-value of 0.03 and positive sign, these results allow me to reject the null hypothesis that $\beta_1 = 0$. The interpretation of this result is that the presence of a progressive candidate increases the expected amount of prediction error by 0.062, ceteris paribus. This result is consistent with my expectation that a wishful thinking bias effect might be measurable in PredictIt's markets due to potential inefficiencies in market structure but is inconsistent with

[5] As mentioned earlier, I also run a Hausman test between the FE and CRE models and the Gibbons et al. (2018) test to see whether the results for these time-varying coefficients are potentially affected by bias from heterogeneous treatment effects. The Hausman test concludes that the FE and CRE coefficients are essentially identical ($p = 0.99$), as should be the case, so I am unable to reject the null hypothesis that the CRE model can be used. I then used the *GSSUtest* command from Gibbons et al. (2018) to test whether the time-variant estimators are biased. For the days ($p = 0.90$) and Google volume ($p = 0.98$), I fail to reject the null hypothesis of no difference between the FE estimator and the heterogeneous effects robust FE estimator at the 5% level; therefore, I have not found evidence of bias from heterogeneous treatment effects.

Table 8: Panel Data Continuous DV Specification Results

	Incremental Models				Full Model
	(FE)	(CRE1)	(CRE2)	(CRE3)	(CRE4)
Progressive (H1)		0.087	0.062*	0.066*	0.062*
		(0.06)	(0.03)	(0.02)	(0.03)
Female (H2)		−0.033	−0.039	−0.039	−0.035
		(0.18)	(0.06)	(0.06)	(0.09)
Minority (H3)		−0.018	0.016	0.030	0.035
		(0.58)	(0.50)	(0.24)	(0.18)
(Log) Enthusiasm (H4)		0.04**	0.009	0.007	−0.001
		(0.00)	(0.52)	(0.65)	(0.96)
Days Until Election (H5b)	0.004**	0.004**	0.004**	0.004**	0.004**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Google Search Volume (H6)	0.000	0.000	0.000	0.000	0.000
	(0.25)	(0.16)	(0.06)	(0.07)	(0.07)
Time-Invariant Predictors?	No	Yes	Yes	Yes	Yes
Election Controls?	No	No	Yes	Yes	Yes
Electorate Controls?	No	No	No	Yes	Yes
Market Controls?	No	No	No	No	Yes
Observations	6,465	6,465	6,465	6,465	6,465

p-values in parentheses

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

Forsythe et al., 1999's findings about wishful thinking bias. The magnitude and significance of the progressive variable through the incremental models was relatively consistent, although it did not attain significance until the second specification. No further evidence in favor of H2, H3, H4, and H6 has been obtained. In particular, it is noteworthy that the coefficients on female and minority are not statistically significant, in contrast with the results in the cross-sectional continuous specification. The change in statistical significance for the progressive, female, and minority variables between specifications could be an indication that they are sensitive to model specification, and the previous significant results could be false positives, or that their relationship with prediction error depends on the timespan of measurement (i.e. at a moment in time vs. an average over time and over the long run vs. the short run).

The interpretation of the H5b coefficient now changes due to the change in the dependent variable. One further day of distance from the election is associated with a predicted increase of 0.004 in prediction error, ceteris paribus, for a maximum of 15 days. The model expects that a prediction 15 days from election day will have an error score 0.056 greater than one made 1 day before election day.[6]

Again, the results for the original specification of this model correspond to the (CRE1) column in Table 8. As can be seen, in that specification the enthusiasm variable was once again measured to be

[6]The result for the Hausman is still as expected (p = 0.99). The Gibbons et al. (2018) test again fails to reject the null hypothesis of no difference between the FE estimator and the robust estimator at the 5% level for the days ($p = 0.91$ and Google volume ($p = 0.93$) variables; therefore, there is still insufficient evidence of bias from heterogeneous treatment effects.

significant at the 1% level. This effect disappears with the inclusion of control variables in subsequent models and in the final model. The days variable was significant in the original specification and remains so in the full model, so no change is noted with regard to that variable.

6.5 Summary and Analysis

The hypothesis that has the strongest support from these specifications is Hypothesis 5b. The coefficient for this hypothesis was strongly statistically significant in both panel specifications, the result is consistent with previous literature, and concerns about unobserved heterogeneity affecting the outcome are minimal. In terms of economic significance, this result has a meaningful magnitude but is unlikely to dramatically change one's interpretation of a forecast. Bettors or other users of prediction market forecasts could use this information to marginally update their expectation about the likelihood of the forecast being wrong depending on when the prediction is made. To aid in the interpretation and visualization of results, I include plots of the relationship between the days variable and prediction error in Figures 1 and Figure 2.

