FLORIDA STATE UNIVERSITY

COLLEGE OF SOCIAL SCIENCES AND PUBLIC POLICY

THE SOCIAL IDENTITY OF PARTISANSHIP:

MEASURING THE "IDENTITY" IN PARTY IDENTIFICATION

By

DENNIS FRANKLIN LANGLEY

A Dissertation submitted to the
Department of Political Science
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2018

Copyright © 2018 Dennis Franklin Langley. All Rights Reserved.

Dennis Franklin Langley defended this dissertation on July 12, 2018.
The members of the supervisory committee were:

Brad Gomez

Professor Directing Dissertation

Ashby Plant

University Representative

Matthew Pietryka

Committee Member

Robert Jackson

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To my parents, for their unfailing love and support.

ACKNOWLEDGMENTS

I thank Brad Gomez and Matthew Pietryka for pushing me to ask questions, think critically, and for training me to be a scientist. I thank Rob Carroll for much needed advice on machine learning. I thank Rachel Tuning for countless coffee trips and sanity checks. Finally, I thank Kevin Fahey and Phil Henrickson for helping me survive graduate school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Party identification is perhaps the central concept in political science. It has appeared in countless theories and empirical analyses across the political science literature. Party identification was originally conceived as a psychological attachment to the political parties, but recent evidence shows that the standard measure for party identification, the NES measure (PID), confounds group identity and group attitude. Recent work has turned to social identity theory in order to develop a measure of party identification, called party social identity (PSI), which more directly accounts for the social identity nature of party identification. This dissertation essentially explores the difference between PID and PSI as measures of party identification.

Chapter One explores social identity theory and its application to political primary elections in the US. I argue that social identity theory explains support for establishment and anti-establishment candidates. Using data from the 2016 Cooperative Congressional Election Study (CCES), I show that party social identity explains variation in candidate preferences in political primaries where the standard measure of party identification cannot. Primary voters with strong psychological attachments to the political parties are more likely to support establishment candidates and are less likely to support anti-establishment candidates. The results demonstrate the importance of the party social identity measure.

The second Chapter explores an important data problem. Party social identity should be used in a wide variety of theoretical settings and empirical analyses. Unfortunately, the measure requires a set of survey questions that only recently have been adopted in political

science and still do not appear on major surveys like the American National Election Studies (ANES) or the CCES. This poses a problem, as the inferences we've drawn in the past are based on a measure that is inadequate. This amounts to a missing data problem; past surveys are "missing" the PSI variable. I describe a process by which scholars can use machine learning to obtain a measure of PSI. I evaluate a number of machine learning models and show that predictions for party social identity can be obtained in the 2016 ANES even though that survey lacked the questions necessary to directly measure party social identity. I then replicate the analysis in Chapter One to validate the predicted measure.

Chapter Three explores an application of the machine learning process evaluated in Chapter Two. A major criticism of the original party identification measure argues that party identification is responsive to the political environment of the time and that changes in party identification are the result of retrospective evaluations. That is, high levels of presidential approval produce aggregate changes in identification toward the incumbent party. I argue that these findings are predicated on a measure that confounds party identification and party attitude. That is, scholars cannot distinguish whether aggregate movement toward the incumbent party is a result of changing identifications or changing attitudes. I use the machine learning process described in Chapter Two to obtain predictions of party social identity in the 2004 ANES Panel Study. I show that retrospective evaluations are shown to produce changes in PID but not in PSI. The results further illustrate the need for scholars to measure party social identity.

CHAPTER 1

PARTY IDENTIFICATION AND CANDIDATE PREFERENCES IN PRIMARY ELECTIONS

Every four years the two major political parties hold presidential primary elections to choose their nominees for the general election. While a great deal of research explains the various factors influencing candidate preferences in general elections, the academic literature regarding preferences in primary elections is less expansive. Existing literature suggests that primaries are less about issues (Gopoian, 1982) and more about electability (Abramowitz, 1989) and sophisticated voting (Abramson et al., 1992). Whereas scholars can use party identification as a strong predictor of general election behavior, scholars cannot use such identity as a predictor of primary election behavior. Primary elections essentially split the sample of voters by party ID, "controlling" for the effects of partisanship. Democrats tend to vote for Democrats in general elections, but such tendencies give no guidance for primary elections.

Given the way the NES measure (PID) captures party identification, scholars cannot use "party identification" in its standard seven-point form to explain preferences in political primaries. Scholars could turn to the collapsed "partisan strength" version to argue that strong identifiers behave differently than weaker identifiers. However, strength of partisanship in this sense is conceptually distinct from a psychological measure of party identity.

The former *could* be tapping into a psychological identification with the party but it could also be a proxy for political awareness, agreement with the party, or a handful of other theoretically-related concepts.

Furthermore, Greene (2002) argues that the NES measure "confounds the empirically and theoretically distinct psychological concepts of attitude and group identity." It measures a broader concept of partisanship, i.e. how political someone is in general, rather than the more specific concept of a person's psychological attachment to a political party. While attitudes toward the parties and identification with the parties are related concepts, it is difficult for scholars to properly assess the effect of party identification on certain political outcomes with a measure that confounds identity and attitude.

Scholars like Greene (2002; 2004) have turned to social identity theory to help solve this problem. A social identity is the aspect of a person's self-concept based upon their perceived group membership (Turner and Oakes, 1986). Even among strong partisans, variation in party social identity exists and party social identity explains variation in party feeling thermometers even while accounting for partisan strength (Greene, 2004). The same is true for ideological social identity: even among strong liberals and conservatives, variation in ideological social identity exists (Devine, 2015). The variation in party social identity, while almost entirely ignored in extant research, can be a useful tool in explaining candidate preferences in political primaries.

According to social identity theory, members of social groups seek to minimize in-group differences and maximize differences with out-groups. Among the most widely-replicated findings of social identity theory is the existence of ingroup bias and favoritism and outgroup

discrimination. These results have strong implications for behavior in political primaries. Strong party identifiers, according to SIT, would likely support primary candidates whom they see as members of their social group– i.e., establishment candidates of their political party. Strong party identifiers should also be more likely to support candidates who have been on the public stage longer or those who have held party leadership positions, as both traits signal a stronger connection to the party group.

Observational data from a nationally-representative survey can be used to test these implications. Using data from the 2016 CCES, I test whether respondents with strong party social identities were more or less likely to support establishment candidates in the primaries. Even controlling for partisan strength (the folded version of the PID variable), I show that strong party social identifiers were more likely to vote for establishment candidates and less likely to vote for anti-establishment candidates in the 2016 primary elections. These findings underscore the importance of a social identity-based measure of party identity by demonstrating its applicability and theoretical relevance in a context where the standard partisanship measure, PID, is not theoretically or empirically useful.

1.1 Participation in Political Primaries

Vote choice in general elections has long been a focus in political science. Primary elections are often just as intense as general elections yet they remain relatively understudied. This lack of attention can pose an important problem: as Gopoian (1982) argues, "primaries represent the first chain linking public opinion with public policy." Furthermore, primary elections are an opportunity for sincere voting. Voters have a variety of choices close

to their own positions, allowing them to select candidates close to their own policy positions. An understanding of vote choice in primary elections is thus of great theoretical importance.

The literature on primary election behavior identifies some important correlates of vote choice. Gopoian (1982) and Marshall (1984), for example, show that issue preferences are less relevant than the personal attributes of the candidates. Norrander (1986) agrees, showing that issues tend to be uncorrelated with candidate preference. Abramowitz (1989) argues that voters base their vote choices on electability, supporting candidates they think will perform well in general elections. Abramson et al. (1992) supports this, arguing that voters update their preferences "more in line with changing perceptions of viability rather than candidate evaluations" and note that voters "support the office seeker who best combines feasibility as a candidate with attractiveness as a potential nominee." Collingwood et al. (2012) demonstrates a similar effect in the 2008 primaries, showing that those who cast votes in elections later on in the process used results from earlier elections to update their preferences.

Scholars have made some use of partisan and ideological identification, but they often occupy secondary roles in explaining vote choice. Wattier (1983), for example, argues that ideological identification is only relevant when voters see the leading candidates as identical in terms of favorability. Norrander (1986) shows that ideological identification is only important in multi-candidate primaries, where it can help voters select a nearby candidate, and not two-candidate primaries, where the candidates should be ideologically near each other anyway. In general, extant literature tends to focus on the attitudes and evaluations held by individuals, rather than aspects of their identity, in explaining behavior in primary elections.

Primary voters, according to this literature, are seen as strategic actors. In order to avoid a negative outcome (i.e., losing the general election), primary voters will select a less preferred candidate in the primary if it means a better chance at winning the general election. But perceptions of things like personal attributes and electability might still be a function of the voter's identity. A voter who strongly identifies with a social group (e.g. a party or an ideological group) would view candidates who belong to that social group more favorably than other candidates. (That is, a voter in the 2016 primaries who strongly identified with the Democratic Party would have viewed Hillary Clinton more favorably than Bernie Sanders by virtue of her connection to the Democratic Party.) Evaluations of the candidate's personal attributes or general election viability are thus a function of the voter's own identity. Scholars might be overestimating the occurrence of strategic behavior by misinterpreting sincere behavior.

One circumstance of studying primary election behavior is the inability to use party identification as an important explanatory variable. The authors of *The American Voter* (Campbell et al., 1960) emphasize "the role of enduring partisan commitments in shaping attitudes toward political objects." Yet primary elections effectively control for party identification by splitting Republican and Democratic voters, leaving scholars unable to examine the role of those partisan commitments. Even open primary elections require voters to choose which party's primary to enter before they make their vote decisions, so the fact that a given respondent is a Republican tells us little about their vote preferences once they've entered the Republican primary. But if party identification plays a primal role in determining political

behavior, as Campbell et al. (1960) suggest, surely its omission poses important theoretical and empirical concerns.

The omission of party identification in models of primary election behavior stems from the design of the survey questions that measure party identification. The NES measure of partisanship is a series of questions, the first of which asks survey respondents whether they think of themselves as Republicans or Democrats. This is an important psychological consideration, to be sure, and likely does tap into the respondent's psychological attachment to the party group. Miller (1991) argues that scholars should use this "root" question as the definition of party identification. However, this question does little more than identity *which* group respondents feel like they belong to; it provides no empirical measure of the *strength* of that identification.

The second question in the standard PID measure asks Republican- and Democratic-identifiers whether they think of themselves as strong or weak identifiers and asks Independents whether they lean closer to one party or the other. The assumption is that this question measures the strength of identification with the parties, that strong identifiers identify with the party more than weak identifiers. However, this assumption may not be valid. The latter part of the measure, which is often used to create a "strength of partisanship" variable (PID strength), does not address identification as a particular psychological attachment consistent with the way the authors of *The American Voter* conceive it. Is a "strong partisan" one who psychologically identifies with the party (a la *The American Voter*), one who likes the party, or one who agrees with the party on a number of issues? And if Respondent A is "stronger" than Respondent B, is it because A identifies more with the party, because A likes the party

more, or because A agrees with the party on more issues? PID strength cannot distinguish between these cases.

The inadequacies of the PID strength measure are particularly apparent in political primary elections. In general elections, PID strength gives scholars some indication of the decisions voters will make: a strong Democrat is more likely to vote for the Democrat than for the Republican. In primary elections, though, the voter is choosing from a slate of candidates from the same party. The fact that the voter identifies with the Democratic Party tells scholars nothing about whether that voter is more likely to support, say, Hillary Clinton or Bernie Sanders. Furthermore, because the folded PID strength measure cannot distinguish between one who strongly likes one party more than the other and one who strongly agrees with one party more than the other, it is not particularly useful in explaining vote choices among primary candidates that are members of the same party and who likely agree on a variety of policy issues.

A scholar interested in the effect of party attitudes on political behavior can forego the NES measure in favor of feeling thermometers. Likewise, a scholar interested in the effect of party agreement on political behavior can ask policy preference questions. But scholars interested in the effect of identity on political behavior are limited to the NES measure. Greene (2000; 2002; 2004) presents an alternative measure of party identification, one without the problems of the NES measure. Pointing to the social psychological roots of *The American Voter*, Greene draws on social identity theory to develop a measure of party social identification.

