Online Appendix

Increasing Inequality: The Effect of GOTV Mobilization on the Composition of the Electorate

Ryan D. Enos, Anthony Fowler, and Lynn Vavreck

This document provides additional information and results that are supplemental to the main paper. Additional details are provided about the data and statistical analyses. Also, additional results are provided regarding the mechanisms behind the main results and the variation of results across different experimental interventions. Please contact the authors with questions or comments.

Table of Contents

I. More Details on Each Analyzed Experiment

Here we provide more details on each of the experiments that we analyzed. Because some of the studies included in our analysis contained unique features, readers may be interested to see how we incorporated those features into our analysis. We also report whether we can exactly replicate the original results reported by each published paper and compare our estimates of the additive coefficient from Table 3 (our estimated effect for respondents with an average propensity level) to the original studies' estimates of the average treatment effect. As we discuss in the text, these two estimates need not be the same. First, the average treatment effect is not the same thing as the effect for the average subject. Second, the treatment effect need not vary linearly across propensity levels (and we do not assume that it does). Rather, we employ our empirical specification to test whether the treatment effect increases or decreases, on average, as propensity increases. We also provide non-parametric analyses which allow us to assess the conditional average treatment effect as each level of propensity.

Gerber and Green (2000)

Gerber and Green (GG00) randomly assigned non-partisan phone calls, direct mail, and door-to-door canvassing across households in New Haven in the run up to the 1998 November general election. Households with 1 or 2 registered voters were randomly assigned to receive 0, 1, 2, or 3 mailings, 0 or 1 phone calls, and 0 or 1 door-to-door canvassing attempts. Some individuals received multiple treatments (e.g. 2 pieces of direct mail and a phone call). For the sake of simplicity, we collapse the direct mail conditions into 2 categories (no mail and at least one piece of mail). This leaves us with one control condition of 11,596 individuals and 7 different treatment conditions: Phone only (n =846), Mail only (7,776), Door only (2,877), Phone+Mail (4,749),

Phone+Door (194), Mail+Door(1,853), and Phone+Mail+Door(1,207). These numbers match up very closely (but not exactly) with those in Table 2 in the original study.

For each of these 7 treatment groups, we test whether the treatment significantly increased voter turnout by comparing turnout treatment group to the control group. We focus only on intention-to-treat effects and do not incorporate data on successful contacts in our main analysis. Simple regressions of turnout on the treatment with controls for prior turnout, age, and party (which are not necessary but increase precision) reveal statistically significant treatment effects for only 4 of the treatment conditions: Mail, Door, Mail+Door, and Phone+Mail+Door. Some of the other treatment combinations mostly likely did exhibit a positive effect, but the sample size of the treatment group was too small to derive precise estimates. For example, we estimate that the Phone+Door intervention increased turnout by 5.6 percentage points, but the small sample size for the treatment group means that the estimate is not precise enough to include in our sample.

While Gerber and Green do not categorize their treatment groups in the same way that we do here, our analysis produces results that are very consistent with the results reported by the original study. We find no detectable effect of phone calls (with a slightly negative point estimate), a small positive effect of direct mail (about 1 percentage point), and a larger positive effect of door-to-door canvassing (3 to 4 percentage points). We also see evidence suggesting that these effects are not additive. For example, The Mail+Door treatment appears to be no more effective than the Door treatment alone, even though the Mail treatment on its own exhibited a positive effect. The average treatment effects (estimated by us and in GG00) line up very closely with our estimates of the "Treatment" coefficients (the effect for citizens with average propensity) in Table 3. Because approximately half of GG00's subjects lived in two-person households (as opposed to one-person households), we cluster the standard errors by household in our analysis of this data.

Gerber, Green, and Nickerson (2003)

Large-scale door-to-door canvassing experiments were conducting in 6 cities, Bridgeport, Columbus, Detroit, Minneapolis, and St. Paul, leading up to the 2001 local elections in those cities. Gerber, Green, and Nickerson (GGN03) show statistically significant treatment effects in Bridgeport, Detroit, and St. Paul. We also find a positive, statistically significant treatment effect in Minneapolis after adding controls and improving the efficiency of the test, so we include that experiment in our analysis as well. Re-analyzing their data, we construct our propensity variable using age, party registration, gender, race, and turnout in 1999, 2000, and a 2001 primary. Not all variables are available for each city. For example, racial data is only available in Raleigh, and a 2001 primary was not conducted in Raleigh.

In each city, households were selected for contact and one registered voter within each household was specifically selected for study. The households were grouped geographically into "walk lists." In some cases, treatment was randomly assigned across all households in the study, and in other cases, the randomization was stratified within walk lists to improve balance. GGN03 estimate the treatment effect in each city with a regression that includes walk list fixed effects. These fixed effects improve efficiency but are not necessary for unbiased results because, even in the cases of stratifications, the probability of treatment was uniform across all walk lists. For our own regression results in Table 3, we do not include fixed effects for walk lists, but the results are nearly identical in either case.

With the replication data from GGN03, we are able to exactly replicate their results shown in Table 2 (of the original study). Moreover, the additive effects shown in our Table 3 line up fairly closely with the average treatment effects estimated in the original study. Because the study only includes 1 registered voter per household, the standard errors do not need to be clustered. Consistent with GGN03, we report heteroskedasticity-robust standard errors.

Nickerson (2006)

Nickerson analyzed 8 different phone-based get-out-the-vote experiments across 6 cities in either the 2000 presidential election or in 2001 local elections. Only one of those experiments – that in Stonybrook, NY in the run up to the 2000 presidential election – showed a statistically significant effect of the treatment. Therefore, we only include this experiment in our analysis. We replicate Nickerson's estimate of the average treatment effect in Stonybrook (8.2 percentage points), and our estimate of the additive coefficient in Table 3 (7.1 percentage points) matches this estimate closely. This particular experiment does not present an ideal opportunity for our analysis for several reasons. First, the sample size is small (680 individuals in the treatment group and 279 in the control group) leading to high standard errors and imprecise estimates. Second, the sample consists of newly registered college students meaning that there is no vote history data available, and little variance in age. As a result of these factors, there is little variance in our estimates of vote propensity for this sample and the standard error on our estimate of the interactive coefficient in this sample is significantly greater than in any other experiment in our analysis. For these reasons, we should not be surprised by the "null" finding in this case and should not put significant weight on this particular experiment.

Nickerson (2007)

This study assesses the effect of get-out-the-vote phone calls to young registrants in selected cities in the 2002 general election. The sample was randomly divided into four equally sized treatment groups: a control group which received no calls, a group that was called by a volunteer phone bank, another group called by a professional phone bank, and a final group called by both phone banks. We treat the three different treatment groups as three separate interventions and

estimate their effects by comparing turnout in each treatment group to that in the control group. Nickerson also embedded several sub-experiments into the study that we ignore or average over. The timing of GOTV calls and the content of conversation were randomized for some subjects, but we focus on the average effects of these three interventions.

Nickerson does not explicitly test for the average intent-to-treat effect for each of these three interventions. However, our own analysis suggests that that all three interventions positive, statistically-significant average effects, although consistent with the original study, the average effect is smallest for the "volunteer only" intervention.

Because the treatment probabilities were constant across sites, site fixed effects are not necessary for unbiased estimates but they could potentially improve efficiency. In Table 3, we report our results without site fixed effects, but the results are unchanged if they are included. Dummy variables for cites are included in our propensity regression, so this geographic information is implicitly already contained in the "propensity" variable.

Gerber, Green, and Larimer (2008)

The main text provides details on Gerber, Green, and Larimer's "neighbors" experiment. As part of that study, the authors also conducted three additional experimental interventions, all delivered via direct mail. A "self" treatment contained everything in the "neighbors" treatment except for information about other individuals. A "Hawthorne" treatment removed vote history information but emphasized that "YOU ARE BEING STUDIED." Finally, a "civic duty" treatment removed the social pressure and monitoring components and provided only the more traditional encouragement: "DO YOUR CIVIC DUTY AND VOTE!" The four treatment groups were approximately equal in size (about 38,200 individuals in each) while the treatment groups was larger (191,243 individuals). We examine each of the four interventions separately. All treatments

exhibited a positive, statistically significant average treatment effect, so all interventions are included in our analysis.

We successfully replicated the original results reported by the authors and generated our propensity variable as described in the main text. Because the treatments occurred at the household level, we cluster our standard errors by household. The authors stratified their randomization by geographic clusters which were determined by mail routes. However, the probabilities of treatment were constant across clusters so simple differences-in-means yield unbiased estimates of the average treatment effects. Cluster fixed effects could potentially be added to our regressions in order to improve efficiency, but our results are nearly identical with our without these fixed effects.

Middleton and Green (2008)

This study is the only study in our analysis that is not an explicitly randomized trial. Middleton and Green take advantage of the quasi-random inability of of MoveOn to treat certain neighborhoods where door-to-door canvassing was planned. They only include in their analysis streets on the border of two neighborhoods where one received treatment and one quasi-randomly did not receive treatment. Therefore, the houses on one side of the street were treated and houses on the other side of the street were not. Even though the unit of analysis is an individual, the randomization effectively took place at the level of a street. With their replication data in hand, our analysis is strikingly similar to that of our other studies despite the fact that this treatment was not explicitly randomized by the researchers. We were able to replicate the authors' original results.

In constructing our propensity scores, we only include treatment observations (as always) and we use the exact same set of pre-treatment variables used by Middleton and Green as control variables. Because the randomization took place at the street level and because the probability of treatment may slightly differ across streets (although not significantly), we include street fixed effects

in our final analysis (M&G call this variable "block_id"). Because the relevant unit receiving treatment is a side of a street, we cluster our standard errors by each group of individuals living on one side of a street (M&G call this variable "block_id_new"). All of these steps are consistent with the original analysis conducted by the authors to estimate the average treatment effect. Our "treatment" coefficient of .016 almost perfectly matched the original study's estimates of the average intent-to-treat effect.

Nickerson (2008)

This study employed door-to-door canvassing experiments in the run up to primary elections in 2002 in Denver and Minneapolis. Households were evenly divided into one of three treatment groups: a no-contact control group, a placebo treatment group where individuals who answered the door were reminded to recycle, and a get-out-the-vote treatment group. For our main analysis, we collapse the placebo and no-contact groups into a single control group. We find no statistically-significant average intent-to-treat effect in Denver but we do find such an effect in Minneapolis, so we confine our analysis to the Minneapolis study.

Nickerson does not report the average ITT effects, because they are less relevant for the specific questions being addressed in the original study. Instead, the original paper focuses on differences in turnout between households and individuals contacted in the GOTV and placebo treatment groups. We take advantage of this novel design to address an additional question about whether high propensity citizens are easier to mobilize conditional on being contacted in the first place. That analysis is described in a subsequent section in the Appendix.

Dale and Strauss (2009)

The original study examines the effect of get-out-the-vote text messages in the 2006 general election. Half of the subjects were randomly assigned to a control condition of no contact. The remaining half received one of four types of texts messages. Among the treatment group, the researchers randomly varied whether they assigned a "civic duty" or a "close election" treatment and they also randomly varied whether they included a phone number for a hotline where subjects could receive information about their polling location, creating four equally sized treatment groups. At the end of the study, the authors successfully matched approximately 8,000 subjects to voter files, leaving one control group of about 4,000 individuals and 4 treatment groups of about 1,000 individuals. When we analyze the average effect of each treatment separately, we only find a statistically significant estimate in one case (close election message with no hotline information). However, the point estimates are similar for all four treatments and the small sample sizes limit our precision. For that reason, we pool all treatments together and estimate the average effect of any text message. Consistent with the original study, we estimate an effect of 3.0 percentage points, and that estimate is statistically distinguishable from zero ($p = .01$). Applying our test to this data, the additive coefficient is nearly identical to this estimate of the average treatment effect.

Gerber, Green, and Larimer (2010)

This study represents an extension of the previous study by the same authors discussed earlier. The authors conducted a direct mail experiment in the run up to local election in Michigan in 2007. They focused on households with only one or two individuals in each household where all individuals had voted in a recent election and had abstained from another recent election. They randomly assigned households into one of four groups: a control group that received no contact (353,341 individuals), a "civic duty" treatment group similar to that in GGL08 (3,238 individuals), a

"shame" treatment group similar to the "self" treatment from GGL08 but where subjects were only informed about the recent election where they abstained (6,325 individuals), and a "pride" treatment group also similar to the "self" treatment but where subjects were only informed about the recent election where they voted (6,307). We replicate positive, statistically-significant treatment effects for all three treatments and apply our test separately to each treatment. Because the treatment was assigned at the household level, our standard errors are clustered by household. The additive coefficients are nearly identical to the average treatment effects reported in the original study.