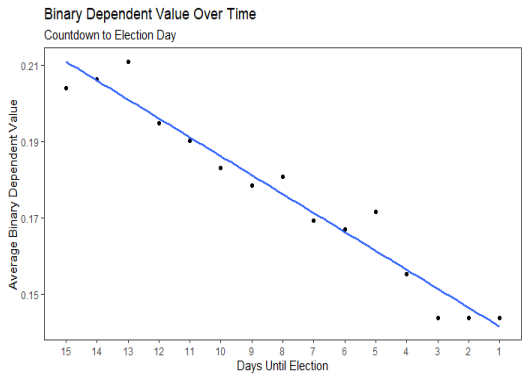

Figure 1: Binary DV Relationship

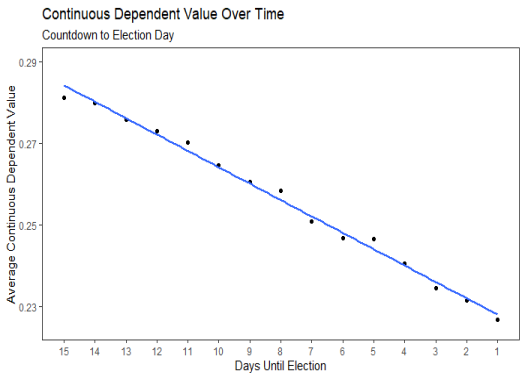

Figure 2: Continuous DV Relationship

As is visible from the figures, the relationship between days until the election and prediction error looks straightforward. The scatter in the binary dependent variable plot looks somewhat noisier, likely because of the discrete nature of the variable and the sharp cutoff at $p_{m,t} > 0.50$, but the relationship still appears to hold. Importantly, the linear functional form of the variable seems appropriately specified over the 15 day range since the observations are largely distributed evenly around the line of best fit.

The significant results for the progressive, female, and minority variables are at similar levels of magnitude but, if they are to believed, are more interesting than the result for 5b: they represent potential evidence about a cognitive bias in traders' decision-making corresponding to candidate characteristics. This could be valuable information when trying to assess the reliability of existing prediction markets or trying to profit from them based on the irrational behavior of other traders. But it is important to keep in mind that the support for these hypotheses is mixed, at best. The coefficient for H1 was only significant in the continuous panel data specification and its magnitude

was highly inconsistent between the binary (-0.023 to 0.079) and continuous (0.021 to 0.062) models. Beyond that, the finding of an effect of wishful thinking bias would go against previous literature. The coefficient for H2 was only significant in the continuous cross-sectional specification and its magnitude was remarkably consistent between binary (-0.045 to -0.056) and continuous (-0.035 to -0.045) models, but no support has been earned for the hypothesis since the sign did not go in the expected direction. The coefficient for H3 was also only significant in the continuous cross-sectional specification and its magnitude was also consistent between binary (0.039 to 0.053) and continuous (0.035 to 0.046) models. The minority result is consistent with the political science literature about misperceptions in electability, while the female result is not; however, neither of these hypotheses have been specifically tested in prediction markets. In any case, concerns about OVB and the inconsistency of significance between specifications remains for all of these judgment bias-related hypotheses, so attempts to distinguish the strength of evidence between results are somewhat futile. Finally, I was unable to observe significant coefficients for Hypothesis 4, Hypothesis 5a, and Hypothesis 6 in any specification, so these hypotheses have earned no support. The lack of support for H4 and H5a are consistent with mixed or null results about the duration of trading and wishful thinking bias from previous research. Hypothesis 6 is relatively novel but its lack of significant results tends to lend support to the literature suggesting that the arrival of new information may not improve forecasts.

7. Robustness Checks

In the following robustness check, I determine whether coding the continuous dependent variable as a Brier score instead of an absolute error score would have changed the conclusions made about statistical significance for the second and fourth specifications in Section 6.