1.2 Social Identity in Political Primaries

Social identity theory is a social psychological concept originating in the early 1970s in the work of Tajfel and Turner (Turner and Oakes, 1986). One's social identity consists of "those aspects of a person's self-concept based upon their group memberships." The basic hypothesis of social identity theory, according to Turner and Oakes, is that "people are motivated to seek positive social identity by comparing in-groups favorably to outgroups." People with strong social identities seek to maximize differences between the ingroup and the outgroup (Greene, 2002). At its core, social identity theory is a story about intergroup behavior. Motivated by this need for positive social identity, people will favor other people whom they perceive to be in the same social group.[1]

This ingroup favoritism has direct implications in political science and especially in studies of behavior in political primaries. Presidential elections, for example, are inter-group conflicts. Voters have clear information about the social groups to which the candidates belong and can act on that group belonging. We know that Republicans tend to vote for Republicans in general elections and social identity theory provides an explanation as to *why* this is the case. Those with strong social identities are motivated to engage in behaviors that favor their ingroup and thus should be more likely to vote for that party or to engage in other forms of participation like volunteering for or donating to the party.

Political primaries, despite being intra-party conflicts, should still elicit ingroup favoritism. Indeed, not all primary candidates are created equal. Take the 2016 Republican presidential primaries for example. Of the 17 total major candidates, three (Donald Trump,

[1]See Huddy (2001) for a more thorough review of social identity theory.

Ben Carson, and Carly Fiorina) were not politicians, some (like Ted Cruz) ran on explicitly anti-establishment principles, and others (like Jeb Bush) positioned themselves squarely as establishment candidates. Even though all of the candidates were nominally Republicans they had markedly different attachments to the party and thus to the social group. According to social identity theory, voters in the primary elections still likely acted on these perceptions of group belonging. Strong party social identifiers should favor the candidates with closer attachments to the party.

However, because of the way the NES measure of partisanship is constructed, scholars do not have a measure of party identity that works for political primaries. As Greene (2002) shows, strength of partisanship alone is a limited measure of party identity. So how, then, should scholars measure party identity? Greene (2000; 2002; 2004) develops exactly such a measure. Drawing from social psychology and organizational studies, Greene shows that the Identification with a Psychological Group (IDPG) scale provides a reliable, valid measure of social identity for use in political science. The original 10-item IDPG scale asks respondents questions like "when someone criticizes this group, it feels like a personal insult" and "this group's successes are my successes." The IDPG scale thus gives a measure of party identification that properly accounts for the distinction between attitude and group identity. Greene (2002) demonstrates that the scale "captures an important group element of party identification not fully measured" by the standard NES measure. Indeed, he shows that party social identity explains candidate feeling thermometer differentials, over-time party support, party activity, and in-party *and* out-party feeling thermometers even while accounting for partisan strength.

Social identity theory not only provides scholars with directly testable implications about political behavior but a method of measuring party identification that properly accounts for the distinction between group attitude and group identity. Voters with strong party identities are motivated to engage in behavior consistent with their attachment to their political party. This should be true even in primary elections, where establishment candidates would be seen as ingroup members. Establishment candidates are generally those seen as elites in the party group, or those who have the support of party elites. Candidates like Hillary Clinton pose themselves as members and representatives of the party itself and are thus seen as members of the social group. Anti-establishment candidates, on the other hand, pose themselves as outsiders wishing to change the party. Establishment and anti-establishment candidates directly parallel the in-group and out-group dynamics of social identity theory. Voters who strongly identify with the party as a social group would support candidates who are clear members of their social group and would oppose outsider candidates they see as a threat.

If strong group identifiers are more likely to favor in-group members and discriminate against out-group members, we should expect that those with strong party identities are more likely to support establishment candidates (regardless of that candidate's ideological position) and are less likely to support anti-establishment candidates in primary elections. In other words, high levels of party social identity should *increase* the likelihood of voting for an establishment candidate and *decrease* the likelihood of voting for an anti-establishment candidate. These effects should occur even when accounting for partisan strength or ideology.

1.3 Data

In this chapter I argue that party social identity (PSI) has a strong effect on vote choice in political primary elections and that this persists even when controlling for strength of partisan identity (PID). To test these hypotheses I draw data from the 2016 Coooperative Congressional Election Study. I make use of both the "Common Content" and a number of supplementary questions I added specifically regarding social identity.

1.3.1 Variables

Party Social Identity. I asked a number of questions to measure the party social identity of respondents. These questions are drawn from those originally tested in Mael and Tetrick (1992) and validated for political science in Greene (2004). Survey space constraints necessitated the use of a four-item scale, restricted only to the respondent's given party. The four questions were asked as follows: "When someone criticizes [this group], it feels like a personal insult;" "When I talk about [this group], I usually say "we" instead of "they";" "[This group]'s successes are my successes;" "When someone praises [this group], it feels like a personal compliment."[2] Responses were on a five-point scale of agreement ranging from "Strongly disagree" to "Strongly agree."

I used the respondent's given political party (Democrats/Republicans) to produce four party questions for each respondent. The party social identity (PSI) scores used in this analysis are summed totals for the four responses, producing a range of 0 (no social identity) to 16 (the strongest social identity). For the purposes of this analysis, respondents who answered

[2]Devine (2015) shows that a subset scale still performs reliably compared to the full 10-item scale. He provides evidence that a simple two-item subset, using just the "success" and "praise" items.

"Independent" for the party question were coded as 0 for the party social identity measure. This is done to illustrate the effect of increased levels of social identity– an Independent and a Republican with no party social identity are empirically identical for these purposes.[3]

Primary Participation. I also make use of the CCES questions regarding participation in the political primaries. Respondents were asked whether they participated and, if so, for which candidate they voted. Responses for Democratic candidates were either Hillary Clinton, Bernie Sanders, or "Another Democrat"; for Republicans, available responses were Donald Trump, Ted Cruz, John Kasich, Marco Rubio, or "Another Republican." I used these responses to create binary variables measuring votes for establishment (Hillary Clinton, John Kasich, or Marco Rubio) or anti-establishment (Bernie Sanders, Donald Trump, or Ted Cruz) candidates.

Clinton and Sanders made their establishment and anti-establishment identities a clear part of their campaign strategies. Kasich's long career in the U.S. House and as Governor of Ohio made him a clear establishment choice. Rubio, despite having entered the U.S. Senate with large Tea Party support, received support from the establishment and was seen as the establishment's best chance of defeating Trump in the primary. Trump and Cruz, meanwhile, were prototypical anti-establishment candidates, often attacking elites in both parties and positioning themselves as outsiders.

Control Variables. To account for alternate explanations of candidate preference in the primaries I include in my analysis measures of partisan and ideological strength, cre-

[3]Greene (2004) offers support for this. He asks respondents about their social identity regarding both their given party *and* independents and shows that independent social identity is neither common nor particularly meaningful.

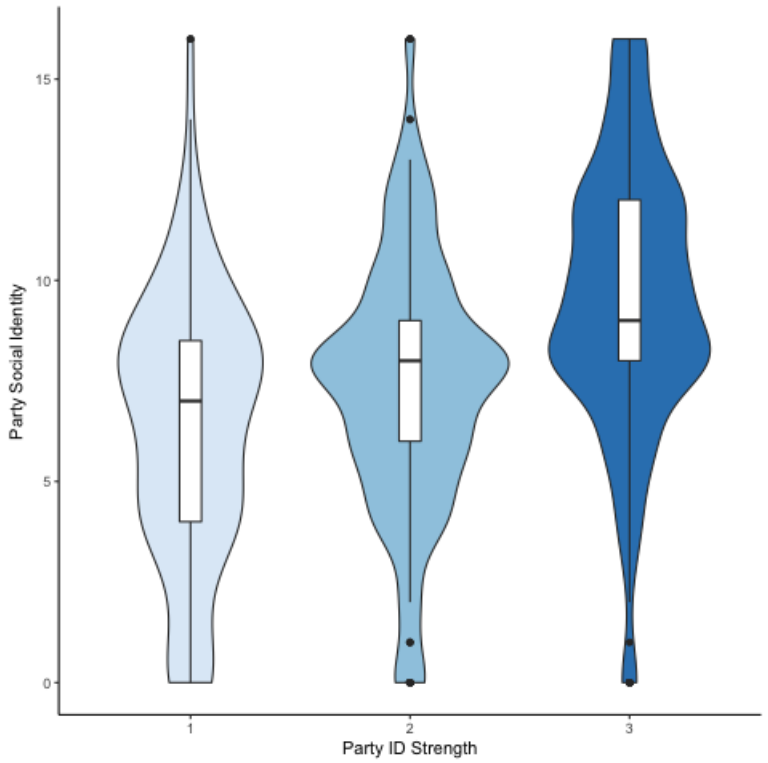

Figure 1.1: Party ID Strength and Party Social Identity

ated by folding the standard partisanship and ideology measures in half, where "Strong Republican" and "Strong Democrat" are both scored highly and "somewhat conservative" and "somewhat liberal" are low scores. I also include a measure of party agreement which measures the extent to which a respondent agrees with their given party on a number of policy issues. I also include controls for gender (a binary variable indicating 1 for female respondents), marriage status (a binary variable indicating 1 for married respondents), race (a binary variable indicating 1 for minority respondents), frequency of church attendance, interest in politics, income, and education.

1.3.2 Measurement Validation

Figure 1.1 is a violin plot showing the covariation between the folded party ID strength variable and party social identity.[4] As Greene (2002) showed, party social identity is correlated with strength of partisanship ($\rho = 0.532$ among Democrats and Republicans) and PSI increases monotonically with strength of party identification. However, there still remains considerable variation in party social identity at each level of party ID strength. The party ID strength variable essentially ignores that variation. A respondent with high social identity may still appear low in partisan strength and vice versa, suggesting that PID strength is a poor measure of group identification.

1.4 Analysis

Table 1.1 presents four logistic regression models: two use vote choice for an establishment candidate as the dependent variable and two use vote choice for an anti-establishment candidate. For each dependent variable I estimate one model without the social identity measure and one including it in order to demonstrate the effects of its inclusion. Political interest and the minority dummy variable are the only statistically significant control variables. Minority voters were more likely to have supported establishment candidates and less likely to have supported anti-establishment candidates. Those with higher political interest were more likely to have supported anti-establishment candidates.

For the first dependent variable, ideology strength is statistically significant and negative. Strong ideological identifiers were less likely to support establishment candidates. This could

[4]This violin plot shows the distribution of PSI at each level of party ID strength. A boxplot of PSI is also shown for each level of party ID strength, indicating the median and the interquartile range.

Table 1.1: Candidate Preference in the 2016 Primary Elections

	Establishment		Anti-Establishment	
	Model 1	Model 2	Model 3	Model 4
Party Social Identity		**0.11** (0.03)		**−0.10** (0.03)
Party ID Strength	**0.56** (0.16)	0.30 (0.17)	**−0.34** (0.15)	−0.09 (0.17)
Ideology Strength	**−0.22** (0.09)	**−0.24** (0.09)	**0.19** (0.09)	**0.21** (0.09)
Party Agreement	0.05 (0.07)	0.05 (0.07)	−0.06 (0.07)	−0.06 (0.07)
Female?	0.01 (0.18)	0.01 (0.19)	0.02 (0.18)	0.02 (0.18)
Minority?	**1.17** (0.21)	**1.20** (0.21)	**−1.15** (0.20)	**−1.17** (0.21)
Married?	−0.07 (0.19)	−0.07 (0.20)	−0.05 (0.19)	−0.05 (0.19)
Church Frequency	0.00 (0.05)	−0.03 (0.06)	−0.06 (0.05)	−0.03 (0.05)
Political Interest	−0.06 (0.12)	−0.10 (0.12)	0.23 (0.12)	**0.26** (0.12)
Income	−0.01 (0.02)	−0.01 (0.02)	0.01 (0.02)	0.01 (0.02)
Education	0.14 (0.10)	0.15 (0.10)	−0.17 (0.09)	−0.17 (0.09)
Intercept	**−1.65** (0.66)	**−1.67** (0.67)	0.56 (0.64)	0.55 (0.64)
Num. obs.	580	580	580	580
Log Likelihood	-349.48	-341.22	-364.68	-357.88
LR test p-value		0.00004805		0.000226

Coefficients significant at the p = 0.05 level are bolded.

be due to perceived conflict between the party and the ideology: strong ideologues wanted a departure from the status quo that establishment candidates represented. PID strength is a statistically significant predictor of voting for an establishment candidate, as previous theory might have predicted. Those with strong partisan identifications (according to the NES measure) are more strongly connected to their political party and are thus more likely to support candidates connected to that party.