Gerber, Huber, and Washington (2010)

In the original study, the authors focused on registered voters in Connecticut who were not registered with a particular party and who reported in a survey that they were politically independent (leaners on the 7 point party identification questions are included in the experiment). While the original authors focus most of their analysis on these leaners or "latent partisans," we also include pure independents in our analysis as they were included in the experiment. In the 2008 presidential primary in Connecticut, voters had to be registered with a party in order to vote in the election, so the subjects would have to register with a party in order to even be eligible to turn out. As such, voter turnout was very low among the control group (2.5%). Among the 2,348 individuals deemed appropriate for the experiment (most of whom lived in one-person households), the authors randomly assigned half of them into the treatment group where they received mail informing them that they would have to identify with a party in order to vote in the primary and providing a blank party affiliation form. As expected, the treatment appeared to increase turnout in the primary election (turnout in the treatment group was 5.9 percentage points). Applying our test to the data, our additive coefficient lines up very closely with the average treatment effect. The minor discrepancy between our estimate and the estimate in Table 3 of the original study is explained by

the fact that the treatment effect was smaller for pure independents whom we include in the analysis but are not included in the original analysis.

Nickerson and Rogers (2010)

This study presents the study of registered Democrats in Pennsylvania in the 2008 presidential primary. Individuals without phone number or on a "do not call" list were excluded for practical purposes. Also for practical purposes, household with more than 3 registered voters were excluded. Additionally, individuals who had voted in multiple recent primaries were dropped because they were deemed to be extremely likely to vote in the absence of any treatment. Subjects were randomly divided into a control group (no contact) or one of three treatment groups where their household received a phone call. In the first group, subjected received a traditional GOTV phone call reminding them about the election and priming their civic duty. Individuals in the second group received the standard GOTV treatment and were also asked whether they planned to vote. Individuals in the final group received these 2 treatments and were also asked three follow-up questions designed to facilitate plan-making. The standard GOTV and the self-prediction treatments had no statistically significant effect over the control condition, so these two treatments are removed from our analysis. We focus solely on the "planning" treatment.

The randomization was stratified by household size. Also, while this fact is not reported in the paper, the probability of treatment varied across household size (with a lower treatment probability for one-person households). Therefore, household fixed effects are necessary for unbiased estimates. For this reason, we include household fixed effects in both the estimation of propensity scores and the subsequent estimation of the interactive effect. As it turns out, results are nearly identical if we exclude the household fixed effects from the second estimation step because

the propensity variable already contains that information. Because the treatment occurs at the household level, we cluster our standard errors by household.

We are unable to exactly replicate the results reported in the original study, but our own estimates of the ITT effects (including fixed effects for household size) are very close to the reported estimates. Similarly, our additive coefficient in Table 3 and its corresponding standard error are nearly identical to the intent-to-treat results reported in the original paper. ‘

II. More Details on the Pooled Analysis

In conducting our pooled analysis, we aggregate the data from all experiments into one regression. The total sample size is 1,167,771 individuals clustered into 877, 787 households. 319,251 individuals received some treatment while 848,521 were set aside in a control group. The totals do not match up with the sums of the treated and control columns in Table 3, because for some for several of the interventions, the control group is unchanged. To be clear, we do not "double count" control individuals in our pooled analysis. For example, the 11,665 control individuals from Gerber and Green (2000) are only included once, along with the individuals from the various treatment groups.

For each individual in this analysis, we have a binary indicator for voter turnout, a binary indicator for experimental treatment. All treatments are treated equally in the analysis in order to obtain average estimates across all experiments. However, we do not assume that treatment effects are homogeneous across setting. We also have the propensity scores calculated previously. These scores are not necessarily comparable across settings. For example, an individual with a score of 0.5 in the GG00 sample is not necessarily comparable to an individual with a score of 0.5 in the GGL08 sample. Similarly, the experimental settings vary across low and high salience elections. To account for these factors, we include dummy variables for each study and interactions of the propensity variable with these dummy variables.

A final complicating factor with the pooled analysis is that two of the experimental studies require additional conditional variables in order to obtain unbiased estimates. In the case of Middleton and Green (2008), we must also include block fixed effects and in the case of Nickerson and Rogers (2010) we much include household size fixed effects. These covariates are included in the pooled analysis but they make almost no difference. We include these factors by creating a unique study dummy variable for each block within the Middleton and Green study and for each

household size within the Nickerson and Rogers study. See the replication data files for the code required to execute this test.

III. Are High-Propensity Citizens Easier to Contact?

Using all experimental data where contact information is available, we test whether high-propensity citizens are easier to contact via door-to-door canvassing or phone calls, helping us to understand whether differential contact rates can partially explain the differential intent-to-treat effects that we find. Each row of Table A1 represents a separate regression, where we regress a dummy variable for household contact on our propensity variable. Only individuals in the treatment group are included in these analyses. The *Propensity* coefficients as the extent to which a single standard deviation increase in propensity corresponds to the probability of household contact. For example, in Gerber and Green's (2000) New Haven study, a standard deviation increase in propensity corresponds to an extra 5 percentage point chance of canvassing contact and an extra 12 percentage point chance of phone contact. On average, high propensity citizens are much easier to contact than low-propensity citizens, suggesting that differential contact rates are one important mechanism behind our empirical results.

Table A1. Are high-propensity citizens easier to contact?

	Study	Propensity	Constant
Door-to-Door Canvassing	GG00	.052(.007)**	.313(.007)**
	GGN03 - Bridgeport	.079(.018)**	.180(.015)**
	GGN03 - Columbus	.038(.013)**	.113(.011)**
	GGN03 - Detroit	.012(.006)*	.157(.007)**
	GGN03 - Minneapolis	.044(.009)**	.103(.009)**
	GGN03 - Raleigh	.019(.010)	.359(.012)**
	GGN03 - St. Paul	.051(.009)**	.127(.011)**
	N08 - Denver	-.005(.015)	.334(.016)**
	N08 - Minneapolis	.061(.021)**	.462(.024)**
Phone Calls	GG00	.121(.007)**	.344(.007)**
	N06 - Albany	−.005(.018)	.616(.017)**
	N06 - Boston	.038(.014)**	.554(.014)**
	N06 - Stonybrook	.020(.012)	.886(.012)**
	N07 - Professional	.031(.003)**	.386(.003)**
	N07 - Volunteer	.060(.003)**	.422(.003)**
	NR10	.030(.002)**	.248(.002)**

*Robust/household-clustered standard errors in parentheses; ** significant at 1%, * significant at 5%*
For each experiment where contact information is available, we test whether contact rates are higher for high-propensity citizens by regressing contact on propensity for those individuals in the treatment group. The table shows that high-propensity citizens are much easier to contact via both door-to-door canvassing and phone calls. For example, in Gerber and Green's (2000) New Haven study, a standard deviation increase in propensity corresponds to an extra 5 percentage point chance of canvassing contact and an extra 12 percentage point chance of phone contact. These results suggest that differential contact rates explain much of the variation in intention-to-treat effects between high and low-propensity voters.

IV. Does Contact Explain the Findings Entirely?

We find that high-propensity individuals are much easier to contact through either phone calls or door-to-door canvassing, suggesting that a significant share of the heterogeneous treatment effects that we identify can be explained by differential contact rates. Here, we test whether differential contact rates can explain all of these patterns. We take advantage of the fact that one of the control groups in Nickerson (2008) received a placebo treatment, so we know which of those individuals would have received a treatment had they been in the GOTV condition. Table A2 shows the results of our empirical test for 4 different samples: all individuals in a household that was contacted in both Minneapolis and Denver and all individuals who were contacted themselves in both Minneapolis and Denver. In each case, the sample sizes are small and the statistical precision is limited, but the interactive coefficients are large and positive in each case. Even conditional on being contacted, the conditional average treatment effects are much greater for high-propensity citizens than they are for low-propensity individuals. This analysis suggests that differential contact cannot entirely explain the patterns uncovered in the paper. Low-propensity individuals are indeed harder to contact, but they are also harder to mobilize even after they have been contacted.

Table A2. Applying Our Test to Households and Individuals Who Received Treatment

	Contacted Households		Contacted Individuals	
	Minneapolis	Denver	Minneapolis	Denver
Treatment	.076	.067	.097	.079
	(.025)**	(.029)*	(030)**	(.034)*
Propensity	.218	.290	.223	.283
	(.025)**	(.015)**	(.026)**	(.019)**
Treatment*Propensity	.041	.025	.039	.030
	(.027)	(.021)	(.034)	(.027)
Constant	.146	.385	.137	.397
	(.016)**	(.021)**	(.019)**	(.024)**
Observations	786	1,124	394	562
R-squared	.389	.380	.397	.361

*Standard errors are in parentheses, clustered by household in the case of contacted household and heterskedasticity-robust in the case of contacted individuals; **p<.01, *p<.05.*

V. Can the Findings be Explained by Data Limitations?

We may worry that our finding is a result of poor data quality from the public voter files typically employed in field experiments. Public voter records are often inaccurate and out of date. For example, an individual may have passed away or moved and could not possibly vote in an upcoming election, but a researcher would have no way of knowing this. We refer to these individuals as "deadweight," because they shouldn't be on the voter file at all, but the researcher hopelessly tries to mobilize them. Deadweight could be particularly troubling for our study if these ineligible individuals tend to be classified as low-propensity. What if the treatment effect is actually homogeneous across the eligible population, but many individuals categorized as low-propensity are actually deadweight, leading us to falsely conclude that GOTV interventions exacerbate the participation gap?[1]

To address this concern, we perform a simple sensitivity analysis. We cannot entirely rule out concerns about deadweight, but we can determine how extensive the problem would have to be in order to drive our results. Pooling all experiments in our sample and including study fixed effects, we estimate an average treatment effect of 3.7 percentage points. Then, we break the sample into 20 subsamples according to each individual's propensity score.[2] As expected, the conditional average treatment effect is larger for the higher propensity subsamples. The largest treatment effect we

[1] From the perspective of the participation gap, the reason that an individual cannot be mobilized is highly relevant. If a person is deceased, then they truly should not be in the sample. However, if the tendency to move and be unreachable by a political campaign is correlated with turnout as well as demographics and policy preferences, then the systematic tendency of GOTV treatments to miss these individuals will increase the participation gap.

[2] Our general results are robust to different numbers of subsamples.

observe for a subsample is s 5.2 percentage points. If the treatment effects were truly homogeneous and our result were driven by deadweight in the voter file, then we could say that the true treatment effect for non-deadweight individuals in each sample would have to be at least 5.2 percentage points. Therefore, the minimum proportion of deadweight in our sample would have to be 30% (1 - .037/.052) in order for us to obtain the results that we do. Put another way, deadweight could only explain our results if deadweight individuals constitute at least 30% of these GOTV samples.

While voter records surely contain errors, this 30% figure is implausibly large.[3] We demonstrate this by estimating the proportion of deadweight on a typical voter file. Campaigns and for-profit data vendors have a strong incentive to identify and remove deadweight from the file because targeting deadweight is costly. According to the data base of Catalist, a widely used political data services company based in Washington D.C., only 4% of the individuals on statewide voter files are classified as "deceased" or having a "bad address." Focusing specifically on those individuals who voted in the most recent election, that number shrinks to 2.5%. Also, large-scale mail based surveys in Florida and Los Angeles County suggest that less than 10% of individuals on voter files are deadweight (Ansolabehere et al. 2010). Our sensitivity analysis indicates that data quality cannot reasonably be argued to explain our results, because the actual amount of bad data is much less than it would have to be in order to pose a threat to our inferences.

[3] This is especially true in cases where the researchers selectively chose their sample to minimize deadweight. For example, Gerber, Green, and Larimer (2008) only include individuals who voted in 2004 in their sample, so the proportion of their sample that could have moved or passed away in the two year interim period is small.

VI. Variation across Different Experimental Interventions

Having seen that GOTV interventions tend to exacerbate the participation gap, on average, we would like to know whether this effect varies across different contexts, settings, or mobilization methods. Our strategy allows us to assess this variation. By conducting our test across many experiments, we hope to identify the types of interventions that are most effective (or least ineffective) in reducing the participation gap. Of course, even though this is the largest analysis of different types of interventions yet undertaken, the sample size prevents us from making strong claims about variation across different interventions. Nevertheless, we hope that our test here will guide future researchers in applying this test to identify the types of treatments that might effectively reduce the participation gap.

First, we test for variation across electoral salience. Arceneaux and Nickerson (2009) argue that high-propensity voters will be easier to mobilize in low salience elections and low-propensity voters will be easier to mobilize in high salience elections. To test this hypothesis, we compare empirical results across elections with different levels of voter turnout. Figure A1 plots the interactive coefficient from our regressions against the constant coefficients from the same regressions. The constant term indicates the predicted probability of turnout for the average subject in the control group. The interactive term indicates the extent to which the treatment exacerbated the participation gap. So, by looking for a relationship between these coefficients, we can see if interventions are more likely to increase the participation gap in low or high salience elections

The hypothesis of Arceneaux and Nickerson is largely confirmed: as electoral salience increases, the exacerbating effect of GOTV interventions decreases. However, looking at the graph, we would only expect an intervention to decrease the participation gap in elections with turnout of 50% or greater. Elections with 50% turnout are rare in the U.S. outside of presidential races. This

analysis suggests that GOTV interventions *can* actually reduce the participation gap in very high-salience elections, but these interventions have the opposite effect in most settings.

