ESSAYS IN ELECTORAL BEHAVIOR AND BAYESIAN DATA ANALYSIS

Thesis by

Gabriel Katz

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2010

(Defended May 11, 2010)

To my parents.

ACKNOWLEDGEMENTS

I would like to begin by thanking Dr. R. Michael Alvarez for his guidance, encouragement and, above all, his patience and understanding during the five years I spent at Caltech. He has been a wonderful teacher and mentor. I would also like to thank Dr. Jonathan N. Katz for introducing me to the world of Bayesian econometrics and sparking my interest in MCMC methods, and Dr. Robert Sherman for his thoughtful advice and, especially, for pointing to me early on the importance of computational methods in empirical work. My research has also benefitted from various comments and suggestions from Dr. Matias Iaryczower and Dr. Jean-Laurent Rosenthal during the course of my studies.

Many thanks go to the faculty and staff of the Division of Humanities and Social Sciences of the California Institute of Technology for providing an excellent environment for graduate studies. In particular, I would like to acknowledge Laurel Auchampaugh for her constant good humor and invaluable assistance.

Finally, and most importantly, I would like to thank my wonderful parents, Regina and Pedro, for their constant source of support and wisdom and for allowing me to be as ambitious as I wanted throughout the years. I also thank Aline for being a true friend, besides my sister. And to Laura, whose quiet patience and unwavering support have been fundamental in allowing me to carry on with my work at difficult times, my everlasting love.

ABSTRACT

The resurgence of Bayesian statistics in political research and, in particular, the rising popularity of Markov chain Monte Carlo (MCMC) methods, has unlocked estimation problems long thought to be considered impossible or intractable. Besides opening new terrain to political methodologists, these developments have allowed scholars to explore new problems or to revisit longstanding puzzles. This dissertation takes advantage of the generality and power of the techniques comprising MCMC methods to address novel substantive and methodological questions about abstention, voter choice and turnout misreporting, areas where substantive controversies remain despite the rich story of academic studies on electoral behavior and the considerable attention that has been paid to them.

The second chapter of the dissertation develops a statistical model to jointly analyze invalid voting and electoral absenteeism, two important sources of abstention in compulsory voting systems that had so far not been simultaneously examined. I illustrate the application of the model using data from Brazilian legislative elections between 1945 and 2006, underscoring relevant differences in the determinants of both forms of non-voting. The third chapter presents a study of voter choice in Chile's 2005 presidential elections, examining substitution patterns in voters' preferences over the competing candidates and highlighting the influence of candidates' entry and exit strategies on the election results, an aspect that has received virtually no attention in previous analyses of Chilean electoral politics. Finally, the fourth chapter develops a model to correct for misclassified binary responses using information from auxiliary data sources, and applies it

to the analysis of voter turnout in the U.S. While the main contribution of the chapter is methodological, the empirical application has clear implications for researchers interested in the influence of race on voting behavior in America.

TABLE OF CONTENTS

Acknowledgements iv

Abstract v

Table of Contents vii

List of Figures x

List of Tables xii

Chapter 1: Introduction 1

Chapter 2: A Statistical Model of Abstention under Compulsory Voting 8

- 2.1 Introduction 8
- 2.2 Alternative explanations of invalid voting and absenteeism 11
- 2.3 A statistical model of abstention under compulsory voting 14
- 2.4 Analyzing invalid voting and electoral absenteeism in Brazil's lower house elections 27
  - 2.4.1 Data and methodology 27
  - 2.4.2 Results of the compositional-hierarchical model 35
  - 2.4.3 Comparison with alternative modeling approaches 47
- 2.5 Concluding remarks 54
- 2.A Appendix: Algorithm implemented to compute causal effects 56
- 2.B Appendix: Alternative strategies to modeling invalid voting and electoral absenteeism 58

Chapter 3: Structural Cleavages, Electoral Competition and Partisan Divide: A Bayesian Multinomial Probit Analysis of Chile's 2005 Election 60

3.1 Introduction 60

3.2 A first look at Chile's 2005 presidential election 66

3.3 A multi-candidate model of vote choice for the 2005 election 72

3.3.1 Data and research design 72

3.3.2 Model specification and empirical strategy 75

3.4 Empirical results 80

3.4.1 Multinomial probit estimates 80

3.4.2 The effect of individual characteristics on vote choice 86

3.4.3 The role of the electoral context: candidate competition and voter calculus 93

3.5 Final remarks 100

Chapter 4: Correcting for Survey Misreports using Auxiliary Information with an Application to Estimating Turnout 102

4.1 Introduction 102

4.2 Correcting for misreporting in binary choice models 109

4.2.1 Defining the problem 109

4.2.2 A Bayesian model to correct for misreporting using auxiliary data 114

4.2.3 Extending the model to account for missing data 122

4.3 A Monte Carlo experiment 127

4.3.1 Comparison of alternative approaches to dealing with misreporting 127

4.3.2 Assessing robustness to the specification of the misreport model 139

4.3.2 Accounting for missing response and covariate values 145

4.4 An empirical application: correcting for misreporting in the

analysis of voter turnout 148
4.4.1 Turnout misreporting in the 1978 – 1990 ANES 150
4.4.2 Correcting for misreporting using a validation sub-sample 161
4.4.3 Correcting for misreporting using an external validation design 169
4.4.4 Accounting for item and unit non-response 173
4.5 Concluding remarks 177
4.A Appendix: Additional results from the Monte Carlo experiment in Section 4.3 180
4.B Appendix: Sample description and variable coding 187

Chapter 5: Concluding remarks 190

Bibliography 192

LIST OF FIGURES

Number *Page*

Chapter 2

2.1Proportion of invalid votes, absenteeism and valid votes in Brazil's lower house elections, 1945 - 2006 15

2.2 Invalid voting and absenteeism by state, as a proportion of the electorate – Lower house elections, 1945 - 2006 29

2.3 Posterior means of the Level-1 residuals from Model 1-a................... 38

2.4 Posterior densities of υ and $w_{i,t}$ under Model 1-b 40

2.5 Posterior means of standardized election residuals and marginal posterior means of weight parameters 41

2.6 Estimated marginal effects of the predictors across models 48

2.7 Actual and expected proportions of invalid voting and electoral absenteeism under alternative modeling strategies 53

Chapter 3

3.1 Distribution of votes among *Concertación* identifiers......................... 70

3.2 Effect of views on democracy and the government and socio-economic variables on choice probabilities................................ 91