However, upon including the party social identity measure, the effect of party ID strength disappears and is replaced by a significant effect of party social identity. In other words, party ID strength by itself is no longer a significant predictor of voting for establishment candidates in primaries once party social identity is accounted for. Also, the coefficient on

party social identity is smaller than the coefficient for PID strength. This is due to the empirical overlap of the two variables; PID strength measures both identity and attitude, while PSI measures only identity. While holding social identity constant, PID strength no longer has an effect on candidate preferences.

Moving to the last two models, we first see that ideological strength is a statistically significant predictor of voting for anti-establishment candidates; those with strong ideological identifications are more likely to have supported anti-establishment candidates. PID strength is statistically significant and negative, in keeping with expectation. Including party social identity has the same result as it did in the first two models. Party social identity is once again statistically significant but in the negative direction. Respondents with strong party social identities are significantly less likely to have voted for anti-establishment candidates in the 2016 primaries. Anti-establishment candidates are seen as coming from outside of the social group, so strong social identifiers are less likely to support those candidates. PID strength is not statistically significant when controlling for party social identity. Likelihood ratio tests indicate that models 2 and 4 fit the data better than models 1 and 3, respectively.

For party social identity to affect support for a candidate, that candidate must have clear ties to the social group. That is, if voters do not know which candidates are most closely associated with the party as a social group, party social identity should not have as large of an effect on candidate preferences. The 2016 primary elections provide an opportunity to test this. The Democratic primary's two candidates, Hillary Clinton and Bernie Sanders, were clearly establishment and anti-establishment, respectively. Their positions were central components of their messaging, and Democratic voters had clear signals about the two

Table 1.2: Candidate Preference among Democrats

	Establishment		Anti-Establishment	
	Model 1	Model 2	Model 3	Model 4
Party Social Identity		**0.15** (0.04)		**−0.15** (0.04)
Party ID Strength	**1.22** (0.34)	**1.02** (0.35)	**−1.08** (0.34)	**−0.87** (0.35)
Ideology Strength	−0.27 (0.14)	−0.27 (0.15)	**0.31** (0.14)	**0.31** (0.15)
Party Agreement	0.25 (0.15)	0.15 (0.15)	−0.22 (0.14)	−0.11 (0.15)
Female?	−0.18 (0.28)	−0.15 (0.29)	0.21 (0.28)	0.18 (0.29)
Minority?	**1.12** (0.32)	**1.23** (0.34)	**−0.97** (0.32)	**−1.07** (0.33)
Married?	−0.02 (0.29)	−0.10 (0.30)	−0.03 (0.29)	0.04 (0.30)
Church Frequency	0.15 (0.09)	0.07 (0.10)	−0.14 (0.09)	−0.06 (0.10)
Political Interest	−0.12 (0.20)	−0.17 (0.21)	0.22 (0.21)	0.28 (0.22)
Income	−0.04 (0.03)	−0.05 (0.04)	0.05 (0.03)	0.05 (0.04)
Education	−0.02 (0.15)	0.03 (0.15)	0.01 (0.15)	−0.04 (0.15)
Intercept	**−3.18** (1.39)	**−3.14** (1.43)	2.01 (1.39)	1.89 (1.44)
Log Likelihood	-152.90	-146.52	-152.96	-146.20
Num. obs.	266	266	266	266
LR test p-value		0.0003541		0.0002368

Coefficients significant at the p = 0.05 level are bolded.

candidates' group belonging. The field of Republican primary candidates was much wider and much more complicated. For example, Marco Rubio (R-FL) entered the Senate in 2010 as an anti-establishment Tea Party candidate, but in the context of the 2016 primary he was relatively close to the establishment.

To illustrate this point, I separate Democratic and Republican voters and replicate the models in Table 1.1. These models are presented in Tables 1.2 and 1.3. Table 1.2 shows the same effect from the larger sample: voters who strongly identify with the Democratic Party were more likely to have voted for Hillary Clinton and less likely to have voted for Bernie Sanders. Democratic voters had clear signals about the extent to which Clinton belonged, and Sanders did not belong, to the party as a social group. Among Republicans, however,

Table 1.3: Candidate Preference among Republicans

	Establishment		Anti-Establishment	
	Model 1	Model 2	Model 3	Model 4
Party Social Identity		−0.01 (0.07)		0.01 (0.07)
Party ID Strength	−0.10 (0.52)	−0.08 (0.54)	0.49 (0.46)	0.46 (0.48)
Ideology Strength	0.33 (0.29)	0.33 (0.29)	**−0.54** (0.27)	**−0.54** (0.27)
Party Agreement	**−0.33** (0.16)	**−0.33** (0.16)	**0.29** (0.14)	**0.29** (0.14)
Female?	0.22 (0.48)	0.21 (0.48)	−0.35 (0.43)	−0.34 (0.44)
Minority?	−0.69 (0.83)	−0.70 (0.84)	−0.21 (0.61)	−0.20 (0.62)
Married?	0.02 (0.50)	0.01 (0.50)	0.10 (0.45)	0.10 (0.45)
Church Frequency	−0.12 (0.14)	−0.11 (0.14)	−0.00 (0.13)	−0.01 (0.13)
Political Interest	−0.03 (0.29)	−0.03 (0.29)	0.33 (0.25)	0.33 (0.25)
Income	0.05 (0.06)	0.05 (0.06)	−0.04 (0.05)	−0.04 (0.05)
Education	0.32 (0.25)	0.32 (0.25)	−0.30 (0.23)	−0.29 (0.23)
Intercept	−1.37 (1.84)	−1.35 (1.85)	−0.38 (1.59)	−0.42 (1.60)
Log Likelihood	-63.76	-63.75	-75.24	-75.21
Num. obs.	148	148	148	148
LR test p-value		0.8864		0.8287

Coefficients significant at the p = 0.05 level are bolded.

the signals were less clear. As table 1.3 shows, party social identity has no strong effect on candidate preferences. I argue that this is simply a reflection of the field of candidates in the Republican primary. Smaller candidate fields should mean clearer distinctions between establishment and anti-establishment candidates which should lead to a stronger relationship between a voter's party social identity and their candidate preference, as seen in the Democratic subsample in Table 1.2.[5]

As a whole, these results demonstrate the power of a social identity-based measure of party identification. Both models show that party social identity is an important factor in vote choice in political primary elections. Those with strong party social identities are more

[5]A Chow test confirms that the results presented in Table 1.1 are not sensitive to any one candidate.

likely to support establishment candidates and are less likely to support anti-establishment candidates. The results also show that the classic NES measure of partisanship is lacking in meaningful ways, both theoretically and empirically. Inclusion of the social identity variables improved model fit in both cases, indicating that they measured a concept not accounted for with the standard partisan and ideological strength variables.

1.5 Discussion

Party identification has long been regarded as a psychological attachment to a political party. While the field of social identity theory wasn't developed until the 1970s, Campbell et al. (1960) clearly conceived of party identification as a social identity and consider political parties to be social groups like racial or religious groups. However, the canonical measure of party ID, the NES measure, does not adequately account for the psychological nature of party ID. While it is capable of separating "strong partisans" from "not very strong partisans," the result is a blunt measure of an important concept. Treating party ID as a psychological attachment to a political party without measuring it as such presents theoretical and empirical problems for researchers exploring the relationship between party identification and political behavior.

In this chapter I argue that the social identity-based measure of party ID first developed by Greene (2002; 2004; 2005) can aid researchers in understanding the link between identity and behavior. The results presented in this chapter indicate that such a measure can be used to help explain vote choice in political primaries, even when accounting for partisan strength. Voters with strong party social identities are more likely to support establishment

candidates and are less likely to support anti-establishment candidates. These findings were not possible without a better measure of party identification.

Political scientists should more thoroughly explore the interaction between party social identity and other correlates of vote choice in primaries. For example, do primary voters rely more on social group belonging or on electability when choosing which candidate to support? Party social identity may also affect perceptions of electability: a voter who is high in party social identity may overestimate an establishment candidate's electability and underestimate an anti-establishment's electability. Party social identity may also affect evaluations of the personal attributes of the candidates: a voter who is high in party social identity may exaggerate the positive attributes of in-group candidates and the negative attributes of out-group candidates. Future scholars should consider party social identity when examining behavior in political primaries.

Further research can demonstrate potential uses for such a measure and help paint a broader picture of the relationship between identity and behavior. For example, are over-time changes in partisanship a function of changes in identity or an artifact of social desirability bias? Implementing a social identity-based measure of party identification in election surveys over time can help researchers distinguish trends in identification with the parties from trends in attitudes toward the parties. Future work also needs to determine the extent to which an individual's party identification is truly stable. Retrospective evaluations are known to produce changes in party identification. But if the existing measure of party identification confounds identity and attitude, perhaps retrospective evaluations only produce changes in attitudes toward the parties. The party social identity measure can help researchers

determine whether certain conditions cause respondents to alter the way they *report* their party identity without changing the underlying identification.

Social identity theory offers an important theoretical contribution to political science. Not only does a measure of party social identity empirically outperform the NES measure of party identification, it offers better theoretical explanations of political behavior. Political scientists must continue to explore the measure of party social identity and its applications. This requires scholars to continue asking survey questions about party social identities, particularly in longitudinal, nationally representative surveys like the ANES.

CHAPTER 2

MEASURING PARTY SOCIAL IDENTITY: AN APPLICATION OF MACHINE LEARNING

Party identification is one of the central concepts of political science. The vast majority of the larger political science literature deals with party identification, either investigating its origins in individual respondents or its effects on individual voting behavior. In *The American Voter* (1960), Campbell et al. pose party identification as a fundamental psychological aspect of a person's identity. While not explicitly, they argue that voters identify with their parties like they identify with their racial and religious groups. The view of party identification as a psychological attachment has received a great deal of scholarly attention, but the measurement of PID has remained constant throughout the literature.

The so-called "Michigan/NES measure" of party identification, PID, is ubiquitous but suffers from some important drawbacks. "At the root of the problem, the Michigan measure confounds the empirically and theoretically distinct psychological concepts of attitude and of group identity" (Greene, 2002, p. 171). It measures a broader concept of partisanship, i.e. how political someone is in general, rather than the more specific concept of a person's psychological attachment to a political party. The distinction between identification and evaluation is prominently highlited in *Partisan Hearts & Minds* (Green, Palmquist and Schickler, 2002). Recent work by Greene (2000; 2002; 2004) has demonstrated the need for

a new measure of party identification based in Social Identity Theory (SIT). Chapter One of this dissertation showed that this new measure of party social identity (PSI) can be used even in cases where the classic NES measure fails, such as in political primary elections. The growing literature on party social identity thus bears the clear recommendation that future scholars should update their surveys in order to measure PSI properly.

However, this offers little advice for scholars concerned with things like time trends in mass party identification, or those scholars who are simply interested in revisiting past work. Scholars cannot go back in time to improve their surveys, so the most they can do is address future surveys. Another solution might be to search past surveys for proxies for party social identity, but this would need to be done on a survey-by-survey basis and any such proxy would likely be inconsistent from survey to survey. Scholars might be interested in revisiting past work to gain a more thorough understanding of party social identity, but data limitations make this difficult.

In this chapter I argue that machine learning can be used to solve this empirical quandary. I evaluate a number of machine learning algorithms in their ability to predict a party social identity based on data from the 2016 Cooperative Congressional Election Study (CCES), which contains both the NES measure of PID and a set of questions that measure PSI. After choosing the best model at predicting PSI, I use that model to predict values of PSI in the 2016 American National Election Survey (ANES), which does not contain PSI survey questions. Finally, I replicate the logistic regression analysis found in Chapter One to assess the validity of the obtained predictions.

I show that these predicted values behave the same way in the ANES data as the observed values in the CCES data. The method presented in this chapter allows scholars to revisit any past data set and obtain predicted values of PSI, which in turn enables scholars to test hypotheses regarding party social identity. Because the predicted values are a valid measure of PSI, scholars can use this method to identify good candidates for future surveys and empirical tests: if the predicted measure of PSI results in some theoretically or empirically interesting finding, direct measurement in a future study should also result in some theoretically or empirically interesting finding.