7.1 Brier Score Robustness Check

Table 9: Brier Score Continuous DV Results

	Cross-sectional Coefficient	Panel Coefficient
Progressive (H1)	0.013	0.055*
	(0.44)	(0.04)
Female (H2)	−0.033*	−0.027
	(0.04)	(0.15)
Minority (H3)	0.036	0.042
	(0.07)	(0.07)
(Log) Enthusiasm (H4)	0.003	−0.007
	(0.77)	(0.58)
(Log) Duration of Trading (H5a)	0.001	
	(0.91)	
Days Until Election (H5b)		0.003**
		(0.00)
Google Search Volume (H6)		0.000
		(0.08)
Observations	570	6,465

p-values in parentheses

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

Upon inspection of the cross-sectional results, all of the coefficient estimates sizes got smaller in magnitude as compared to the main specification. This is to be expected: since all of the absolute error score observations are between 0 and 1, the Brier score's squaring function will make each dependent variable observation smaller. The coefficient estimate is the mean change of an observation's dependent variable associated with a one-unit change in an independent variable after partialing out other effects, so it will look smaller in the Brier specification. What matters more for measuring an effect is how much standard errors change in comparison. While standard errors would also be expected to get smaller, their change may be disproportionate to that of the coefficient depending on how the dependent variable observations are distributed. As mentioned in Section 4, this is because the Brier score's squaring function has an unequal effect on error score observations on the spectrum from 0 to 1; the Brier transformation has a much bigger impact on error scores closer to 0 than on scores closer to 1. In general, how much standard errors change between the absolute and Brier specifications depends on how error scores are distributed and if they are skewed closer to 0 or 1.

These observations about Brier scoring will help in interpreting the findings of the robustness check and understanding why significance levels could change between specifications. In general, it appears that the p-values of each coefficient grew, indicating comparatively weaker evidence in favor of the claim that the effect is significant. The p-value for female indicates that the coefficient retained its significance, while the p-value for the minority coefficient grew by about 0.03 and is now above the 5% significance threshold. In other words, if the Brier score dependent variable had been used in the main specification, I would not have rejected the null hypothesis for the minority variable. Since these rather small changes in p-values can lead to different conclusions about statistical significance, this is further suggestive evidence about how sensitive statistical significance can be to choices made about model specifications and the coding of variables. Skepticism is therefore warranted in interpreting significant results based on a single or limited set of specifications, especially in the absence of pre-registration plans or other researcher commitments.

In the panel data results, the coefficient magnitudes once again decreased, as expected. The p-values increased for some and decreased for other coefficients but they did not cross any significance thresholds. The coefficients for the progressive and days until election variables retained their statistical significance. While this alteration would not have changed any conclusions in the main specification, the warning against making sweeping conclusions based on significance thresholds from one or a few specifications remains.

In the following robustness check, I determine whether omitting the 139 markets that were dropped in the panel data analysis would have changed the conclusions made about statistical significance for the cross-sectional specifications in Section 6.

7.2 Omission of Observations Robustness Check

Table 10: Cross-sectional Results with Omitted Observations

	Binary Coefficient	Continuous Coefficient
Progressive (H1)	−0.018	0.022
	(0.69)	(0.29)
Female (H2)	−0.049	−0.049*
	(0.22)	(0.02)
Minority (H3)	0.054	0.061*
	(0.24)	(0.01)
(Log) Enthusiasm (H4)	−0.003	0.014
	(0.91)	(0.26)
(Log) Duration of Trading (H5a)	0.005	0.022*
	(0.84)	(0.04)
Observations	431	431

p-values in parentheses

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

There were no major changes to report from the omission of the 139 markets on the binary spec-

ification. This was already the most underpowered specification, so it is unsurprising that significant results were not observed after the sample size declined. The coefficient estimates for some variables changed by a somewhat large amount; these changes are better addressed in the next paragraph.