3.3 Predicted vote-shares in the two hypothetical three-candidate races .. 97

Chapter 4

4.1 Estimated posterior densities for β_2 across models 133

4.2 Marginal covariate effects... 136

4.3 Marginal covariate effects under different specifications of the misreport model.. 138

4.4 Marginal covariate effects when x_1 is omitted from the misreport model... 142

4.5 Estimated turnout from self-reported vs. validated responses 151

4.6 Coefficients of the probit models for self-reported vs. validated turnout. 154
4.7 Marginal effects of race on turnout. 156
4.8 Determinants of misreporting. 158
4.9 Estimated probability of misreporting for the respondents in the 1978 – 1990 ANES validated studies. 162
4.10 Predicted and actual misreport rates, by election-year. 163
4.11 Posterior distributions of selected coefficients of the misreport model. 165
4.12 Posterior summaries for selected parameters under an internal validation design. 166
4.13 Marginal effect of race on turnout estimated under our proposed method. 168
4.14 Posterior densities of β under an external validation design. 170
4.15 Sensitivity analysis for the external validation design. 172
4.16 Posterior summaries for $\beta_{Non-white}$ with list-wise deletion versus Bayesian imputation. 176

LIST OF TABLES

Number *Page*

Chapter 2
2.1 Summary statistics – Independent variables.. 31
2.2 Estimated posterior means and 90% credible intervals for fixed effects under alternative distributional assumptions for the error terms.. 36
2.3 Posterior means of variance-covariance components under Model 1-b 43
2.4 Effect of a one-unit change in the predictors of invalid voting and absenteeism under Model 1-b 46
2.5 Estimates of the Posterior Predictive Loss for alternative modeling strategies.. 52

Chapter 3
3.1 Vote choice by respondents' views and characteristics 66
3.2 Posterior means and 95% credible intervals for the parameters of the multinomial probit model 82
3.3 Marginal effect of individual-specific variables on voter choice 87
3.4 Expected vote-shares of the candidates under alternative electoral scenarios.. 94

Chapter 4
4.1 Marginal covariate effects under alternative specifications of the misreport model.. 144
4.2 Posterior means and 95% credible intervals for β with missing data 146
4.A.1 Misreport probabilities under Design B 180
4.A.2 Posterior means and 95% credible intervals – Design A 181

4.A.3 Posterior means and 95% credible intervals – Design B 183

4.A.4 Posterior means and 95% credible intervals for $\overline{\pi^{1|0}}$ and $\overline{\pi^{1|0}}$ 185

4.B.1 Vote misreporting in the 1978 -1990 ANES 187

4.B.2 Rates of non-response for the variables included in the voter turnout models 189

Chapter 1

Introduction

During the past decade, the vast improvements in computing power and the development of flexible and freely available statistical software have led to a growing interest in Markov chain Monte Carlo (MCMC) simulations for estimation and inference in the social sciences (Jackman, 2004). In fact, as argued by Jackman (2000b), the Bayesian paradigm, in particular Bayesian simulation based on MCMC algorithms, has "…the potential to become the unifying principle for social scientific statistical practice in the early century" (p. 310). In political science, the ability of Bayesian simulation to estimate complex models, avoiding the need for post-estimation steps or simulation procedures that rely on asymptotic normality to characterize uncertainty in the quantities of interest, providing a straightforward approach for incorporating prior (e.g., historical) information about model parameters and a simple way of handling missing data as part of the estimation process has, to a large extent, revolutionized the scope and nature of empirical research (Gill, 2000; Jackman, 2009). The increasing adoption and application of MCMC methods has allowed methodological specialist and applied researchers to explore new terrains and to address many longstanding problems in the discipline (Jackman, 2000b).

The potential influence and significance of these methodological advances can hardly be understated in electoral politics research, an area where serious scientific work is clearly on the upswing and where, despite enjoying a rich and dynamic history and benefiting from high-quality data and sophisticated theories, many of the major questions are not yet settled (Niemi and Weisberg, 2001; Converse, 2006). This dissertation takes advantage of the theoretical and practical advantages of the Bayesian framework – in particular, its ability to deal with small sample sizes, to incorporate historical information and to simplify the estimation of complex models - to analyze electoral data and to investigate and test different models of electoral behavior. In particular, the chapters that comprise this dissertation rely on MCMC methods to address substantive and methodological questions about electoral abstention, vote choice and turnout misreporting that have not been previously considered in the literature, combining macro- and micro-level data from different polities, periods and types– i.e., presidential, congressional - of elections.

Chapter 2 proposes a model to analyze the determinants of abstention in compulsory voting systems. Although mandatory voting has been found to be an effective mechanism for increasing voter turnout (Hirczy, 1994; Fornos1996), compelling citizens to go to the polls does not automatically mean that they will cast a vote for one of the candidates. Individuals can cast invalid votes, i.e., blank or null ballots, and thus their right not to vote remains intact (Lijphart, 1997). In addition, since mandatory voting does not generate universal compliance (Power and Roberts, 1995), illegal abstention constitutes a second form of non-voting. While invalid voting and absenteeism can thus be seen as “functional equivalents” of abstention under compulsory voting (Power and Roberts, 1995), previous studies in this area have not considered the correlation between both variables and ignored the compositional nature of the data, discarding helpful information that may contribute to better understand abstention and its causes and potentially leading to unfeasible and/or erroneous results (Zellner, 1971; Katz and King, 1999). In order to overcome these problems, Chapter 2 develops a statistical model to jointly analyze the determinants of invalid voting and electoral absenteeism, accounting for the compositional structure of the data, combining information at different levels of aggregation (e.g., individual, district-level and national-level data), and addressing robustness concerns raised by the use of small sample sizes typically available for countries with mandatory voting. In this setting, the Bayesian approach provides two main advantages. First, unlike with alternative estimation techniques, inference about the parameters of interest (e.g., fixed effects) does not depend on the accuracy of the point estimates of the variance-covariance parameters: they are based on their posterior distribution given only the data, averaging over the uncertainty for all the parameters in the model (Goldstein, 1995). Taking into account the uncertainty in the estimation of the random parameters is especially important in small datasets, where the variance parameters are usually imprecisely estimated (Bryk and Raudenbush, 2002). Also, with small sample sizes, outlying data points can seriously distort estimates of location (e.g., means or regression coefficients). Bayesian simulation methods are particularly well suited for fitting outlier-resistant regression models such as the Student-t regression model implemented in this chapter, allowing us to easily estimate the degrees of freedom parameters along with location and scale parameters even with moderately sized data, propagating the uncertainty in the former into inferences about the parameters of interest, and providing a valuable tool with which to assess the sensitivity of inferences to prior distributional assumptions (Gelman, Carlin, Stern and Rubin, 2004). The model is used to explore the causes of both sources of abstention in Brazil, the country with the largest electorate in the world subject to mandatory voting provisions. The results show considerable differences in the determinants of both forms of non-voting: while invalid voting was strongly positively related both to political protest and to the existence of important informational barriers to voting, the influence of these variables on absenteeism is less evident. Comparisons based on posterior simulations indicate that the compositional-hierarchical model developed in this chapter fits the dataset better than several other modeling approaches and leads to different substantive conclusions regarding the effect of different predictors on the both sources of abstention.

