Rating Rankings:
Effect of Instant Run-off Voting on participation and civility

Preliminary and incomplete
2020-10-08

Current Version: `http://eamonmcginn.com/papers/IRV_in_Minneapolis.pdf.`

By Eamon McGinn[*]

American elections almost always use a First Past the Post (FPTP) system for electing candidates. Moving to a ranked choice/instant run-off vote (IRV) approach has been suggested by a number of advocacy groups to improve the electoral system. Currently, IRV is only used in a small number of municipal elections in the United States but this number has grown significantly over the last ten years. There are mixed findings in the literature on the benefits of IRV for voters and politicians, this makes informed debate around its adoption challenging. Analysis of the Minneapolis-St. Paul Metro Area, which has strong natural experiment characteristics, indicates that the introduction of IRV caused a 9.6 percentage point increase in turnout for Mayoral elections, on average. The effect on turnout is larger for precincts that have higher poverty rates. Text based sentiment analysis on mayoral debates, a new approach in this area, in a broader set of cities across the U.S. indicates that the introduction of IRV improved the civility of debates with candidates substituting negative or neutral words for positive words.
JEL: C55, D72, D78, H11, H73
Keywords: voter turnout, civility, instant run-off, difference-in-differences, sentiment analysis, Minneapolis

Since the mid-20th century, American elections have almost universally used a First Past The Post (FPTP) system for electing candidates (Douglas, 2014). In this system, the candidate who gets the most votes is elected. There has recently been concerns raised that FPTP elections may adversely affect political equality and fair representation (FairVote, 2019). Moving to an instant run-off vote (IRV) approach[1] — where voters rank multiple candidates and the lowest ranked candidate is iteratively removed until the winner has a majority of votes — has been suggested as a way to address some of the concerns with FPTP.

IRV is not an exotic import to the U.S.; IRV was invented in the U.S. and saw early success at the start of the 20th Century but essentially died out by mid-Century. Currently, IRV is only used in a small number of municipal elections in the U.S. Internationally, IRV is used in Australia and Ireland as well as for some smaller elections in the U.K. and New Zealand.

In the U.S. the number of municipalities using IRV has grown significantly over the last ten years. Recent success has been seen with the use of IRV being endorsed by The New York Times (2018) and IRV has recently been adopted in New York City for specific elections. Increased use of IRV is also being seen in Canada, although to a lesser extent then in the U.S.[2]

The resurgence is based on a range of expected benefits including: ensuring majority support for elected candidates, reducing costs of running elections, increasing civility between candidates, reducing conflict within the electorate, reducing strategic effects for voters and increasing diversity of candidates and elected representatives[3]. These direct benefits have also been expected to translate into an increase in voter turnout.

The growth in IRV means that there is a need to understand whether the purported benefits are being realized. A sizable literature has developed analyzing the various purported benefits. There is mounting evidence in this literature that IRV has not been living up to expectations in many areas. Voters tend to find IRV more challenging, which is particularly true for minority groups. IRV does not ensure majority support for candidates and seems to not affect the outcome of most elections in a meaningful way. IRV has also not reduced administrative costs. IRV has resulted in more diversity of candidates and elected officials but seems to have done this at the cost of greater racial polarization[3]. These findings are also reflected in the broader economic community as leading economists remain mixed in their view of whether IRV is superior to other methods (IGM Forum, 2018).

---

[1]IRV is also known as Alternative Voting, Ranked Choice Voting and Preferential Voting and is occasionally misnamed Single Transferrable Vote (STV). Strictly, STV refers to a similar approach for multi-member electoral districts. For consistency, IRV is used throughout this paper to avoid any potential confusion with other voting systems.

[2]Local elections in Cambridge and Kingston, Ontario will likely chance to IRV in 2022.

[3]For detailed references on each of these items see Section II

This paper seeks to address two particular areas of potential benefit of IRV that are important and which have weak methodologies in the existing literature: increasing turnout and improving civility. Turnout is clearly important to elections: higher levels of turnout improve the mandate of elected officials and ensure that all parts of the community contribute to political outcomes. Civility has been growing in importance due to the perception that elections are becoming more acrimonious and that this both reduces the time spent discussing matters of substance as well as reducing voter's interest in the election process.

While this paper doesn't focus on the theoretical reasons for the relationship between IRV, turnout and civility, some intuition is possible. Under IRV, voters can vote for their most preferred candidate without worrying about whether their vote will be 'wasted', which could lead to higher turnout. Also, under IRV, negative campaigning could turn voters away from placing a candidate second or third in their ranking and so candidates could use more positive communication styles.

Turnout is a focus because even at the national level, institutional variables are often regarded as the most powerful determinants of voter turnout; a recent meta-analysis finds that there are relatively few studies that analyze institutional impacts on voter turnout in subnational elections (Cancela and Geys, 2016). This view is supported by McDaniel (2016), Donovan, Tolbert and Gracey (2016) and McDaniel (2019) who consider that there is a lack of systematic and rigorous analysis of voter turnout in the context of IRV rules and that the findings to-date are mixed. Most existing research indicates that IRV has lowered (or at least had no impact on) turnout and there is evidence that this is particularly true for minority and disadvantaged groups. However, the most sophisticated analyses in this area, most remaining unpublished, tend to use data for many cities and a matched difference-in-differences (DID) approach[3]. This approach is problematic as the common trends assumption is unlikely to hold across the range of cities in the analyses. For example, McDaniel (2019), which is the most recent and thorough of these papers, notes that "visual inspection of the data does suggest that the comparison group election data may present a violation of the parallel trends assumption." Given the likely violation of the common trend assumption, the conclusions from these previous papers may not be reliable. This paper contributes to the literature by focusing on an in-depth analysis of a single metro area where the common trends assumption does hold and can be more carefully analyzed.

Civility is a focus because recent advances in econometric technique allow for new approaches to the analysis of speech and text (Gentzkow, Kelly and Taddy, 2019). This can provide a more precise quantification of the effect of IRV on civility during campaigns than previous research, which have used surveys or interviews. In terms of civility, previous surveys of candidates and voters as well as quantitative analysis of campaign materials indicate that candidates were less likely to make

negative attacks on their opponents following the introduction of IRV. However, relying on survey information can be problematic. For example, Palfrey and Poole (1987) are able to compare survey results to actual voting behavior and find that approximately 40% of non-voters in their sample inaccurately reported that they had voted. This paper is the first to use modern, natural language processing techniques to analyses the civility of debates and is also the first to make use of the large amount of transcribed debates that are now available from data sources like YouTube. The use of natural language processing techniques is a major contribution of this paper as it ensures impartiality and allows the analysis of civility to be done a scale not possible when using manual techniques.

The results of the analysis indicate that, in the Minneapolis-St. Paul Metro Area, the introduction of IRV caused a 9.6 percentage point increase in turnout for mayoral elections. The effect on turnout is larger for precincts that have higher poverty rates. This result is in line with the expectations of proponents of IRV, that it will increase turnout, but is somewhat counter to the bulk of the literature, which generally find no effect or that the introduction of IRV results in a reduction in turnout, particularly for minority groups. Text based sentiment analysis of mayoral debates in a broader set of cities across the U.S. indicates that the introduction of IRV improved the civility of debates. The improvement in civility is due to candidates substituting negative or neutral words for more positive words throughout the debate. Again, this is in line with the expectations of proponents of IRV but also aligns with existing survey and hand-classification based research which tends to find that IRV improves civility.

This paper also makes a methodological contribution by calculating confidence intervals according to the approach of Conley and Taber (2011). Conley and Taber's approach allows for a small number of policy changers (treatment units) by using information from the larger sample of non-changing groups (control units). This is in contrast to regular inference in DID, which assumes that the number of treatment groups is large. This is an important contribution of this paper as previous research has used regular or clustered standard errors to conduct inference, likely overstating the statistical significance of the treatment effect due to the small number of areas that have used IRV in the U.S.

Section I gives a summary of approaches to aggregating preferences and the recently renewed interest in the adoption of IRV in the U.S. Section II provides a summary of the literature. The results of the literature review lead to the development of a research design and econometric specification in Section III. The data sources are set out and summarized in Section IV. The main results of the analysis are presented in Section V with further robustness checks provided in VI. Section VII concludes.

I. Background

This section first provides some genreal background on FPTP and IRV, provides some formal defintions and then presents some recent background on the transition towards IRV in Minneapolis in particular.

A. Background on IRV

The main two approaches for aggregating preferences in the U.S. and of relevance to this paper are FPTP and IRV. IRV is also known as Alternative Voting, Ranked Choice Voting and Preferential Voting and is occasionally misnamed Single Transferrable Vote (STV). For consistency, IRV is used throughout this paper to avoid any potential confusion with other voting systems.

Generally speaking, FPTP awards the election to whichever candidate receives the most votes whereas IRV iteratively removes the candidate with the least number of votes until one candidate has a majority of votes. An important difference between the two systems is that FPTP only requires one candidate to be identified on the ballot as the most preferred whereas IRV requires either a partial or full ranking of all candidates on the ballot.

IRV and FPTP are two of a large range of voting rules that have been proposed and used in elections. Common electoral systems in use around the world today for single member constituencies include FPTP, Two-Round Systems (TRS), and IRV. Other systems that have been proposed and studied theoretically but are rarely used in practice include Borda Count (Black, 1976), Kemeny-Young (Kemeny, 1959) and Majority Judgement (Balinski and Laraki, 2011).

The existence of many different electoral systems may be related to theoretical results from Arrow (1950) who shows that no election system can satisfy a small and reasonable set of desirable criteria at the same time[4], and Gibbard (1973) and Satterthwaite (1975) who show that all voting systems are open to manipulation through tactical voting.[5] That is, all voting systems have flaws.

FPTP is used in over 40 countries, including major democracies such as the U.S., Canada, UK and India. Historically, IRV is far less common and is used in state and federal elections in Australia; in presidential elections in Ireland; and by some jurisdictions in the U.S., United Kingdom, and New Zealand. A similar method to IRV that is used for multi-member constituencies, Single Transferrable Vote (STV), is also used in national elections in the Republic of Ireland and Malta;

---

[4]Specifically, in elections involving three or more options, no ranked choice electoral system can aggregate preferences in a way that is complete and transitive while maintaining the properties of unrestricted domain, non-dictatorship, Pareto efficiency, and independence of irrelevant alternatives.

[5]Specifically, for any deterministic ordinal electoral system that chooses a single winner, one of three situations must hold: the rule is dictatorial; the rule limits the possible outcomes to two alternatives only; or the rule is susceptible to tactical voting

it is also used in some elections in Australia, Northern Ireland, Scotland, New Zealand and the United States.

IRV was invented by a Harvard professor, W.R. Ware, in 1871 as a modification to Thomas Hare's earlier proposals (Reilly, 2004). In the U.S., IRV was popular in the first half of the 20th Century. Various preferential voting systems were used for local elections in around two dozen cities over the course of the early 20th Century (Reilly, 2004). But reform movements, often sponsored by party machines, led to their replacement, in virtually all cases, by plurality systems (Reilly, 2004). IRV, in particular, was repealed shortly after being passed in cities and states in nearly two-thirds of the jurisdictions in which it originally passed; including Boulder, CO; Cincinnati, OH; and Ann Arbor, MI (Rhode, 2018). By 1962, only Cambridge, MA retained the system (Amy, 1996). This means that, since the mid-20th Century, American elections have almost universally used a FPTP system for electing candidates (Amy, 1996).

There is no consensus on whether IRV or FPTP is preferable. For example, the IGM Economic Experts Panel economists remain mixed in their view of whether IRV is superior to other voting methods (IGM Forum, 2018). A range of authors including Kimball and Anthony (2016), Rhode (2018), Reilly (1997), Sutherland (2016) and Fraenkel and Grofman (2004) do identify a core set of expected benefits of IRV, which can be summarized as:

- Ensuring majority support for winners: with reference to Definition I.5, it's clear that, in theory, the winner should have majority support.

- Reducing costs of run-off elections: with reference to Definition I.5, by moving progressively through preferences stated by voters, IRV removes the need for multiple rounds of elections and should conceptually reduce costs.

- Encouraging collaboration and civility among competing candidates: negative campaigning could turn voters away from placing a candidate second or third in their ranking and so candidates could use more positive communication styles.

- Reducing conflict and tension between ethnic groups: voters from minority ethnic groups can vote for their preferred candidate first, who may have little chance of reaching majority support, while still expressing preferences over other candidates and influencing the final result.

- Allowing voters to provide a more complete and non-strategic report of their preferences: Bartholdi and Orlin (1991) show that under STV, which is essentially identical to IRV from a strategic point of view, strategic manipulation of a vote is an NP complete problem even in an election for a single seat.

- Providing incentives for more and more diverse candidates to run for office: candidates who have a low chance of being elected can choose to run without fear that they will 'split the vote' of other similar candidates and cause an opponent to be elected.

On the basis of the perceived benefits, proponents of IRV often argue that it will increase voter participation and engagement (McDaniel, 2019). Richie, Kleppner and Bouricius (2000) state that "the combination of better choices, less money in politics, clearer mandates and less negative campaigning could lead to higher voter turnout and increased overall participation in politics."

Interestingly, based on a phone survey of voter preferences, application of IRV to the national presidential election for 2016 indicates that Hilary Clinton would likely have won the presidency with 54% of the final round vote (Trump receiving 46% of the final round vote). This result is closer to the national popular vote where Clinton received 48.2% and Trump 46.1% (Igersheim et al., 2018) than that delivered by the Electoral College.

Recently, concerns have been raised that FPTP elections may adversely affect political equality and fair representation (FairVote, 2019). Since 2000, likely spurred on by the highly contested Presidential election, there has been renewed interest in implementing IRV in the U.S. (Reilly, 2004). In the early 2000s, newspaper endorsements for using IRV came from USA Today; major articles and commentary about IRV appeared in publications, such as the New York Times, the Wall Street Journal, and the Washington Post. Support for IRV was also given by groups, such as the League of Women Voters (Richie and Hill, 2001). A major impediment to implementation of IRV in the early 2000s was that the voting machines used in most U.S. jurisdictions were not equipped to handle ranked ballots (Reilly, 2004).

Following two decades of increasing momentum and improving voting machine technology, as of early 2018, around a dozen different municipalities are actively utilizing some form of IRV to elect officials. The majority of these cities first implemented IRV in the years since 2010 (Rhode, 2018). This includes in state and congressional elections in Maine (since 2018) and in local elections in 11 cities, including San Francisco (since 2004), Oakland (since 2010), Berkeley (since 2010) and Santa Fe (since 2010). Recently, using IRV for elections in the U.S. Congress has been endorsed by The New York Times (2018) and has recently been adopted in New York City for specific elections.

B.   Approaches to Aggregating Preferences

Formal definitions of FPTP and IRV are rare in the literature but do provide some clarification of the concepts. This is important as the use of IRV is uncommon,

and small differences in its application can affect the outcomes of a vote[6]. To provide a formal defintion of IRV and FPTP, some primitives need to be set out starting with a voters, candidates and the preferences of voters.