Next, we test for variation across the strength of treatment. We quantify severity using our treatment coefficient, the effect of the treatment for the average subject in the sample. Figure A2 plots the interactive coefficient against the additive coefficient from our analyses. We see that as the severity of the experimental treatment increases, the exacerbating effect also increases. One possible explanation is that the most effective treatments tend to have a psychological or social pressure component to them, as opposed to the more traditional GOTV messages. This psychological component which makes these interventions so effective may have a particularly concentrated effect among high-propensity citizens. As a result, the most effective interventions have the unexpected consequence of exacerbating the participation gap.

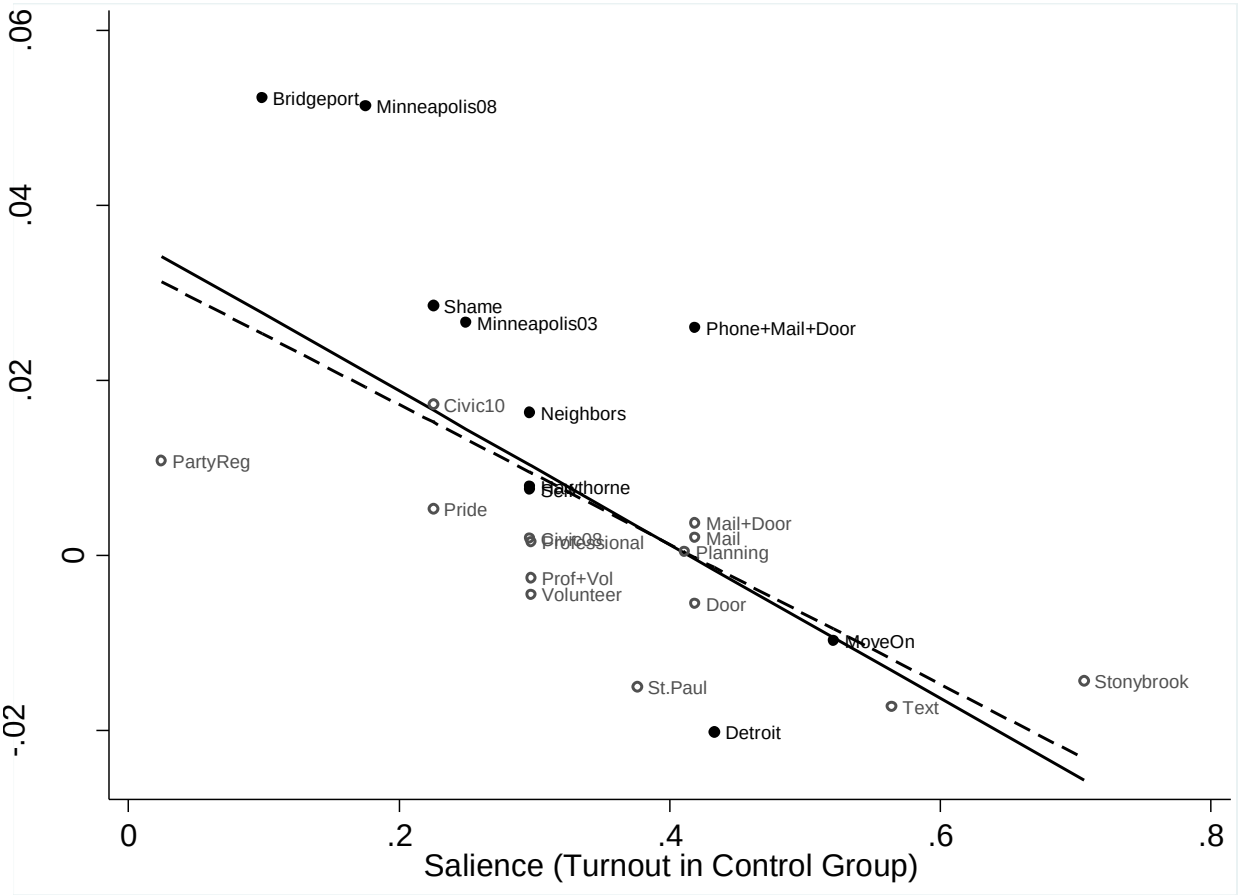
Only two interventions in our analysis demonstrate statistically significant evidence that the participation gap was reduced. What might explain the difference in these two cases? One intriguing similarity between the two experiments with negative interaction effects is that they both targeted citizens in communities with large African American populations. One explicitly targeted African Americans (Middleton and Green 2008) and the other was set in the largely African American city of Detroit[4] (Gerber, Green, and Nickerson 2003).

We tested for the possibility that African Americans respond differently to GOTV experiments by examining field experiments for which both Blacks and non-Blacks are identified in the experimental population. Most public voter files do not identify the race of the voters. As such, there are only three studies in our sample that could be used for this purpose: Dale and Strauss (2009); the Gerber, Green, and Nickerson (2003) study in Raleigh; and Nickerson and Rogers

[4] Detroit was 83% African American in 2010.

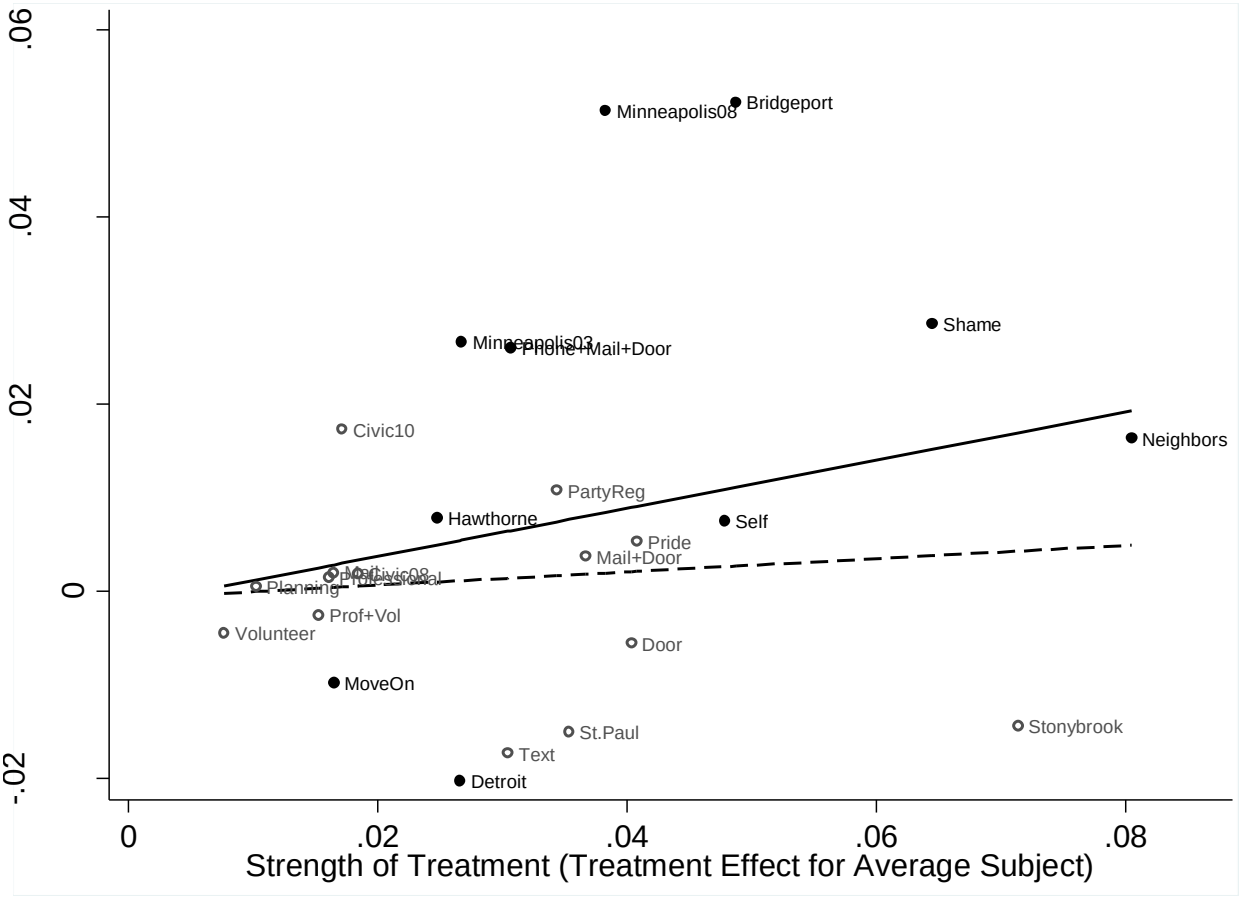
(2010). In these studies, Black citizens do appear to respond differently to GOTV efforts than non-Black citizens. In two studies, we find that high-propensity Blacks are actually *demobilized* by the treatment and in the other they were mobilized much less than whites or low-propensity Blacks. If GOTV efforts are demobilizing likely African American voters, this undermines the purpose of many efforts, in addition to presenting serious ethical concerns. Of course, this is a preliminary analysis on only three existing data sources so we draw no strong conclusions about these mechanisms. Moreover, we would expect to see some negative interactions by chance alone, so we should not draw strong conclusions from the few cases where we see this.

Figure A1. Variation across Electoral Salience



The figure assesses the salience hypothesis of Arceneaux and Nickerson (2009) that GOTV treatments will mobilize high-propensity citizens in low-salience elections and low-propensity citizens in high-salience elections. The y-axis is the multiplicative coefficient from these analyses, indicating the extent to which the treatment effect changes as propensity increases. The x-axis is the average level of turnout in the control group, a proxy for the salience of the election. Black, solid circles denote cases where the interactive coefficient is statistically significant ($p < .05$) and gray, hollow circles denote cases where the interactive coefficient is not statistically significant. The solid line represents a linear fit where all studies are weighted equally, and the dashed line indicates a linear fit where the studies are weighted by their sample sizes. The Arceneaux/Nickerson hypothesis is largely confirmed: as salience increases the effectiveness of GOTV treatments for low-propensity citizens increases relative to high-propensity citizens. However, significant variation remains in the data, suggesting that some treatments may be more or less effective in reducing the participation gap. Moreover, this analysis predicts that GOTV treatments will tend to exacerbate the participation gap in any setting where the average level of turnout is less than 50% -- essentially all electoral settings in the U.S. outside of presidential races.

Figure A2. Variation across Strength of Treatment



The figure presents the regression results from Table 3 graphically. The y-axis is the multiplicative coefficient, indicating the extent to which the treatment effect changes as the propensity variable increases. The x-axis is the additive coefficient, the treatment effect for the average citizen in the sample. Solid circles indicate that the interactive coefficient is statistically significant ($p < .05$). As before, the solid line represents a linear fit with all experiments weighted equally and the dashed line indicates a linear fit where experiments are weighted by sample size. We see that the interactive effect tends to be larger for experiments with larger average effects.

VII. More Detailed Results

The following graphs present the non-parametric results of the paper in more detail. For each experiment (or for each set of experiments that share a single control group), we present three plots. The first is a histogram representing the distribution of the propensity variable. The second are kernel regression of turnout and the propensity variable for the control and treatment groups, similar to the top panel of Figure 2. The final plot represents the difference between the kernel regressions for the treatment group(s) and the control group, indicating the conditional average treatment effect at each level of propensity, similar to the bottom panel of Figure 2.