Chapter 3, coauthored with R. Michael Alvarez, implements a Bayesian multinomial probit model to analyze voter choice in Chile's historical 2005 election.[1] For the first time since the re-establishment of democracy, the right-wing *Alianza por Chile*, one of the coalitions that has dominated contemporary politics in Chile, presented two presidential candidates who adopted electoral strategies and platforms appealing to different groups of voters. In the context of a fragmented and polarized political scene, there is little consensus among scholars about whether the presence of two viable conservative candidates bolstered *Alianza*'s support or, on the contrary, actually damaged the coalition's electoral chances. The lack of rigorous empirical studies of the 2005 Chilean presidential election, however, has prevented addressing this issue. Unlike other polytomous choice models that rely on the independence of irrelevant alternatives property, our multinomial probit model (MNP) allows us to answer this question by accounting for possible substitution patterns in voters' electoral preferences over the candidates. Given the computational complexity of fitting the MNP, though, the model has seen relatively few applications in the political science literature. Most of them have resorted to maximum likelihood estimation, relying on asymptotic normality in making

[1] A paper based on the material in Chapter 2 has been published in *Electoral Studies* 28(2), 177 – 189, 2009.

inferences about the error variance and covariance parameters (e.g., Alvarez and Nagler, 1995; Alvarez, Nagler and Bowler, 2000; Dow and Endersby, 2004). However, as shown by McCulloch and Rossi (1994), asymptotic approximations are quite problematic in the context of the multinomial probit model. In a series of experiments examining the sampling distributions of MLE estimates for a three-choice multinomial probit model, the authors found that, even with as many as 1,000 observations per parameter – many more than is usually the case in most political science applications - there was considerable skewness in the sampling distributions of the error variance-covariance parameters, concluding that "…asymptotic theory may be of little use for the MNP model" (p. 219).[2] In this regard, the main advantage of the Bayesian approach based on MCMC methods is that it allows obtaining arbitrarily precise approximations to the posterior densities, without relying on large-sample theory (McCulloch and Rossi, 1993; Jackman, 2004). In addition, it avoids direct evaluation of the likelihood function and the resulting convergence problems exhibited by maximum likelihood optimization, and is computationally more efficient than simulation-based methods of classical estimation when dealing with a relatively large number of alternatives (Kim, Kim and Heo, 2003; Train, 2003). Hence, the Bayesian approach overcomes some of the main criticisms that

[2] As noted by Jackman (2000a), part of the problem stems from the normalization employed to identify the MNP model, which leads to estimating bounded functions of variance parameters, such as variance ratios and correlations. Since there Is not much information about these parameters even in large sample, the "boundedness" the estimated parameters is likely to stop asymptotic normality from "kicking in". See also McCulloch and Rossi (1994, pp. 221- 222).

have been leveled against the use of MNP in electoral studies (Dow and Endersby, 2004). Furthermore, since comparison of different models that can be used to operationalize alternative sets of hypothesis can be easily achieved using Bayes factors (Quinn and Martin, 1998), the Bayesian framework is particularly well suited to examine the relative validity of the various competing explanations that have been traditionally proposed to account for voters' behavior in Chile (Valenzuela, 1999; Torcal and Mainwaring, 2003).

Chapter 4, which is coauthored with Jonathan N. Katz, addresses the issue of measurement error in survey data.[3] In particular, we focus on the problem of misclassified binary responses, which has been a major concern in the political science literature analyzing voter behavior, especially voter turnout. The chapter develops a parametric model that corrects for misclassified binary responses, allowing researchers to continue to rely on the self-reported turnout data commonly used in political science research while improving the accuracy of the estimates and inferences drawn in the presence of turnout misreporting. In order to do so, our model resorts to information on the misreporting patterns obtained from auxiliary data sources such as internal or external validation studies, matched official records, administrative registers, and possibly even aggregate data. While incorporating this information into the analysis of the sample of interest using frequentist methods is far from straightforward (Prescott and Garthwaite, 2005), this can be easily accomplished within the Bayesian framework *via* MCMC simulations, avoiding the need

[3] A shortened version of Chapter 4 is forthcoming in the *American Journal of Political Science* 54(3), July 2010.

for complex numerical methods to approximate analytically intractable posterior distributions. In addition, our approach also allows us to simultaneously address another important problem with (political) survey data, namely missing outcome and/or covariate values, using fully Bayesian model-based imputation. Compared to alternative imputation techniques, Bayesian methods allow easily estimating standard errors in multiparameter problems and handling “nuisance” parameters, and have been shown to be particularly efficient when data loss due to missing observations is substantial (Ibrahim, Chen and Lipsitz 2002). Using Monte Carlo simulations, we show that, even with small rates of misclassification, our proposed solution improves estimates and inference with respect to standard models ignoring misreporting, and it also outperforms other methods proposed in the literature when misreporting is associated with the covariates affecting the true response variable. While our model is in fact fully generally, we illustrate its application in the context of estimating models of turnout using data from the American National Election Studies. We show that substantive conclusions drawn from models ignoring misreporting can be quite different from those resulting from our model.

Finally, Chapter 5 concludes.

Chapter 2

A Statistical Model of Abstention under Compulsory Voting

2.1 Introduction

The desire to provide a political system with popular legitimacy and to increase the representativeness of elected public officers have often been asserted as major arguments justifying the imposition of compulsory voting provisions (Verba, Nie and Kim, 1978; Hill, 2002). Twenty-four countries, comprising approximately 20% of the world's democracies, employ mandatory voting to some extent (Australian Joint Standing Committee on Electoral Matters, 2000). Although compulsory voting has been found to be an effective mechanism for increasing voter turnout (Hirczy, 1994; Lijphart 1997; Fornos, 1996), compelling voters to go to the polls does not automatically mean that they will cast a vote for one of the candidates. Citizens can cast invalid votes, i.e., blank or null ballots, and thus their right not to vote remains intact (Lijphart, 1997); in fact, a long-standing feature of compulsory voting systems is a higher rate of invalid ballots (Hirczy, 1994). In addition, since mandatory voting does not generate universal compliance (Hirczy, 1994; Power and Roberts, 1995), illegal abstention constitutes a second form of non-voting.