In the continuous specification, coefficient magnitudes for the progressive and female variables hardly changed at all, while the changes for the minority, enthusiasm, and duration of trading variables were relatively big. Larger changes between the main and robustness specifications are suggestive that the magnitude of the relationship between the dependent and independent variable may have changed due to selection bias. With that in mind, the statistical significance of the female and minority variables did not change, so selection bias would not have changed the conclusions about those effects. The magnitude and p-value of the progressive variable did not change almost at all; that suggests that the finding of statistical significance for this variable in the final panel data specification was likely driven by increased power, not from selection bias, since the omission of markets by itself did not move this variable toward statistical significance. The magnitude of enthusiasm changed somewhat, but this had no effect on statistical significance. Finally, I note that the coefficient for duration of trading becomes significant at the 5% level after the omission of 139 markets. This is both intuitive and interesting: it makes sense that the magnitude and significance of this variable would be so strongly affected by selection bias because the criteria for omission of markets was the number of days; it is also interesting because it suggests that evidence of the hypothesized relationship exists only after a certain threshold of days is passed. While this variable does not appear in the panel specifications, this finding mirrors the statistically significant results for the days until election variable. Since the duration of trading variable was not significant elsewhere, I will give a sense of the magnitude of the coefficient here. Given the logarithmic form of the independent variable, I will be careful with interpretation. Strictly speaking, the coefficient indicates that a 1-unit increase in the log of the number of days in a market is associated with an increase of 0.022 in the predicted error score, ceteris paribus. To convert this into more meaningful terms, I will use examples from the distribution of the data. The median, 75^{th} percentile, and 95^{th} percentile number of days are 138, 267, and 476 in the newly restricted dataset; the log of each score is 4.93, 5.59, and 6.17, respectively. Therefore, my interpretation is that the predicted error score is 0.015 higher for a market at the 75^{th} percentile and 0.027 higher at the 95^{th} percentile, each in comparison to the median. This coefficient is not significant in the main models, so its interpretation and meaningfulness should be viewed skeptically.

7.3 Enthusiasm as a Bad Control Robustness Check

Table 11: All Results with Enthusiasm Omitted

	Cross-sectional Results		Panel Data Results	
	Binary Coefficient	Continuous Coefficient	Binary Coefficient	Continuous Coefficient
Progressive (H1)	−0.024	0.024	0.077	0.062*
	(0.52)	(0.16)	(0.09)	(0.02)
Female (H2)	−0.056	−0.045*	−0.045	−0.035
	(0.11)	(0.02)	(0.21)	(0.08)
Minority (H3)	0.039	0.045*	0.053	0.035
	(0.31)	(0.04)	(0.22)	(0.18)
(Log) Duration of Trading (H5a)	−0.007	−0.004		
	(0.65)	(0.67)		
Days Until Election (H5b)			0.005**	0.004**
			(0.00)	(0.00)
Google Search Volume (H6)			0.000	0.000
			(0.18)	(0.07)
Observations	570	570	6,465	6,465

p-values in parentheses

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

Interestingly, there are no noteworthy changes in coefficient magnitudes and p-values at all between the main specifications and these robustness specifications. This is despite the relatively strong correlation observed between the progressive and enthusiasm variables. This suggests that there is no significant bias induced in the coefficients of my independent variables of interest by the inclusion of enthusiasm as a potential mediator variable.

7.4 Competitiveness Controls Robustness Check

Since this robustness check involves two alternative specifications and significant results were never observed for the time-invariant variables in the binary models, I only present results for the continuous variables in Table 12.[7] In the columns under the *Omitting Competitiveness* subheader, I present results for the continuous cross-sectional and panel models omitting the competitiveness dummies as controls entirely. In the columns under the *Alternative Measure* subheader, I present results for the models using the past margin of victory as an alternative measure for expected level of competitiveness.

The conclusions for the cross-sectional models are essentially identical to those from the main models. The female and minority variables retain their significance at the 5% level and similar magnitudes. This suggests that the significant results for these variables are not particularly dependent on the inclusion or specification of variables measuring competitiveness, which seems consistent with

[7]However, I did confirm that the conclusions from the binary models would not have changed under these specifications.

Table 12: Alternative Competitiveness Specifications

	Omitting Competitiveness		Alternative Measure	
	Cross-sectional Coefficient	Panel Coefficient	Cross-sectional Coefficient	Panel Coefficient
Progressive (H1)	0.007	0.047	0.006	0.045
	(0.74)	(0.16)	(0.77)	(0.17)
Female (H2)	-0.040^{*}	−0.015	-0.043^{*}	−0.016
	(0.04)	(0.51)	(0.03)	(0.48)
Minority (H3)	0.046^{*}	0.037	0.047^{*}	0.037
	(0.04)	(0.22)	(0.04)	(0.22)
(Log) Enthusiasm (H4)	0.025	0.017	0.024	0.017
	(0.05)	(0.29)	(0.06)	(0.29)
(Log) Duration of Trading (H5a)	−0.017		-0.018	
	(0.14)		(0.11)	
Days Until Election (H5b)		0.004^{**}		0.004^{**}
		(0.00)		(0.00)
Google Search Volume (H6)		0.000		0.000
		(0.07)		(0.07)
Observations	570	6,465	570	6,465

p-values in parentheses

$^{*}\ p < 0.05, \quad ^{**}\ p < 0.01,$

the evidence cited in the literature review that these types of candidates do not tend to be less electable or otherwise be associated with the competitiveness of elections. The time-variant variables are unaffected by the alternative specifications, as they should be. By comparison, the progressive variable completely loses its significance in the panel data models. This is suggestive that the result for this variable is sensitive to the coding and specification of the competitiveness variable(s) and could be adversely impacted by the presence of measurement error or other issues. The impact of the competitiveness controls on the progressive variable is interesting given research about the relationship between ideology and competitiveness, such as in the median voter theorem (Holcombe, 2006).