$\mathbb{I}$, the set of voters, indexed $i = 1, ..., \bar{i}$.
$\mathbb{H}$, the set of candidates, indexed $h = 1, ..., \bar{h}$.
Each $i \in \mathbb{I}$ has strict preferences over the elements of $\mathbb{H}$ denoted $\succ_i$.

It is assumed that $\bar{h} \leq \bar{i}$. Further, preferences for all $i \in \mathbb{I}$ are assumed to be strict, so that there is no indifference and they are also assumed to be complete and transitive. Strict preferences are used for simplifying the definitions and also because they reflect the common approach in voting systems where rules for completing ballots don't normally allow for voters to express indifference between candidates.

Turning to how individual preferences are aggregated, some preliminaries are required. First, the set of preference relations on $\mathbb{H}$ is denoted $\mathcal{R} = \{\succ_1, ..., \succ_{\bar{i}}\}$ With this, we are able to define the voting function that is used to aggregate voter preferences:

Definition I.1. A <u>voting function</u> $f : \mathcal{R} \times 2^{\mathbb{H}} \to \mathbb{H}$ with $f(\cdot, H) \in H \;\; \forall \;\; H \in 2^{\mathbb{H}}$ assigns a chosen element $h \in H$ for any profile of voter preferences.

That is, a voting function is a social choice function that, when given a preference profile of voters and a set of candidates, will assign a candidate to all voters whose preferences are included in the preference profile. Within the voting function, the way that the element is selected could be by various voting rules such as FPTP or IRV. These voting functions are defined as follows.

Definition I.2. A candidate h^* has a <u>plurality</u> among candidates $H \subseteq \mathbb{H}$ if:

$$|\{i | i \in \mathbb{I}; h^* \succ_i h \;\; \forall \;\; h \in H, h^* \neq h\}| \geq |\{i | i \in \mathbb{I}; h' \succ_i h \;\; \forall \;\; h \in H, h' \neq h\}|$$

for all $\;\; h' \in H, h^* \neq h'$.

Definition I.3. A candidate h^* has a <u>majority</u> if:

$$\frac{|\{i | i \in \mathbb{I}; h^* \succ_i h \;\; \forall \;\; h \in H, h^* \neq h\}|}{|\mathbb{I}|} > \frac{1}{2}$$

Definition I.4. In a <u>First Past the Post (FPTP)</u> voting function, $f(\mathcal{R}, H) = h$ if one candidate h has a plurality where $H \subseteq \mathbb{H}$. If multiple candidates $H' \subseteq H$

---

[6]Bartholdi and Orlin (1991) do provide formal definitions based on pseudocode algorithms.

have a plurality then $f(\mathcal{R}, H) = h'$ for an arbitrary $h' \in H'$.

Definition I.5. In an <u>Instant Runoff Vote (IRV)</u> voting function, $f(\mathcal{R}, H^k) = h$ if candidate h has a majority where $H^k \subseteq \mathbb{H}$. If no candidate has a majority then define:

$$\hat{h} = \arg\min_h |\{i | i \in \mathbb{I}; h \succ_i h' \ \forall \ h' \in H, h \neq h'\}|$$

If $|\hat{h}| = 1$ then set $H^{k+1} = H^k \setminus \hat{h}$ otherwise set $H^{k+1} = H^k \setminus \hat{h}'$ for an arbitrary $\hat{h}' \in \hat{h}$. Continue this process until some $h \in H^{k+j}$ has a majority, where $j \in \mathbb{Z}_+$.

For FPTP and IRV voting functions, the need to rely on arbitrary selections is included to deal with the potential case where there are ties. In practice, this arbitrary selection could be one of many methods, such as giving one voter the deciding vote, running another election or even using a coin toss.

C.   *Implementation of IRV in Minnesota*

Within Minnesota, IRV has been adopted by the two municipalities that make up the core of the Minneapolis-St. Paul Metro Area[7]. Minneapolis-St. Paul, often referred to as the Twin Cities, is a large metropolitan area of around 4 million people. The core of the Metro Area is based on the municipalities of Minneapolis, the most populous city in Minnesota and seat of Hennepin County; and St. Paul, the state capital of Minnesota and seat of Ramsey County. Minneapolis has a population of around 422,000 whereas St. Paul has a population of around 310,000. Both Minneapolis and St. Paul are independent municipalities with defined borders, electing separate councils and mayors. Municipal elections are, however, coordinated across Minnesota, meaning that municipal elections occur at the same time in both Minneapolis and St. Paul.

In this paper, the Minneapolis-St. Paul Metro Area (also referred to as the Twin Cities area) is defined as the seven counties that make up the Metropolitan Council[8]. In the data used in this paper, these seven counties contain 152 municipalities.

On November 7, 2006, voters in Minneapolis approved the use of IRV to elect the city council and mayor by a 65% majority[9]. This was the third effort at a charter amendment after failed attempts in 1999 and 2001 (Schultz and Rendahl, 2010). On November 3, 2009, IRV was implemented for the first time as a voting mechanism in Minneapolis (Schultz and Rendahl, 2010). In its first implementation

---

[7]The potential for using IRV is also under consideration in Rochester, MN and is expected to be implemented in St Louis Park, MN in 2019.

[8]The seven counties in the Metropolitan Council's Twin Cities Metropolitan Area are: Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington counties.

[9]Note that IRV is not used for elections for the Minneapolis School Board, or the county, state or federal offices.

in Minneapolis, the election involved a popular incumbent, Mayor R. T. Rybak, who received 74% of the first preference vote.

In general, Minneapolis aimed to be the "Gold Standard" (Chadha, 2019) of IRV implementation and sought to achieve cost reductions and increased turnout, ensuring that elected candidates receive a majority of votes and encouraging support for third parties (Schultz and Rendahl, 2010).

Similarly, in 2009, Saint Paul voters approved the use of IRV in elections for city council and mayor. The first mayoral election using IRV was held in 2013 and, as with Minneapolis, the first outing involved an incumbent mayor, Chris Coleman, who received 78% of the first preference vote.

This means that the following voting systems have been used in the Twin Cities for recent elections:

TABLE 1—TWIN CITIES' MAYORAL ELECTIONS

	Minneapolis	St. Paul
2001	FPTP	FPTP
2005	FPTP	FPTP
2009	IRV	FPTP
2013	IRV	IRV
2017	IRV	IRV

Note: Only Mayoral elections shown

Source: Author's calculations

From an administrative point of view, the initial implementation in Minneapolis was largely seen as a success with the major issue being that hand-counting and verification in Minneapolis took a long time (Aiba et al., 2012).

The 2013 Minneapolis election was also seen as a success, with turnout over 80,000, the highest for a municipal election in 12 years. In the city's most ethnically diverse ward, voters understood and appreciated IRV. Moreover, 75% of voters ranked two choices, and 63% ranked all three available choices in the council race (Chadha, 2019).

In the 2013 Minneapolis election, there was no incumbent candidate and 35 candidates ran for election. Because voters were able to rank up to three candidates on their ballot, voting went to the 33rd round and took two days of tabulation to find the winner. Competition and debate in the 2013 Minneapolis election was noted as being particularly civil. In 2018, the Minneapolis city clerk noted that "the tone and rhetoric of the [2013] campaign was much more focused on policy and issues and priorities, what you would do as mayor, and less about the traditional attack ads against their opponents... at one point in the 2013 mayoral debate,

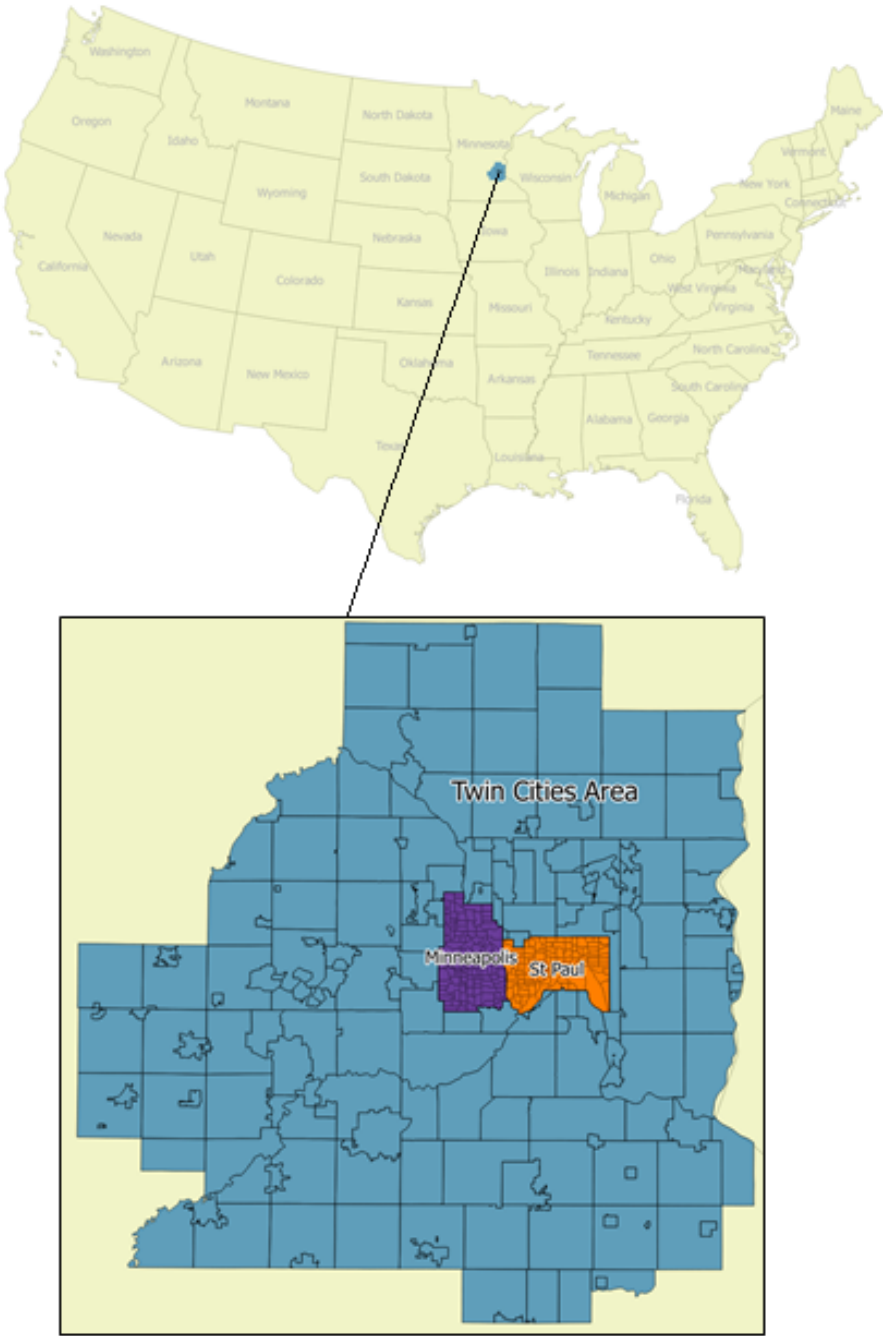FIGURE 1. MINNEAPOLIS-ST. PAUL METRO AREA MUNICIPALITIES SHOWING MINNEAPOLIS AND ST. PAUL'S
VOTING PRECINCTS

Note: The Minneapolis municipality is shown in purple and the St. Paul municipality is shown in orange.
Heavy lines within the Metro Area show municipalities while lighter lines within Minneapolis and St.
Paul show voting precincts
Source: Based on data from Minnesota Legislature (n.d.)

the candidates linked arms and literally sang 'Kumbaya', refusing to insult one another out of fear of losing voters" (News Centre Maine, 2018). Commentators generally considered that this civility could be attributed to candidates not wanting to alienate voters who may rank them second or third on the ballot. The 2013 mayoral campaign was also seen as enabling more politically diverse candidates to run in the election. An example is the candidacy of Cam Winton who ran as an independent on a moderate-conservative platform — which isn't common in Minneapolis mayoral elections.

St. Paul had a similar experience to Minneapolis on its second outing with IRV: the incumbent mayor decided not to run, resulting in a contest between six candidates. The successful candidate, Melvin Carter, is the first African American mayor of the city.

II.   Evidence in the Empirical Literature

This section provides a detailed review of previous analyses of the performance of IRV, first covering general findings then considering turnout and civility in more detail.

A.   General Findings

In terms of a basic understanding of IRV, Neely, Blash and Cook (2005) conduct a survey of voters in San Francisco in 2014 and find high levels of understanding of IRV with a majority (61%) preferring IRV over the runoff system that was previously used in San Francisco. Based on a phone survey of voters in a number of locations, Douglas (2014) reports that over 90% of those surveyed considered the instructions on how to complete their IRV ballot were either somewhat or very easy to understand. Donovan, Tolbert and Gracey (2019) find that voters in IRV cities are more likely to report that understanding voting instructions was "very or somewhat difficult" compared to voters in plurality voting systems. Neely and Cook (2008) show evidence that minority groups, such as Latinos and the elderly, may find it particularly challenging to make use of IRV ballots. These results are supported in follow up research by Neely and McDaniel (2015) who identify that, under IRV in San Francisco, voided ballots are consistently more common in precincts where more African-American citizens reside, and are often observed at higher rates in precincts that contain more Latino, elderly, foreign-born, and less wealthy residents. It is worth noting that very similar results are also found by Sinclair and Alvarez (2004) when looking at voting via punch cards in Los Angeles, which may indicate that previous findings are not solely related to the presence of IRV but may be a general feature of 'complicated' voting procedures.

Adequate understanding of IRV may, however, not be sufficient to achieve support for its use. Using an experimental survey, Nielson (2017) found that, while participants were largely able to understand the rules of IRV, voting in an IRV election did not have any measurable impact on attitudes toward elections, nor

did participation in an IRV election lead to an increase in preference for IRV elections. However, in Minneapolis Chadha (2019) reports that a plurality of voters (41%) preferred IRV with older, more educated, wealthier and white voters preferring IRV more strongly.

Further, due to exhausted ballots (where the candidates who a voter ranks are not elected), when examining results from four IRV elections Burnett and Kogan (2015) calculate that the winner never receives a majority of the total votes cast. A similar finding is made by Endersby and Towle (2014). One contributor to the prevalence of exhausted ballots is an apparent preference among many voters for only ranking a small number of candidates. Undertaking a review of ballots from Ireland, Laver (2004) finds that the modal number of preferences recorded was three while the median was four (in elections with between 9 and 14 candidates).

With respect to cost reduction, Rhode (2018) uses a matched DID approach to compare expenditure between IRV and FPTP cities and finds no statistically significant change in expenditure. In fact, Rhode (2018) finds that in the years before IRV was implemented, the cities that would implement it spent more per election cycle than control cities, $3.39 compared to $1.24. Anthony et al. (2019) conduct a survey of election administrators in Maine and conclude that "most municipal clerks in our sample are not enthusiastic about implementing ranked choice voting and do not want to continue its use in Maine."