Previous research on compulsory voting systems has focused either on the determinants of electoral absenteeism (Hirczy, 1994; Fornos, Power and Garand, 2004) or on the determinants of invalid voting (McAllister and Makkai, 1993; Power and Garand, 2007). The common approach of studies in this area has been to treat the proportion of invalid votes or electoral absenteeism as the dependent variable and regress each on a set of explanatory variables. This standard procedure exhibits two main shortcomings. First, it does not take into account the connection between both sources of non-voting and the relationship between their determinants. Since, under compulsory voting, invalid voting and electoral absenteeism can be seen as "functional equivalents" of abstention, jointly modeling them may contribute to a better understanding of abstention and its causes. Moreover, without a model for exploring the interrelation between these two sources of abstention, helpful information from an inferential standpoint maybe discarded because the correlation between them is assumed to be zero, and changes in the standard error estimates that might result from a bivariate model could substantially modify the conclusions drawn from separate univariate analyses (Zellner, 1971; Thum, 1997). Second, the prevailing modeling strategy ignores the "compositional" nature of the data (Aitchison, 1986), i.e., the fact that the proportions of invalid ballots, electoral absenteeism and votes for candidates or parties among the electorate cannot be negative and that must sum one. Ignoring these non-negativity and unit-sum constraints might lead to unfeasible results, such as negative percentages of invalid ballots or sums of proportions greater or less than one (Katz and King, 1999).

This chapter develops a statistical model to address these problems, jointly analyzing the determinants of invalid voting and electoral absenteeism in district-level elections. While national-level studies have the advantage of allowing more countries in the analysis, they are generally based on a small number of observations and may fail to capture the contextual and "neighborhood" effects that might have considerable influence in local (e.g., legislative) elections (King, 1997; Katz and King, 1999). In addition, given the absence of survey data covering large historical periods in many of the countries with compulsory voting, most of which are recently democratized Latin American nations (International Institute for Democracy and Electoral Assistance, IDEA, 2007), district-level elections allow studying both sources of abstention at the lowest possible level of aggregation.

However, analyzing district-level elections introduces an additional methodological challenge. The proportion of invalid votes and absenteeism may be influenced not only by local variables but also by country-level factors affecting all districts in a given election (Power and Roberts, 1995), violating the standard assumption of independent and identically distributed errors. Ignoring the hierarchical structure of the data and simply pooling national- and district-level variables may thus result in inefficient parameter estimates and negatively biased standard errors, potentially leading to “spuriously significant” statistical effects (Antweiler, 2001; Maas and Hox, 2004; Franzese, 2005).

Drawing on the literature on compositional data (Aitchison and Shen, 1980; Aitchison, 1986; Katz and King, 1999), and on multi-level modeling (Goldstein, 1995; Bryk and Raudenbush, 2002; Gelman and Hill, 2007), the model presented here relates both sources of abstention in compulsory-voting systems, accounting for the compositional and hierarchical structure of the data and addressing robustness concerns raised by the use of small samples that are typical in the literature. I illustrate the use of the model analyzing data on invalid voting and electoral absenteeism in Brazil’s lower house elections at the state level. Brazil has the largest electorate in the world subject to compulsory voting and has experienced considerable variations in institutional, political and socioeconomic conditions across history and between states, therefore providing an illuminating case to examine rival explanations of invalid voting and absenteeism. The percentage of blank and null ballots in the country has been historically larger and more volatile than in most other democracies with compulsory voting (Instituto Universitario de Pesquisas de Rio de Janeiro, IUPERJ, 2006; IDEA, 2007), and absenteeism has remained relatively high despite mandatory voting.

Power and Roberts (1995) used ordinary least square pooled time-series regressions to separately analyze the determinants of the two sources of abstention in legislative elections between 1945 and 1990, combining country-level and state-level predictors by assigning the national variables to each state. I extend the period of analysis to include all the elections held up to 2006 and compare the results of the model developed in this chapter with those obtained from alternative modeling strategies that that fail to account for the compositional and/or the hierarchical structure of the data. Based on posterior simulations, I show that the compositional-hierarchical model leads to different substantive conclusions and fits the data better than these alternative modeling approaches.

The remainder of the chapter is organized as follows. Section 2.2 briefly reviews alternative theories for explaining invalid voting and absenteeism under compulsory voting systems. Section 2.3 presents the compositional-hierarchical model developed in this chapter to analyze the determinants of invalid voting and absenteeism at the district level. Section 2.4 applies the model to analyze 16 lower house elections in Brazil and compares the performance of the compositional-hierarchical model with three competing approaches. Finally, Section 2.5 concludes.

2.2 Alternative explanations of invalid voting and absenteeism

Drawing on the literature on voter turnout in industrialized democracies, three basic explanations, focusing on socioeconomic factors, on institutional variables, and on “protest voting", have been proposed to account for invalid voting and absenteeism in compulsory voting systems (McAllister and Makkai, 1993; Power and Roberts, 1995; Fornos et al., 2004; Power and Garand, 2007).

Some scholars have argued that the high rate of blank and null ballots in polities with mandatory voting reflects the alienation of citizens from the political system and is the consequence of mobilizing disinterested and poorly informed citizens who would otherwise abstain (Jackman, 2001). Previous analyses (1993; Power and Roberts, 1995; Power and Garand, 2007) found that socioeconomic variables such as urbanization, literacy and education levels substantially affect the percentage of blank and null ballots cast through their effect on the perceived efficacy, access to information and development of political skills among the electorate. Although the literature on electoral behavior has also found a strong correlation between these variables and political participation in voluntary voting settings (Verba et al., 1978; Powell, 1986; Rosenstone and Hansen, 1993), empirical evidence from countries with mandatory voting (Power and Roberts, 1995; Fornos et al., 2004) suggest that the impact of socioeconomic factors on electoral absenteeism in these countries is quite moderate.

Other authors have underscored the role of the institutional context and design in explaining invalid voting and absenteeism. For instance, Blais and Dobrzynska (1998) and Kostadinova (2003) concluded that a higher number of political parties depress turnout by increasing the unpredictability of electoral and policy outcomes, and the same would apply for highly disproportional systems that punish minor parties and reduce voters' perceived efficacy (Jackman, 1987; Jackman and Miller, 1995). In the same direction, McAllister and Makkai (1993) and Power and Roberts (1995) provide evidence that institutional factors such as district magnitude and ballot structures have a considerable impact on invalid voting in mandatory voting settings.

Finally, an alternative explanation can be traced to the literature on protest voting (Kitschelt, 1995; Lubbers and Scheepers, 2000). A protest vote can be defined as a vote primarily cast to express discontent with politics, rather than to affect public policies (Van der Brug and Fennema, 2003). In a system of compulsory voting, citizens' discontent with the political establishment would translate into higher null and blank ballots and illegal abstention (Derks and Deschouwer, 1998). This interpretation has often been quoted in Brazil and Latin America to explain temporary increases in invalid voting and absenteeism (Moisés, 1993; Jocelyn–Holt, 1998; Escobar, Calvo, Calcagno and Minvielle, 2002).