8. Conclusion and Limitations

As alluded to in earlier sections, the purpose of this thesis is to contribute to the extant body of literature about the efficacy of prediction markets as a tool for forecasting political outcomes under conditions of questionable market efficiency. It will be helpful to summarize my findings and how they relate to existing literature. Although there are four main models, it is important to note that these specifications are not necessarily equally valid or convincing: I would argue that the continuous specifications are superior to the binary ones because they allow for more dependent variable variation and do not rely on the sharp threshold of 0.50 to determine whether a prediction is completely right or wrong. For these reasons, the lack of any statistically significant results for the time-invariant variables in the binary specifications is not especially surprising. But distinguishing between the merits of the cross-sectional and panel data models is more ambiguous. The cross-sectional models benefit from the inclusion of more markets but the panel data models have many more observations, though the observations are clustered within panels, which limits power. Perhaps the more meaningful difference between these models is that they provide evidence about the hypotheses in different settings: in the cross-sectional models, prediction error is calculated as an average across time, whereas in the panel data models, prediction error is calculated at a moment in time. I would therefore argue that neither type of model is inherently superior, but they both provide evidence about whether significant results are robust to being measured under different circumstances.

With that in mind, the progressive coefficient for Hypothesis 1 was significant in only one specification and insignificant in the other main specifications. Its inconsistent magnitude between specifications and the previous research evidence showing that wishful thinking bias does not tend to affect market-level outcomes is cause for additional skepticism about the meaningfulness of this result. The female coefficient for Hypothesis 2 was also only significant in one specification and benefited from a fairly consistent magnitude between models, although the observed sign went against the hypothesized direction in all of them, so no support for the hypothesis has been obtained. While the result went against my prior from political science research, it does not directly contradict previous research from prediction markets as I was unable to find anyone who had tested this specific hypothesis. The coefficient for H3 was also only significant in one specification and fairly consistent in magnitude and sign in the main specifications, although it lost significance in the continuous specification during the Brier score robustness check. Importantly, this result was consistent with my prior and also did not directly contradict any previous results from prediction market research. Although the insignificant coefficients for these hypotheses are unsurprising in the binary specifications, the observation of significant coefficients for these variables in only one of the two continuous specifications makes the evidence of their effects fairly weak. No significant coefficients were measured for Hypothesis H4, so I have gathered no evidence in favor of the enthusiasm effect. The lack of evidence for this effect is not surprising given the null results about wishful thinking bias in previous prediction market research. In summary, the evidence obtained about the judgment bias-based hypotheses was either nonexistent or

not especially strong as it suffered from sensitivity to specification and concerns about OVB. Further research into these possible effects is needed before making stronger claims about their existence and magnitude.

The results for the information flow-based hypotheses are much clearer. Hypothesis 5b earned strong support due to its consistent significance at the 1% level between both panel data models, despite the lack of dependent variable variation in the binary model that can make it difficult to measure significant effects. In addition, concerns about OVB for this coefficient are minimal due to the short duration of the panel and the correlated random effects specification, which partials out all sources of time-invariant unobserved heterogeneity. Furthermore, this result is specifically consistent with findings of previous research from Restocchi et al. (2018) about a reduction in bias as election day gets closer over short time spans. By contrast, the result of Hypothesis 5a earned no support due to the lack of significant coefficients; this is unsurprising given the mixed or null results about duration of trading over longer timespans from prior research. Finally, no evidence was obtained for Hypothesis 6 since no statistically significant coefficients were observed in either panel specification. This hypothesis has not been specifically tested in previous literature, but the null finding suggests that the heterogeneous interpretation of information and recency bias, among other efficiency issues, might cause traders to fail to benefit from new information.