Considering the diversity of candidates, John, Smith and Zack (2018) use a DID strategy to analyse how the representation of women and minorities changed in San Francisco following the introduction of IRV. They find that IRV resulted in an increase in the share of minority candidates and an increase in the share of female victors. However, there was no statistically significant effect on minority victors or female candidature. These results broadly align with earlier cross sectional analysis of across the U.S. by Trebbi, Aghion and Alesina (2008), finding that the electoral rule adopted by a city is associated with the representational ratio of minorities.

Despite an apparent increase in candidate diversity, McDaniel (2018) uses ecological inference and DID on data from Oakland and San Francisco to show that the introduction of IRV has very little, if any, effect on racially polarized voting and does not contribute to any moderation of racial conflict or competition. There is even some evidence to show that it may contribute to higher levels of racial polarization.

Felsenthal, Maoz and Rapoport (1993) who, by recalculating the results of elections under different voting rules, find that there is no significant difference among a number of voting procedures (including IRV) in terms of preserving the social preference ordering nor of electing the candidates who ought to be elected.

B.   Participation

A number of papers estimate that the introduction of IRV results in a reduction in turnout. Cook and Latterman (2011) note that turnout in San Francisco's 2011 Municipal election (which used IRV) was considerably lower than in either the midterm election of 2010 or the 2008 presidential election. Looking at voting in Oakland, CA, Holtzman and Nall (2012) find that IRV had a negative effect on turnout in 2010 relative to 2006. Using ecological inference and controlling for relevant socioeconomic covariates McDaniel (2016) finds that use of IRV in San Francisco decreased turnout for black (18 percentage points) and white voters (16 percentage points). In a recent working paper, McDaniel (2019) uses a DID of matched cities throughout the U.S. and finds a significant decrease in voter turnout of approximately 3–5 percentage points in IRV cities after the implementation of IRV. Despite using national data, this result may reflect the fact that much of McDaniel's treatment and control group is drawn from California, making it similar to previous analyses.

Analysis of implementation of IRV in the U.S. in the early 20th century generally concludes that there was no meaningful effect on voter turnout Amy (1996). Similar findings have been made for Canada in the same time period (Jansen, 2004). Using more up to date data and approaches (a DID of matched cities throughout the U.S.), Kimball and Anthony (2016) come to a similar conclusion: IRV does not appear to have a strong impact on voter turnout in municipal elections. Kimball and Anthony (2016) do observe higher rates of spoiled ballots in the IRV elections than in the plurality election.

There are a small number of papers that find an increase in turnout. Looking at a particular race, the Assessor Recorder race in San Francisco in 2005, Jerdonek (2006) estimates that turnout increased by an average of 2.7 times following the introduction of IRV. Robb (2011) finds a positive correlation between the use of IRV and higher turnout rates in San Francisco with turnout being significantly higher after the 2004 implementation of IRV than it was before. Although unable to make inferences about IRV, Sutherland (2016) does note that cities that adopt 'nonpartisan alternative variable' voting approaches (which includes IRV) had the highest level of turnout.

There has been a range of findings on turnout as seen in Minneapolis. Schultz and Rendahl (2010), looking at trends over time, find that IRV did not increase voter participation compared to participation in previous municipal elections, though it did not appear to be a cause of non-voting. Further, McDaniel (2019) finds no statistically significant effect for turnout in Minneapolis following the introduction of IRV.

Many studies also consider the effect that IRV has on specific groups within the community. The majority of studies find that IRV has disproportionate effects on certain groups, particularly minorities or at-risk groups. Cook and Latterman

(2011) find that precincts with higher proportions of Asian and Pacific Islander, Latino, and older voters were disproportionately likely to make mistakes on ballots in San Francisco. Similarly, in Oakland, CA, Holtzman and Nall (2012) find that Asian and Latino turnout declined during the transition to IRV.

In contrast, using a more sophisticated matched DID approach, Kimball and Anthony (2016) find similar levels of socioeconomic and racial disparities in voter participation in FPTP and IRV elections. Similarly, McDaniel (2016) found no statistically significant difference in Asian or Latino turnout associated with IRV in San Francisco. More complicated is Donovan, Tolbert and Gracey (2016) who find that older voters are significantly less likely to report understanding IRV systems but find no racial disparities.

Considering Minneapolis in particular, for the 2013 election (which used IRV), Kimball and Anthony (2016) do find that voter participation in the mayoral contest was higher in the wards with the highest share of white voters. This finding aligned with earlier, more simple analysis from Jacobs and Miller (2014) who found that, in the 2013 elections, voters who were more affluent and white turned out at a higher rate, completed their ballots more accurately, and were more likely to use all three opportunities to rank their most preferred candidates compared to voters living in low-income neighborhoods and in communities of color. These two findings are in contrast to FairVote Minnesota's determination that there were no significant discrepancies across demographic groups in understanding and casting of ranked choice ballots in Minneapolis in 2013 (Mauter, 2014).

Much of the available research to date is based on surveys and analysis of trends over time. Relying on survey information can be problematic. For example, Palfrey and Poole (1987) are able to compare survey results to actual voting behavior and find that approximately 40% of non-voters in their sample inaccurately reported that they had voted when, in fact, they did not vote.

The most sophisticated analyses in this area, many of which are working papers, tend to use data for many cities and use a matched DID approach.[10] This approach is problematic as the common trends assumption is unlikely to hold across the range of cities in the analyses. For example, McDaniel (2019), which is the most recent and thorough of these papers, notes that "visual inspection of the data does suggest that the comparison group election data may present a violation of the parallel trends assumption." Given the likely violation of the common trend assumption, the conclusions from these previous papers may not be reliable. This paper contributes to the literature by focusing on an in-depth analysis of a single metro area where the common trends assumption does hold and can be more carefully analyzed.

---

[10]In this approach the reference group is made of cities that continue to use FPTP voting.

C.   Civility

The decision to go negative in a campaign has been studied both theoretically and in practice. Theoretically, Skaperdas and Grofman (1995) construct a model of negative campaigning with the main conclusions being that the front runner is less likely to go negative, 3rd party spoiler candidates are likely to go positive, and stronger candidates are unlikely to go negative against weaker candidates. In practice, Damore (2002) analyses information from media campaigns and finds that candidates who are trailing in pre-campaign polls are more likely to attack, attacks are more likely to occur as election-day approaches, and that candidates respond to attacks by their opponents with attacks of their own. Peterson and Djupe (2005) undertake a text analysis of all Senate primaries in 1998 and identify that negativity is a function of the timing of the race, the status of the Senate seat, and the number and quality of the challengers.

The relationship between civility and turnout is unclear. For example, Djupe and Peterson (2002) present evidence based on analysis of newspaper articles that shows that campaign negativity boosts turnout in primary elections for Senate candidates with negative campaigns receiving more media coverage. This is different from earlier work from Ansolabehere et al. (1994) who find that the exposure to negative advertisements lowered intentions to vote by 5%, consistent across election data and experiments.

Considering how negativity is related to IRV, there have been a number of surveys of both candidates and voters. FairVote (2014) conducted a phone survey across IRV and FPTP locations and found that both voters and candidates reported less negativity in IRV locations than FPTP locations. Donovan, Tolbert and Gracey (2016) also use a survey across both IRV and FPTP locations to analyse perceptions of civility and find that both voters and candidates are likely to view campaigns as less negative when conducted under IRV. Similar results are also reported in Douglas (2014). From the same survey data, Tolbert (2014) identified that IRV elections increase: 1) perceptions of the fairness of the election; 2) the frequency of candidates praising or cooperating with their opponents; 3) general interest in the election; 4) usefulness of campaign information; and 5) satisfaction with the choice of candidates. In a phone survey focused on California, John (2015) and finds that candidates spent less time criticizing opponents in IRV cities than in cities that did not use IRV and that respondents reported less negative campaigns in IRV cities. These findings are essentially repeated in John and Douglas (2017).

Looking only at candidates' perceptions of civility across a range of IRV and FPTP cities, Donovan (n.d.) notes that IRV candidates were more likely to hire staff, more likely to spend money on internet ads, less likely to report spending funds on radio and TV ads, spend less time on the phone and less time meeting with staff, spend more time knocking on doors, and were less likely to report that

their rivals described them in negative terms.

In contrast, Neely, Blash and Cook (2005) use a survey of voters in San Francisco in 2004 and find that they were split on whether the campaigns were more or less negative compared to past elections (14% said more negative, 15% said less negative). These perceptions are, however, at odds with surveys of candidates and quantitative analysis of campaign mailers from San Francisco. Robb (2011) shows that there was a considerable increase in activity of cooperative campaigning in the first year of IRV and a decrease in negative campaigning.

Focusing on Minneapolis itself, analysis of civility has been more difficult to conduct. For example, Schultz and Rendahl (2010) find that, in the 2009 Minneapolis election, campaigns focused heavily on voting method and less on the issues, making inferences about civility challenging. Relying on interviews with candidates, Mauter (2014) generally found that the introduction of IRV helped to create a more inclusive election in 2013 in Minneapolis and that campaigns mostly did not use negative campaigning. Quoting an interview with candidate Shultz shows some of the logic behind this finding:

> "For instance where Mark Andrew said about Betsy Hodges, she has the disease of a small vision. I mean that was a big deal. In any other political context, it would have been nothing, but the fact that it was so jarring because it was really one of the only instances we had in the campaign we could point to of actual negativity happening. . . (Mauter, 2014)."

The extant analyses of civility in IRV elections use either opinion polls or subjective personal identification of tone, rather than objective measures. This paper contributes by using a modern, natural language processing approach that impartially analyses the civility of debates. This is an important improvement in ensuring the veracity of the conclusions on civility.

III. Research Design and Econometric Specification

The section sets out the different approaches used to address how IRV has affected voter participation and candidate civility.

A. Participation

Participation focuses on voter turnout, which is the percentage of eligible voters who actually cast a vote. Turnout is an important measure of the performance of a democratic election as higher levels of turnout can potentially indicate an election that is more closely followed by the electorate, more important for the electorate and ultimately may result in representatives who have greater support among the electorate.

Turnout is seen as an outcome that could be improved by IRV as, in FPTP voting,

if a voter believes that their preferred candidate does not have a plurality then there may be little incentive to cast a 'wasted' vote for a losing candidate. In contrast, in an IRV system, this voter can cast their first preference for their most preferred candidate without the concern that their vote will be 'wasted'.

Elements of how IRV was introduced in the Minneapolis-St. Paul Metro Area allow for identification of the effect of introducing IRV on voter turnout. In particular, there was a staggered introduction of IRV between Minneapolis and St. Paul while the rest of the Metro Area retained the existing FPTP approach to elections over the same period. Further, IRV was only introduced for metropolitan level elections. In other elections (such as state and federal), the existing FPTP system was retained. These circumstances create a strong case for a natural experiment in the introduction of IRV in Minneapolis and St. Paul.

Analyzing the impact of the introduction of IRV on turnout is possible using a staggered DID approach of the form shown below. Where the analysis will focus on municipalities in the Minneapolis-St. Paul Metro Area during the transition period from FPTP into IRV:

$$(1) \qquad Turnout_{it} = \beta_0 + \beta_1 IRV_{it} + \gamma_t + \delta_i + \beta_2 NME_{it} + \beta_3 X_{it} + \epsilon_{it},$$

where i indexes municipalities (such as Minneapolis, St. Paul, St. Louis Park, etc.) and t indexes the election period.

In this specification, turnout is measured as the share of registered voters who vote (normally referred to as Registered Voter turnout). This is in line with common approaches from the literature. In a meta-analysis of work on voter turnout, Stockemer (2017) finds that about two-thirds of existing studies use Registered Voter turnout[11].

For fixed effects, γ_t is a time fixed effect while δ_i is a municipality fixed effect. NME is an indicator variable for non-mayoral municipal elections. This variable accounts for cases where a municipal election is held for lower level offices – for example, St. Paul's 2015 election was just for City Council positions, not for mayor.The variable is included as the turnout for non-mayoral municipal elections is likely to be lower than for mayoral municipal elections and, due to differences in timing between different municipalities, this would not be captured in year fixed effects. X is a vector of covariates, such as demographic characteristics, if available. The possibility that the error term is correlated within a metropolitan area is accounted for through the use of clustered standard errors (Bertrand, Duflo and Mullainathan, 2004). In this specification β_1 will indicate whether the use of IRV resulted in an increase in voter turnout.

---

[11]One-third of the studies use Voting Age Population turnout.

Due to the small number of treated units, confidence intervals computed using the approaches described by Conley and Taber (2011) are also be reported. Conley and Taber's approach allows for a small number of policy changers (treatment units) by using information from the larger sample of non-changing groups (control units). This is in contrast to regular inference in DID, which assumes that the number of treatment groups is large[12]. This is an important contribution of this paper as previous research has used regular or clustered standard errors to conduct inference, likely overstating the statistical significance of the treatment effect due to the small number of areas that have used IRV in the U.S.

A key assumption here is that Minneapolis and St. Paul follow the common trends assumption. This assumption requires that the trend that control units show over time would have applied to the treated units if they had not been treated. The assumption is critical as ensures that control units provide an appropriate counterfactual for the treated units. The fact that the two areas are part of a single conurbation involving municipalities that were not treated is likely to support a common trends assumption, but this will be covered in more detail in the following data section.

B. Civility

Civility focuses on the messages and communication style of candidates during the campaign. In particular, based on the literature review, the expectation is that the candidates should use more positive communication styles under IRV. Civility is important as previous literature has found that it can have a meaningful impact on participation of voters during elections (see Ansolabehere et al. (1994) for example).

This analysis focuses on communication during mayoral debates, which is anticipated to improve after the introduction of IRV. The mechanism is that, under IRV, a candidate can still benefit from being a voter's second or third choice. A candidate may consider that negative campaigning will turn voters away from placing them second or third, potentially leading the candidate to make fewer or less negative statements throughout the debate.

Analysis of the change in civility is based on a text-as-data approach (Gentzkow,

---

[12]As there are only two treated units, the confidence intervals reported are generally estimated using the approach described under Proposition 1 in Conley and Taber (2011), which uses a direct calculation. Where this is not the case it is explicitly noted in the table notes. The approach to calculating the Conley and Taber confidence intervals used in this paper differs slightly from Conley and Taber's approach and code. In this approach, the set of control observations in the data is limited to those that have observations available for the years where the treated units receive treatment. This alteration from Conley and Taber's approach is required to account for the fact that each municipality only has one observation per year and the panel is not balanced. Conley and Taber's approach requires either a balanced panel or multiple observations per time period to allow re-weighting of control observations to match treated observations. The algorithm developed also allows for corrections due to the presence of heteroskedasticity in the data. A detailed review of the data indicated that heteroskedasticity based on municipality population is not present and that attempting to correct for the presence of such potential heteroskedasticity does not affect the overall results of the Conley and Taber confidence intervals.

Kelly and Taddy, 2019). Sentiment analysis of debate transcripts using standard lexicons is able to provide a quantitative measure of the tone of language used in a debate that is summarized in a single numerical score for each debate. The resulting score can then be used as the dependent variable in further regressions. The sentiment analysis process is described further in Section IV.B.