Although the socioeconomic, institutional and protest approaches are usually presented as competing rather than complementary explanations, previous research (Power and Roberts, 1995; Fornos et al., 2004) has shown that fusing them in a combined model helps to better understand the phenomena under study. However, since these approaches are grounded in the literature on political participation in developed democracies, where invalid voting has not received much academic attention (Power and Garand, 2007), past work has made no theoretical distinctions regarding the effect of the different sets of variables on invalid voting and electoral absenteeism. The underlying assumption in previous analyses has been that the same basic causal mechanisms account for both forms of non-voting (Power and Roberts, 1995). Furthermore, from a methodological perspective, they failed to examine the potential interactions between the determinants of these two sources of abstention, implicitly assuming that the effects of the relevant predictors on invalid voting are independent of their impacts on absenteeism. The statistical model presented in the next section allows me to test these assumptions.

2.3 A statistical model of abstention under compulsory voting

The model used to analyze the determinants of invalid voting and absenteeism at the district level is grounded in the literature on "compositional data" (Aitchison and Shen, 1980; Aitchison, 1986; Katz and King, 1999) and on Bayesian hierarchical modeling (Lindley and Smith, 1972; Gelman and Hill, 2007), although it is modified and adapted to the problem under study.

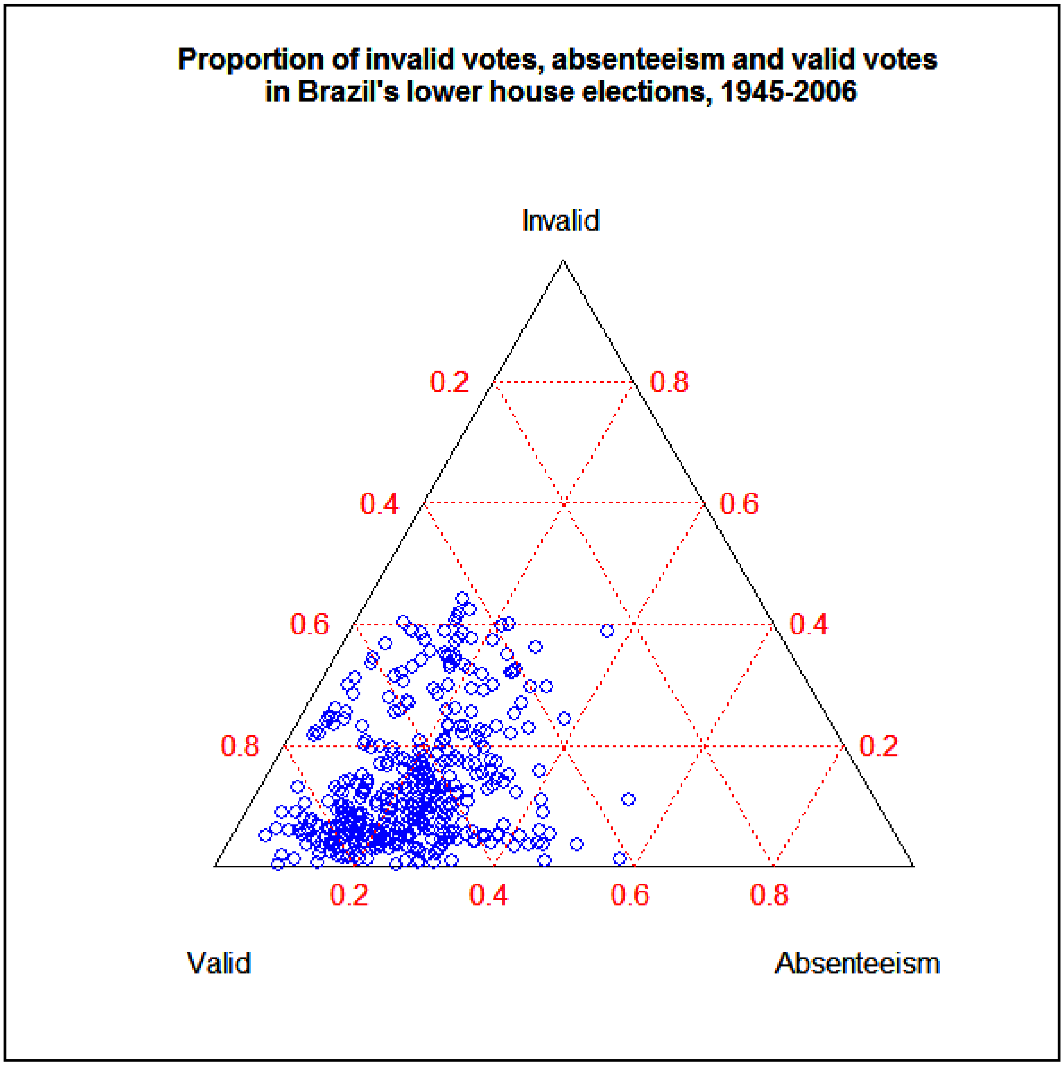
Let $P^I_{i,t}$, $P^A_{i,t}$ and $P^V_{i,t}$ denote the proportion of invalid votes, electoral absenteeism and valid votes (i.e., votes for candidates or parties) among the electorate in district i at election t, $i = 1,2 \ldots. n$, , $t = 1,2 \ldots T$. For all i and t, $P^I_{i,t}$, $P^A_{i,t}$ and $P^V_{i,t}$ must satisfy the following non-negativity and unit-sum constraints (Katz and King, 1999):

$$P^s_{i,t} \in [0,1], \qquad s = I, A, V \tag{2.1}$$

$$P^I_{i,t} + P^A_{i,t} + P^V_{i,t} = 1 \tag{2.2}$$

These constraints determine that $P^I_{i,t}$, $P^A_{i,t}$ and $P^V_{i,t}$ fall in the simplex space. Figure 2.1 illustrates the simplex sample space using a ternary plot for lower house elections in Brazil between 1945 and 2006. Each circle in the figure indicates the values of P^I, P^A and P^V in a particular district for a given election.

Figure 2.1



Note: The diagonal lines parallel to the triangle's left side indicate the proportion of electoral absenteeism, measured on the scale in the triangle's base. The diagonal lines parallel to the right side mark the proportion of valid votes, measured on the scale in the triangle's left side, and the dashed horizontal lines indicate the proportion of invalid votes, measured on the triangle's right side.

A model aimed at analyzing the determinants of abstention in compulsory voting systems must take the constraints defined in (1) and (2) into account. Neither the standard approach of regressing invalid voting and absenteeism independently on a set of predictors nor estimating a system of seemingly unrelated equations satisfies these constraints, even if eventually the point predictions obtained happen to fall within the boundaries of the simplex (Katz and King, 1999). In order to address this problem, I adapt Aitchison's (1986) and Katz and King's (1999) models for compositional data using a Bayesian implementation of a bivariate mixed model for invalid voting and electoral absenteeism.