Figure A3. Gerber and Green (2003)

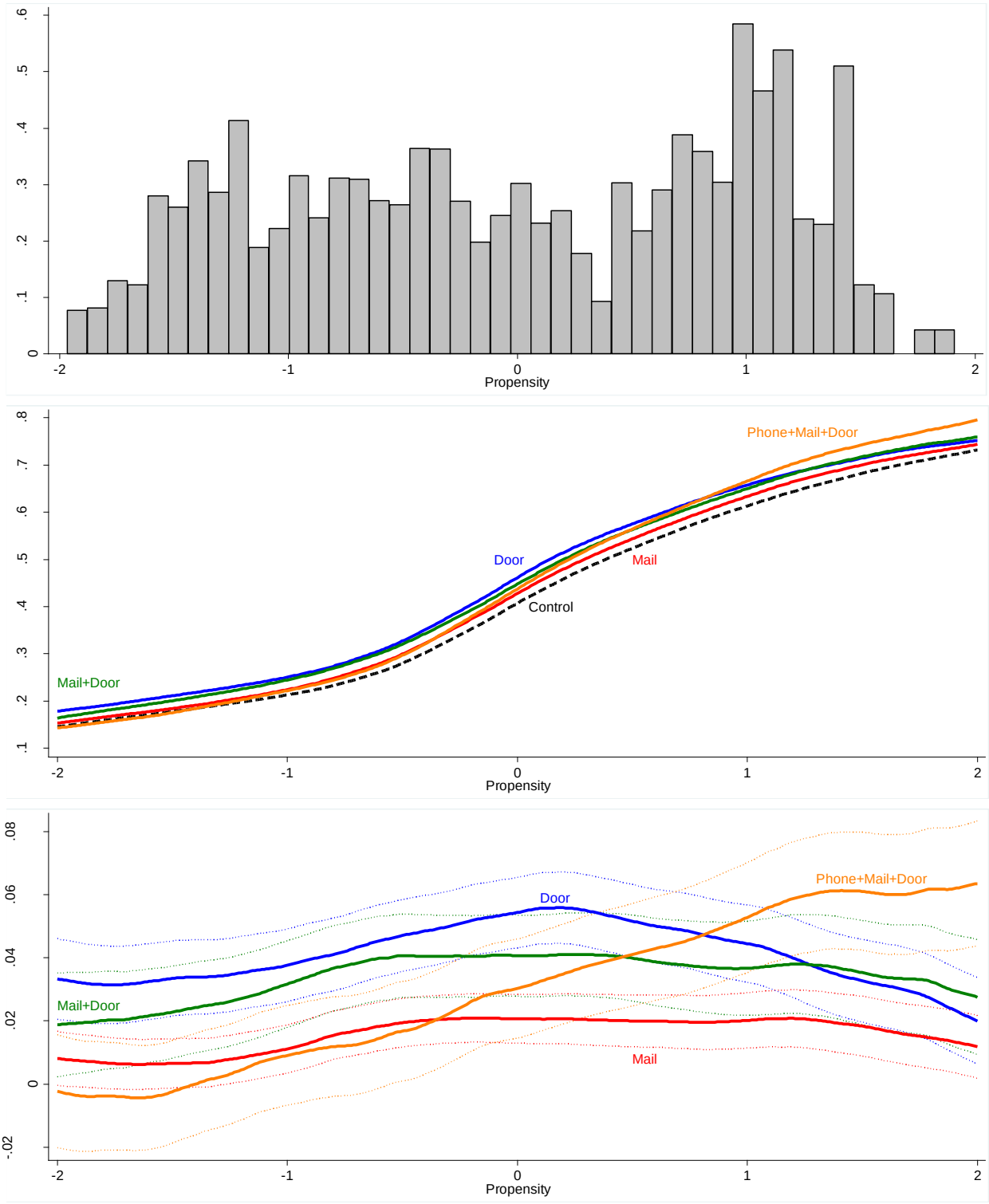


Figure A4. Bridgeport (Gerber, Green, and Nickerson 2003)

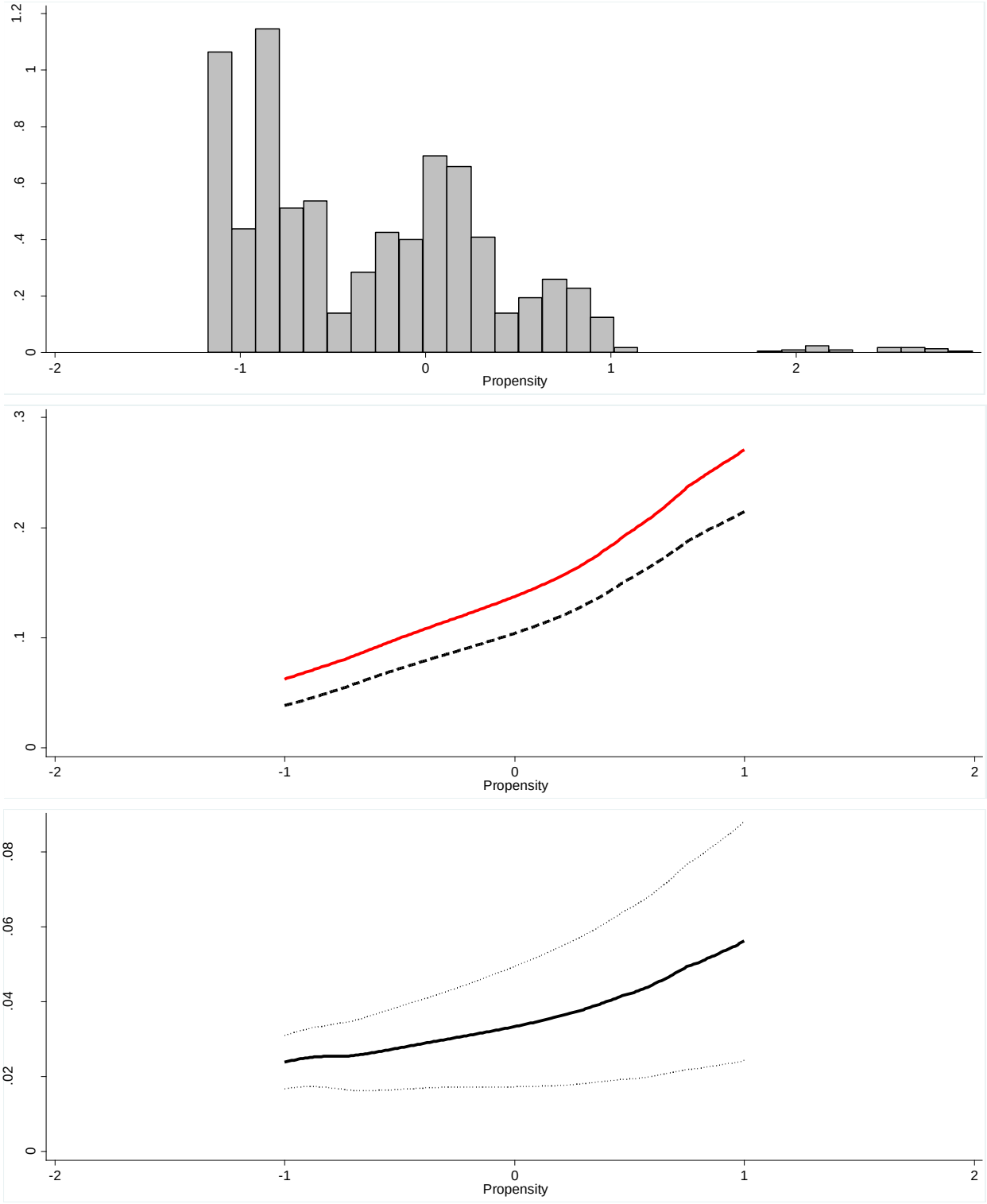


Figure A5. Detroit (Gerber, Green, and Nickerson 2003)

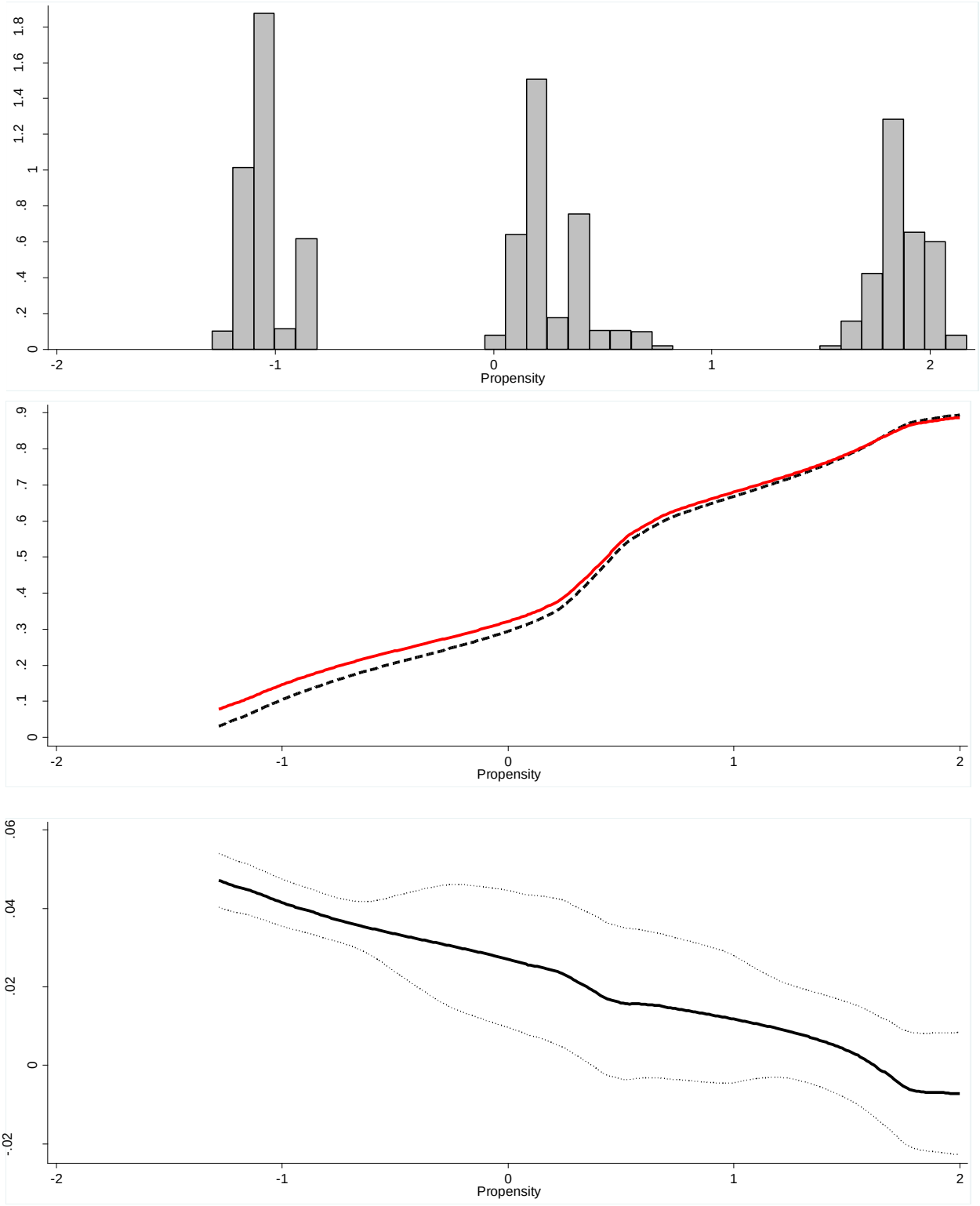
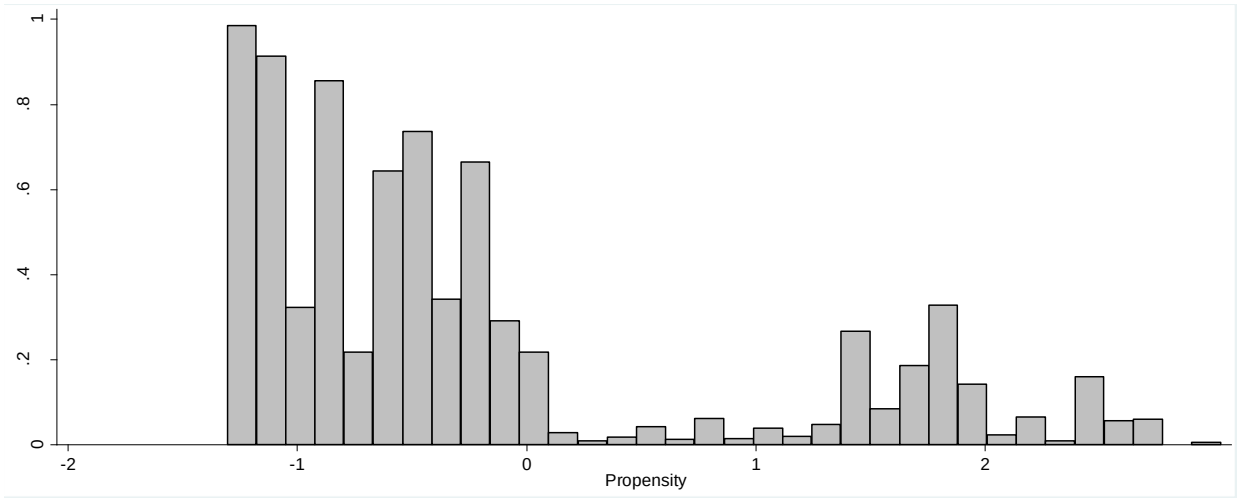
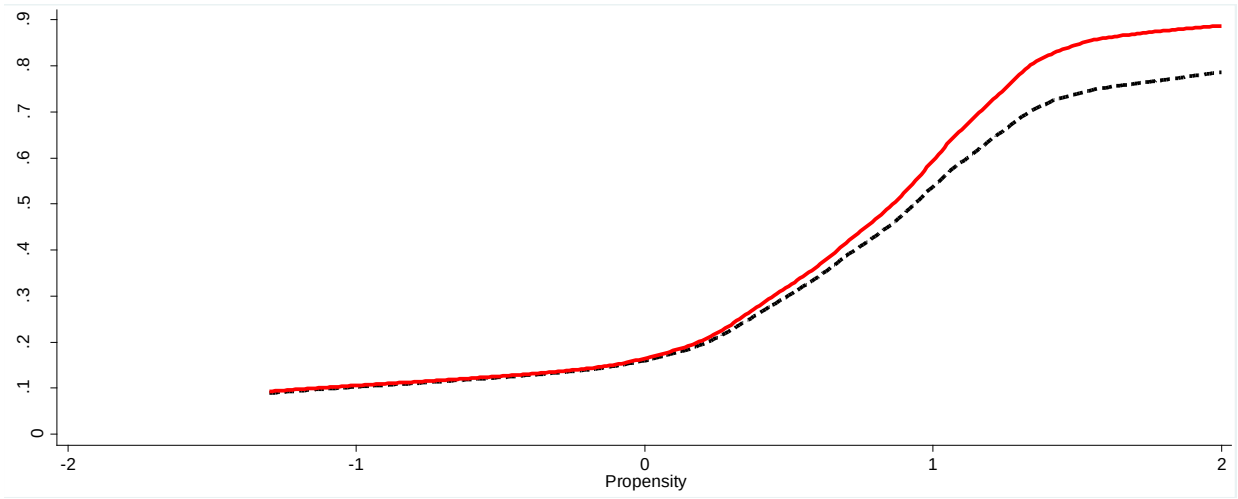