I hope that two takeaways are clear to readers. The first is that the maximal efficiency of markets and the rationality of market participants are not guaranteed properties of prediction and other price-driven markets. Indeed, even published research that concludes that the markets are efficient tends to admit that the average trader does suffer from judgment bias, and that the conclusions of efficiency are dependent on there being enough so-called marginal traders to debias the outcome. While my results regarding market efficiency are certainly not dispositive due to mixed conclusions, the appearance of statistically significant results for three out of four judgment bias-related independent variables of interest in at least one specification is suggestive that further research into these or similar hypotheses has potential. In combination with the prior from existing research that PredictIt's markets could be inefficient due to their fee structure and other restrictions, future authors who are able to address some of the limitations of this thesis could more convincingly identify an effect of a systematic bias in PredictIt's markets.

Given the different findings between specifications, the next takeaway follows naturally: it is essential to exercise skepticism about research claims for which the only evidence is statistical significance from a regression. If I had not committed to model specifications beforehand, it would have been possible for me to choose which specifications to present in this thesis on the basis of which ones produce consistently statistically significant results and omit the ones that do not. This sort of unethical behavior is suspected to be quite common in social sciences, as concluded by researchers such as Simmons et al. (2011). But the prevalence of this behavior can be mitigated with the adoption of practices like filing a pre-registration report specifying methodological choices before gaining access to the data. In my case, it was not possible to file such a report due to time constraints,

which I consider to be a limitation of this analysis. This limitation is mitigated by my documented commitment to certain methodological choices as agreed upon with my supervisor before running any regressions or analyses. Another promising method that future researchers studying this and other topics could employ to enhance the believability of their results is something called a multiverse analysis. In these types of analyses, instead of only presenting one or a handful of specifications in a research paper, researchers instead run and present results from all reasonable combinations of model specifications (Steegen et al., 2016). For example, in addition to my main specifications I could have run countless possible combinations of my panel data regressions with 5, 10, 15, 20, etc. market-day observations per market while varying the inclusion and omission of the full set of control variables based on different theories of OVB and potential mediator bias. My specifications give a sense of the range of possible results, but a set of multiverse analysis regressions would produce a comprehensive distribution of coefficient estimates and p-values that more rigorously demonstrate how sensitive the main results are to reasonable changes in specification.

Beyond limitations and research opportunities relating to transparency, future researchers can also improve upon my analysis with more traditional augmentations as well. PredictIt's markets are ongoing and hundreds more have closed or will soon close due to results from the 2020 U.S. elections. Including these observations in an analysis would certainly improve the power to detect effects. Extending the research by analyzing data from other prediction markets and betting websites, such as BetFair, would enhance the external validity of research like this, since the generalizability of results may be limited due to the particularities of PredictIt's platform. Improving the measurement of certain independent variables, such as by employing independent coders for the manual coding of competitiveness priors or developing quantitative estimates for the competitiveness of each election, could also be useful. Future researchers can and should make their own judgment about the robustness of my identification strategy and decide whether there are some material omissions in control variables or changes to be made in functional form. My identification strategy for the time-invariant independent variables of interest could potentially be improved with altered model specifications or with the use of convincing instrumental variables to debias the coefficients. Given that limitations on the number of traders, investment amounts, and information available in certain markets is the rationale for why systematic bias may be measurable, introducing interaction terms between the time-invariant variables and certain indicia of market inefficiency could be another approach to measuring systematic bias. I was not able to obtain trader-level data that would have allowed me to use investment amounts or the number of traders for the interaction terms in time for this thesis, but that data can be obtained for future research; alternatively, it would be possible to use trade volume as a proxy for the number of traders, investment amounts, and information available. In particular, time-variant versions of these interaction terms would allow for the use of fixed effects to debias the coefficients of interest more convincingly than ordinary OLS. Finally, there are countless other judgment bias and information flow related hypotheses that could be tested with this or an expanded dataset. Among others, these include whether there is a systematic prediction error associated with religious minority

candidates against whom there are documented biases in the U.S., namely Muslims and atheists, once there are a sufficient number of candidates who identify as such. Another is whether there is an association between the number of media mentions or positive/negative sentiment on social media and prediction error on a given day, as an alternative variable to Google search volume. I encourage future researchers to explore these possibilities while maintaining a commitment to methodological transparency and openness.

References

Abadie, A., Athey, S., Imbens, G., & Wooldridge, J. (2017). When should you adjust standard errors for clustering?