Let $Y_{i,t}^{I}=\ln\left(P_{i,t}^{I}/P_{i,t}^{V}\right)$ and $Y_{i,t}^{A}=\ln\left(P_{i,t}^{A}/P_{i,t}^{V}\right)$ denote the log-ratios of the proportion of invalid votes and absenteeism relative to valid votes, respectively.[4] Note that, unlike the baseline composites $P_{i,t}^{I}, P_{i,t}^{A}$ and $P_{i,t}^{V}$, $Y_{i,t}^{I}$ and $Y_{i,t}^{A}$ are unbounded and unconstrained. The variables of interest for the analysis, $P_{i,t}^{I}, P_{i,t}^{A}$, are obtained from $Y_{i,t}=\left[Y_{i,t}^{I}, Y_{i,t}^{A}\right]$ through the additive logistic transformations:

$$P_{i,t}^{I}=\frac{\exp\left[Y_{i,t}^{I}\right]}{1+\exp\left[Y_{i,t}^{I}\right]+\exp\left[Y_{i,t}^{A}\right]} \tag{2.3}$$

$$P_{i,t}^{A}=\frac{\exp\left[Y_{i,t}^{A}\right]}{1+\exp\left[Y_{i,t}^{I}\right]+\exp\left[Y_{i,t}^{A}\right]} \tag{2.4}$$

[4] Due to the logarithmic transformations involved, the baseline composites are assumed to be strictly positive. Although this poses no problem for this type of electoral data, alternative models based on Box-Cox transformations (Rayens and Srinivasan, 1991) have been proposed to deal with the problem of null composites.

Since the $Y_{i,t}^{s}$, $s = I, A$, are defined over the whole real line, it is possible to model $Y_{i,t} = \left[Y_{i,t}^{I}, Y_{i,t}^{A} \right]$ using a normal/independent distribution (Andrews and Mallows, 1974; Liu, 1996; Seltzer, Novak, Choi and Lim, 2002) that assigns weight parameters to each observation in the sample, as in a Weighted Least Squares analysis:

$$Y_{i,t} = \mu_{i,t} + \frac{\varepsilon_{i,t}}{\sqrt{w_{i,t}}} \tag{2.5}$$

where $\mu_{i,t} = \left[\mu_{i,t}^{I}, \mu_{i,t}^{A} \right]'$, $\varepsilon_{i,t} = \left[\varepsilon_{i,t}^{I}, \varepsilon_{i,t}^{A} \right]' \sim \mathrm{N}(0, \Sigma)$, $w_{i,t}$ is a positive random variable with density $p\left(w_{i,t} \middle| \upsilon \right)$, and υ a scalar or vector-valued parameter. The main advantage of assuming a normal/independent distribution is that, due to the unconstrained properties of Σ, the model now allows for any pattern of dependency between $P_{i,t}^{I}$ and $P_{i,t}^{A}$.[5] In addition, besides including the bivariate normal as a particular case (when $w_{i,t} = 1 \; \forall i, t$), the normal/independent distribution also provides a group of thick-tailed distributions often useful for robust inference and identification of outliers (Seltzer, Novak, Choi and Lim, 2002; Rosa, Padovani and Gianola, 2003), particularly when the number of districts or elections in the sample is relatively small.

---

[5] This is, in fact, the key advantage of assuming a scale mixture of multivariate normals *vis-à-vis* alternative statistical models for compositional data, such as the Dirichlet distribution (Johnson and Kotz, 1972) and the S$^-$ distribution (Barndorff-Nielsen and Jørgensen, 1991).

The focus of the model lies in the specification of $\mu_{i,t}$. Since $\mu^{I}_{i,t}$ and $\mu^{A}_{i,t}$ are unbounded, it is possible to reparametrize them as linear functions of regressors. As mentioned in the introduction, it seems plausible that the proportion of invalid votes and electoral absenteeism in a district is influenced not only by district-level variables but also by national conditions that vary across elections. Moreover, the impact of district-level variables on invalid voting and absenteeism might itself be mediated by these country-level factors. In order to account for these possibilities, I use a hierarchical random-coefficients model for the components of $\mu_{i,t}$. The first-level equations model $\mu^{I}_{i,t}$ and $\mu^{A}_{i,t}$ as functions of district-level variables measured at a particular election. The second-level equations specify the first-level coefficients as functions of country-level variables measured contemporaneously with the district level variables, plus zero-expectation random effects assumed to be constant across all districts in a given election, accounting for election-to-election variability beyond that explained by national-level variables. In addition, I also introduce zero-mean random intercepts in order to account for time-constant heterogeneity across districts. This modeling strategy strikes a balance between a completely pooled approach, which ignores the clustered nature of the data and the potential variability between districts and elections, and local regressions that would be highly unstable given the paucity of the data typically available for analyzing countries with compulsory voting, most of them recently democratized Latin American nations (Browne and Draper, 2001; Gelman and Hill, 2007).

Letting $x_{i,t}$ and z_t represent $(1\times K)$ and $(1\times L)$ row vectors of district-level and country-level variables, respectively, the specification adopted is then:

$$\mu_{i,t} = X_{i,t}\beta_t + \lambda_i \tag{2.6}$$

$$\beta_t = Z_t\delta + \eta_t \tag{2.7}$$

where

$X_{i,t}$ is a $2\times 2(K+1)$ matrix, $X_{i,t} = \left[I_2 \middle| x_{i,t} \otimes I_2\right]$,

β_t is a $2(K+1)\times 1$ vector, $\beta_t = \left[\beta^I_{0,t} \quad \beta^A_{0,t} \quad \beta^I_{1,t} \quad \beta^A_{1,t} \quad ... \quad \beta^I_{K,t} \quad \beta^A_{K,t}\right]'$,

Z_t is a $2(K+1)\times 2(L+1)(K+1)$ block diagonal matrix: $Z_t = I_{2(K+1)} \otimes \left[1 \middle| z_t\right]$,

δ is a $2(K+1)(L+1)\times 1$ vector, $\delta = \left[\delta^I_{0,0} \quad \dots \quad \delta^I_{0,L} \quad \delta^A_{0,1} \quad \dots \quad \delta^A_{0,L} \quad \delta^I_{1,0} \quad \dots \quad \delta^A_{K,L}\right]$,

$\eta_t = \left[\eta^I_{0,t}, \eta^A_{0,t}, \eta^I_{1,t}, \eta^A_{1,t}, ...\eta^I_{K,t}, \eta^A_{K,t}\right]' \sim N\left(0, \Omega_\eta\right)$ and $\lambda_i = \left[\lambda^I_i, \lambda^A_i\right]' \sim N\left(0, \Omega_\lambda\right)$ are election- and district- random effects.[6]

From (2.5) - (2.7), the model can be written as:

$$Y_{i,t} = X_{i,t}Z_t\delta + X_{i,t}\eta_t + \lambda_i + \frac{\varepsilon_{i,t}}{\sqrt{w_{i,t}}} \tag{2.8}$$

with error terms $\varepsilon_{i,t}$ and random the effects η_t and λ_i assumed mutually independent.