Similarly to the analysis of participation, this structure allows for the use of a staggered DID approach of the form shown below. Where the analysis will focus on municipalities across the U.S.:

$$(2) \qquad Sentiment_{it} = \beta_0 + \beta_1 IRV_{it} + \gamma_t + \delta_i + \beta_2 X_{it} + \epsilon_{it},$$

where i indexes municipalities (such as Minneapolis, St. Paul, San Francisco, Oakland, etc.) and t indexes the election period.

The dependent variable for sentiment will be measured in a number of different ways, explained in more detail in IV.B. Regardless of the approach to measurement, a higher value will indicate higher levels of civility.

For fixed effects, γ_t is a time fixed effect while δ_i is a municipality fixed effect. X is a vector of covariates, such as demographic characteristics, if available. It is assumed that the error term $\epsilon_{i,t}$ $N(0, \sigma_i^2)$ and that errors can be correlated within a metropolitan area through the use of clustered standard errors ($Bertrand, Dufloand Mullainathan$, 2004). Due to the small number of treated units, confidence intervals computed using the approaches described by Conley and Taber (2011) will also be reported. In this specification, β_1 will indicate whether the use of IRV resulted in an improvement in debate civility.

As with participation, the key assumption here is that treated municipalities follow a common trend in sentiment with areas that were untreated.[13] The relevant scarcity of debate recordings means that, unlike participation, the data for analysis of civility is based on a group of cities from across the U.S. As a result, the common trends assumption may be weaker than in the analysis of participation. This will be covered in more detail in the following section.

IV. Data

A. Participation

The main piece of data required to analyze participation is voter turnout for each election in the Minneapolis-St. Paul Metro Area at the municipality level.

---

[13]As stated previously: this assumption requires that the trend that control units show over time would have applied to the treated units if they had not been treated. The assumption is critical as ensures that control units provide an appropriate counterfactual for the treated units.

Data on voting is available from the Minnesota Secretary of State (n.d.) for every metropolitan area in the state of Minnesota and is available in detailed machine readable formats for elections going back to 1992. The lowest level of data available is at the precinct level[14] and, although data varies from year to year, there is generally data available for each precinct in each election on the number of registered voters as well as the number of ballots received. The aforementioned two pieces of data allow for turnout to be calculated.

Reporting complete data to the Secretary of State is not compulsory and so, to ensure complete data coverage for treated locations, voting data for Minneapolis and St. Paul was sourced from their respective electoral agencies, the City of Minneapolis (n.d.) and Ramsey County (n.d.).

Voting data from the year 2000 onwards was compiled and resulted in around 47,000 precinct level observations. Removing data for school board districts, which are not of interest in this paper, results in around 44,000 remaining observations spread across 2,600 metropolitan areas. A density plot of the turnout rate is shown in the figure below.

This main data source was merged with other supporting information. First, data on the type of each election was included (including date, whether it was a Presidential election, a mid-term election, a state election, a municipal election and/or a mayoral election). Next, data from the 2010 census sourced from the Minnesota Legislature (n.d.) was included. The census data is at the voting district level (which is a higher level of aggregation than the precinct level) and includes information such as population and race. After merging in census data, the number of observations reduced to around 39,000 as some precincts did not have identifying information in the census data. Finally, socioeconomic data at the County level sourced from the Minnesota Department of Health (n.d.) was included. This data covers the number of households, the percentage of elderly and young, unemployment, the number of households on food stamps, average per capita incomes, poverty rates and the number of school students. This County level socioeconomic data only covers the years 2006-2016, resulting in a reduction in the number of observations down to around 23,000. In the results section, the sensitivity of the results to these data reductions is tested.

With this combined data set, the data was then aggregated up to the municipal level, as this is the level of treatment, reducing the observations to 220. The subset of municipal areas within the Twin Cities metropolitan area was identified.

---

[14]Precincts roll up into wards and then into Municipal areas. Municipal areas themselves roll up into counties which then roll up into state level data. The Minneapolis-St. Paul Metro Area is used throughout the analysis and is defined as the seven counties that form the Metropolitan Council's Area: Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington counties. The Minneapolis-St. Paul Metro Area does not elect representatives as a single geography to any level of representation. The Minneapolis-St. Paul Metro Area also does not map neatly to any Federal level election area as it includes parts of five different congressional districts.
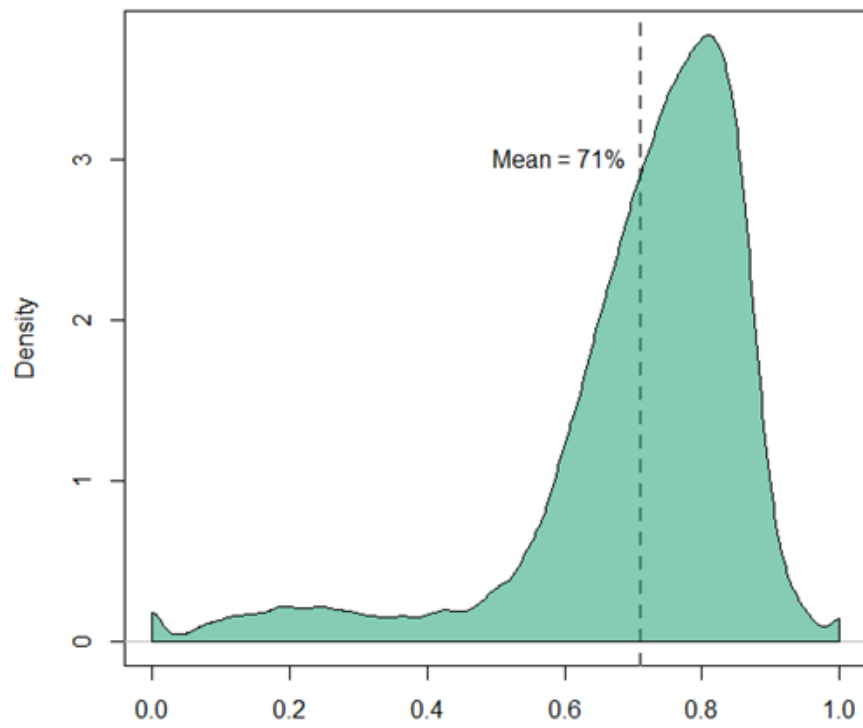
FIGURE 2. KERNEL DENSITY PLOT OF VOTER TURNOUT AT PRECINCT LEVEL

Note: Bandwidth = 0.0125

Source: Author's calculations based on data from Minnesota Secretary of State (n.d.)

The Metropolitan Area was defined to include the seven counties of Hennepin, Anoka, Washington, Ramsey, Carver, Scott and Dakota that together form the Metropolitan Council. Attention was restricted to these locations as they are likely to present the best comparator group for Minneapolis and St. Paul. Again, in the results section, the sensitivity of the results to this data reduction is tested. This resulted in a final set of 168 observations for use in the regression analysis.

The following table provides a summary of the socioeconomic characteristics of the treated and control municipal areas within the Minneapolis-St. Paul Metro Area. In summary, municipal areas in the treatment group are located in counties that have slightly higher populations, higher rates of unemployment and poverty and lower household incomes.

TABLE 2—SUMMARY OF SOCIO-ECONOMIC CHARACTERISTICS IN TWIN CITY MUNICIPALITIES

	All Data		Municipal Elections		Other Elections		Treated		Control	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Total Votes	$16,240.49$	$33,949.14$	$5,748.95$	$12,048.80$	$22,710.28$	$40,859.98$	$96,637$	$62,498.91$	$7,511.73$	$9,586.02$
Turnout	0.52	0.28	0.20	0.10	0.72	0.11	0.50	0.24	0.52	0.28
Municipal Election	0.38	0.49	1	0	0	0	0.37	0.50	0.38	0.49
State General Election	0.62	0.49	0	0	1	0	0.63	0.50	0.62	0.49
Mid Term Election	0.31	0.46	0	0	0.50	0.50	0.32	0.48	0.31	0.46
Presidential Election	0.31	0.46	0	0	0.50	0.50	0.32	0.48	0.31	0.46
Mean Per Capita Income	$51,525.16$	$7,554.29$	$51,187.49$	$7,115.37$	$51,733.39$	$7,834.67$	$51,129.46$	$7,373.11$	$51,568.12$	$7,593.16$
Household Income	$65,259.98$	$11,714.51$	$65,666.24$	$11,442.88$	$65,009.45$	$11,919.55$	$57,394.38$	$5,426.23$	$66,113.96$	$11,902.92$
Unemployment (%)	4.97	1.49	5.37	1.71	4.72	1.28	5.12	1.50	4.95	1.49
Poverty (%)	11.08	4.20	11.06	4.43	11.10	4.06	13.92	2.36	10.77	4.24
Households on Food Stamps	$32,689.51$	$22,867.06$	$32,513.99$	$22,992.72$	$32,797.74$	$22,885.15$	$40,544.17$	$16,993.94$	$31,836.71$	$23,295.01$
Population	$710,598.90$	$426,805.30$	$683,677.20$	$428,324.90$	$727,200.60$	$426,811.10$	$793,793.20$	$335,023.70$	$701,566.30$	$435,436.70$
Households	$291,158.70$	$182,394.20$	$279,881.80$	$183,541.00$	$298,112.80$	$182,104.20$	$326,193.90$	$143,085.80$	$287,354.90$	$186,101.50$
School Students	$92,137.21$	$60,339.76$	$99,632.97$	$54,946.58$	$87,514.82$	$63,217.19$	$104,404.80$	$50,996.68$	$90,805.30$	$61,247.07$
Young (%)	29.89	3.19	29.88	3.35	29.90	3.10	28.53	1.15	30.04	3.30
Elderly (%)	16.72	2.83	16.55	2.89	16.83	2.81	18.18	1.46	16.57	2.90

Notes: Observation level is the municipality

Source: Author's calculations

To successfully apply the research design requires that the common trends assumption holds between the treatment group (made up of Minneapolis and St. Paul) and the control group (other municipal areas). The likelihood of the common trends assumption holding can be tested by considering turnout in elections that should not be affected by the policy. The plots below show turnout levels over time for treated and control groups in mid-term and presidential elections respectively. Overall, it appears that both groups follow roughly the same time-trend and there appears to be little direct effect from the policy on turnout, particularly so for mid-term elections in Figure 3. This gives support to the common trends assumption being applicable for the case of municipal elections.

It's also possible to graphically see the likely treatment effect in the raw data. Looking at trends across time, as shown in the charts above, is challenging as year on year variation tends to dominate. However, simplifying into 'before' and 'after' groups clarifies the graphical analysis. Figure 5 shows turnout rates for municipal elections, while Figure 6 shows turnout rates for all other elections. Figure 5 indicates that the introduction of IRV may have led turnout to decrease more slowly than it otherwise would have, while Figure 6 – where there should be no treatment effect – indicates that some of this may have been due to an underlying increase in turnout in treated municipalities. Together this graphical analysis suggests that, in the raw data, the treatment effect of IRV is likely to be positive.

B. Civility

For Civility, the goal is to have a measure of civility in mayoral debates for a large range of cities across the U.S.

To do this, the main piece of data that is used are transcripts from mayoral debates. These transcripts were predominantly sourced from YouTube videos of mayoral debates. This paper is the first to use modern, natural language processing techniques to analyses the civility of debates and is also the first to make use of the large amount of transcribed debates that are now available from data sources like YouTube. The use of natural language processing techniques is a major contribution of this paper as it ensures impartiality and allows the analysis of civility to be done a scale not possible when using manual techniques.

A structured process was followed to identify and gather the transcripts. First, the largest 100 municipalities in the U.S., measured by population, were identified. These municipalities account for around 62 million people which is about 19% of the entire U.S. population. For each of these cities, a manual search on YouTube was undertaken to identify relevant videos. The search was based on the phrase "[City Name] Mayoral Debate", so that, for example, searching for New York City was done using the phrase "New York Mayoral Debate". Search filters were applied to return only videos longer than 20 minutes. This process identified a total of 459 relevant videos from 78 municipalities covering years from 1988 to
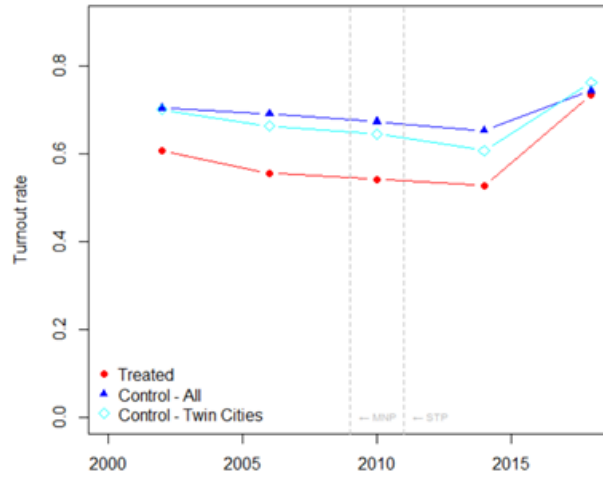
FIGURE 3. TURNOUT RATE IN UNAFFECTED ELECTIONS – MID-TERMS

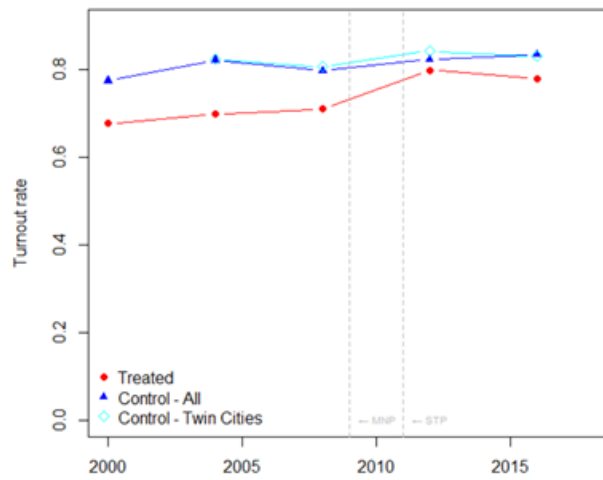*Source:* Author's calculations

FIGURE 4. TURNOUT RATE IN UNAFFECTED ELECTIONS – PRESIDENTIAL
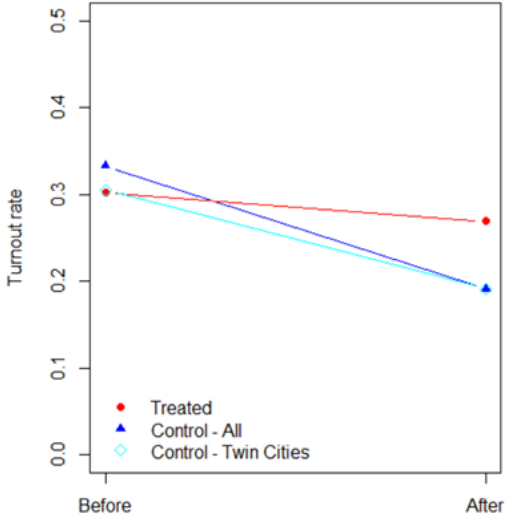
Source: Author's calculations

FIGURE 5. TURNOUT RATE IN MUNICIPAL ELECTIONS

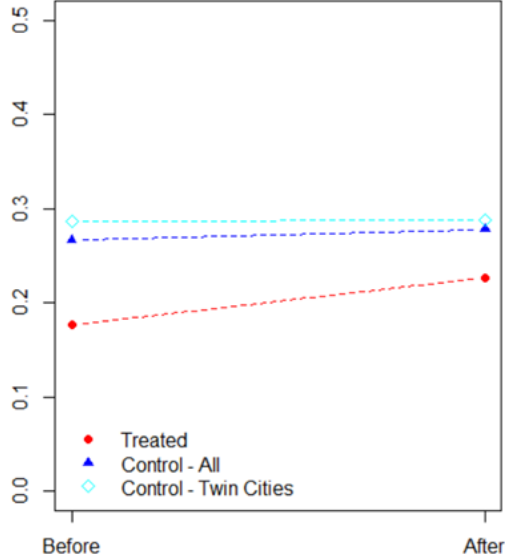*Source:* Author's calculations

FIGURE 6. TURNOUT RATE IN NON-MAYORAL, STATE AND NATIONAL ELECTIONS

Source: Author's calculations

upcoming elections in 2020.