Appleman, H. S. (1960). A Fallacy in the Use of Skill Scores. *Bulletin of the American Meteorological Society*, *41*(2), 64–67. https://doi.org/10.1175/1520-0477-41.2.64

Barrow, B. (2017). Georgia special election shapes up as referendum on trump. *Associated Press*. https://apnews.com/article/75e4d661a17743768202f23cd05d7a9e

Berg, J., Forsythe, R., Nelson, F., & Rietz, T. (2003). Chapter 80 results from a dozen years of election futures markets research. *Handbook of Experimental Economics Results*, *1*. https://doi.org/10.1016/S1574-0722(07)00080-7

Berg, J. E., & Rietz, T. A. (2019). Longshots, overconfidence and efficiency on the iowa electronic market. *International Journal of Forecasting*, *35*(1), 271–287. https://doi.org/https://doi.org/10.1016/j.ijforecast.2018.03.004

Berlemann, M., & Schmidt, C. (2001). *Predictive accuracy of political stock markets: Empirical evidence from an european perspective* (Dresden Discussion Paper Series in Economics No. 05/01). Technische Universität Dresden, Faculty of Business and Economics, Department of Economics. https://EconPapers.repec.org/RePEc:zbw:tuddps:0501

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.

Bronner, L., & Bacon Jr., P. (2020). What defines the sanders coalition? *FiveThirtyEight*. https://fivethirtyeight.com/features/what-defines-the-sanders-coalition/

Chatla, S., & Shmueli, G. (2013). Linear probability models (lpm) and big data: The good, the bad, and the ugly. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2353841

Christiansen, J. (2007). Prediction markets: Practical experiments in small markets and behaviours observed. *Journal of Prediction Markets*, *1*, 17–41. https://doi.org/10.5750/jpm.v1i1.418

Dawid, A. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, *77*, 605–610.

Dawood, Y. (2014). Campaign finance and american democracy. *Annual Review of Political Science*, *18*, 150403170711000. https://doi.org/10.1146/annurev-polisci-010814-104523

Doherty, C., Kiley, J., & Johnson, B. (2017). Political typology reveals deep fissures on the right and left. *Pew Research Center*.

Doherty, D., Dowling, C., & Miller, M. (2019). Do local party chairs think women and minority candidates can win? evidence from a conjoint experiment. *The Journal of Politics*, *81*, 000–000. https://doi.org/10.1086/704698

Dolan, K. (2013). Gender stereotypes, candidate evaluations, and voting for women candidates: What really matters? *Political Research Quarterly*, *67*, 96–107. https://doi.org/10.1177/1065912913487949

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417. http://www.jstor.org/stable/2325486

FEC. (2020). Contribution limits. *Federal Election Commission*. https://www.fec.gov/help-candidates-and-committees/candidate-taking-receipts/contribution-limits/

Foran, C. (2018). Ex-naacp leader ben jealous aims to become maryland's first black governor. *CNN*. https://edition.cnn.com/2018/06/27/politics/ben-jealous-maryland-governor-bernie-sanders/index.html

Forsythe, R., Nelson, F., Neumann, G., & Wright, J. (1992). Anatomy of an experimental political stock market. *American Economic Review*, *82*, 1142–61.

Forsythe, R., Rietz, T. A., & Ross, T. (1999). Wishes, expectations and actions: A survey on price formation in election stock markets. *Journal of Economic Behavior Organization*, *39*(1), 83–110. https://EconPapers.repec.org/RePEc:eee:jeborg:v:39:y:1999:i:1:p:83-110

Gibbons, C., Serrato, J., & Urbancic, M. (2018). Broken or fixed effects? *Journal of Econometric Methods*, *8*. https://doi.org/10.1515/jem-2017-0002

Hayek, F. A. (1945). The use of knowledge in society. *The American Economic Review*, *35*(4), 519–530. http://www.jstor.org/stable/1809376

Head, M., Holman, L., Lanfear, R., Kahn, A., & Jennions, M. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*, e1002106. https://doi.org/10.1371/journal.pbio.1002106

Holcombe, R. (2006). *Public sector economics: The role of government in the american economy*.

How to Trade, P. (2020). How to trade on predictit. *PredictIt*. https://www.predictit.org/support/how-to-trade-on-predictit

Jiang, W. (2016). Stock market valuation using internet search volumes: Us-china comparison.