---

[6] Throughout this chapter, $\otimes$ denotes the left Kronecker product.

In order to estimate the model, I employ a fully Bayesian strategy, treating all unknown quantities as random and specifying prior distributions for all the parameters. The Bayesian approach straightforwardly accommodates problems with small samples typically available for countries with mandatory voting, since it does not rely on asymptotic results for inference (Thum, 2003; Jackman, 2004). In particular, unlike alternative estimation techniques (e.g., Full or Restricted Maximum Likelihood), inference about the fixed effects does not depend on the accuracy of the point estimates of the variance-covariance parameters: they are based on their posterior distribution given only the data, averaging over the uncertainty for all the parameters in the model (Goldstein, 1995; Bryk and Raudenbush, 2002). Taking into account the uncertainty in the estimation of the random parameters is especially important in small datasets, where the variance parameters are usually imprecisely estimated (Bryk and Raudenbush, 2002).[7,8]

Assuming conditional independence throughout, the model can be specified in a Bayesian context as

---

[7] In the context of frequentist estimation techniques, this uncertainty can be taken into account through bootstrapping (Goldstein, 1995) or simulation (King, Tomz and Wittenberg, 2000). However, the fact that the Bayesian approach directly takes into account the uncertainty in variance components makes it particularly appropriate for this kind of analysis.

[8] In addition, as shown by Browne and Draper (2001), Maximum Likelihood methods are susceptible to convergence problems in two-level random-coefficients regression models with few higher-level units.

$$Y_{i,t} \sim N\left(X_{i,t}\beta_t + \lambda_i,\ \frac{1}{w_{i,t}}\Sigma \right),\ i = 1,...,n,\ t = 1,...,T \tag{2.9}$$

$$\beta_t \sim N\left(Z_t\delta, \Omega_\eta \right),\ t = 1,...,T \tag{2.10}$$

$$\lambda_i \sim N\left(0, \Omega_\lambda \right),\ i = 1,...,n \tag{2.11}$$

with conjugate priors for the fixed effects and the precision matrices:

$$\begin{aligned}
&\delta \sim N\left(\delta_0, \Omega_\delta\right), \\
&\Sigma^{-1} \sim Wishart(P, \rho_P), \quad |P| > 0, \rho_P \geq 2 \\
&\Omega_\eta^{-1} \sim Wishart(Q, \rho_Q), \ \ |Q| > 0, \rho_Q \geq 2\left(K+1\right) \\
&\Omega_\lambda^{-1} \sim Wishart(R, \rho_R), \ \ |R| > 0, \rho_R \geq 2
\end{aligned} \tag{2.12}$$

and $p\left(w_{i,t} \middle| \upsilon\right)$ depending on the particular normal/independent distribution adopted for the level-1 errors. Routine sensitivity analyses can be performed in order to examine the effect of the hyperparameters on the model fit.

The joint posterior density of all the unknown parameters of the model, $f\left(\beta,\lambda,\delta,\Sigma,\Omega_\eta,\Omega_\lambda,w,\upsilon \middle| Y\right)$, is intractable analytically, but inference on the parameters of interest can be performed by Markov chain Monte Carlo (MCMC) simulations, using Gibbs sampling to repeatedly draw samples from each unknown parameter's full conditional posterior distribution in order to form the marginal distributions used for Bayesian inference (Gelfland and Smith, 1990; Casella and George, 1992). In order to implement the Gibbs sampler, I subdivide the entire set of unknowns in such a way that it is possible to sample from the conditional posterior of each subset of unknowns given the other subsets and the data. This leads to an iterative scheme whereby, given an arbitrary set of starting values, samples are drawn from each full conditional posterior given the data and the most recently sampled values for the other unknowns (Gelfland, Hills, Racine-Poon, and Smith, 1990; Seltzer et al., 2002). Under mild regularity conditions (Geman and Geman, 1984), samples from these complete conditionals approach samples from the marginals for a sufficiently large number of iterations. The power and simplicity of the Gibbs sampler in handling complex hierarchical models involving covariates makes it an attractive option against alternative Bayesian/empirical Bayesian methodologies that must often rely on "…a number of approximations whose consequences are often unclear under the multiparameter likelihoods induced by the modeling" (Gelfland et al., 1990, p. 978).

Given $w=\left(w_{i,1},...,w_{n,T}\right)'$ the full conditional posterior densities of $\{\beta_t\},\{\lambda_i\},\delta,\Sigma,\Omega_\eta$ and Ω_λ are:

$$\begin{aligned}
&\beta_t \,|\, Y,\lambda,\delta,\Sigma,\Omega_\eta,\Omega_\lambda,w,\upsilon \sim N\left(b_t,B_t\right), \quad t=1,...T,\\
&b_t=\left[\sum_{i=1}^{n} w_{i,t}X_{i,t}{}'\Sigma^{-1}X_{i,t}+\Omega_\eta^{-1}\right]^{-1}\left[\sum_{i=1}^{n} w_{i,t}X_{i,t}{}'\Sigma^{-1}\left(Y_{i,t}-\lambda_i\right)+\Omega_\eta^{-1}Z_t\delta\right],\\
&B_t=\left[\sum_{i=1}^{n} w_{i,t}X_{i,t}{}'\Sigma^{-1}X_{i,t}+\Omega_\eta^{-1}\right]^{-1}
\end{aligned} \tag{2.13}$$

$$\begin{aligned}
&\lambda_i \,|\, Y,\beta,\delta,\Sigma,\Omega_\eta,\Omega_\lambda,w,\upsilon \sim N\left(d_i,D_i\right), \quad i=1,...n,\\
&d_i=\left[\Sigma^{-1}\sum_{t=1}^{T} w_{i,t}+\Omega_\lambda^{-1}\right]^{-1}\left[\Sigma^{-1}\sum_{t=1}^{T} w_{i,t}\left(Y_{i,t}-X_{i,t}\beta_t\right)\right],\\
&D_i=\left[\Sigma^{-1}\sum_{t=1}^{T} w_{i,t}+\Omega_\lambda^{-1}\right]^{-1}
\end{aligned} \tag{2.14}$$

$$\delta \left| Y, \beta, \lambda, \Sigma, \Omega_\eta, \Omega_\lambda, w, \upsilon \sim N\left(\left[\sum_{t=1}^{T} Z_t\, '\Omega_\eta^{-1} Z_t + \Omega_\delta^{-1} \right]^{-1} \left[\sum_{t=1}^{T} Z_t\, '\Omega_\eta^{-1} \beta_t + \Omega_\delta^{-1} \delta_0 \right], \left[\sum_{t=1}^{T} Z_t\, '\Omega_\eta^{-1} Z_t + \Omega_\delta^{-1} \right]^{-1} \right) \right.$$