2.1 Measuring Party Identification

Party identification is perhaps the foundational concept at the core of political science research. It can be found in practically every political science article, whether as a variable of interest or as a control variable. As a concept, party identification has received a great deal of scholarly attention. The classic view of party identification is that it is a long-term psychological attachment to a political party. It develops early in life, largely as the result of family and social influences, and is relatively stable. In other words party identification is exogenous (Campbell et al., 1960).

One of the first major challenges to this conception of party identification comes from Fiorina (1981). Fiorina argues that rather than being an exogenous reflection of one's social identity, party identification is instead endogenous to the particulars of the political environment at the time. Because individuals are utility maximizers, individuals will update

their party identification in accordance with the outcomes they desire. In other words, party identification is a "running tally" of one's evaluations of the two political parties.

I argue that the distinction between identification as a psychological attachment and identification as a running tally is less to do with the underlying theoretical concept of party identification and more to do with the empirical conflation of identity and attitude (Greene, 2002). That is, it may simply be the *measure* of party identification that operates like a running tally rather than the underlying *concept* of party identification. As Greene argues, "much of the controversy surrounding partisanship stems from our attempts to examine a complex, multifaceted psychological concept with a rather blunt instrument." (Greene, 2002, p. 171). The seemingly contrary findings found in the larger literature on party identification can be viewed as the result of relying on the NES measure as the measure of party identification. In other words, party ID sometimes looks like an identity and other times like an attitude because the NES measure confounds identity and attitude.

The American Voter (1960) articulated party identification as a social identity years before social identity theory was developed in social psychology. Though the authors do not explicitly argue this, they essentially state that people identify with political parties the way they identify with racial or religious groups. Developments within the field of social psychology (specifically, in social identity theory) have improved on the ability to measure various social identities (Greene, 2004). Despite these developments, the measure of party identification (which itself was originally conceived as a social identity) has largely failed to update.

Recent work by Greene (2000; 2002; 2004) shows that a measure of party identity based in social identity theory allows scholars to measure party identification directly in a way which does not confound identity and attitude. Party social identity is a key element in evaluations of both the in-party and the out-party (measured as feeling thermometers), ideological self-placement (individuals with stronger social identities tend to identify as more ideologically extreme), and various forms of partisan activity (including the frequency with which one votes for a member of the respondent's given political party) (Greene, 2004). These effects hold even controlling for party ID strength. Measuring party social identity not only improves the predictive power of models of political behavior but improves our theoretical understanding of how party identification affects political behavior.

The implication for future work is that scholars need to take more care with their measurement of party identification. The simple solution is to include questions on party social identity on every survey. This would allow scholars to measure party identity directly, without relying on the NES measure which conflates identity and attitude (Greene, 2002). Greene (2004) and Chapter One of this dissertation both demonstrate the need for a direct measure of party social identity.

Given the extensive use of party identification in the literature, scholars should also be concerned about whether and how previous work should be reinterpreted. Indeed, Chapter One shows that inclusion of a party social identity variable actually removes the statistical significance of party ID strength in explaining candidate preferences in primary elections, leading to different inferences. While I certainly do not expect that including a direct measure of party social identity will radically undermine *all* empirical analyses, revisiting previous

work should still be of great interest to scholars. The inferences scholars make about the effect of party identification are directly tied to the way scholars measure party identification. If that measure is less than ideal, the inferences based on that measure might be suspect.

2.2 Machine Learning

Party social identity should be used in a wide variety of theoretical settings and empirical analyses. Unfortunately, the measure requires a set of survey questions that only recently have been adopted in political science and still do not appear on major surveys like the American National Election Studies (ANES) or the Cooperative Congressional Election Studies (CCES). This poses a problem, as the inferences we've drawn in the past are based on a measure that Greene (2000; 2002; 2004) shows is inadequate. This amounts to a missing data problem; past surveys are "missing" a measure of party social identity.

The challenge is that we cannot simply add new survey questions to past surveys. Scholars are essentially stuck with the data contained in these past surveys. It is trivial to add more questions to future surveys, but the task is a bit more complicated when we must make do with existing data. Fortunately, machine learning can aid us in this task. We cannot go back in time and precisely measure party social identity, but we can learn from existing surveys how to predict party social identity. Scholars can use these predicted values to reevaluate past analyses with the original data. These predictions, while 'artificial' in a sense, will be close to what they would have been if we had observed them directly.

Machine learning is a process of using statistical techniques to "learn" from data. There exist numerous different machine learning models, but the general process is the same. Con-

sider some data set, D_1, that contains some predictor variables X_1 and an outcome variable Y_1. A machine learning algorithm can learn from the data and determine a rule by which X_1 can predict Y_1. Machine learning algorithms in general produce a rule that minimizes the prediction error, or the difference bewteen the observed Y_1 and the predicted Y_1. The power of machine learning lies not just in its ability to learn from the data, but in the ability to apply the learned rule to new data. That is, the machine learning model learned from D_1 can be applied to some new data set D_2, which contains new observations of the same predictor variables X_2 but which does not observe Y_2, in order to predict values of Y_2.

Many modern day services and websites make use of machine learning algorithms. Consider Netflix recommendations, for example. When a user watches some shows A, B, and C, Netflix will recommend a new show D to watch. Behind the scenes, a machine learning algorithm learns from what other people watch, determines that users who watch shows A, B, and C also typically watch and like D, and recommends D to the user. Email spam filters make use of machine learning algorithms as well. Those algorithms view millions of emails that are classified as either "spam" or "not spam" and learns features that separate "spam" from "not spam." The filter then views a new email, determines whether it looks more like other "spam" or "not spam" emails, and classifies it accordingly.

In practice, there is a large selection of machine learning models to choose from. In this chapter, I discuss and evaluate three basic categories of machine learning models. The first group consists of linear regression-based models. The regression model attempts to find a linear combination of the predictor variables that achieves the smallest prediction error. Some of these models, like ridge and lasso regression, use different error calculations to shy

away from complicated model outputs. Another class of machine learning models consists of non-linear regressions. These models move beyond attempting to find linear combinations of the predictor variables. With k-nearest neighbors, for example, the prediction for a given observation is a weighted average of the outcome values of the k nearest observations.

The last group of models to discuss are decision tree-based models. A regression tree takes a vector of predictor variables and maps it to a predicted value. The "tree" takes the form of a series of nodes, each of which splits into a number of branches. At each node, the value of one variable determines which branch is followed. This process continues, node after node and branch after branch, until a terminal node, a leaf, is reached. This leaf returns a prediction. Consider a real estate agent selling a home. The agent needs to predict a selling price for the home, so they consider all the features of the home: number of rooms and bathrooms, square footage, etc. The value of each feature determines which branch to follow, and eventually the agent arrives at a prediction for the selling price.

One downside of tree-based models is that they are susceptible to overfitting. That is, a tree with a number of nodes equal to the number of observations can perfectly predict the data, but might perform extremely poorly at predicting new data. To avoid overfitting, many ensemble methods exist that combine a series of trees and aggregate their predictions. Random forest, for example, fits a series of trees on random subsets of the data. The predictions of these trees are then aggregated to produce the final predictions. Boosted trees use the residuals of each tree in the next tree, and cubist builds each new tree as a weighted average of the previous trees.[1]

[1]See Kuhn and Johsnon (2013) for a more thorough description of each of these methods.

Machine learning has seen recent applications in political science. Carroll and Kenkel (2016) build a measure of relative power between states, the Dispute Outcome Expectations (DOE) score, using machine learning. They argue that the existing measure, the capability ratio, is as good as random guessing at predicting military dispute outcomes. They replicate 18 recent empirical studies to show that their measure performs better than the capability ratio in improving both in-sample and out-of-sample predictions (Carroll and Kenkel, 2016). Barrilleaux and Rainey (2014) use random forests to assess variable importance in predicting governors' opposition to Medicaid expansion under Obamacare. They show that those decisions were largely a function of political variables rather than economic conditions.

In Chapter One I argued that party social identity is better than party ID strength at explaining candidate preference in primary elections and that future scholars should include measures of party social identity in their surveys. In this chapter I argue that machine learning can be used to obtain predictions for party social identity even in surveys that lack questions about party social identity. I test a number of machine learning models against each other using 2016 CCES data and select one that offers good predictive power while being easy for future scholars to implement. I then use that model to predict party social identity in the 2016 ANES and show that these predictions, even though they are not directly obtained from respondents, are a valid proxy for direct measures.

2.3 Analysis

The goal of this chapter is to describe a process by which scholars can predict values of party social identity (PSI) despite not observing them directly. This process requires two

things. First, I need to learn how to predict PSI. The machine learning models I test will learn how to take the given predictor variables to accurately predict PSI. So I need a data set that contains both a direct measure of PSI and a large set of predictor variables. Second, I need to establish a metric by which to compare the different machine learning models I test. Ultimately, I want to find best machine learning model at predicting PSI in the CCES data.

2.3.1 Data

Data for this chapter come from two sources: the 2016 CCES and ANES surveys. Florida State University purchased space on the 2016 CCES to ask a number of extra survey questions on a subset of 1,000 observations from the larger CCES sample. This FSU module contains survey questions used to measure PSI. The CCES also contains a robust set of predictor variables, including some that are related to political behavior (party ID strength (PID), ideological strength, party agreement, and political interest) and some standard demographic variables (dummy variables for female, minority, and married status; frequency of church attendance; income; and education). It is something of a conscious decision on my part to use only the variables contained in regression analysis of Chapter One. In practice, any variable that is found in both the CCES and the ANES can be used to build these models. The intention of the present analysis is to show the process by which we can predict party social identity; future work can and should explore a larger set of covariates to improve even further the prediction of party social identity.

2.3.2 Learning

In order to know which machine learning model to select, I first need to establish a metric by which I can compare competing models. The outcome variable to be predicted, party social identity, is a 0-16 scale. As such, a common measure of predictive power is the root mean squared error (RMSE) of the model. If a machine learning model produces a small RMSE, its predictions of the outcome variable were close (on average) to the observed values of the outcome variable. As the goal of this process is predictive in nature, I am only concerned with predictive performance. As such, the model I end up choosing should produce a small RMSE.

I also want to select a model with good out-of-sample prediction properties. Machine learning models can sometimes be overfit, meaning they rely too heavily on the training data and perform poorly on new data. In other words, not only do I want a model that predicts CCES data well, I want a model that I am confident will produce accurate predictions in the ANES data. Cross-validation is a method for assessing how well a given machine learning model generalizes to new data. A model that performs well in cross-validation will be less susceptible to overfitting and therefore will perform better with new data. Ideally, I would split the data into three sets: a training set, a validation set, and a test set. I would use the training set to fit the models, the validation set to estimate the prediction error for each model, and the test set to assess the generalization error of the chosen model (Hastie, Tibshirani and Friedman, 2009, p. 222).

Unfortunately, this cross-validation process requires quite a lot of data. With only 1,000 observations in the CCES data, I have a relatively small data set so I need a slightly differ-

ent strategy. Hastie, Tibshirani, and Friedman (2009) describe k-fold cross-validation as a process to estimate generalized prediction error of the machine learning models. This generalized prediction error is an indication of how well the given machine learning model will perform with new data. A model with low generalized prediction error is an ideal choice for predicting party social identity in new data sets like the 2016 ANES.

K-fold validation is an iterative process with k iterations. In this analysis I use $k = 5$. I split the full CCES data set into 5 random, approximately equal subsets of data, called "folds." For each fold, I follow the same process: First, I set that fold as the test set and combine the remaining 4 folds as the training set. Next, I fit each machine learning model using the training set. Finally, I use the test set to calculate the prediction error for the model. The 5-fold cross-validation process then produces 5 estimates of the prediction error for each model. I average the 5 estimates for each model to produce a final cross-validation estimate of the prediction error for each model. This process allows me to avoid overfitting and ensure that the model I eventually choose will perform well with future data sets.