Figure A6. Minneapolis (Gerber, Green, and Nickerson 2003)





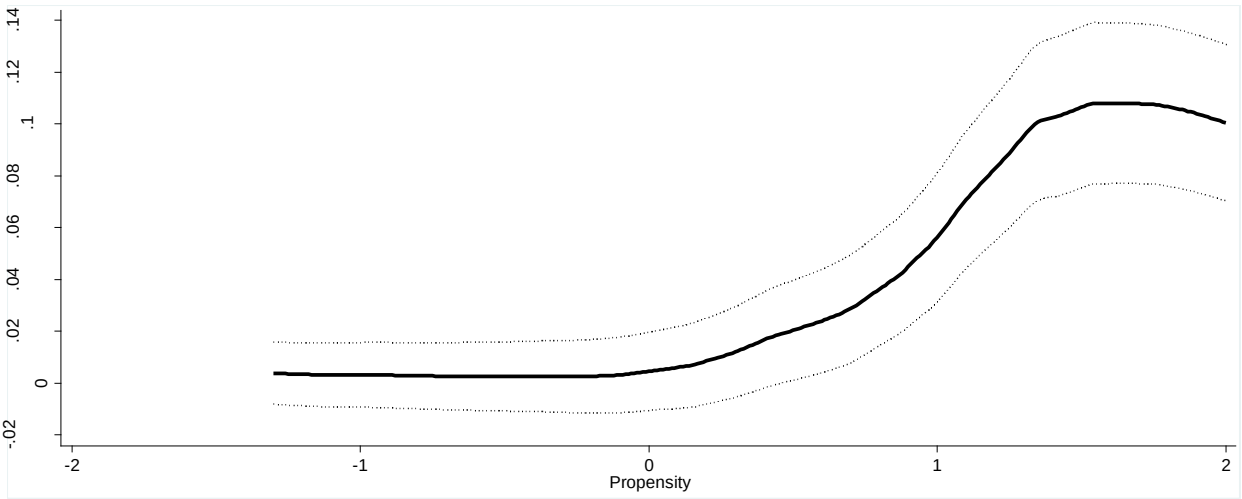
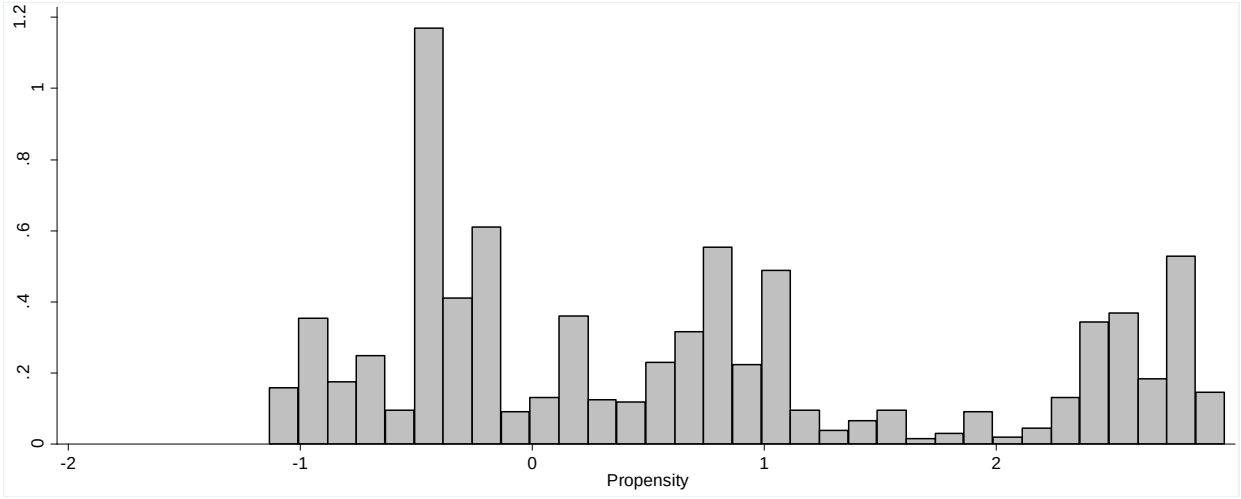
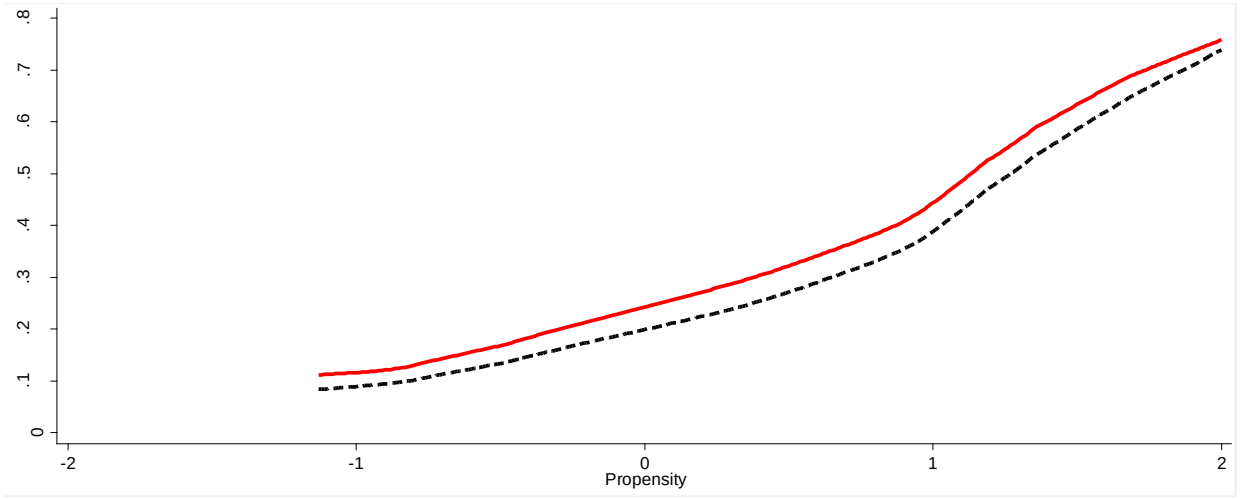


Figure A7. St. Paul (Gerber, Green, and Nickerson 2003)





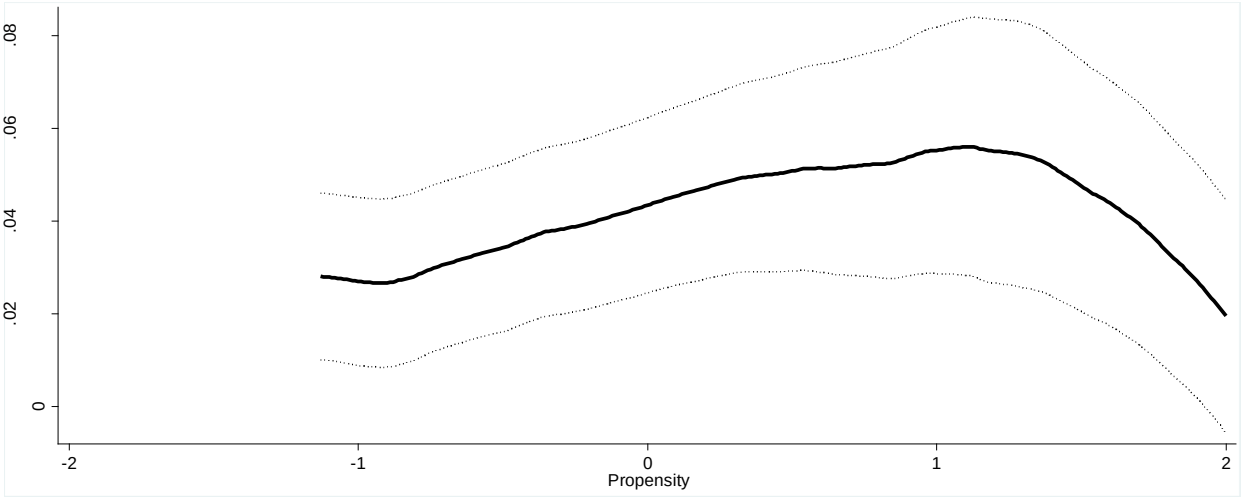


Figure A8. Stonybrook (Nickerson 2006)

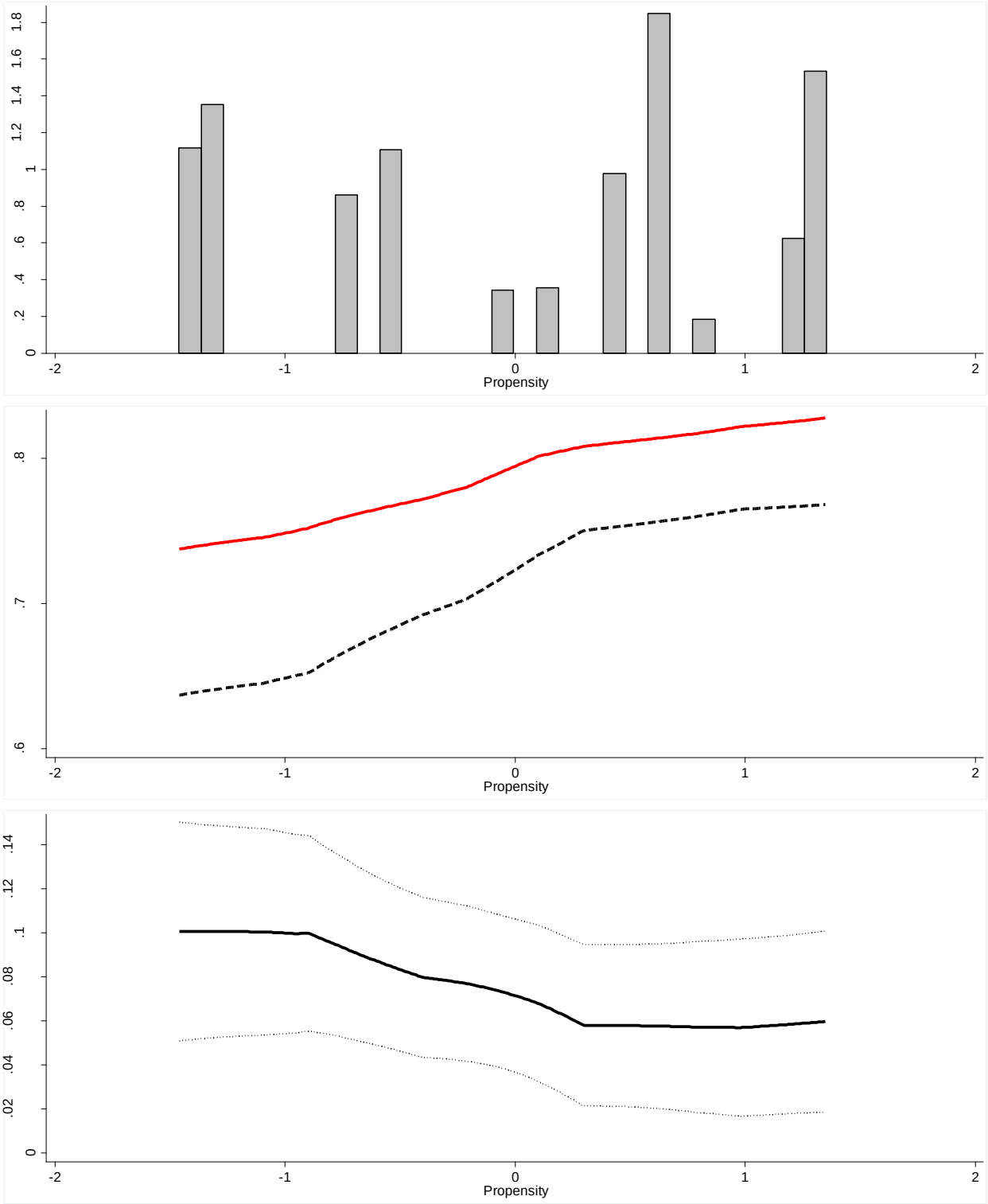
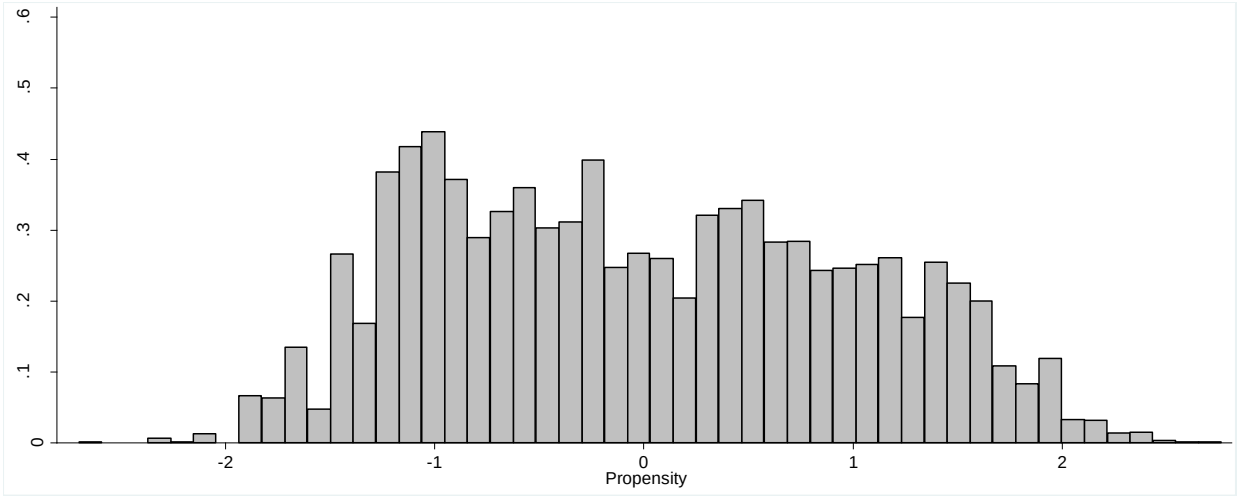
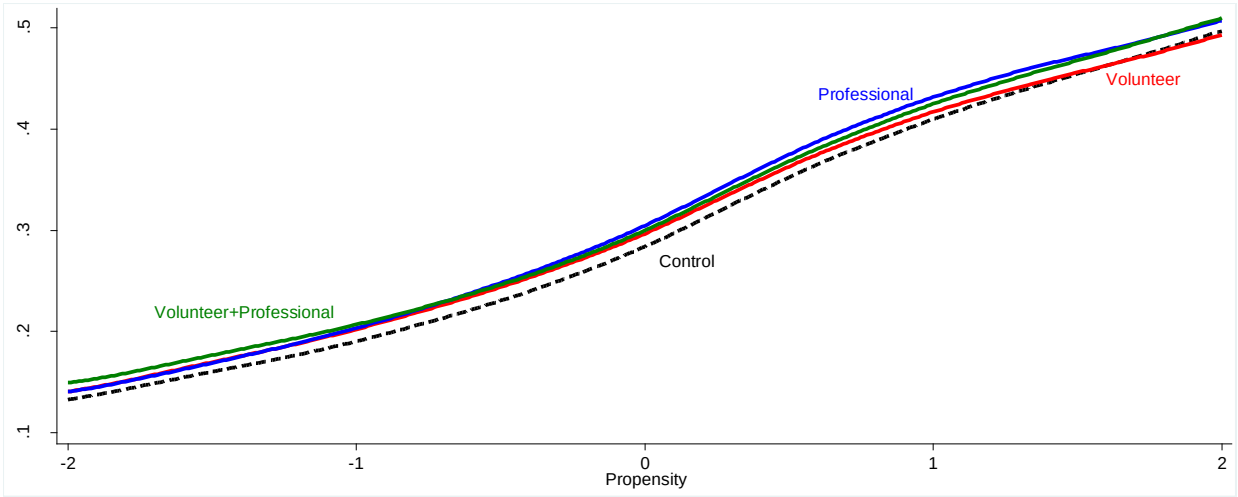