A Python and R script was then used to check for the availability of subtitles for each video. Most of these subtitles are generated by Google's automatic speech recognition technology (Harrenstien, 2009)[15]. In total, 329 of the 459 videos had transcripts available, creating a total word count of around 3 million words. These videos covered 70 municipalities. At an average speaking pace of 120 words per minute, this results in just over 400 hours of debate video for which transcripts are available.

This process did not identify any debate recordings for Minneapolis prior to the introduction of IRV. As a result, two additional recordings were sourced from PBS's video archive for the 2005 mayoral race (PBS, 2020). Additional research, including contacting previous candidates, organizations that hosted debates and Minneapolis Public Radio, did not identify further transcripts that could be added to the database.

As there is no thorough data available on the quality of Google's automated transcripts, the quality was verified by hand for all debates in Minneapolis in 2013. This hand verification indicated that the quality of the machine generated transcripts was good, with only minor corrections identified.

Some supporting socioeconomic data was sourced from the American Community Survey (ACS) (United States Census Bureau, 2019). Data was matched between the debate transcripts and the socioeconomic data using the main county in which the municipality is located. The ACS data was only available from 2010 to 2017 and, due to the presence of fixed effects, can only be used in regressions for cities that have debate transcripts available for multiple years. This restriction reduces the dataset to a final 227 debates from 37 cities. The remaining cities are reported in Appendix A.

The text from the transcripts was cleaned by converting it to lower case, removing excess spaces and whitespace, and removing punctuation and special characters. This created clean text data that could be used in the sentiment analysis.

The sentiment analysis was undertaken using both the AFINN and Bing lexicons. The AFINN lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment (Nielsen, 2011). The AFINN lexicon contains around 2500 words. The Bing lexicon categorizes words into positive and negative categories, in this analysis these are coded as +1 and -1, respectively (Hu and Liu, 2004). The Bing lexicon contains around 7000 words. There are about 1300 words that appear in both the AFINN and Bing lexicons and only 17 where there is a disagreement between the two lexicons on whether the word is positive or negative in senti-

---

[15]As at 2020, captions on YouTube's desktop site can be accessed by clicking the ellipses below a video and then clicking on the "Open Transcript" option.

ment. To apply the lexicons, each word in each speech was checked against the lexicon and assigned its relevant score. There were only two words in the 2013 Minneapolis debates that were assigned a score of +5 using AFINN: these were 'outstanding' and 'superb'. Words with an AFINN of +4 in the 2013 Minneapolis debates include 'awesome', 'fabulous' and 'win'. On the negative side, there were no words used in the 2013 Minneapolis debates with an AFINN of -5 and the only word used with an AFINN of -4 was 'hell'.

The results of the sentiment analysis were summarized in a number of measures: the total number of words in the sentiment lexicon used during a debate; the average score of words in the lexicon used during a debate; and the average score of all words used in the debate. Table 3 provides a summary of the sentiment analysis for both the AFINN and Bing lexicons.

Table 3—Summary of Sentiment Analysis

	Min	Median	Mean	Max	SD
Word Count	286	$9,689$	$9,503.7$	$24,326$	$4,438.3$
AFINN Word Count	9	486	489.3	$1,291$	234.7
AFINN Total Score	-118	340	377.3	$1,452$	229.2
Average AFINN Score	-1.2	0.8	0.8	1.4	0.3
AFINN Per Word	-0.04	0.04	0.04	0.1	0.02
Bing Word Count	1	468	475.4	$1,312$	225.4
Bing Total Score	-13	174	187.8	620	107.7
Average Bing Score	-1	0.4	0.4	0.7	0.2
Bing Per Word	-0.003	0.02	0.02	0.04	0.01
Stop Word Count	2	$5,857$	$5,694.0$	$14,538$	$2,672.9$

Source: Author's calculations

In this table, Word Count is the total number of words spoken in a debate, AFINN/Bing Word Count is the number of words spoken that appear in the AFINN/Bing lexicon, AFINN/Bing Total Score is the sum of AFINN/Bing values for all words spoken in the debate, Average AFINN/Bing score is AFINN/Bing Total Score ÷ AFINN/Bing Word Count and AFINN/Bing Per Word is AFINN Total Score ÷ Word Count. In summary, an average debate has around 9,700 words (at around 120 words per minute, this is a speaking time of around 80 minutes). Of these, around 490 words are in the AFINN lexicon and 470 in the Bing lexicon. The mean Bing Total Score of 188 indicates that, on average, there are more positive words said than negative words, but both the Average AFINN (mean 0.8) and Average Bing (0.4) indicate that the average word used, given that it is in the lexicon, is positive but not resoundingly so.

Stop Word Count records the frequency of use of common English language words. The list of words is based on a database compiled by Silge and Robinson (2016)

but has been edited to remove words that appear in the AFINN and Bing lexicons. Words appearing in the NRC lexicon were also excluded (Mohammad and Turney, 2010). The NRC lexicon is not used in this paper as it generates categorical rather than quantitative sentiment assignments, but the presence of a word in the NRC lexicon still indicates that it contains important linguistic information and so should therefore be excluded from the stop word list.

Although all measures will be reported in Section V, 'AFINN/Bing Per Word' is the preferred measure as it provides both an indication of the frequency of word usage as well as the sentiment of word usage. AFINN is preferred to Bing as it provides a classification of the intensity of sentiment as well as the direction of sentiment. The correlation coefficient between the total AFINN score of a debate and the total Bing score is around 0.95.

As with the analysis of Participation, the application of the research design requires that the common trends assumption holds. In this case the common trends assumption would mean that municipalities that adopt IRV would follow the same trend in civility of debates as municipalities that don't adopt IRV.

Visual inspection of the common trends assumption is more difficult with the data used for civility because of the intermittent timing and availability of debate transcripts for treated and control cities. Plotting the average AFINN per word does, however, provide some indication that the common trends assumption may be reasonable. Figure 7 shows that treated municipalities are broadly within the same range of AFINN scores over time and that there does not appear to be a divergence between treated and control municipalities. This is true for treated municipalities both before and after they receive treatment.

It's again possible to graphically see the likely treatment effect in the raw data. As an example, Figure 8 shows a simplified case where the data is assigned into 'before' and 'after' groups and only focuses on the cities of Minneapolis and St. Paul in the treated group. Figure 8 indicates that the introduction of IRV may have led to an improvement in the sentiment of debates in the Twin Cities, as the AFINN score per word increases and this is counter to the decreasing trend in the control cities. However, the standard errors in this simple example are large, indicating likely challenges in achieving statistical significance if the analysis of civility was limited to the Twin Cities alone in the treated group.

V.   Main Results

A.   Participation

For Participation, there are several sets of results presented below with a range of robustness and specification tests also shown.

For all models presented below, the results are presented using clustered standard errors with the cluster being defined at the municipal level. Clustered standard
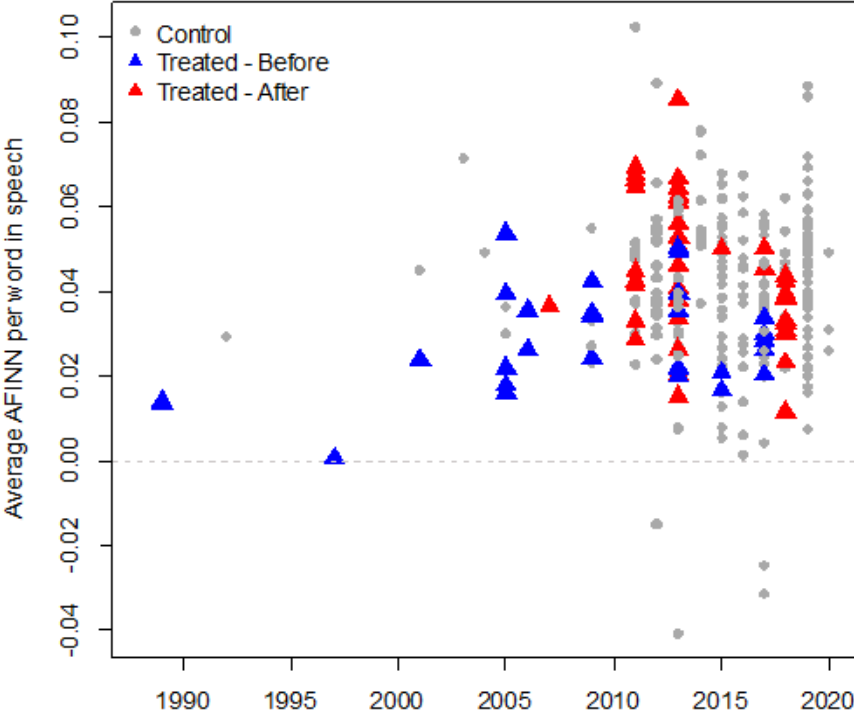
FIGURE 7. AVERAGE AFINN PER WORD IN SPEECHES OVER TIME FOR TREATED AND CONTROL CITIES
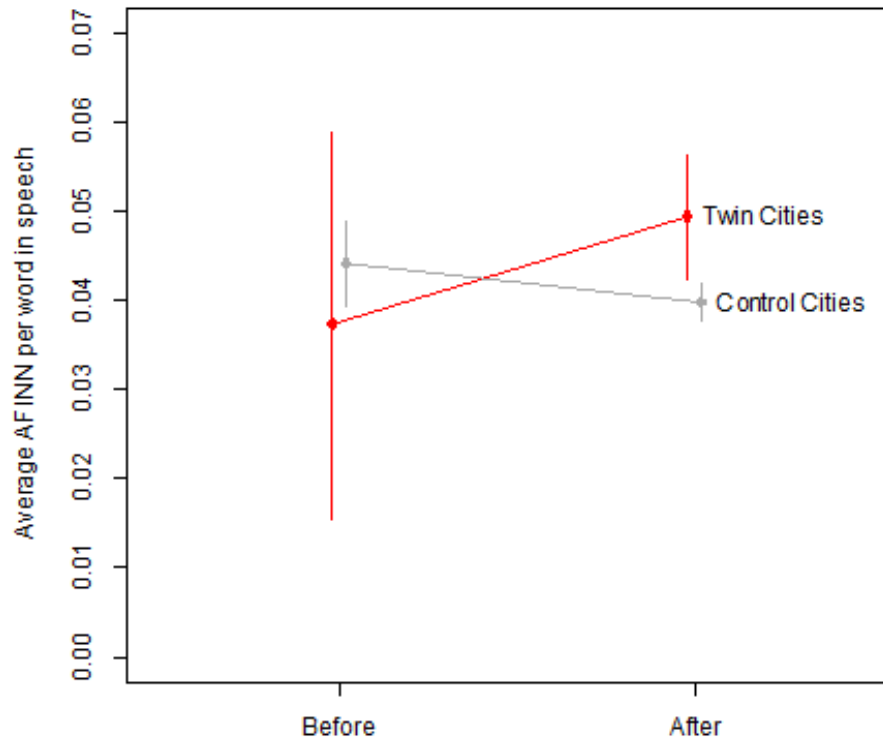
Source: Author's calculations

FIGURE 8. AVERAGE AFINN PER WORD IN THE TWIN CITIES AND CONTROL CITIES, GROUPED INTO
BEFORE AND AFTER

Source: Author's calculations

errors have been used to address issues identified in Bertrand, Duflo and Mullainathan (2004). Clustering at the municipal level allows for the error term to be correlated within a municipality but independent between municipalities. This assumption seems reasonable as turnout within a municipality is not likely to be affected by turnout in other municipalities but there is likely to be some form of correlation over time within the same municipality.

The first table, Table 4, presents the main results for participation and shows two alternative models to demonstrate how different model specifications affect the results. Column 1 shows the results when including only year and municipality fixed effects while Column 2 introduces socioeconomic covariates. This results in relatively minor changes to the estimated covariate of interest and no meaningful changes to statistical significance when considering the clustered standard errors. The introduction of covariates does shift the Conley Taber confidence interval so that it excludes zero.

The parameter estimate for IRV has a straightforward interpretation: its value of 0.096 indicates that the introduction of IRV results in a 9.6 percentage point increase in turnout, on average, while holding other factors constant in Minneapolis and St. Paul mayoral elections. The size of these parameter values suggests significance in practical terms as well as in statistical terms.

As explained in Section IV.A, the inclusion of socioeconomic covariates required significant reductions in the data available for the regression due to both difficulties in matching some locations and the limited number of years for which socioeconomic data was available. A further reduction in the data occurs as the results above focus on other municipalities in the Twin Cities as being the relevant control group. These reductions in data could have the potential to affect the estimated treatment effect and the precision with which it is estimated. Table 5 reports results at each of the stages of data reduction. Overall, the estimated treatment effect does vary but only changes from around 6.6 percentage points to 9.6 percentage points.

Although the data is available at the precinct level, the treatment is applied at the municipal level, which means that the results above are presented at the municipal level. Table 6 and Figure 9 show versions of the results at the precinct level as well as at the municipal level and verify that aggregation does not result in a major change to the estimated treatment effect. This is unsurprising as the municipal level results should be some weighted average of the precinct level results. However, the fact that the results are similar in nature is useful in the following subgroup analysis as looking at the precinct level data allows for far more variation in the underlying social and economic variables than aggregating up to the municipal level.