I test a total of 14 different candidate models. In choosing the candidate models I follow the advice of Wu et al. (2007), Kuhn and Johnson (2013), and Fernandez-Delgado et al. (2014). These 14 candidate models are a small selection from the total number of machine learning models out there but they nonetheless cover a wide range of bases and offer a robust set of options. To serve as a baseline, I set one model as a constant-only ordinary least squares model. This model essentially uses the mean value of party social identity to predict each respondent's value of party social identity. I set a second model as an ordinary least squares using just party ID strength as a predictor and a third model as an ordinary

least squares with the full set of predictors. The remaining candidate models use the full set of predictor variables. The fourteen candidate models are as follows:

- Linear Regression Models

    1. A constant-only ordinary least squares null model as a baseline for evaluating the other candidate models.
    2. An ordinary least squares model using just party ID strength (PID) as a predictor, to test whether simply using party ID strength serves as an adequate replacement for measuring party social identity.
    3. An ordinary least squares model using all predictors, which might be a naive but computationally simple method of predicting PSI.
    4. A partial least squares model, which attempts to find linear combinations of the predictor variables to achieve a smaller prediction error.
    5. A ridge regression model, a penalized model to help protect against over-fitting.
    6. A lasso regression model, a penalized model to help protect against over-fitting.
    7. An elastic net model, which combines the ridge and lasso penalties.

- Non-linear Regression Models

    8. A k-nearest neighbors model, where the outcome is a weighted average of the outcomes of the k nearest neighbors.
    9. A support vector machine, a variation on linear regression that weights large residuals differently than small residuals.
    10. An averaged neural net model.

- Regression Tree Models

    11. A Classification and Regression Tree (CART) model. Tree-based models require little pre-processing, are quick to implement, and produce interpretable results (Breiman et al., 1984; Hastie, Tibshirani and Friedman, 2009).
    12. A random forest model, an ensemble method in which many independent trees are made and then predictions are aggregated.
    13. Boosted trees, another ensemble tree method where the residuals of one tree are used in the next tree.
    14. Cubist, a similar model where each new tree is a weighted mixture of the previous trees.

Some of these models are chosen for their raw predictive power. Others are chosen because they are easy to implement. Again, the ideal model will possess good out-of-sample properties and achieve a low generalized RMSE. Several of these models rely on tuning parameters (e.g. the penalty rate in the ridge, lasso, and elastic net models, or the complexity parameter in CART). I use a separate k-fold cross-validation process to estimate appropriate values for these tuning parameters. I then take those values for the tuning parameters into another k-fold cross-validation process to estimate the RMSE for each model. This ensures that the final estimates for generalized prediction error are robust to tuning parameters.

For each candidate model I calculate its out-of-sample prediction error (RMSE), the standard deviation of its predictions, a measure of the computation time to complete each model, and a measure of the proportional reduction in error (PRE) of the candidate model relative to the baseline null model.[2] The RMSE value for each model is the average of the 5

[2]The specific values for computational time are more a reflection of the computer on which this analysis was performed rather than the data itself. These values should only be interpreted relative to each other.

Table 2.1: Out-of-sample Performance of the Candidate Models

	RMSE	SD	Time (s)	PRE
Null Model	4.4135	0.0965	0.5819	0.0000
K-nearest neighbors	3.4688	0.1590	0.6158	0.2140
Partial least squares	3.2969	0.1641	0.6576	0.2530
Ridge regression	3.2526	0.1584	0.6393	0.2630
Elastic Net	3.2526	0.1584	0.6549	0.2630
Party ID strength	3.2411	0.1350	0.6034	0.2656
Ordinary least squares	3.2407	0.1412	0.6161	0.2657
Lasso regression	3.2349	0.1374	0.6510	0.2671
Support vector machine	3.2123	0.1061	1.0716	0.2722
Neural network	3.1153	0.1007	1.3815	0.2941
Cubist	3.0859	0.1136	0.9166	0.3008
CART	3.0675	0.1151	0.6766	0.3050
Random forest	3.0436	0.1081	3.4485	0.3104
Boosted Trees	3.0418	0.1273	0.7177	0.3108

RMSE values obtained in each iteration of the k-fold cross-validation process. A PRE of zero indicates the model performs exactly as well as the null model, larger positive coefficients indicate the percentage improvement over the null model, and negative coefficients would indicate a decrease in predictive power compared to the null model.

2.3.3 Results

Cross-validation. Table 2.1 shows the results of the cross-validation process. As expected, the null model (which just uses the mean for PSI to predict each value of PSI) performs the worst. Using just party ID strength to predict party social identity, as theory expects those to be highly correlated, gives a 26.6 % increase in predictive power relative to the null model. However, half of the candidate models perform even better than using party ID strength alone. All four of the tree-based methods (random forest, classification

and regression tree, boosted trees, and cubist) outperform the null model and the party ID strength model. The boosted trees model performs the best in terms of raw RMSE performance, while CART gives a slight improvement over boosted trees for computational time and retains interpretability.

These results show the potential for improving our understanding of party identification. If the PID-only model performed the best, scholars could simply use party ID strength as a proxy for party social identity. This is essentially what the literature already does. Even though party social identity and the NES measure of party identification are theoretically distinct, a well-performing PID-only machine learning model would indicate that the two are empirically close enough to be a suitable solution. However, not only do several machine learning models outperform the PID-only model, several of the better models are quite flexible in their applications. Random forest, for example, is an excellent "off-the-shelf" model. It requires no data pre-processing, is robust against outliers, and can automatically identify nonlinearities and interactions in the data without requiring scholars to explicitly model them. This flexibility also makes machine learning in general a better approach than methods like errors-in-variables, as machine learning requires no modeling assumptions or identifiability conditions.

Before proceeding with the analysis, I should note a few things. First, it does appear that the PID strength variable offers a great deal of the predictive power of these models. Even the boosted trees model offers only a 6% improvement over the PID-only model. The predictor variables related to political behavior (party ID strength (PID), ideological strength, party agreement, and political interest) are also relatively stable, as opposed to evaluation-based

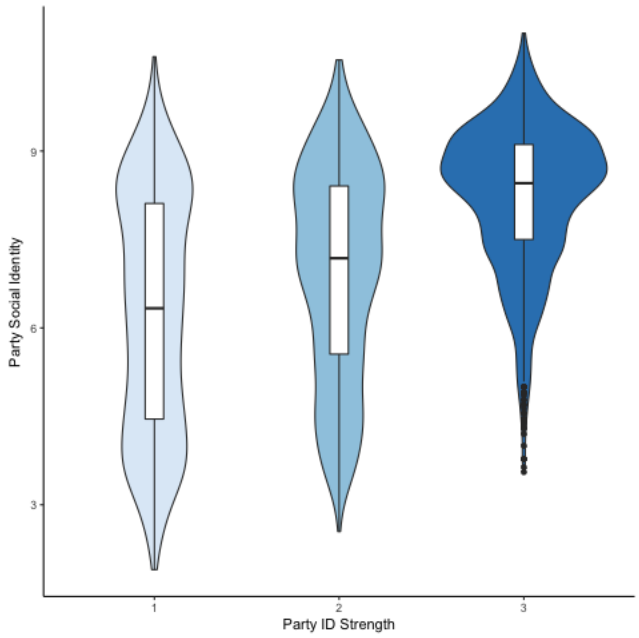

Figure 2.1: Party ID Strength and Party Social Identity, KNN Predictions

survey responses that might be more dynamic, and this may bias the final predictions to be more stable than they should be. Future work should indeed examine a larger set of predictor variables. A data set that contains perhaps a dozen (or more) politics-related variables can help reduce the reliance on the PID variable and obtain predictions that are less likely to be biased towards being stable.

I should also note that the ultimate goal is to optimize predictive performance. Readers should not assume that good predictive performance (measured by a low RMSE) is indicative of other properties. Some of the machine learning models have low RMSE but produce predicted variables with extremely low variation. For example, CART predicts only three values for the outcome variable. Consider KNN and Boosted Trees, the worst and best machine learning models. Figures 2.1 and 2.2 show violin plots between party ID strength and

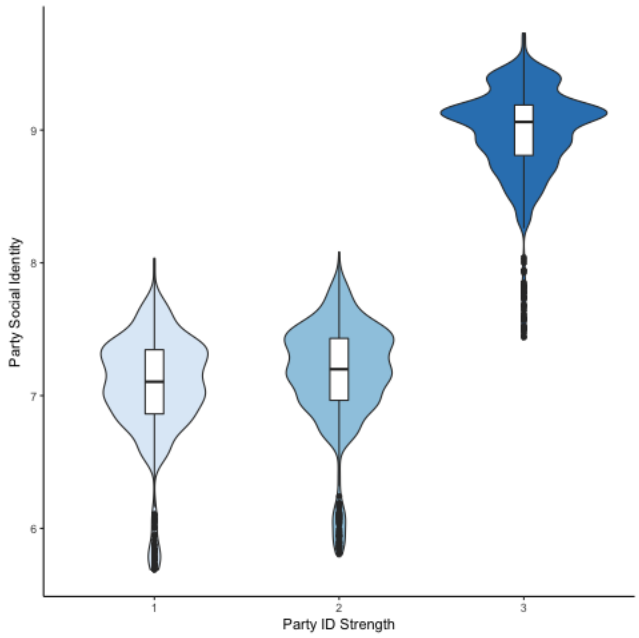

Figure 2.2: Party ID Strength and Party Social Identity, Boosted Trees Predictions

the PSI predictions from KNN and boosted trees, respectively.[3] KNN's predicted variable shows much more variation at each level of party ID strength than the boosted trees model, even though KNN was the worst machine learning model and boosted trees was the best. Ultimately we want to be certain that the predicted value of PSI is close to what it would have been if it had been directly observed, which means we want the machine learning model with the lowest RMSE regardless of other properties of the resulting predicted variable.

Most of these machine learning models tend to avoid predicting observations at the extreme ends of the scale. This is partially a function of the relatively small sample size and partially a function of the high correlation between party ID strength and PSI. Few observations in the CCES data had very high or very low values of PSI, so the machine

[3]These violin plots show the distribution of PSI at each level of party ID strength. A boxplot of PSI is also shown for each level of party ID strength, indicating the median and the interquartile range.

learning models avoid predicting very high or very low values. Scholars interested in applying and evaluating this method may benefit from training machine learning models on much larger data sets, both with more observations and with more predictor variables. With more data, machine learning models can better identify patterns and thus produce a wider range of predicted values. However, as the goal is predictive performance, concerns about other properties of the variable are secondary.

Chapter 1 Replication. For the rest of this analysis, I proceed with the boosted trees model. Boosted trees provides the best prediction performance for relatively little computational cost. While other analyses may prefer a simple model that produces easy-to-interpret results, the aim here is to maximize predictive power. As such, boosted trees gives us the best possible method for predicting PSI in the 2016 ANES data set. Having selected an appropriate machine learning model, I apply it to the ANES data in order to obtain predicted values of party social identity for those respondents. I use bootstrapping to obtain 1,000 predictions for each observation; I use as my final prediction the mean of these bootstrapped predictions.[4] Even though PSI was not directly observed in the ANES data, machine learning nonetheless produces predictions of what values we would have observed. The next step is to show that these predicted values are not only accurate (i.e. close to what they would have been if they had been measured directly), but valid. That is, this predicted measure of PSI in the ANES data should behave similarly to the observed measure of PSI

[4]The final predicted variable is highly correlated with party agreement (0.94761821), party (0.86042318) and ideological (0.33368668) strength, and political interest (0.22269358), as we would expect. It is also strongly correlated with education (0.12793660) frequency of church attendance(0.13933668).

Table 2.2: Candidate Preference in the 2016 Primary Elections ANES Data

	Establishment		Anti-Establishment	
	Model 1	Model 2	Model 3	Model 4
Partisan Social Identity		0.13 (0.10)		−0.14 (0.10)
Party ID Strength	**0.38** (0.07)	0.19 (0.16)	**−0.28** (0.07)	−0.08 (0.15)
Ideology Strength	**−0.45** (0.06)	**−0.46** (0.06)	**0.43** (0.06)	**0.43** (0.06)
Party Agreement	0.00 (0.00)	−0.00 (0.01)	−0.00 (0.00)	0.01 (0.01)
Female?	**0.41** (0.10)	**0.41** (0.10)	**−0.41** (0.10)	**−0.42** (0.10)
Minority?	**1.02** (0.12)	**1.02** (0.12)	**−0.91** (0.12)	**−0.90** (0.12)
Married?	0.05 (0.11)	0.05 (0.11)	−0.14 (0.11)	−0.14 (0.11)
Church Frequency	−0.05 (0.03)	−0.05 (0.03)	0.02 (0.03)	0.02 (0.03)
Political Interest	**0.14** (0.07)	0.14 (0.07)	−0.03 (0.07)	−0.02 (0.07)
Income	**0.02** (0.01)	**0.02** (0.01)	−0.01 (0.01)	−0.01 (0.01)
Education	**0.21** (0.05)	**0.21** (0.05)	**−0.19** (0.05)	**−0.19** (0.05)
Intercept	**−2.06** (0.26)	**−2.69** (0.55)	**1.36** (0.25)	**2.04** (0.54)
Log Likelihood	-1162.92	-1162.07	-1206.83	-1205.78
Num. obs.	1882	1882	1882	1882

Coefficients significant at the p = 0.05 level are bolded.

in the CCES data. To this end I replicate the logistic regression analysis from Chapter One using ANES data instead of CCES data.