Juenke, E. G., & Shah, P. (2016). Demand and supply: Racial and ethnic minority candidates in white districts. *The Journal of Race, Ethnicity, and Politics*, *1*(1), 60–90. https://doi.org/10.1017/rep.2015.2

Kennedy, C., Blumenthal, M., Clement, S., Clinton, J., Durand, C., Franklin, C., McGeeney, K., Miringoff, L., Olson, K., Rivers, D., Saad, L., & Wlezien, C. (2018). An evaluation of the 2016 election polls in the united states. *Public Opinion Quarterly*, *82*. https://doi.org/10.1093/poq/nfx047

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*, 443–77. https://doi.org/10.1162/003355397555253

Lawless, J., & Pearson, K. (2008). The primary reason for women's underrepresentation? reevaluating the conventional wisdom. *The Journal of Politics*, *70*, 67–82. https://doi.org/10.1017/S002238160708005X

Mackinnon, J., & Webb, M. (2016). Wild bootstrap inference for wildly different cluster sizes. *Journal of Applied Econometrics*, *32*, n/a–n/a. https://doi.org/10.1002/jae.2508

McCarthy, J. (2019). Less than half in u.s. would vote for a socialist for president. https://news.gallup.com/poll/254120/less-half-vote-socialist-president.aspx

Mercier, B., Celniker, J., & Shariff, A. (2020). Overestimating explicit prejudice causes democrats to believe disadvantaged groups are less electable. https://doi.org/10.31234/osf.io/s52qz

Perticone, J. (2018). There's a 'stock market for politics' where users can make money on washington's chaos. *Business Insider*. https://www.businessinsider.com/predictit-is-a-stock-market-for-politics-where-users-can-make-money-2018-5?r=US

Phillips-Wren, G., Power, D., & Mora, M. (2019). Cognitive bias, decision styles, and risk attitudes in decision making and dss. *Journal of Decision Systems*, *28*, 63–66. https://doi.org/10.1080/12460125.2019.1646509

Rakich, N., & Conroy, M. (2020). Progressive groups are getting more selective in targeting incumbents. is it working? *FiveThirtyEight*. https://fivethirtyeight.com/features/progressive-groups-are-getting-more-selective-in-targeting-incumbents-is-it-working/

Restocchi, V., Mcgroarty, F., & Gerding, E. (2018). The temporal evolution of mispricing in prediction markets. *Finance Research Letters*, *29*. https://doi.org/10.1016/j.frl.2018.08.003

Rhode, P. W., & Strumpf, K. (2013). *The long history of political betting markets: An international perspective*.

Rothschild, D., & Sethi, R. (2013). Trading strategies and market microstructure: Evidence from a prediction market. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2322420

Silver, N. (2016). 2016 election forecast. *FiveThirtyEight*. https://projects.fivethirtyeight.com/2020-election-forecast/senate/

Silver, N. (2020a). 2020 election forecast. *FiveThirtyEight*. https://projects.fivethirtyeight.com/2020-election-forecast/senate/

Silver, N. (2020b). How fivethirtyeight's house, senate and governor models work. *FiveThirtyEight*. https://fivethirtyeight.com/methodology/how--house-and-senate-models-work/

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *20*, 1–8. https://doi.org/10.5334/jopd.aa

Skelley, G. (2019). What we know about andrew yang's base. *FiveThirtyEight*. https://fivethirtyeight.com/features/what-we-know-about-andrew-yangs-base/

Smith, D. (2019).

i prefer non-religious': Why so few us politicians come out as atheists. *The Guardian*. https://www.theguardian.com/world/2019/aug/03/athiesm-us-politics-2020-election-religious-beliefs

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712. https://doi.org/10.1177/1745691616658637

Stershic, A., & Gujral, K. (2020). Arbitrage in political prediction markets. *The Journal of Prediction Markets*, *14*. https://doi.org/10.5750/jpm.v14i1.1796

Terms, & Conditions, P. (2020). Terms and conditions. *PredictIt*. https://www.predictit.org/terms-and-conditions

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, *8*(3), 201–207. https://doi.org/https://doi.org/10.1016/0165-1765(81)90067-7

Tziralis, G., & Tatsiopoulos, I. (2007). Prediction markets: An extended literature review. *Journal of Prediction Markets*, *1*, 75–91. https://doi.org/10.5750/jpm.v1i1.421

Who will win, P. (2020). Who will win the 2020 u.s. presidential election? *PredictIt*. https://www.predictit.org/markets/detail/3698/Who-will-win-the-2020-US-presidential-election

Williams, L. V. (2007). Introduction to the first issue from the editor. *Journal of Prediction Markets*, *1*. https://doi.org/doi:10.5750/jpm.v1i1.416

Wooldridge, J. (2012). *Introductory econometrics: A modern approach* (Vol. 20).