A number of subgroups within the data were analyzed using data at the precinct level. The literature review indicated that minorities or at-risk groups (such

TABLE 4— MAIN RESULTS: PARTICIPATION

	Dependent variable: Turnout	
	(1)	(2)
IRV	0.086**	0.096**
	(0.041)	(0.045)
Non-Mayoral Election Dummy	✓	✓
Year FE	✓	✓
Municipality FE	✓	✓
Covariates		✓
Conley-Taber 90% CI	(-0.003, 0.185)	(0.005, 0.197)
Observations	194	194
R^2	0.958	0.962

Note: Data in this regression is aggregated to the municipality level, the level where treatment is applied. Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Column 1 shows the results of a basic difference-in-differences regression; Column 2 introduces a range of socioeconomic covariates. Confidence intervals estimated using the approach of Conley and Taber (2011) are reported in the footer of the table.
*p<0.1; **p<0.05; ***p<0.01
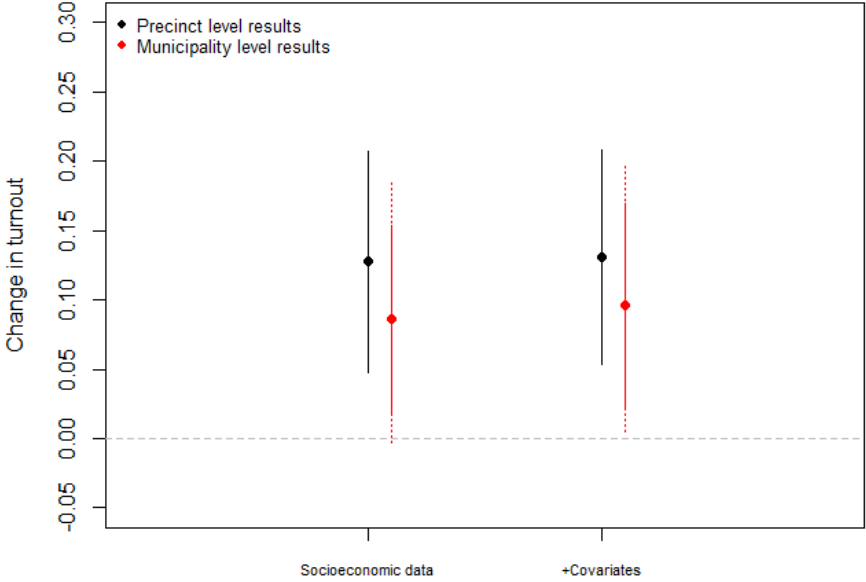
FIGURE 9. ESTIMATED TREATMENT EFFECTS AND CONFIDENCE INTERVALS

Note: Solid lines show 95% confidence intervals using clustered standard errors while dashed lines show 95% confidence intervals using Conley Taber confidence intervals. For the precinct level results, the Conley Taber confidence intervals are smaller than the clustered standard error confidence intervals.
Source: Author's calculations

TABLE 5—EFFECT OF DATA REDUCTIONS ON TREATMENT EFFECT

	Dependent variable: Turnout				
	FE	Socioeconomic data	+Covariates	Twin Cities	+Covariates
	(1)	(2)	(3)	(4)	(5)
IRV	0.096***	0.066*	0.071*	0.086**	0.096**
	(0.020)	(0.034)	(0.039)	(0.041)	(0.045)
Non-Mayoral Election Dummy	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Covariates			✓		✓
Conley-Taber 90% CI	(0.033, 0.206)	(-0.02, 0.176)	(-0.024, 0.176)	(-0.003, 0.185)	(0.005, 0.197)
Observations	611	294	294	194	194
Adjusted R^2	0.899	0.926	0.929	0.958	0.962

Note: Data in this regression is aggregated to the municipality level, which is the level where treatment is applied. Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Column 1 shows the results of a basic difference-in-differences regression using the full set of available data at the state level; Column 2 shows the same regression but only using data for which socioeconomic variables are available; Column 3 includes these covariates; Column 4 restricts the data further to only show data for municipalities within the Minneapolis-St. Paul Metro Area; Column 5 introduces a range of socioeconomic covariates. Confidence intervals estimated using the approach of Conley and Taber (2011) are reported in the footer of the table. *p<0.1; **p<0.05; ***p<0.01

as those of lower socioeconomic status) may be important in determining the effect of voting policies. Investigation of subgroups based on indicators of food stamp usage and differences in race did not identify any statistically significant differences between groups. Table 7 shows the results when splitting the sample depending on poverty rates. The results indicate that areas with higher indicators for poverty saw turnout increase substantially more than areas with lower levels of poverty indicators. This suggests that the introduction of IRV affected turnout particularly strongly for lower income voters.

The results presented in this section provide evidence that the transition to IRV resulted in an increase in turnout in the Minneapolis-St. Paul Metro Area. Further, the results indicate that the increase in turnout was most prominent in lower income areas.

B. Civility

For civility, the first two tables show the main results when using the AFINN and Bing lexicons, respectively. The columns in each table are the same and align with the summary statistics reported in Section IV.B. In this table, Word Count is the total number of words spoken in a debate, AFINN/Bing Word Count is the number of words spoken that appear in the AFINN/Bing lexicon, AFINN/Bing Total Score is the sum of AFINN/Bing values for all words spoken in the debate,

Table 6—Effect of Geographic Aggregation on Treatment Effect

	Dependent variable: Turnout			
	Precinct level		Municipal level	
	(1)	(2)	(3)	(4)
IRV	0.127***	0.131***	0.086**	0.096**
	(0.049)	(0.047)	(0.041)	(0.045)
Non-Mayoral Election Dummy	✓	✓	✓	✓
Year FE	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓
Covariates		✓		✓
Conley-Taber 90% CI	(0.061, 0.197)	(0.066, 0.199)	(-0.003, 0.185)	(0.005, 0.197)
Observations	5,850	5,850	194	194
Adjusted R^2	0.904	0.904	0.950	0.953

Note: Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level Columns 1 and 3 show the results of a basic difference-in-differences regression; Columns 2 and 4 introduce a range of socioeconomic covariates. Columns 1 and 2 present results where the observation is the (Precinct, Year) pair while Columns 3 and 4 present results where the observation is the (Municipality, Year) pair. A single municipality contains many precincts. Confidence intervals estimated using the approach of Conley and Taber (2011) are reported in the footer of the table. For Columns 1 and 2, the Conley and Taber confidence intervals are calculated using the simulation approach of Proposition 2, due to the large number of treated and control observations. *p<0.1; **p<0.05; ***p<0.01

TABLE 7—SUBGROUP ANALYSIS - POVERTY RATES

	Dependent variable: Turnout	
	(Lower Poverty)	(Higher Poverty)
IRV	0.074*	0.243***
	(0.039)	(0.017)
Non-Mayoral Election Dummy	✓	✓
Year FE	✓	✓
Municipality FE	✓	✓
p-value for test of difference		0.000
Average Turnout	66.6	61.8
Observations	1,819	2,853
R^2	0.912	0.912

Note: Data in this regression is at the precinct level. Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Data has been split into a 2-quantile. The 'Lower Half' includes precincts with low rates of poverty indicators while the 'Upper Half' includes precincts with high rates of poverty indicators. In Panel B, the uneven number of observations in each column comes about due to allocation of observations with the same observed poverty rates. *p<0.1; **p<0.05; ***p<0.01

Average AFINN/Bing score is AFINN/Bing Total Score ÷ AFINN/Bing Word Count and AFINN/Bing Per Word is AFINN Total Score ÷ Word Count.

The results in Table 8, for AFINN, indicate that the change in the total number of words, the number of words in the AFINN lexicon and the total AFINN score are not statistically significant at conventional levels. There is a statistically significant increase in both the Average AFINN Score (10% level of significance) and AFINN Per Word (5% level of significance). Overall, this indicates that, although the total amount spoken during the debate doesn't change much following the introduction of IRV, there is a statistically significant change in the type of language used following the introduction of IRV. Using Conley-Taber confidence intervals to address the small number of treated units in the data set indicates that the 90% Confidence interval does not include zero for AFINN per word, whic is the preferred measure.

The results indicate that the length of the debates and number of relevant words used remains unchanged while there is an improvement in the rate of positive sentiment words. This leads to an increase in both the average score of words from the lexicon as well as an increase in the average score of each word in the debate. This suggests that the introduction of IRV results in a substitution of negative or neutral words towards more positive words. This provides support for the proposition that the introduction of IRV leads to an improvement in the civility of mayoral debates.

Table 8—Main Results – Civility – AFINN

	Dependent variable:				
	Word Count	AFINN Word Count	AFINN Total Score	Average AFINN Score	AFINN Per Word
	(1)	(2)	(3)	(4)	(5)
IRV	4,998.009	557.615	866.205	1.138*	0.079**
	(13,181.570)	(798.810)	(752.079)	(0.591)	(0.039)
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Covariates	✓	✓	✓	✓	✓
Conley-Taber 90% CI	(-18536.146, 35310.002)	(-902.217, 2250.681)	(-560.202, 2363.529)	(-0.958, 1.983)	(0.004, 0.14)
Observations	156	156	156	156	156
Adjusted R^2	0.124	0.099	0.104	0.171	0.197

Note: Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Word Count is the total number of words spoken, AFINN Word Count is the number of words spoken that appear in the AFINN lexicon, AFINN Total Score is the sum of AFINN values for all words spoken, Average AFINN score is (3)÷(2) and AFINN Per Word is (3)÷(1). Covariates are County population and County population over 65. Confidence intervals estimated using the approach of Conley and Taber (2011) are reported in the footer of the table. These are based on the simulation approach of Proposition 2. *p<0.1; **p<0.05; ***p<0.01

The results in Table 9, for the Bing lexicon, provide broadly similar results as were seen when using the AFINN lexicon, but statistical significance is not achieved in any measure. An advantage of the Bing lexicon is the ease of interpretting results,

noting that the lack of statistical significance reduces the weight that should be placed on this interpretation. The change in Bing Total Score of around +280 indicates a net increase of 280 positive words used per speech. Comparing this to the change in Bing Word Count indicates that speakers are saying about 65 more negative words and 345 more positive words, which results in a ratio of around 84:16.

Table 9—Main Results – Civility – Bing

	Dependent variable:				
	Word Count	Bing Word Count	Bing Total Score	Average Bing Score	Bing Per Word
	(1)	(2)	(3)	(4)	(5)
IRV	4,998.009	409.054	279.704	0.468	0.028
	(13,181.570)	(698.689)	(387.290)	(0.365)	(0.017)
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Covariates	✓	✓	✓	✓	✓
Conley-Taber 90% CI	(-18536.146, 35310.002)	(-876.543, 1862.992)	(-311.652, 978.376)	(-0.115, 0.993)	(-0.001, 0.059)
Observations	156	156	156	156	156
Adjusted R^2	0.124	0.124	0.105	0.083	0.193

Note: Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Word Count is the total number of words spoken, Bing Word Count is the number of words spoken that appear in the Bing lexicon, Bing Total Score is the sum of Bing values for all words spoken, Average Bing score is (3)÷(2) and Bing Per Word is (3)÷(1). Covariates are County population and County population over 65. Confidence intervals estimated using the approach of Conley and Taber (2011) are reported in the footer of the table. These are based on the simulation approach of Proposition 2. *p<0.1; **p<0.05; ***p<0.01

The results in this section provide evidence that changing the electoral system to IRV improves the civility of the debate between candidates. The analysis indicates that candidates in IRV elections tended to substitute negative or neutral words for more positive words during the debate. These results align with the previous literature based on voter and candidate surveys and interviews, which generally find perceived improvements from the introduction of IRV.

VI. Robustness Tests

A. Participation

A range of additional robustness checks were also undertaken. Table 10 presents the results of a placebo test where the dependent variable is changed by randomly reallocating treatment across all observations (while holding the total rate of treatment constant). This creates a treatment variable where there is not expected to be a genuine treatment effect and the turnout rate should not, theoretically, be affected by the randomized IRV variable. The treatment is not found to be statistically significant at conventional levels in either specification. This

placebo test provides supporting evidence that the treatment effect estimated in the main results is a genuine effect and not a chance result of noise in the data.

Table 10—Placebo Test - Participation

	Dependent variable: Turnout	
	(1)	(2)
Randomised IRV	−0.031	−0.028
	(0.034)	(0.034)
Non-Mayoral Election Dummy	✓	✓
Year FE	✓	✓
Municipality FE	✓	✓
Covariates		✓
Observations	194	194
Adjusted R^2	0.949	0.951

Note: Data in this regression is aggregated to the municipality level, which is the level where treatment is applied. Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level Column 1 shows the results of a basic difference-in-differences regression; Column 2 introduces a range of socioeconomic covariates. In both data sets, the treatment has been randomly reallocated within the data. $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Next, reported in Table 11, three alternative model specifications were tested. The first model replaces the single IRV variable with three variables to indicate the year of treatment. This model is to test the possibility that the treatment effect may be due to an initial boost in turnout and not a genuine ongoing effect. The next model uses a lagged dependent variable (LDV) approach. LDV adjusts for pre-treatment outcomes and is considered more appropriate than DID in cases where the common trends assumption does not hold (O'Neill et al., 2016). The last model uses data that has been matched on pre-treatment outcomes[16]. Matching on pre-treatment outcomes is another approach that can be used if the parallel trends assumption does not hold. Applying DID to the matched data then allows for control of unobserved time-invariant factors (O'Neill et al., 2016).

Columns 1 and 2 of Table 11 show that the treatment effect measured in the

[16]In particular, the matching is done using the turnout in 2006 and 2008. Matching is carried out using the algorithms provided in Ho et al. (2007) using "full matching," which offers variable numbers of matches in each subclass (Hansen, 2004). The absolute value of the standardised mean difference post matching for pre-treatment outcomes is less than 0.1, which is the rule of thumb for a successful match provided in Flury and Riedwyl (1986)

Table 11—Alternative model specifications

	Year of Treatment		Lagged Dependent Variable		Matched on Pre-treatment outcomes	
	Dependent variable: Turnout					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant			-0.328^{***} (0.102)			
IRV_{1st}	0.062^{**} (0.026)	0.063^{**} (0.029)				
IRV_{2nd}	0.099 (0.065)	0.113 (0.073)				
IRV_{3rd}	0.087^{**} (0.036)	0.072^{*} (0.039)				
IRV			0.088^{***} (0.034)	0.115^{***} (0.042)	0.054 (0.034)	0.056 (0.037)
Non-Mayoral Election Dummy	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓	✓
Presidential Election FE			✓	✓		
Mid Term FE			✓	✓		
Lagged Outcomes			✓	✓		
Matched Data					✓	✓
Covariates		✓		✓		✓
Observations	194	194	194	194	194	194
Adjusted R^2	0.950	0.952	0.944	0.947	0.976	0.978

Note: Data in this regression is aggregated to the municipality level, which is the level where treatment is applied. Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Columns 1 and 2 shows the results of a model that includes dummies for the year of treatment; Columns 3 and 4 show the results of a model that includes lagged dependent variables for 2006 and 2008; Columns 5 and 6 show the results of a model that weights the data to match pre-treatment outcomes. For further information on columns 3-6, please refer to O'Neill et al. (2016). *p<0.1; **p<0.05; ***p<0.01

main results is present when the treatment is broken out by year and and that there doesn't appear to be a strong trend over time. Columns 3 and 4 of Table 11 indicate that, after controlling for lagged dependent variables, roughly the same result as in the main results is found (although larger and more precisely estimated). Columns 5 and 6 do show the same type of results as the main results – an increase in turnout – but the parameter is not statistically significant at conventional levels. This is likely because the matching algorithm puts very high weight on a small number of observations.