To account for alternate explanations of candidate preference in the primaries I include in my analysis measures of party and ideological strength, created by folding the standard partisanship and ideology measures in half, where “Strong Republican” and “Strong Democrat” are both scored highly and ”somewhat conservative” and “somewhat liberal” are low scores. I also include a measure of party agreement which measures the extent to which a respondent agrees with their given party on a number of policy issues. I also include controls for gender (a binary variable indicating 1 for female respondents), marriage status (a binary variable indicating 1 for married respondents), race (a binary variable indicating 1

for minority respondents), frequency of church attendance, interest in politics, income, and education.

Table 2.2 presents four logistic regression models: two use vote choice for an establishment candidate as the dependent variable and two use vote choice for an anti-establishment candidate. For each dependent variable I estimate one model without the party social identity predicted variable and one including it in order to demonstrate the effects of the inclusion of that variable. The female and minority dummy variables are statistically significant in all four models, as is the education variable. All three are positively associated with support for establishment candidates and negatively associated with anti-establishment candidates. Political interest and income are both statistically significant predictors of support for establishment candidates (except for political interest in model 2, which is significant at the $p = 0.10$ level), but neither variable reaches statistical significance in the models for anti-establishment support.

Ideological strength is also statistically significant in all four models. Voters who are more extreme on the ideological scale are less likely to support establishment candidates and more likely to support anti-establishment candidates. Models 1 and 3 both show party ID strength to be a statistically significant predictor of support for establishment and anti-establishment candidates, respectively. Respondents with stronger party identifications are more likely to support establishment candidates and are less likely to support anti-establishment candidates. This evidence supports the inference that variation in party identification explains support for establishment and anti-establishment candidates.

However, Models 2 and 4 cast doubt on these inferences. After including the party social identity variable, party ID strength is no longer statistically significant in either model. Although party social identity itself is not a significant predictor of candidate preferences as it was in the results presented in Table 1.1, including the variable nonetheless changes the inferences that scholars can make based on the results. These results indicate the potential danger of relying on the NES measure of party identification when making inferences about the effect of party identification on political behavior.

2.4 Discussion

Recent work (Greene, 2000, 2002, 2004), including Chapter One of this dissertation, has shown that the measure of party identification can be improved by accounting for its nature as a social identity. The suggestion from this work is to include survey questions to measure party social identity in all future surveys. However, reevaluating past research is limited by the inability to retroactively ask survey questions. Given that the inferences scholars draw are based on the way party identification is measured, improving that measure may change the inferences scholars can draw. Scholars who wish to revisit previous analyses are left with few options: hope to find suitable proxies in past surveys or argue that new analyses can still offer insight on previous findings despite decades of changes in political and economic conditions.

In this chapter I offer machine learning as an alternative method by which scholars can directly revisit past empirical analyses. With a fairly straightforward machine learning process, researchers can obtain predicted values for party social identity within whichever

dataset they are interested in analyzing. I argue that those predicted values work as a valid alternative to direct measurement. Scholars should certainly update their surveys to include direct measurement of party social identity, but the process demonstrated in this chapter allows researchers to reexamine past analyses directly and reevaluate inferences.

The process and results presented in this chapter are not meant to be the final word in applying machine learning to predicting party social identity, but rather a first look at what the process would look like. Scholars should continue to build upon this process to develop better methods for predicting party social identity. Until a measure of party social identity makes its way into larger surveys like the full CCES or the ANES, the process described in this chapter is the next best method for testing hypotheses involving party social identity.

Future work can examine a more robust set of predictor variables and a larger data set than the CCES to attempt even further reduction in RMSE. In practice, there are dozens and perhaps even hundreds of variables in common between the CCES and the ANES that could improve the out-of-sample performance of these machine learning models. The results presented here are meant to be a first glance at how machine learning can be used to improve our ability to evaluate inferences. Future work should continue to use machine learning and explore its applications. For example, scholars can use this method to examine whether party identification is indeed declining over time at the aggregate level or if the observed decline is instead a function of changing attitudes towards the parties.

CHAPTER 3

EXAMINING CHANGES IN PARTY IDENTIFICATION: REVISITING RETROSPECTIVE VOTING

Recent research has identified problems with the way political scientists measure party identification. Though it was originally conceived as a psychological attachment to a political party (Campbell et al., 1960), the current standard of measuring party identification, the NES measure, "confounds the empirically and theoretically distinct psychological concepts of attitude and of group identity" (Greene, 2002, p. 171). Recent work by Greene (2000; 2002; 2004) has demonstrated the need for a new measure of party identification based in Social Identity Theory (SIT). Such a measure of party social identity provides a valid measure of party identification that does not suffer from the same drawbacks of the NES measure.

The suggestion for future scholars is thus to revise their surveys to include a direct measure of party social identity. But scholars should still be interested in revisiting past surveys. Chapter One of this dissertation shows that inclusion of a measure of party social identity can reduce or eliminate the statistical significance of the NES measure of party identification. At best, analyses using the NES measure may overestimate the effect of party identification. At worst, scholars draw invalid inferences about the effect of party identification.

Unfortunately, the task of adding survey questions to past surveys is indeed quite impossible. We are left analyzing the data that exists. Fortunately, machine learning can aid us in this process. Chapter Two of this dissertation shows a process by which scholars can use machine learning to construct predictions for the values of party social identity scholars would have observed if they had measured it directly. As the replication analysis in Chapter Two shows, the predicted values calculated with machine learning serve as a valid proxy for direct measurement. While this process is not ideal, it still offers a way for scholars to reexamine empirical work to more fully explore the empirical implications of an improved measure of party identification.

There are a number of potential applications of a predicted measure of party social identity. If the NES measure of party identification confounds attitude and identity, then observed changes in an individual's party identification could be attributed to changes in the individual's underlying psychological attachment to the party group or changes in the individual's attitudes toward the two parties. A proper measure of party social identity would help scholars explain changes in the NES measure of party identification.

In this chapter I use panel data from the 2000-2002-2004 ANES Panel Study to test whether changes in the NES measure of party identification are due to changes in attitudes or changes in group identity. The machine learning process described in Chapter Two allows me to obtain predictions of party social identity in the 2002 and 2004 waves of the Panel Study. I show that changes in the NES measure of party identification are a function of changes in attitudes, even controlling for changes in group identity, but changes in group identity are not a function of changes in attitudes. In other words, evidence suggests that

movement in the NES measure is a function of changes in group identity *and* attitudes. The findings further underscore the need for a better measure of party identification.

3.1 Changes in Party Identification

Party identification is perhaps *the* central concept in political science. The concept of party identification is found in theories across the greater political science literature. Empirically, party identification appears in the vast majority of political science analyses. Whether as a primary explanatory variable or as a control variable, the empirical effects of party identification are regularly estimated. It is important, then, that scholars properly measure party identification. Our inferences on the effect of party identification on a phenomenon of interest are only valid if the measure of party identification is itself valid. If our measure of party identification does not truly measure party identification, or if it measures party identification and other correlates, findings based on that variable may be suspect. It is thus important for scholars to revisit previous empirical work with this better measurement strategy in mind.

As with any variable, the validity of a party identification measure is a function of its conceptualization. The original conception of party identification described it as a psychological attachment to a political party and viewed it as a fundamental aspect of a person's identity. While Campbell et al. (1960) did not explicitly articulate party identification as a social identity, as social identity theory had not yet been developed, they clearly viewed it as akin to racial or religious identity. In this conception, a resondent's observed value of party identification is a reflection of the psychological attachment that respondent has to the

political party. The first major challenge to this conception of party identification instead posed it as an affective evaluation of the two parties. Rather than being a psychological aspect reflecting a person's attachment to a political party, party identification in this sense is a "running tally" of evaluations of the party's actions (1981). A respondent's observed value of party identification, then, is a product of evaluations of the parties.

Fiorina (1981) offers retrospective evaluations as a causal explanation for changes in party identification. Individual-level changes in party identification are a function of individual-level changes in evaluations of the current administration. This pattern exists in the aggregate as well. MacKuen et al. (1989) show evidence at the aggregate level that high levels of presidential approval produce aggregate movement toward the incumbent party. Thus, at both the individual and aggregate level there is evidence that changes in party identification are a function of changes in attitudes.

However, these findings should be taken in the context of the NES measure of party identification. Greene (2002) argues that the NES measure confounds party identity and party attitudes. Political scientists debate whether party identification is a group identity or a group attitude using an empirical measure that confounds group identity and group attitude. This poses an important problem for drawing inferences about the causes of changes in party identification. For example, recent polling data[1] and academic work (Twenge et al., 2016) suggests that identification with the two major parties is on decline while identification as an Independent is on the rise. Given how the NES measure conflates identity and attitude, it remains unclear whether these results are changes in the way people identify

[1] http://www.people-press.org/interactives/party-id-trend/

or if they are instead a reflection of changing attitudes toward the parties. It could be the case that people largely remain psychologically attached to the parties but hold more negative evaluations of the political parties, or that people are simply becoming less attached to the parties themselves. These situations are observationally equivalent but tell markedly different stories.

It is important, then, to reevaluate an individual-level model of changes in party identification. Panel data provides an excellent premise for such a test. Indeed, Fiorina's (1981) work makes use of panel data to examine causes of changes in an individual's party identification. Weinschenk (2010) uses the 2000-2002-2004 ANES Panel Study to replicate Fiorina's (1981) findings. Using individual-level panel data allows us to explain changes in an individual's party identification, rather than changes in aggregate levels of party identification.

In order to test whether changes in party identification are the result of changes in group identity or in group attitude, though, we need separate measures of group identity and group attitude. Greene's work (2000; 2002; 2004) shows that a measure of party social identity is a measure of group identity distinct from group attitude. Chapter One of this dissertation demonstrates that party social identity explains variation in candidate preferences in political primaries even while controlling for partisan strength, the folded version of the party identification variable. Unfortunately, it is difficult to add survey questions to surveys which were administered decades ago, and obtaining new panel data that contains the necessary variables can be difficult and costly.

Fortunately, machine learning can aid us in this task. Even if we do not directly measure party social identity in a suitable panel data set, I can use machine learning to predict the

values of party social identity we would have observed in the 2004 Panel Study. Chapter Two of this dissertation shows a process by which scholars can use machine learning to construct predictions for the values of party social identity scholars would have observed if they had measured it directly.

Machine learning has seen recent applications in political science. Carroll and Kenkel (2016) build a measure of relative power between states, the Dispute Outcome Expectations (DOE) score, using machine learning and show that their measure performs better than the capability ratio in improving both in-sample and out-of-sample predictions (Carroll and Kenkel, 2016). Barrilleaux and Rainey (2014) use random forests to assess variable importance in predicting governors' opposition to Medicaid expansion under Obamacare. They show that those decisions were largely a function of political variables rather than economic conditions. Chapter Two of this dissertation uses machine learning to replicate the findings of Chapter One in the larger 2016 ANES sample. As a whole, these works demonstrate the usefulness of machine learning in a variety of political science applications.

The general machine learning procedure is straightforward. First, I use the 2016 Cooperative Congressional Election Study (CCES) data set which contains both the NES measure of party identification and a set of questions that measure party social identity. I use a boosted trees machine learning model, based on the results of Chapter Two, to predict values of party social identity based on a set of predictor variables. I then use that boosted trees model to obtain predicted values for party social identity in the 2002 and 2004 panels of the 2000-2002-2004 American National Election Studies (ANES) Panel Study. While this is not the ideal solution—direct observation of party social identity would be ideal—this nonethe-

less provides an opportunity to reexamine the causes of changes in party identification with a measure of party social identity in hand.