Figure A9. Nickerson (2007)





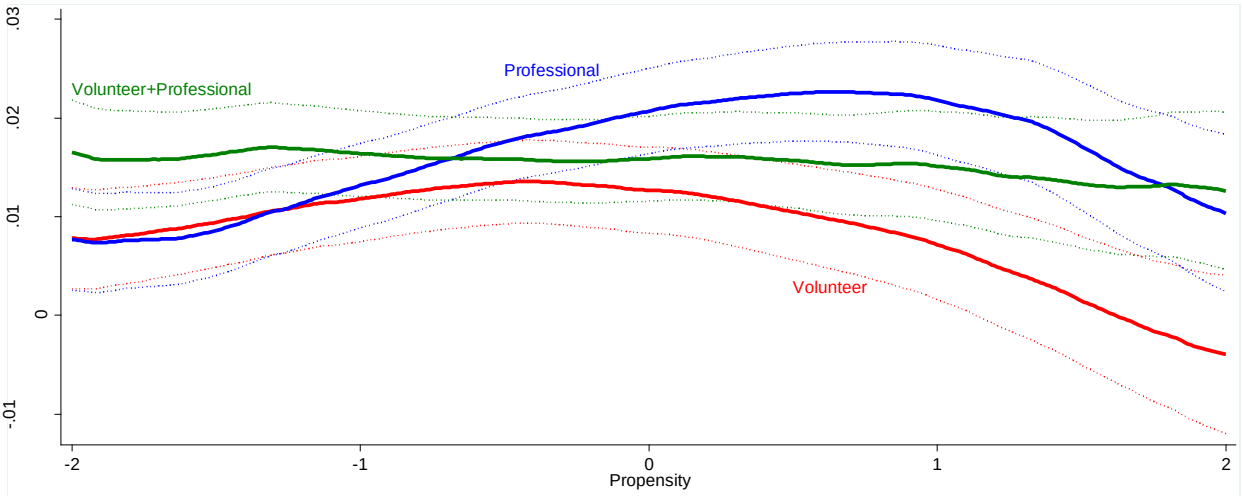


Figure A10. Gerber, Green, and Larimer (2008)

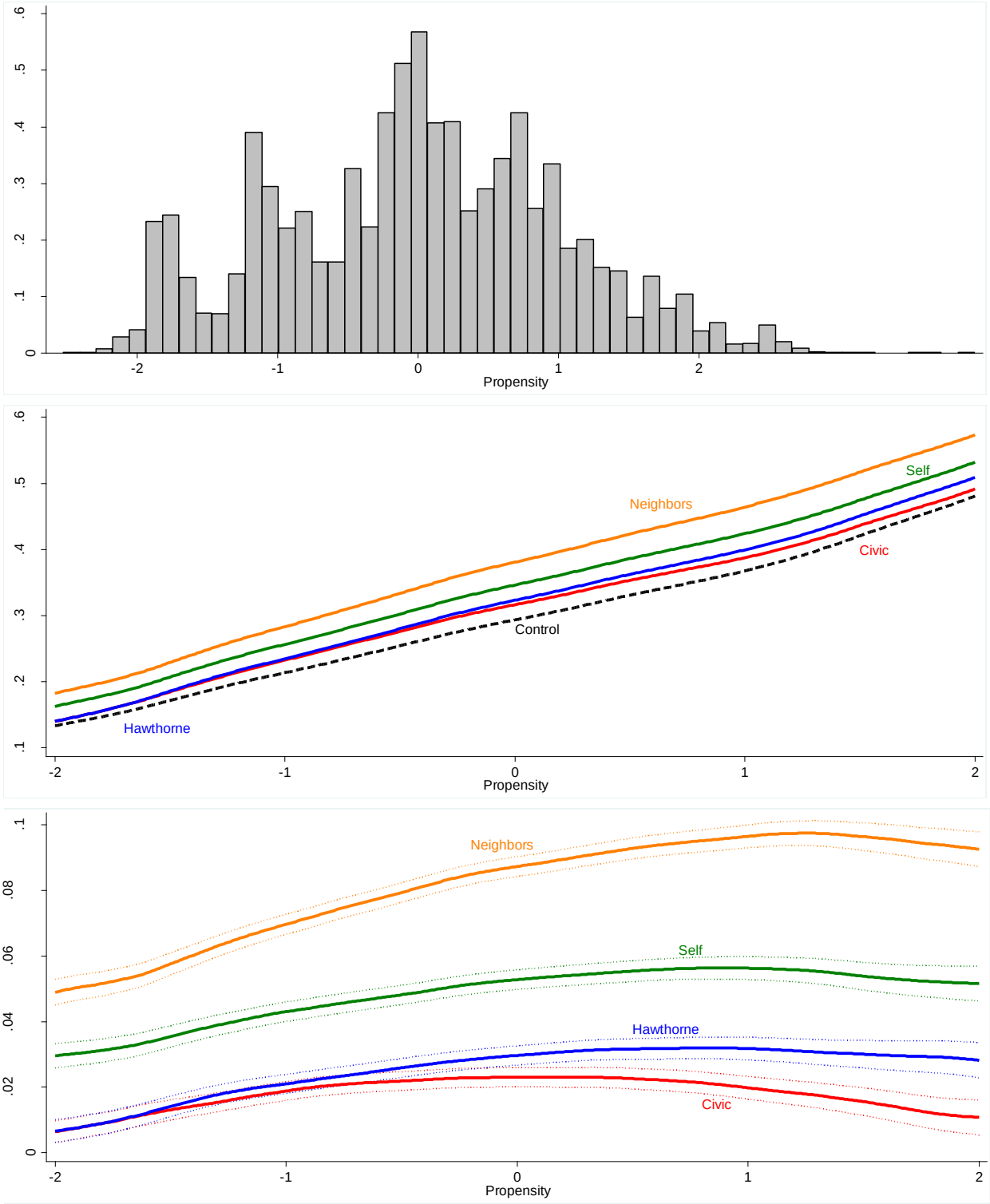
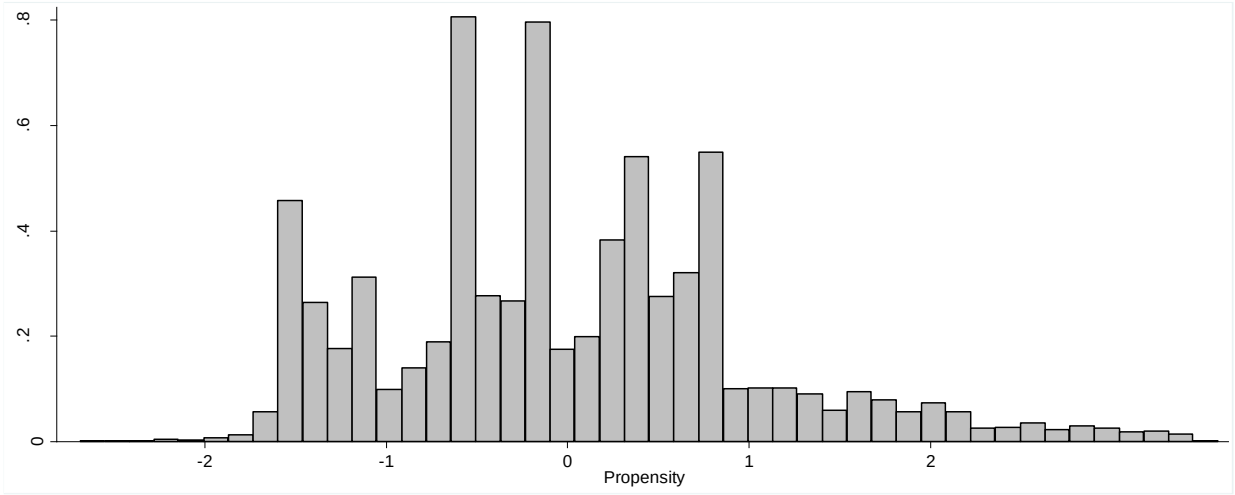
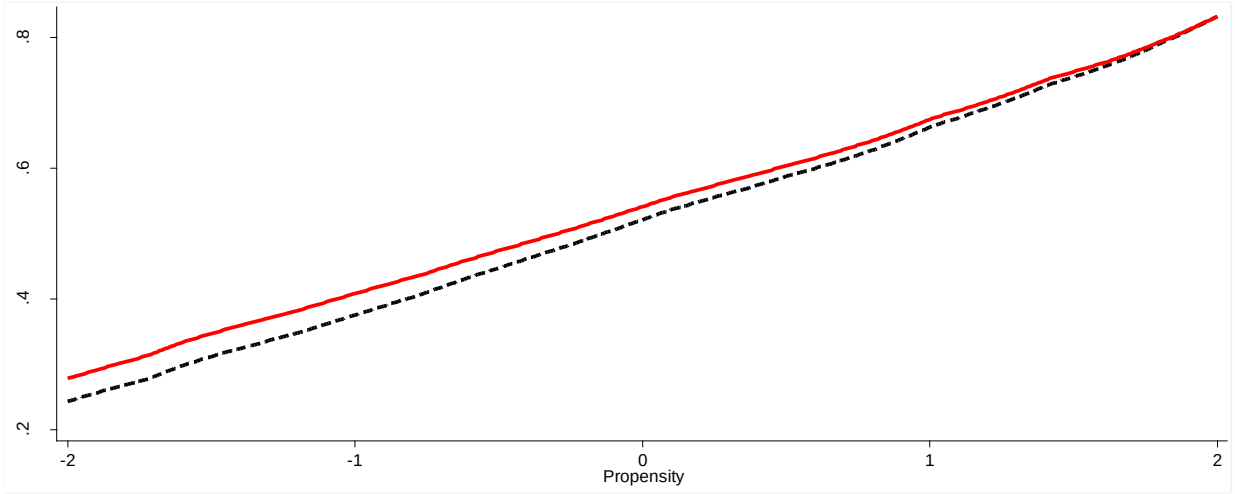