A final check, Table 12, corrects for the potential presence of an Ashenfelter dip – whereby the treated units implement IRV to address a decline in voter turnout prior to the policy change (Ashenfelter and Card, 1984). This is unlikely to be the case in this data as the adoption of IRV occurred due to a popular vote and not at the discretion of an individual or small group. Moreover, it was also implemented after a lag of up to 3 years. To correct for the potential presence of an Ashenfelter dip, simple linear models were used to predict turnout at the precinct level for the year 2005 in Minneapolis and 2008 in St. Paul. These models were fitted using data from pre-2005 and pre-2008, respectively, and the predicted values were used to replace the actual turnout values. In general, the predicted values were higher than the actual turnout values.

TABLE 12— CORRECTION FOR POTENTIAL ASHENFELTER DIP

	Dependent variable: Turnout	
	(1)	(2)
IRV	0.085**	0.095**
	(0.042)	(0.046)
Non-Mayoral Election Dummy	✓	✓
Year FE	✓	✓
Municipality FE	✓	✓
Covariates		✓
Observations	194	194
Adjusted R^2	0.950	0.953

Note: Data in this regression is aggregated to the municipality level, whic is the level where treatment is applied. Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Column 1 shows the results of a basic difference-in-differences regression; Column 2 introduces a range of socioeconomic covariates.
*p<0.1; **p<0.05; ***p<0.01

The results in Table 12 are very similar in size and statistical significance to those in the main results, indicating that the role of a potential Ashenfelter dip is not responsible for the results.

The robustness checks presented demonstrate that this finding is stable across a range of model specifications and analysis approaches.

B. Civility

The following tables focus on providing some additional analysis and robustness checks on the main results above. Table 13 shows the effect of introducing socioeconomic covariates. Columns 2 and 4 of Table 13 are identical to column 5 in Table 8 and Table 9, respectively, while Columns 1 and 3 show the same models but do not include the socioeconomic covariates. The results indicate that the addition of these covariates does not result in a change in sign or inference of any of the estimated coefficients. Rather, the covariates lead to an increase in the size of the estimated effect of the introduction of IRV and an increase in the standard error of the estimate. This indicates that the inclusion or exclusion of socioeconomic variables is not fundamental to the nature of the findings but does affect the precise values of parameter estimates.

TABLE 13—EFFECT OF ADDITION OF COVARIATES ON ESTIMATES

	Dependent variable:			
	AFINN Per Word		Bing Per Word	
	(1)	(2)	(3)	(4)
IRV	0.017**	0.079**	0.002	0.028
	(0.008)	(0.039)	(0.003)	(0.017)
Year FE	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓
Covariates		✓		✓
Observations	331	156	331	156
Adjusted R^2	0.227	0.197	0.229	0.193

Note: Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Covariates are County population and County population over 65.
*p<0.1; **p<0.05; ***p<0.01

The analysis above uses the AFINN and Bing sentiment lexicons with the preferred results being from the AFINN lexicon. These two lexicons are useful for

analysis of political debate as they include words used in general speech rather than being targeted at a particular type of speech. There are, however, a wide range of lexicons available for sentiment analysis, some of which are targeted at financial and economic discussions. Loughran and McDonald (2011) lexicon was developed from a review of words used in 10-K filings and has been used in other analysis of sentiment by economists, such as Shapiro and Wilson (2019), who analyze the sentiment of speeches by the Federal Open Market Committee. Loughran and McDonald compiled their own lexicon as they found that existing lexicons available in 2011 didn't apply satisfactorily to the type of language used in financial discussion. To address this, they reviewed and classified common words used in 10-K filings and demonstrated that this approached produced results with greater predictive power than existing lexicons. Loughran and McDonald's lexicon may not be as suitable for analyzing political speeches as candidates will likely try to convey their message in simple language rather than the technical and audience specific language used in financial statements. However, as a robustness check, Table 14 reproduces the main results when using the positive and negative words included in the Loughran and McDonald lexicon (with positive coded as +1 and negative coded as -1 to allow direct comparison to the Bing results in Table 9).

Table 14—Robustness Check – Civility – Loughran and McDonald Lexicon (L&M)

	Dependent variable:				
	Word Count	L&M Word Count	L&M Total Score	Average L&M Score	L&M Per Word
	(1)	(2)	(3)	(4)	(5)
IRV	4,998.009	223.838***	112.625***	0.933*	0.020
	(13,181.570)	(31.400)	(20.329)	(0.511)	(0.016)
Year FE	✓	✓	✓	✓	✓
Municipality FE	✓	✓	✓	✓	✓
Covariates	✓	✓	✓	✓	✓
Conley-Taber 90% CI	(-18536.146, 35310.002)	(-600.232, 789.071)	(-85.916, 407.476)	(-0.204, 1.791)	(-0.006, 0.041)
Observations	156	156	156	156	156
Adjusted R^2	0.124	0.113	0.268	0.145	0.137

Note: Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Word Count is the total number of words spoken, LM Word Count is the number of words spoken that appear in the Loughran and McDonald (2011) lexicon, LM Total Score is the sum of LM values for all words spoken, Average LM score is (3)÷(2) and LM Per Word is (3)÷(1). Covariates are County population and County population over 65. Confidence intervals estimated using the approach of Conley and Taber (2011) are reported in the footer of the table. These are based on the simulation approach of Proposition 2.
*p<0.1; **p<0.05; ***p<0.01

Overall, the results are more statistically significant than seen in the results when using either the AFINN or the Bing lexicons and the sign and magnitude of the coefficients are similar. All of the Conley Taber confidence intervals do, however, include zero, indicating that the inference is not robust to accounting for the small

number of treated units in the analysis. These results indicate that the nature of the results isn't wholly a result of the particular lexicon used for the analysis and that the results are seen even when using more audience specific lexicons.

Table 15, below, presents the results of a placebo test. In this analysis, Column 1 shows the results where the dependent variable is the number of stop words used in a debate. Stop words are common English language words that do not convey much topical information but are important for sentence construction. They include words such as 'those', 'into', 'the' and 'now'. The initial stop word database was sourced from Silge and Robinson (2016) but was adjusted to remove any words appearing in the AFINN, Bing and NRC lexicons. This was done to ensure that words relevant to the results in Table 8 and Table 9 were not included in the placebo test. As the remaining stop words are frequently used to construct English sentences regardless of the context, it should be the case the frequency of their use should not be affected by the introduction of IRV.

The other columns of Table 15 show results when the dependent variable is set to be the rate of usage of particular words that, intuitively, shouldn't be affected by the introduction of IRV. For example, the dependent variable in Column 2 is the rate at which the word 'mayoral' is used within a debate. These words were selected based on the manual review of 2013 Minneapolis debates as words that were likely to appear in debates but which, on their own, do not necessarily contain any positive or negative intention. For each of Columns 1 to 5, the IRV variable is not statistically significant at conventional levels, indicating that the introduction of IRV did not result in a change in the frequency of the use of that particular word. This gives support to the earlier results by providing some evidence that the effect is not the results of noise in the data or a random occurrence.

VII.   Conclusion

In recent years, IRV has seen a resurgence in use in a number of metropolitan areas in the U.S. Proponents of IRV have claimed a range of benefits including: ensuring majority support, reducing costs, increasing civility, reducing conflict, reducing strategic effects and increasing diversity. These direct benefits have also been expected to translate into an increase in voter turnout.

A sizable literature has developed analyzing the various purported benefits of IRV and there is mounting evidence that IRV has not been living up to expectations in many areas. This paper focuses on two particular areas of potential benefit of IRV that have mixed results and weak methodologies in the existing literature: increasing turnout and improving civility.

For turnout, this paper focuses on a DID research design limited to the Twin Cities Metropolitan Area. Limiting analysis to the Twin Cities Metropolitan Area helps ensure that the common trends assumption underlying the DID approach

Table 15—Placebo Test – Sentiment Analysis

	Dependent variable:		
	Stop Word Rate	'Mayoral' Rate	'Voter' Rate
	(1)	(2)	(3)
IRV	0.001	0.001	−0.001
	(0.115)	(0.001)	(0.001)
Year FE	✓	✓	✓
Municipality FE	✓	✓	✓
Covariates	✓	✓	✓
Observations	156	156	156
Adjusted R^2	−0.156	0.167	0.273

Note: Standard errors reported in parentheses are robust to heteroskedasticity and are clustered at the municipality level. Stop Words Rate is defined as the number of words appearing in a standard list of common English language words – based on a database compiled by Silge and Robinson (2016) – divided by the total number of words spoken. Covariates are County population and County population over 65. $^*p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

is likely to be supported. This approach is in contrast to previous papers that have not had a clear research design or, when they do, they look at a broad range of treated cities without a clear motivation for establishing the control group. The results of the analysis indicate that, in the Minneapolis-St. Paul Metro Area, the introduction of IRV caused a 9.6 percentage point increase in turnout, on average. This result is statistically significant at conventional levels using clustered standard errors and 90% Conley-Taber confidence intervals that exclude zero. The effect on turnout is larger for precincts that have higher poverty rates.

For civility, previous research has essentially used surveys or interviews whereas new techniques based on natural language processing allow for a more precise quantification of the effect of IRV on civility during campaigns. Analysis of the sentiment of language used during mayoral debates indicates that the introduction of IRV improved the civility of debates. The improvement in civility is due to candidates substituting negative or neutral words for more positive words throughout the debate.

The findings on turnout could be extended by applying a similar DID approach in other cities that have both a long history of IRV and staggered geographic introduction (San Francisco, for example) in order to determine whether Minneapolis and St. Paul present an unusual case[17] or whether the results here have external validity. Analysis of civility will naturally improve over time as more debates are recorded and transcribed on the internet. The text-as-data approach also opens up the possibility for analyzing the effect of IRV on other outcomes, such as the topic discussed in debates and whether winners of IRV elections speak differently than winners of FPTP elections. These questions require different methodologies than what is used in this paper[18].

The positive findings in this paper indicate that the introduction of IRV is performing better than the previous literature would suggest. To the extent that turnout can be seen as a barometer of the overall value of a vote (Downs, 1957), these results also suggest that IRV is having a positive effect on the perceived value of voting to the voter.

In practice, these findings suggest that there is genuine value being created by the recent increase in municipalities using IRV, and also provides evidence of additional benefits for municipalities that are considering changing their electoral system. Perhaps the benefits are enough to alter the cost-benefit calculation for politicians and voters who are weighing up a change to IRV.

---

[17]For example, using a structural model, Kawai, Toyama and Watanabe (2020) estimate that Minnesota has the highest level of perceived efficacy of voting among all U.S. states. In contrast, California has the 5th lowest levels of perceived efficacy of voting among the states.

[18]The appropriate methodologies are explained in Gentzkow, Kelly and Taddy (2019).

REFERENCES

Aiba, Satomi, Alex Farley, Jonathan Genki Fujimura, Katey Houck, Philip Jones, Heather Krawiecki, and Derek Schlieps. 2012. "Building a Better Democracy: Electoral Reform in Washington State."

Amy, Douglas J. 1996. "The forgotten history of the single transferable vote in the United States." *Representation*, 34(1): 13–20.

Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does attack advertising demobilize the electorate?" *American political science review*, 88(4): 829–838.

Anthony, Joseph, Amy Fried, Robert Glover, and David C Kimball. 2019. "Ranked Choice Voting in Maine from the Perspective of Election Authorities."

Arrow, Kenneth J. 1950. "A difficulty in the concept of social welfare." *Journal of political economy*, 58(4): 328–346.

Ashenfelter, Orley, and David Card. 1984. "Using the longitudinal structure of earnings to estimate the effect of training programs." National Bureau of Economic Research.

Balinski, Michel, and Rida Laraki. 2011. *Majority judgment: measuring, ranking, and electing.* MIT press.

Bartholdi, John J, and James B Orlin. 1991. "Single transferable vote resists strategic voting." *Social Choice and Welfare*, 8(4): 341–354.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How much should we trust differences-in-differences estimates?" *The Quarterly journal of economics*, 119(1): 249–275.

Black, Duncan. 1976. "Partial justification of the Borda count." *Public Choice*, 1–15.

Burnett, Craig M, and Vladimir Kogan. 2015. "Ballot (and voter)"exhaustion" under Instant Runoff Voting: An examination of four ranked-choice elections." *Electoral Studies*, 37: 41–49.

Cancela, João, and Benny Geys. 2016. "Explaining voter turnout: A meta-analysis of national and subnational elections." *Electoral Studies*, 42: 264–275.

Chadha, Anita. 2019. "Rank Choice Vote: An idea whose time has come." *National Social Science Journal*, 52: 36–42.

City of Minneapolis. n.d.. "Election Results." *Available at http://vote. minneapolismn. gov/ results/ index. htm (2020-06-15).*

Conley, Timothy G, and Christopher R Taber. 2011. "Inference with "difference in differences" with a small number of policy changes." *The Review of Economics and Statistics*, 93(1): 113–125.

Cook, Corey, and David Latterman. 2011. "Ranked choice voting in the 2011 San Francisco municipal election."

Damore, David F. 2002. "Candidate strategy and the decision to go negative." *Political Research Quarterly*, 55(3): 669–685.

Djupe, Paul A, and David AM Peterson. 2002. "The impact of negative campaigning: Evidence from the 1998 senatorial primaries." *Political Research Quarterly*, 55(4): 845–860.

Donovan, Todd. n.d.. "Candidate Perceptions of Campaigns under Preferential and Plurality Voting." *Electoral Systems, Electoral Reform and Implications for Democratic Performance.*

Donovan, Todd, Caroline Tolbert, and Kellen Gracey. 2016. "Campaign civility under preferential and plurality voting." *Electoral Studies*, 42: 157–163.

Donovan, Todd, Caroline Tolbert, and Kellen Gracey. 2019. "Self-Reported Understanding of Ranked-Choice Voting." *Social Science Quarterly*, 100(5): 1768–1776.

Douglas, Andrew. 2014. "Ranked Choice Voting and Civility:New Evidence from American Cities." FairVote: The Center for Voting and Democracy.

Downs, Anthony. 1957. "An economic theory of political action in a democracy." *Journal of political economy*, 65(2): 135–150.

Endersby, James W, and Michael J Towle. 2014. "Making wasted votes count: Turnout, transfers, and preferential voting in practice." *Electoral Studies*, 33: 144–152.

FairVote. 2014. "Ranked Choice Voting in Practice: Candidate Civility in Ranked Choice Elections, 2013 2014 Survey Brief." FairVote: The Center for Voting and Democracy.

FairVote. 2019. "Monopoly Politics 2020." *Available at* `https://www.fairvote.org/monopoly_politics#overview` *(2020-06-15).*

Felsenthal, Dan S, Zeev Maoz, and Amnon Rapoport. 1993. "An empirical evaluation of six voting procedures: do they really make any difference?" *British Journal of Political Science*, 23(1): 1–27.

Flury, Bernhard K, and Hans Riedwyl. 1986. "Standard distance in univariate and multivariate analysis." *The American Statistician*, 40(3): 249–251.

Fraenkel, Jon, and Bernard Grofman. 2004. "A neo-Downsian model of the alternative vote as a mechanism for mitigating ethnic conflict in plural societies." *Public Choice*, 121(3-4): 487–506.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature*, 57(3): 535–74.