3.2 Analysis

Data for this chapter come from two sources: the 2016 CCES and ANES surveys. Florida State University purchased space on the 2016 CCES to ask a number of extra survey questions on a subset of 1,000 observations from the larger CCES sample. This FSU module contains survey questions used to measure PSI. The CCES also contains a robust set of predictor variables, including some that are related to political behavior (party ID strength, ideological strength, party agreement, and political interest) and some standard demographic variables (dummy variables for female, minority, and married status; frequency of church attendance; income; and education). It is something of a conscious decision on my part to use only the variables contained in regression analysis of Chapter One. In practice, any variable that is found in both the CCES and the ANES can be used to build these models. The intention of the present analysis is to show the process by which we can predict party social identity; future work can and should explore a larger set of covariates to improve even further the prediction of party social identity.

The methodological approach I use in this paper follows Weinschenk (2010). Weinschenk offers several different models that show the effect of retrospective evaluations on the present value of party identification as a replication of Fiorina's (1981) original work on retrospective voting. However, I am less interested in examining what causes an individual's present value of party identification as I am in explaining what causes *changes* in one's value of party

identification. As such, I follow Weinschenk's "alternative model of party identification" (Weinschenk, 2010, p. 487).

The dependent variable is the change in the value of the respondent's party identification (PID). The PID variable is coded such that 0 = strong Democrat and 6 = strong Republican. The dependent variable is the respondent's 2004 PID value minus the respondent's 2002 PID value. A positive change means movement toward Republican identification, whereas a negative change means movement toward Democratic identification.

The 2002 and 2004 waves of the panel study asked a number of questions on retrospective evaluations (including a measure of presidential approval, an evaluation of one's personal finances, and an evaluation of the national economy) and attitudes about foreign policy (approval of George W. Bush's handling of the war on terror, an evaluation of whether or not the U.S. war in Afghanistan was worth the cost, and whether the war in Iraq was worth the cost.) To include these evaluations in my model I follow Weinschenk's (2010) coding scheme. If a respondent's evaluation on a given question becomes more positive from 2002 to 2004, they are coded as +1 (regardless of the magnitude of the change). If a respondent's evaluation becomes more negative, they are coded as −1. Respondents whose evaluations stay the same are coded as 0.

Both waves of the panel study also include feeling thermometers towards Democrats and Republicans. I include two measures of changes in these variables, where the respondent's change in the evaluation is their 2004 thermometer value minus their 2002 value. Finally, to control for changes in the underlying psychological identification with the party groups, I include a measure of the change in each respondent's party social identity. If the respondent

exhibits a stronger attachment to the Republican Party, the change in party social identity will be positive; if they exhibit stronger attachment to the Democratic Party, their value for this variable will be negative.

The expectation is that a positive value for any of these change variables, with the exception of the feeling thermometer of Democrats, should lead to changes in the party identification variable toward the Republican Party. I estimate two models with change in party identification as the dependent variable: one that does not include the change in social identity variable and one that does. The goal is to show whether changes in the party identification measure are a function of changes in the underlying group identity, changes in evaluations and attitudes, or both.

I also estimate one model with change in social identity as the dependent variable. While Fiorina (1981) and others (Weisberg, 1980) do indeed acknowledge the long-term stable component of the party identification measure, the lack of a party social identity measure means scholars have not yet been able to explicitly examine changes in that component (Weinschenk, 2010). By examining changes in party social identity I can determine whether changing evaluations are causing changes in the individual psychological attachments to the parties. Such an analysis would not be possible without a direct measure of party social identity. I use the same change-in-evaluation variables as independent variables and I also include a measure of

Table 3.1 presents the three OLS regression models. In each of the three models, the dependent variable is the difference between the respondent's 2004 value of PID or PSI and the respondent's 2002 value. Positive values of the dependent variable mean movement in the

Table 3.1: Change in PID/PSI as a Function of Change in Evaluations

	Model 1 ΔPID	Model 2 ΔPID	Model 3 ΔPSI
Bush Job Approval	**0.25** (0.08)	**0.11** (0.05)	−0.04 (0.18)
Personal Finances	−0.09 (0.06)	−0.03 (0.04)	−0.00 (0.12)
National Economy	0.03 (0.06)	0.06 (0.04)	−0.21 (0.13)
Bush Terrorism Evaluation	0.05 (0.07)	0.03 (0.04)	−0.05 (0.15)
Afghanistan Evaluation	0.16 (0.09)	0.03 (0.06)	0.12 (0.19)
International Reputation	−0.02 (0.09)	−0.05 (0.06)	0.21 (0.20)
Evaluation of Democrats	**−0.01** (0.00)	−0.00 (0.00)	−0.01 (0.01)
Evaluation of Republicans	**0.01** (0.00)	**0.00** (0.00)	−0.00 (0.01)
Party Social Identity		**0.23** (0.01)	
Party identification			**2.67** (0.08)
Intercept	0.11 (0.09)	−0.05 (0.06)	**0.40** (0.20)
R^2	0.08	0.64	0.63
Num. obs.	840	840	840
RMSE	1.14	0.72	2.47

Coefficients significant at the p = 0.05 level are bolded.

Republican direction. Each of the independent variables is coded such that a more positive evaluation in 2004 relative to 2002 means a positive value of the independent variable. With the exception of the Democratic thermometer variable, each beta coefficient should be positive.

Model 1 shows that an increase in evaluation of Bush's job performance is a statistically significant predictor of changes in the PID variable. Both of the feeling thermometer variables are statistically significant, though their effect sizes are relatively small. The Afghanistan evaluation variable is significant at the p = 0.10 level. The inference scholars may draw from these results is that changes in these evaluations produce changes in party identification. However, it remains unclear whether and to what extent changes in PID are changes in the

psychological attachment to the parties or if the observed movement in the PID variable is simply a function of changes in the attitude component of the PID variable.

I explicitly test this by including change in party social identity as a separate independent variable. By accounting for variation in the identity component of PID, any remaining effects would provide evidence of movement in the attitude component of PID. Indeed, Model 2 shows that changes in evaluations of Bush' job performance are still a statistically significant predictor of changes in the PID variable. However, the effect size is less than half of what it was in Model 1. This suggests that the PID variable measures more than just party identification.

Model 3 uses change in party social identity as the dependent variable. If PSI is meant to be the long-term stable component of the PID variable, it should be relatively immune to changes in evaluations. Model 3 confirms this suspicion: none of the independent variables are statistically significant predictors of changes in PSI. The change in party identification variable is statistically significant, as we would expect given the machine learning process by which I obtained the PSI predictions. While an ideal analysis would directly measure party social identity, Chapter Two shows that the machine learning predicted values of PSI behave as we would expect directly-observed values to behave.

3.3 Discussion

Party identification is perhaps the central concept in political science. A large part of the political science literature involves exploring either the causes of party identification or the causes of *changes* in party identification. This work is complicated by recent evidence

that the standard measure of party identification, the NES measure, confounds group identity with group attitude. Without a proper measure of group identity distinct from group attitude, it becomes difficult to attribute movement in the NES party identification measure, PID, to the attitude component or the identity component. The lack of a group identity measure also means scholars cannot empirically examine changes in group identity. Scholars interested in examining causes of changes in group identities should directly measure party social identity, PSI, as their dependent variable. Causal explanations of group identity require a measure of identity separate from group attitude.

Using machine learning and the 2016 CCES, I obtain predicted values of party social identity in the 2000-2002-2004 ANES Panel Study. Panel data allows for a direct examination of changes in party identification and a measure of party social identity serves as a direct measure of group identity. I show that retrospective evaluations are associated with changes in the PID variable, even when controlling for changes in PSI, but that retrospective evaluations are not associated with changes in the PSI variable. In other words, movement in the PID variable indicates changing attitudes *and* changing identifications.

These results largely comport with Greene's assertion that the PID variable confounds group identity and group attitude (Greene, 2002). PID is less a direct measure of identification and more a summary of how political a respondent thinks they are. The results presented in this chapter should not be taken as a refutation of Fiorina's theoretical contributions, but rather as an illustration of the difference between PID and PSI as measures of party identification. PID and PSI are highly correlated, to be sure, and PID may suffice

as a general control variable. But insofar as researchers are interested in testing hypotheses about party identification, they should take care with their choice of measure.

Future surveys must continue to measure party social identity and scholars should continue to evaluate the performance of the PSI measure, particularly in relation to the PID measure. What this chapter offers is empirical evidence that the two measures behave differently. If PID and PSI are both meant to measure party identification, scholars should examine why the two measures behave differently. Furthermore, a more complete theoretical discussion of the two measures is necessary. What does PID measure, if not just party identification? Clearly PID has an attitudinal component, but what other considerations are on a respondent's mind when they answer the PID survey questions? Further research is needed to fully explore these questions.

CHAPTER 4

CONCLUSION

Party identification is a central concept in political science. A large majority of empirical analyses include party identification as an explanatory variable or a control variable and many political science theories at least acknowledge, if not attempt to explain, the influence of party identification. Because of the central role party identification plays in political science, proper measurement of party identification is a prime concern. Recent evidence in the last decade has raised concerns with the standard measure of party identification, the NES measure. This measure, PID, confounds group identity and group attitude. That is, a respondent who strongly psychologically identifies with one party over the other is observationally equivalent to a respondent who has more positive attitudes toward one party over the other. Given that many hypotheses in political science deal with specifically with party identification, a measure of party identification that cannot distinguish group identity from group attitude is inadequate.

Early conceptions of party identification clearly articulated it as a fundamental psychological aspect of a person's identity and defined it as a psychological attachment to the political party. Though the field of social identity theory had not yet been developed, scholars at the time argued that voters identify with their parties like they identify with their racial or religious gropus, with parties being social groups just like racial or religious groups. Despite advancements in the field of social identity theory, political scientists have largely relied on

the same measure of party identification described in 1960. The PID measure served its purpose at the time, but modern political science requires an up-to-date measure.

Political scientists often use "partisanship" and "party identification" interchangeably; a "partisan" is one who identifies with a political party. I argue, based on the evidence presented in this dissertation and other recent work, that these terms should refer to different concepts. Party identification refers specifically to a person's psychological attachment to a political party. Partisanship, on the other hand, should refer to a more generalized concept of how biased a respondent is in favor of one party or another. A respondent can be considered partisan simply because they consistently vote for one party over the other, regardless of their underlying identity. Party identification should be treated as a contributor to partisanship, not as a synonym for partisanship.

The PID measure serves not as a measure of party identification per se but of partisanship. Given how the standard NES measure confounds group identity and group attitude, PID gives a general indication of a respondent's bias towards one party or another. If the political scientist wants a single general control variable that summarizes a given respondent's predilection towards one party, PID is a well-established measure. A veritable mountain of empirical evidence suggests that PID predicts a wide variety of political behaviors. But scholars interested in party identification per se need a better measure. In order to test hypotheses regarding party identification, researchers must have a direct measure of party identification that is not confounded by related but distinct concepts.

Party social identity, PSI, is a measure of party identification that is grounded in social identity theory. Evidence suggests that PSI does not confound identity and attitude the

way PID does, indicating that PSI is a superior measure of party identification. Scholars interested in testing hypotheses regarding party identification should begin measuring party social identity in their surveys. Even though the original scale on which the party social identity measure is based consists of ten survey questions, evidence suggests that a two-item subset serves as a reliable, valid measure. Scholars need only add two survey questions in order to measure party social identity. This is a simple remedy to an important problem.

This dissertation offers both empirical and theoretical contributions to the discussion on party identification. Chapters One and Three both demonstrate that PID and PSI behave differently as measures of party identification. Not only does PSI explain candidate preferences in political primary elections, it outperforms PID's explanatory power. Furthermore, retrospective evaluations are shown to be associated with changes in PID, even controlling for PSI, but not with changes in PSI. These results indicate the empirical necessity for regular measurement of PSI, including in panel surveys.

The procedure described in this dissertation allows scholars to use machine learning as a temporary solution, generating predictions for PSI in a wide variety of surveys. Future researchers should examine a larger data set with a more robust set of predictors and continue to explore avenues for more precise predictions. The predictions produced by machine learning can be used to test hypotheses and identify opportunities for future research projects.