Figure A11. Move On (Middleton and Green 2008)





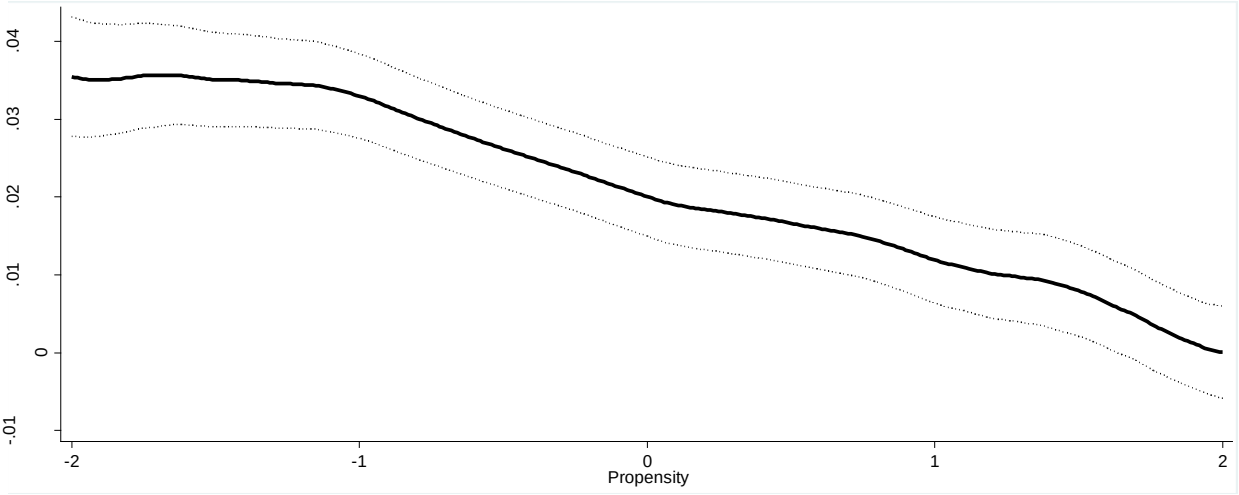
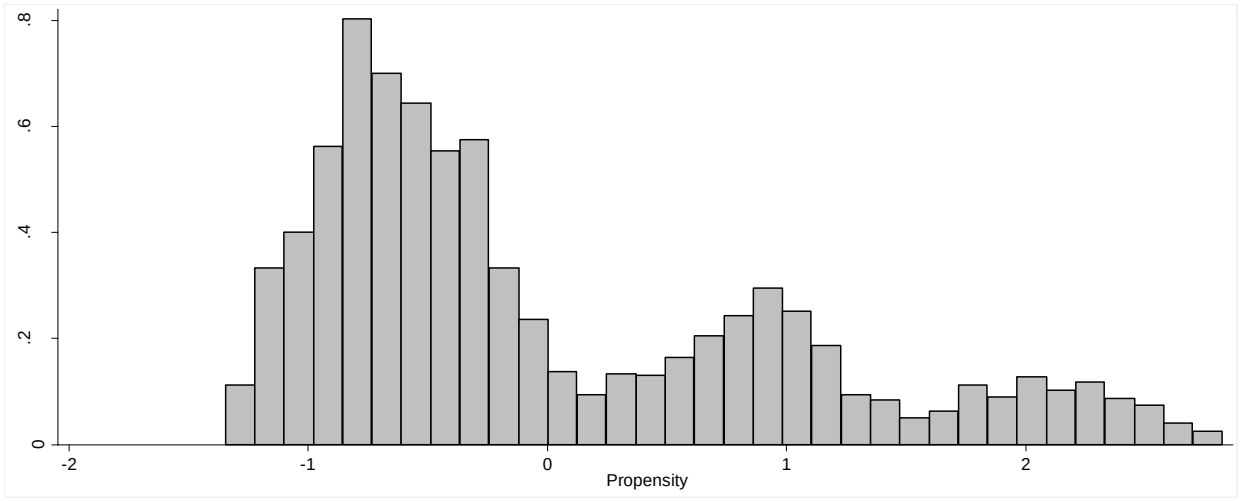
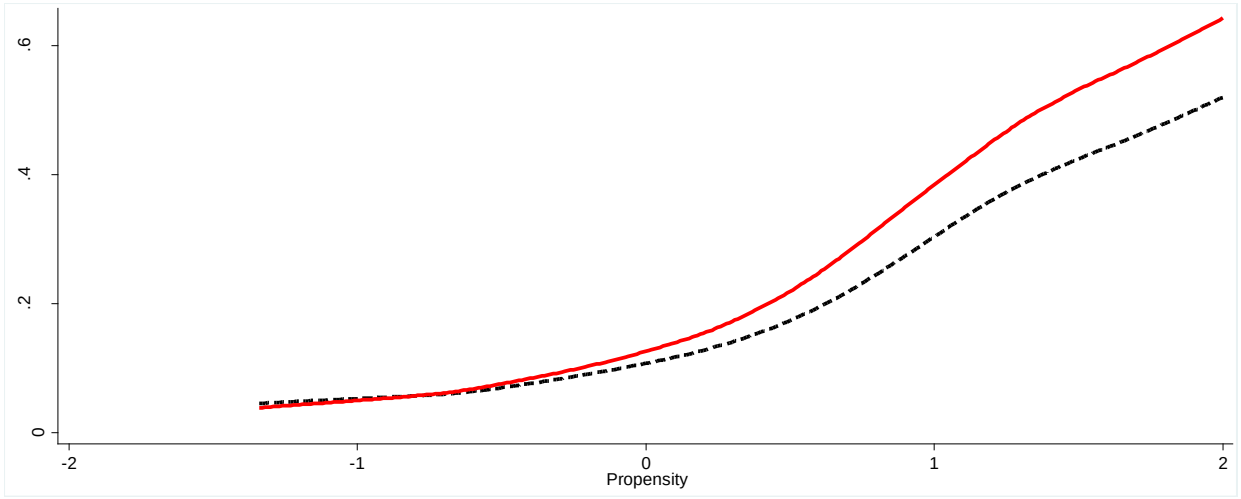


Figure A12. Minneapolis (Nickerson 2008)





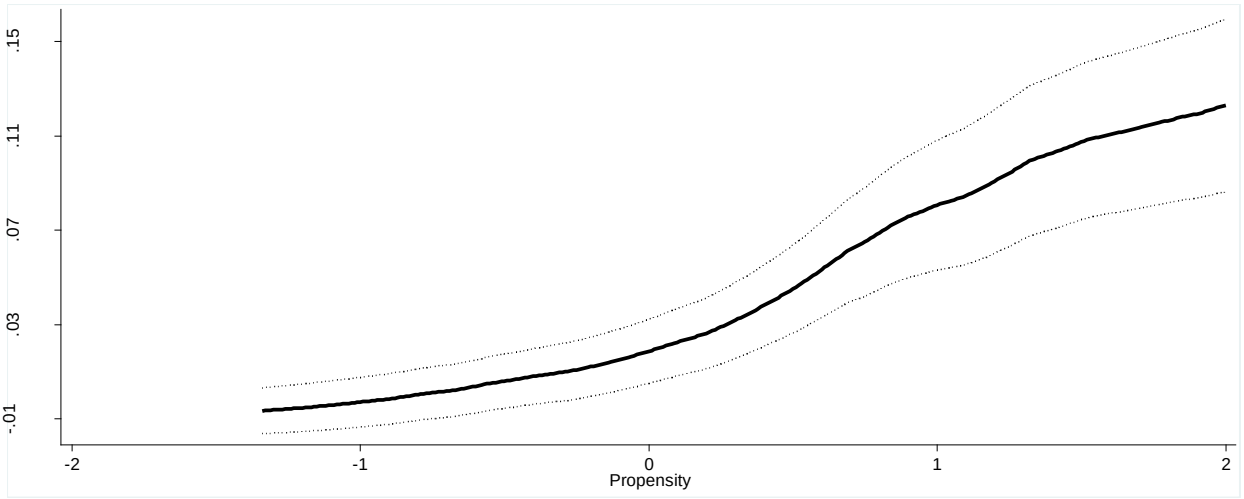


Figure A13. Text Message (Dale and Strauss 2009)

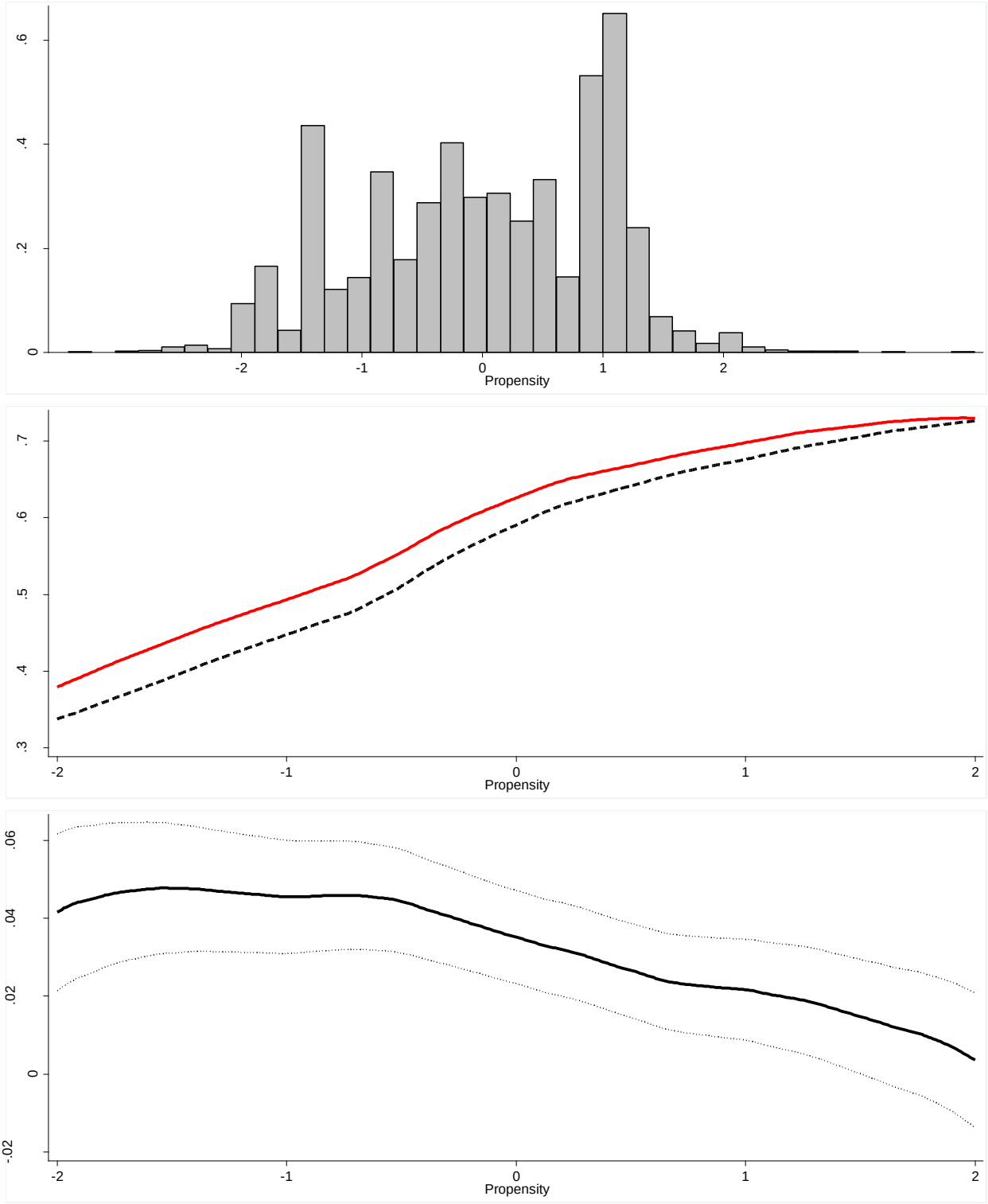
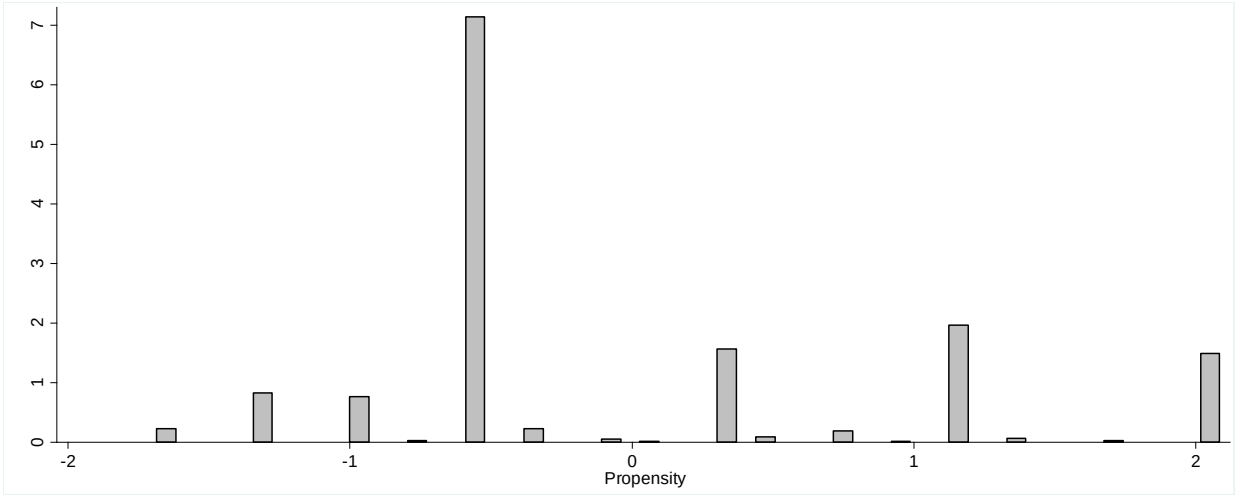
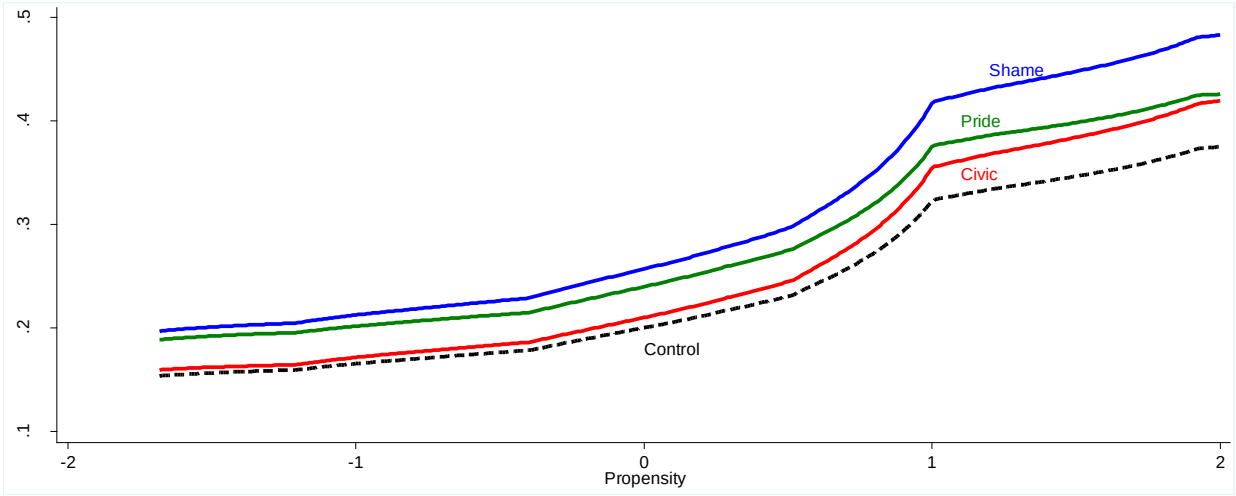