Gibbard, Allan. 1973. "Manipulation of voting schemes: a general result." *Econometrica: journal of the Econometric Society*, 587–601.

Hansen, Ben B. 2004. "Full matching in an observational study of coaching for the SAT." *Journal of the American Statistical Association*, 99(467): 609–618.

Harrenstien, Ken. 2009. "Official Google Blog: Automatic captions in YouTube." *https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html*, (Accessed on 06/20/2020).

Ho, Daniel E, Kosuke Imai, Gary King, and Elizabeth A Stuart. 2007. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." *Political analysis*, 15(3): 199–236.

Holtzman, Alexander, and Clayton Nall. 2012. "The Unanticipated Inequalities of Electoral Reform: Racial and Ethnic Disparities in Voting Behavior under Oakland's Ranked Choice Voting Program."

Hu, Minqing, and Bing Liu. 2004. "Mining and summarizing customer reviews." 168–177.

Igersheim, Herrade, François Durand, Aaron Hamlin, and Jean-François Laslier. 2018. "Comparing Voting Methods: 2016 US Presidential Election."

IGM Forum. 2018. "Ranked-Choice Voting." *Available at http://www.igmchicago.org/surveys/ranked-choice-voting/ (2020-06-15).*

Jacobs, Lawrence, and Joanne Miller. 2014. "Rank Choice Voting and the 2013 Minneapolis Elections."

Jansen, Harold J. 2004. "The political consequences of the alternative vote: Lessons from Western Canada." *Canadian Journal of Political Science/Revue canadienne de science politique*, 37(3): 647–669.

Jerdonek, Christopher. 2006. "Ranked Choice Voting and Voter Turnout in San Francisco's 2005 Election."

John, Sarah. 2015. "Ranked Choice Voting in Practice: Candidate Civility and Voter Understanding Evidence from California City Elections Held in November 2014." FairVote: The Center for Voting and Democracy.

John, Sarah, and Andrew Douglas. 2017. "Candidate Civility and Voter Engagement in Seven Cities with Ranked Choice Voting." *National Civic Review*, 106(1): 25–29.

John, Sarah, Haley Smith, and Elizabeth Zack. 2018. "The alternative vote: Do changes in single-member voting systems affect descriptive representation of women and minorities?" *Electoral Studies*, 54: 90–102.

Kawai, Kei, Yuta Toyama, and Yasutora Watanabe. 2020. "Voter turnout and preference aggregation." *American Economic Journal: Microeconomics*.

Kemeny, John G. 1959. "Mathematics without numbers." *Daedalus*, 88(4): 577–591.

Kimball, David C, and Joseph Anthony. 2016. "Voter participation with ranked choice voting in the United States."

Laver, Michael. 2004. "Analysing structures of party preference in electronic voting data." *Party politics*, 10(5): 521–541.

Loughran, Tim, and Bill McDonald. 2011. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *The Journal of Finance*, 66(1): 35–65.

Mauter, Erica L. 2014. "Ranked Choice Voting in Minneapolis 2013 Elections."

McDaniel, Jason. 2018. "Does More Choice Lead to Reduced Racially Polarized Voting? Assessing the Impact of Ranked-Choice Voting in Mayoral Elections." *California Journal of Politics and Policy*, 10(2).

McDaniel, Jason A. 2016. "Writing the rules to rank the candidates: Examining the impact of instant-runoff voting on racial group turnout in San Francisco mayoral elections." *Journal of Urban Affairs*, 38(3): 387–408.

McDaniel, Jason A. 2019. "Electoral Rules and Voter Turnout in Mayoral Elections: An Analysis of Ranked-Choice Voting." Philadelphia.

Minnesota Department of Health. n.d.. "Minnesota State, County, and Community Health Board Vital Statistics Trend Reports." *Available at* `https://www.health.state.mn.us/data/mchs/genstats/trends/index.html` *(2020-06-15)*.

Minnesota Legislature. n.d.. "Geographic Information Services — Data Center." *Available at* `https://www.gis.leg.mn/html/download.html` *(2020-06-15)*.

Minnesota Secretary of State. n.d.. "Election Results." *Available at* `https://www.sos.state.mn.us/elections-voting/election-results` *(2020-06-15)*.

Mohammad, Saif, and Peter Turney. 2010. "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon." 26–34.

Neely, Francis, and Corey Cook. 2008. "Whose votes count? Undervotes, overvotes, and ranking in San Francisco's instant-runoff elections." *American Politics Research*, 36(4): 530–554.

Neely, Francis, and Jason McDaniel. 2015. "Overvoting and the equality of voice under instant-runoff voting in San Francisco." *California Journal of Politics and Policy*, 7(4).

Neely, Francis, Lisel Blash, and Corey Cook. 2005. *An assessment of ranked-choice voting in the San Francisco 2004 election.* Public Research Institute, San Francisco State University Daly City, CA.

News Centre Maine. 2018. "Ranked Choice voting, cartoon characters, and 'Kumbaya'." *Available at* `https://www.newscentermaine.com/article/news/local/now/ranked-choice-voting-cartoon-characters-and-kumbaya/97-537411043` *(2020-06-15).*

Nielsen, FÅ. 2011. "AFINN. Informatics and mathematical modelling." *Technical University of Denmark.*

Nielson, Lindsay. 2017. "Ranked Choice Voting and Attitudes toward Democracy in the United States: Results from a Survey Experiment." *Politics & Policy*, 45(4): 535–570.

O'Neill, Stephen, Noémi Kreif, Richard Grieve, Matthew Sutton, and Jasjeet S Sekhon. 2016. "Estimating causal effects: considering three alternatives to difference-in-differences estimation." *Health Services and Outcomes Research Methodology*, 16(1-2): 1–21.

Palfrey, Thomas R, and Keith T Poole. 1987. "The relationship between information, ideology, and voting behavior." *American journal of political science*, 511–530.

PBS. 2020. "PBS Video." *Available at* `https://www.pbs.org/search/` *(2020-06-15).*

Peterson, David AM, and Paul A Djupe. 2005. "When primary campaigns go negative: The determinants of campaign negativity." *Political Research Quarterly*, 58(1): 45–54.

Ramsey County. n.d.. "Election Results." *Available at* `https://www.ramseycounty.us/residents/elections-voting/information-reference/election-results` *(2020-06-15).*

Reilly, Ben. 1997. "The alternative vote and ethnic accommodation: new evidence from Papua New Guinea." *Electoral Studies*, 16(1): 1–11.

Reilly, Benjamin. 2004. "The global spread of preferential voting: Australian institutional imperialism?" *Australian Journal of Political Science*, 39(2): 253–266.

Rhode, Christopher. 2018. "The Cost of Ranked Choice Voting." FairVote: The Center for Voting and Democracy.

Richie, Rob, and Steven Hill. 2001. "Renewed Momentum for Voting System Reform." *National Civic Review*, 90(2): 183–188.

Richie, Rob, Caleb Kleppner, and Terrill Bouricius. 2000. "Instant runoffs: A cheaper, fairer, better way to conduct elections." *National Civic Review*, 89(1): 95–110.

Robb, Denise. 2011. *The effect of instant runoff voting on democracy.* University of California, Irvine.

Satterthwaite, Mark Allen. 1975. "Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions." *Journal of economic theory*, 10(2): 187–217.

Schultz, David, and Kristi Rendahl. 2010. "Evaluating ranked choice voting in the 2009 Minneapolis elections: A report for the Minneapolis elections department."

Shapiro, Adam Hale, and Daniel Wilson. 2019. "Taking the fed at its word: A new approach to estimating central bank objectives using text analysis." Federal Reserve Bank of San Francisco.

Silge, Julia, and David Robinson. 2016. "tidytext: Text mining and analysis using tidy data principles in R." *Journal of Open Source Software*, 1(3): 37.

Sinclair, DE "Betsy", and R Michael Alvarez. 2004. "Who overvotes, who undervotes, using punchcards? Evidence from Los Angeles County." *Political Research Quarterly*, 57(1): 15–25.

Skaperdas, Stergios, and Bernard Grofman. 1995. "Modeling negative campaigning." *American Political Science Review*, 89(1): 49–61.

Stockemer, Daniel. 2017. "What affects voter turnout? A review article/meta-analysis of aggregate research." *Government and Opposition*, 52(4): 698–722.

Sutherland, James D. 2016. *Municipal institutions, electoral system design, and voter turnout in large American cities.* Northeastern University.

The New York Times. 2018. "A Congress for Every American. The New York Times." *Available at* `https://www.nytimes.com/interactive/2018/11/`

`10/ opinion/ house-representatives-size-multi-member. html` *(2020-06-15).*

Tolbert, Caroline. 2014. "Experiments in Election Reform: Voter Perceptions of Campaigns Under Preferential and Plurality Voting."

Trebbi, Francesco, Philippe Aghion, and Alberto Alesina. 2008. "Electoral rules and minority representation in US cities." *The Quarterly Journal of Economics*, 123(1): 325–357.

United States Census Bureau. 2019. "American Community Survey." *Available at* `https: // www. census. gov/ programs-surveys/ acs` *(2020-06-15).*

Appendix A – Municipalities remaining in final civility data set

Following preparation of the data for sentiment analysis, the list of cities remaining in the data are:

- Albuquerque, New Mexico
- Austin, Texas
- Boston, Massachusetts
- Buffalo, New York
- Charlotte, North Carolina
- Chicago, Illinois
- Cincinnati, Ohio
- Cleveland, Ohio
- Colorado Springs, Colorado
- Denver, Colorado
- Detroit, Michigan
- Fort Wayne, Indiana
- Honolulu, Hawaii
- Houston, Texas
- Jacksonville, Florida
- Jersey City, New Jersey
- Lexington, Kentucky
- Lubbock, Texas
- Madison, Wisconsin

- Miami, Florida
- Minneapolis, Minnesota
- Nashville, Tennessee
- New York, New York
- Oakland, California
- Orlando, Florida
- Philadelphia, Pennsylvania
- Phoenix, Arizona
- Pittsburgh, Pennsylvania
- Portland, Oregon
- Raleigh, North Carolina
- San Antonio, Texas
- San Diego, California
- San Francisco, California
- Seattle, Washington
- St. Paul, Minnesota
- Toledo, Ohio
- Tulsa, Oklahoma

FIGURE A1. LOCATION OF CITIES INCLUDED IN SENTIMENT ANALYSIS

Note: Map does not show Honolulu, which is included in the data

Source: Map data from Google

Appendix B – Final set of stop words

The final set of stop words included in the analysis are:.

a	away	could	first	herself
a's	b	couldn't	five	hi
able	be	course	followed	him
about	became	currently	following	himself
above	because	d	follows	his
according	become	definitely	for	hither
accordingly	becomes	described	former	how
across	becoming	despite	formerly	howbeit
actually	been	did	forth	however
after	before	didn't	four	i
afterwards	beforehand	different	from	i'd
again	behind	do	further	i'll
against	being	does	furthermore	i'm
ain't	believe	doesn't	g	i've
all	below	doing	get	ie
allows	beside	don't	gets	if
almost	besides	done	getting	immediate
along	between	down	given	in
already	beyond	downwards	gives	inasmuch
also	both	during	go	inc
although	brief	e	goes	indeed
always	but	each	going	indicate
am	by	edu	gone	indicated
among	c	eg	got	indicates
amongst	c'mon	eight	gotten	inner
an	c's	either	h	insofar
and	came	else	had	instead
another	can	elsewhere	hadn't	into
any	can't	entirely	happens	inward
anybody	cannot	especially	hardly	is
anyhow	cant	et	has	isn't
anyone	cause	etc	hasn't	it
anything	causes	even	have	it'd
anyway	certainly	ever	haven't	it'll
anyways	changes	every	having	it's
anywhere	co	everybody	he	its
apart	com	everyone	he's	itself
appear	come	everything	hello	j
are	comes	everywhere	hence	just
aren't	concerning	ex	her	k
around	consequently	exactly	here	keep
as	consider	example	here's	keeps
aside	considering	except	hereafter	kept
ask	contain	f	hereby	know
asking	containing	far	herein	knows
associated	contains	few	hereupon	known
at	corresponding	fifth	hers	l

last	none	r	specify	thus
lately	noone	rather	specifying	to
later	nor	rd	still	together
latter	normally	re	sub	too
latterly	not	really	such	took
least	nothing	regarding	sup	toward
less	now	regardless	sure	towards
lest	nowhere	regards	t	tried
let	o	relatively	t's	tries
let's	obviously	respectively	take	truly
likely	of	s	taken	try
little	off	said	tell	trying
look	often	same	tends	twice
looking	oh	saw	th	two
looks	ok	say	than	u
ltd	okay	saying	thanx	un
m	old	says	that	under
mainly	on	second	that's	unless
many	once	secondly	thats	until
may	one	see	the	unto
maybe	ones	seeing	their	up
me	only	seem	theirs	upon
mean	onto	seemed	them	us
meanwhile	or	seeming	themselves	use
merely	other	seems	then	used
might	others	seen	thence	uses
more	otherwise	self	there	using
moreover	ought	selves	there's	usually
most	our	sent	thereafter	uucp
mostly	ours	serious	thereby	v
much	ourselves	seriously	therefore	value
must	out	seven	therein	various
my	outside	several	theres	very
myself	over	shall	thereupon	via
n	overall	she	these	viz
name	own	should	they	vs
namely	p	shouldn't	they'd	w
nd	particular	since	they'll	wants
near	particularly	six	they're	was
nearly	per	so	they've	wasn't
necessary	perhaps	some	think	way
need	placed	somebody	third	we
needs	plus	somehow	this	we'd
neither	possible	someone	thorough	we'll
never	presumably	something	thoroughly	we're
nevertheless	probably	sometime	those	we've
new	provides	sometimes	though	went
next	q	somewhat	three	were
nine	que	somewhere	through	weren't
nobody	quite	soon	throughout	what
non	qv	specified	thru	what's

whatever	yet	downing	make	presenting
when	you	downs	making	presents
whence	you'd	early	man	put
whenever	you'll	end	member	puts
where	you're	ended	members	room
where's	you've	ending	men	rooms
whereafter	your	ends	mr	seconds
whereas	yours	face	mrs	sees
whereby	yourself	faces	needed	showed
wherein	yourselves	felt	needing	showing
whereupon	z	find	newer	shows
wherever	zero	finds	newest	side
whether	she's	furthered	number	sides
which	he'd	furthering	oldest	smaller
while	she'd	furthers	open	smallest
whither	he'll	gave	opened	state
who	she'll	generally	opening	states
who's	shan't	give	opens	thing
whoever	mustn't	group	order	things
whole	when's	grouped	ordered	thinks
whom	why's	grouping	ordering	thoughts
whose	how's	groups	orders	today
why	area	high	part	turn
will	areas	higher	parted	turned
with	asked	knew	parts	turning
within	asks	large	place	turns
without	back	largely	places	wanted
won't	began	latest	point	ways
would	beings	lets	pointed	wells
wouldn't	cases	longer	pointing	year
x	differ	longest	points	years
y	downed	made	presented	youngest