The theoretical implications of this dissertation are also evident. Scholars need to be clear when they discuss party identification, particularly when developing theories specifically pertaining to party identification as either a dependent or independent variable. Political primary elections illustrate the need for conceptual clarity and proper measurement. If

voters are choosing between members of different parties, a general measure of partisanship can suffice. But if voters are selecting between two (or more) members of the same political party, as is the case in primary elections, partisanship is less ideal than a pure measure of party identification. Furthermore, one of the long-standing debates in the political science literature regards changes in party identification. This debate, I argue, is largely a reflection of the empirical measure of party identification, PID, rather than the concept of party identification. Proper measurement of party identification leads to different causal inferences, as shown in Chapter Three of this dissertation.

Ultimately, scholars need to measure party social identity, especially on large, repeated, nationally-representative surveys like the Cooperative Congressional Election Study and the American National Election Study. With only two survey questions, scholars can obtain a valid, reliable measure of party social identity. This measure, while relatively easy to implement, provides enormous theoretical and empirical benefits.

APPENDIX A

IRB APPLICATION AND APPROVAL

Human Subjects Application For Full IRB and Expedited Exempt Review

1. Project Title and Identification

1.1 Project Title

2016 Cooperative Congressional Election Study, FSU Team Module

Project is: Collaborative Faculty Research

1.2 Principal Investigator (PI)

Name(Last name, First name MI): Gomez, Brad T.	Highest Earned Degr Doctorate
Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human subjects or human subjects records: FSU Training Module	Occupational Position Faculty

1.3 Co-Investigators/Research Staff

Name(Last name, First name MI): Ansolabehere, Stephen D.; Co-Investigator	Highest Earned Degr Doctorate
Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human subjects or human subjects records:	Occupational Positior Faculty
Name(Last name, First name MI): Ahler, Douglass J.; Co-Investigator	Highest Earned Degr Doctorate

Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human subjects or human subjects records: CITI	Occupational Positior Faculty
Name(Last name, First name MI): Kim, Minjung ; Co-Investigator	Highest Earned Degr Master's Degree
Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human subjects or human subjects records: CITI	Occupational Positior Student
Name(Last name, First name MI): Langley, Dennis ; Co-Investigator	Highest Earned Degr Master's Degree
Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human subjects or human subjects records: CITI	Occupational Positior Student
Name(Last name, First name MI): Macdonald, David ; Co-Investigator	Highest Earned Degr Master's Degree
Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human	Occupational Positior

subjects or human subjects records: CITI	Student

1.4 Faculty Advisor/Department Chair/Dean Information

Name(Last name, First name MI): Barrilleaux, Charles J.; Chair	Highest Earned Degre
Mailing Address:	
University Department: POLITICAL SCIENCE	
The training and education completed in the protection of human subjects or human subjects records:	Occupational Position

1.5 Does this project/study involve collaboration with any sites and/or personnel outsi

The Florida State University

Office of the Vice President For Research

Human Subjects Committee

APPROVAL MEMORANDUM

Date: 10/5/2016

To: Brad Gomez

Address:

Dept.: POLITICAL SCIENCE

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research

2016 Cooperative Congressional Election Study, FSU Team Module

The application that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and one member of the Human Subjects Committee. Your project is determined to be Expedited per 45 CFR § 46.110(7) and has been approved by an expedited review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to

weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 10/3/2017 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the Chair of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is FWA00000168/IRB number IRB00000446.

Cc: Charles Barrilleaux, Chair

HSC No. 2016.19100

BIBLIOGRAPHY

Abramowitz, Alan I. 1989. "Viability, Electability, and Candidate Choice in a Presidential Primary Election: A Test of Competing Models." *The Journal of Politics* 51(4):977–992.

Abramson, Paul R., John H. Aldrich, Phil Paolino and David W. Rohde. 1992. ""Sophisticated" Voting in the 1988 Presidential Primaries." *American Political Science Review* 86(1):55–69.

Aldrich, John H. and R. Michael Alvarez. 1994. "Issues and the Presidential Primary Voter." *Political Behavior* 16(3):289–317.

Barrilleaux, Charles and Carlisle Rainey. 2014. "The Politics of Need: Examining Governors' Decisions to Oppose the Obamacare Medicaid Expansion." *State Politics and Policy Quarterly* 14:437–460.

Bartels, Larry M. 1985. "Expectations and Preferences in Presidential Nominating Campaigns." *The American Political Science Review* 79(3):804–815.

Bartels, Larry M. 2000. "Partisanship and Voting Behavior, 1952-1996." *American Journal of Political Science* 44(1):35–50.

Bartels, Larry M. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.

Basinger, Scott J. and Howard Lavine. 2005. "Ambivalence, Information, and Electoral Choice." *The American Political Science Review* 99(2):169–184.

Beck, Paul Allen, Russell J. Dalton, Steven Greene and Robert Huckfeldt. 2002. "The Social Calculus of Voting: Interpersonal, Media, and Organizational Influences on Presidential Choices." *The American Political Science Review* 96(1):57–73.

Bloom, Pazit Bennun, Gizem Arikan and Marie Courtemanche. 2015. "Religious Social Identity, Religious Belief, and Anti-Immigration Sentiment." *American Political Science Review* 109(2):203–221.

Breiman, Leo, Jerome Friedman, Charles J. Stone and R.A. Olshen. 1984. *Classification and Regression Trees*. Chapman and Hall, New York.

Campbell, Angus, Philip E. Converse, Warren E. Miller and Donald E. Stokes. 1960. *The American Voter*. University of Chicago Press.

Carroll, Robert J. and Brenton Kenkel. 2016. "Prediction, Proxies, and Power." Found at `http://doe-scores.com/doe.pdf`.

Conover, Pamela J. and Stanley Feldman. 1981. "The Origin and Meaning of Liberal/Conservative Self-Identifications." *American Journal of Political Science* 25(4):617–645.

Dennis, Jack. 1988. "Political Independence in America, Part I: On Being an Independent Partisan Supporter." *British Journal of Political Science* 18(1):77–109.

Devine, Christopher J. 2015. "Ideological Social Identity: Psychological Attachment to Ideological In-Groups as a Political Phenomenon and a Behavioral Influence." *Political Behavior* 37:509–535.

Fernandez-Delgado, Manuel, Eva Cernadas, Senen Barro and Dinani Amorim. 2014. "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" *Journal of Machine Learning Research* 15:3133–3181.

Finkel, Steven E. and Karl-Dieter Opp. 1991. "Party Identification and Participation in Collective Political Action." *The Journal of Politics* 53(2):339–371.

Fiorina, Morris. 1981. *Retrospective Voting in American Elections*. Yale University Press.

Gopoian, J. David. 1982. "Issue Preferences and Candidate Choice in Presidential Primaries." *American Journal of Political Science* 26(3):523–546.

Gorenendyk, Eric. 2012. "Justifying Party Identification: A Case of Identifying with the "Lesser of Two Evils"." *Political Behavior* 34:453–475.

Green, Donald, Bradley Palmquist and Eric Schickler. 2002. *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. Yale University Press.

Greene, Steven. 2000. "The Psychological Sources of Partisan-Leaning Independence." *American Politics Quarterly* 28(4):511–537.

Greene, Steven. 2002. "The Social-Psychological Measurement of Partisanship." *Political Behavior* 24(3):171–197.

Greene, Steven. 2004. "Social Identity Theory and Party Identification." *Social Science Quarterly* 85(1):136–153.

Greene, Steven. 2005. "The Structure of Partisan Attitudes: Reexamining Partisan Dimensionality and Ambivalence." *Political Psychology* 26(5):809–822.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Huddy, Leonie. 2001. "From Social to Political Identity: A Critical Examination of Social Identity Theory." *Political Psychology* 22(1):127–156.

Huddy, Leonie, Lilliana Mason and Lene Aarøe. 2015. "Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity." *American Political Science Review* 109(1):1–17.

Huddy, Leonie and Nadia Khatib. 2007. "American Patriotism, National Identity, and Political Involvement." *American Journal of Political Science* 51(1):63–77.

Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. "Affect, not Ideology: A Social Identity Perspective on Polarization." *Public Opinion Quarterly* 76(3):405–431.

Klar, Samara. 2014*a*. "Identity and Engagement among Political Independents in America." *Political Psychology* 35(4):577–591.

Klar, Samara. 2014*b*. "Partisanship in a Social Setting." *American Journal of Political Science* 48(3):687–704.

Klar, Samara and Yanna Krupnikov. 2016. *Independent Politics: How American Disdain for Parties Leads to Political Inaction*. Cambridge University Press.

Kuhn, Max and Kjell Johnson. 2013. *Applied Predictive Modeling*. Springer.

Loren Collingwood, Matt A. Barreto and Todd Donovan. 2012. "Early Primaries, Viability, and Changing Preferences for Presidential Candidates." *Presidential Studies Quarterly* 42(2):231–255.

MacKuen, Michael B., Robert S. Erikson and James A. Stimson. 1989. *Macropartisanship*. Vol. 42.

Mael, Fred A. and Louis E. Tetrick. 1992. "Identifying Organizational Identification." *Educational and Psychological Measurement* 52:813–824.

Marshall, Thomas R. 1984. "Issues, Personalities, and Presidential Primary Voters." *Social Science Quarterly* 65(3):750–760.

Miller, Arthur H. and Martin P. Wattenberg. 1983. "Measuring Party Identification: Independent or No Partisan Preference?" *American Journal of Political Science* 27(1):106–121.

Miller, Warren E. 1991. "Party Identification, Realignment, and Party Voting: Back to the Basics." *American Political Science Review* 85(2):557 – 568.

Monardi, Fred M. 1994. "Primary Voters as Retrospective Voters." *American Politics Quarterly* 22(1):88–103.

Mondak, Jeffery J., Matthew V. Hibbing, Damarys Canache, Mitchell A. Seligson and Mary R. Anderson. 2010. "Personality and Civic Engagement: An Integrative Framework for the Study of Trait Effects on Political Behavior." *American Political Science Review* pp. 1–26.

Mullainathan, Sendhil and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31(2):87–106.

Norrander, Barbara. 1986. "Correlates of Vote Choice in the 1980 Presidential Primaries." *The Journal of Politics* 48(1):156–166.

Petrocik, John R. 1974. "An Analysis of Intransitivities in the Index of Party Identification." *Political Methodology* 1(3):31–47.

Stimson, James A., Michael B. MacKuen and Robert S. Erikson. 2002. *The Macro Polity.* Cambridge University Press.

Theodoridis, Alexander George. 2013. "Implicit Political Identity." *PS: Political Science & Politics* 46(3):545–549.

Thornton, Judd R. 2011. "Ambivalent or Indifferent? Examining the Validity of an Objective Measure of Partisan Ambivalence." *Political Psychology.* 32(5):863–884.

Thornton, Judd R. 2014. "Getting Lost on the Way to the Party: Ambivalence, Indifference, and Defection with Evidence from Two Presidential Elections." *Social Science Quarterly* 95(1):184–201.

Turner, John C. and Penelope J. Oakes. 1986. "The Significance of the Social Identity Concept for Social Psychology with Reference to Individualism, Interactionism, and Social Influence." *British Journal of Social Psychology* 25:237–252.

Twenge, Jean M., Nathan Honeycutt, Radmila Prislin and Ryne A. Sherman. 2016. "More Polarized but More Independent: Political Party Identification and Ideological Self-Categorization Among U.S. Adults, College Students, and Late Adolescents, 1970-2015." *Personality and Social Psychology Bulletin* 42(10):1364–1383.

Wattier, Mark J. 1983. "The Simple Act of Voting in 1980 Democratic Presidential Primaries." *American Politics Quarterly* 11(3):267–291.

Weinschenk, Aaron C. 2010. "Revisiting the Political Theory of Party Identification." *Political Behavior* 32:473–494.

Weisberg, Herbert F. 1980. "A Multidimensional Conceptualization of Party Identification." *Political Behavior* 2(1):33–60.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg. 2007. "Top 10 algorithms in data mining." *Knowledge and Information Systems* 14:1–37.

BIOGRAPHICAL SKETCH

I received my Bachelor's Degree in Political Science at Louisiana State University. I came to the FSU Department of Political Science in 2012 in pursuit of a Master's Degree and a Doctor of Philosophy in Political Science. I received my Master's Degree in 2014. In 2015, I began working toward a Master's Degree in Statistics, which I finished in 2017.

My substantive research interests primarily focus on political behavior and political psychology. I am largely interested in the vote decisions people make and the attitudes people hold. I am also interested in statistical methods, including machine learning applications in political science.