Figure A14. Gerber, Green, and Larimer (2010)





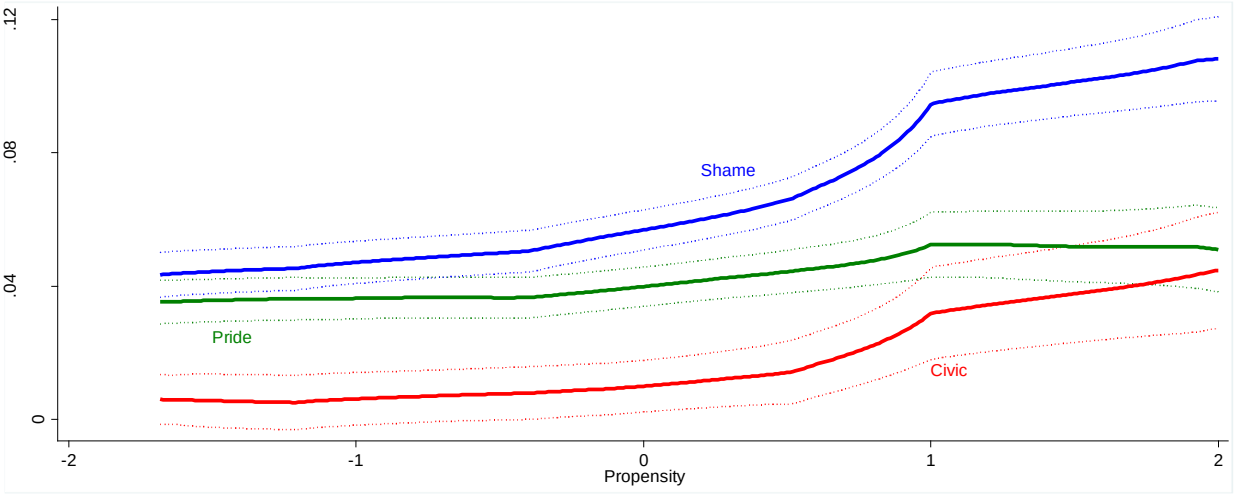


Figure A15. Party Registration (Gerber, Huber, and Washington 2010)

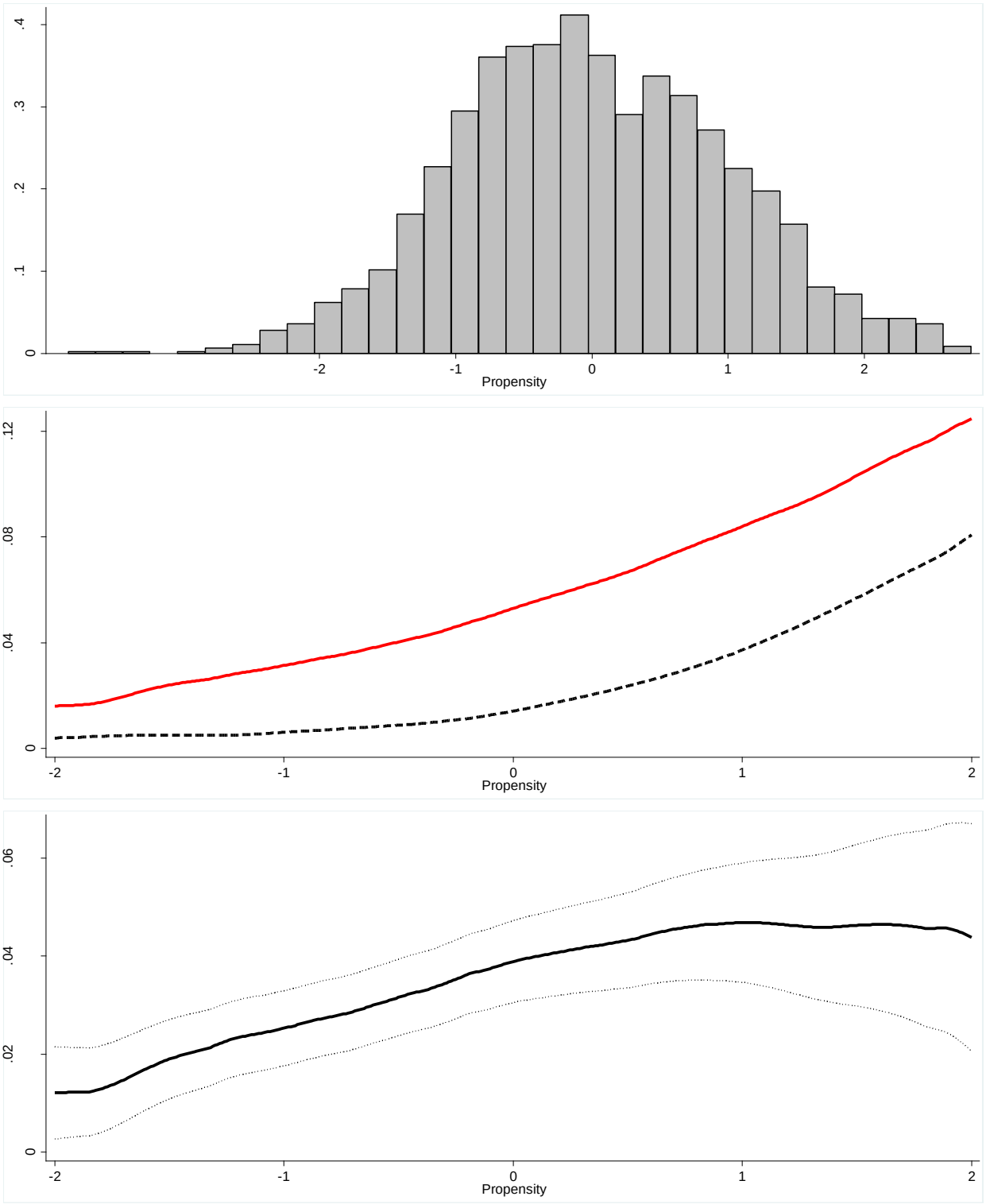
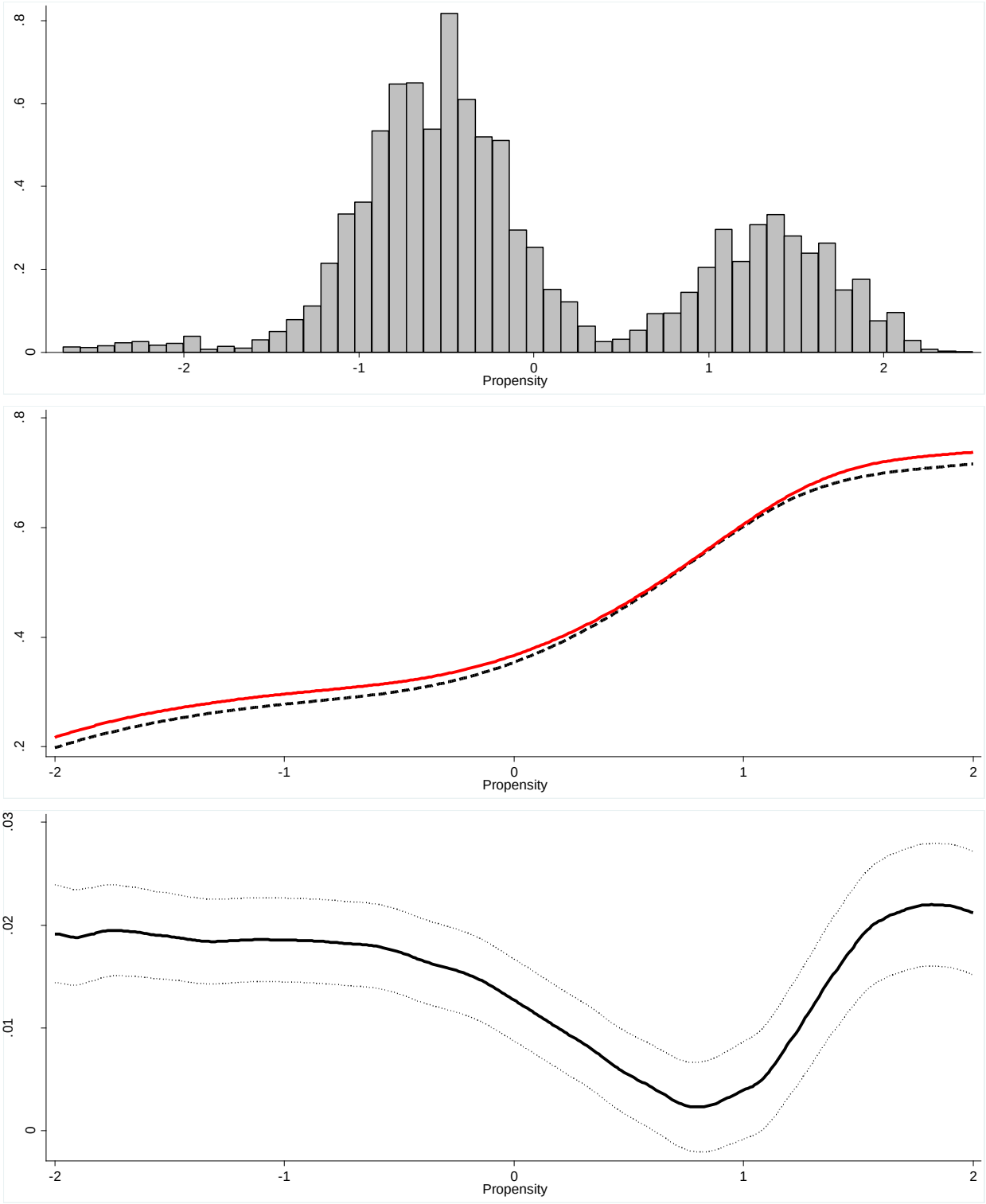


Figure A16. Planning (Nickerson and Rogers 2010)



Reference

Ansolabehere, Stephen, Eitan Hersh, Alan Gerber, and David Doherty. 2010. Voter Registration List Quality Pilot Studies. *Caltech/MIT Voting Technology Project* http://vote.caltech.edu/drupal/node/335.