(2.15)

$$\Sigma^{-1} \left| Y, \beta, \lambda, \delta, \Omega_\eta, \Omega_\lambda, w, \upsilon \sim Wishart\left(\left[\sum_{t=1}^{T} \sum_{i=1}^{n} w_{i,t} \left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i \right)\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i \right)' + P^{-1} \right]^{-1}, \rho_P + nT \right) \right. \quad (2.16)$$

$$\Omega_\eta^{-1} \left| Y, \beta, \lambda, \delta, \Sigma, \Omega_\lambda, w, \upsilon \sim Wishart\left(\left[\sum_{t=1}^{T} \left(\beta_t - Z_t\delta \right)\left(\beta_t - Z_t\delta \right)' + Q^{-1} \right]^{-1}, \rho_Q + T \right) \right. \quad (2.17)$$

$$\Omega_\lambda^{\ -1} \left| Y, \beta, \lambda, \delta, \Sigma, \Omega_\eta, w, \upsilon \sim Wishart\left(\left[\sum_{i=1}^{n} \lambda_i \lambda_i\, ' + R^{-1} \right]^{-1}, \rho_R + n \right) \right. \quad (2.18).$$

To complete the specification for a Gibbs sampling scheme, the full conditional posterior distributions of w and υ are required. For each element of w the fully conditional posterior density is:

$$w_{i,t} \left| Y, \beta, \lambda, \delta, \Sigma, \Omega_\eta, \Omega_\lambda, \upsilon \propto w_{i,t} \,\text{ex}\, \left\{ \text{p} - \frac{\omega_{i,t}}{2} \left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i \right)'\, \Sigma^{-1} \left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i \right) \right\} \times p\left(w_{i,t} \middle| \upsilon \right) \right. \quad (2.19).$$

For υ, the density is:

$$\upsilon \left| Y, \beta, \delta, \gamma, \Sigma, \Omega_\eta, \Omega_\lambda, w, \tau \propto p(\upsilon) \prod_{i=1}^{n} \prod_{t=1}^{T} p\left(w_{i,t} \middle| \upsilon \right) \right. \quad (2.20).$$

From (2.13) – (2.20), it is clear that, assuming Normal level-1 residuals (i.e., if all the $w_{i,t},\ i=1,...,n,\ t=1,...,T$, have degenerate distributions at 1), the conjugacy of the prior distributions at each stage of the hierarchy leads to closed-form full conditional distributions for each parameter of the model, and it is thus straightforward to sample from them in order to obtain the marginal distributions. However, the assumption of Normal level-1 residuals makes inferences vulnerable to the presence of outliers (Andrews and Mallows, 1974; Pinheiro, Liu and Wu, 2001). Assuming a bivariate Student-t prior for $Y_{i,t}$ allows for the possibility of extreme observations, attenuating the influence of outliers (Berger, 1985; Gelman, Carlin, Stern and Rubin, 2004) and providing a valuable tool with which to assess the sensitivity of inferences to prior distributional assumptions (Carlin and Louis, 1996; Thum, 1997).

A bivariate Student t prior for $Y_{i,t}$ can be obtained from the normal/independent distribution by assuming $w_{i,t} \mid \upsilon \sim Gamma\left(\upsilon/2, \upsilon/2\right)$, $w_{i,t} > 0, \upsilon > 0$.[9] The fully conditional posterior densities (2.19) and (2.20) then become:

$$w_{i,t} \mid Y, \beta, \lambda, \delta, \Sigma, \Omega_\eta, \Omega_\lambda, \upsilon \sim Gamma\left(\frac{\upsilon}{2}+1,\ \frac{1}{2}\left[\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i\right)' \, \Sigma^{-1}\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i\right) + \upsilon\right]\right) \qquad (2.19')$$

$$\upsilon \mid Y, \beta, \lambda, \delta, \Sigma, \Omega_\eta, \Omega_\lambda, w \sim \left[2^{\upsilon/2}\Gamma\left(\frac{\upsilon}{2}\right)\right]^{-nT} \upsilon^{\frac{\upsilon nT}{2}} \exp\left\{-\frac{\upsilon}{2}\left(\sum_{i=1}^{n}\sum_{t=1}^{T} w_{i,t} - \log\left(w_{i,t}\right)\right)\right\} \qquad (2.20').$$

[9] I use the parametrization of the gamma distribution found in Rosa et al. (2003).

While it might be argued that working directly with a bivariate Student t density for $\left[\varepsilon_{i,t}^{I}, \varepsilon_{i,t}^{A}\right]'$ would be preferable to adding nT parameters to the model, the conditioning feature of the Gibbs sampler makes the augmentation of the parameter space quite natural (Carlin and Louis, 1996). In addition, this specification allows obtaining estimates of the weight parameters $w_{i,t}$, which can be useful to identify possible outliers (West, 1984; Congdon, 2003; Rosa et al., 2003). Note that, from (2.19'),

$$E\left(w_{i,t} \middle| Y, \beta, \lambda, \gamma, \Sigma, \Omega_{\eta}, \Omega_{\lambda}, \upsilon\right) = \frac{\upsilon + 2}{\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i\right)' \Sigma^{-1}\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i\right) + \upsilon} \qquad (2.21),$$

so that for a large enough υ, $E\left(w_{i,t} \middle| Y, \beta, \lambda, \gamma, \Sigma, \Omega_{\eta}, \Omega_{\lambda}, \upsilon\right) \rightarrow 1$, and approximately normal tails are obtained for the level-1 errors. However, for low values of υ, the expected value of $w_{i,t}$ decreases as $\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i\right)' \Sigma^{-1}\left(Y_{i,t} - X_{i,t}\beta_t - \lambda_i\right)$ increases. Therefore, the weight assigned to each observation in calculating posterior distributions of fixed-effects and level-1 regression parameters will depend on the posterior probabilities of the possible values of υ.[10] Although (2.20') does not have a closed form, this conditional posterior distribution can be approximated by discretizing the density along a grid of values and then sampling from the resulting discrete distributions. When the points in the grid are spaced

[10] A detailed discussion of this point is provided in Seltzer et al. (2002).

closely together, the discrete distribution of υ provides an accurate approximation to the full conditional distribution (Draper, 2001; Seltzer and Choi, 2002; Seltzer et al., 2002).[11]

The two variants of the model (with bivariate normal or bivariate Student-t level-1 errors) can be compared using standard Bayesian criteria for model selection such as the Deviance Information Criterion (DIC) or Bayes factors (Spiegelhalter, Best, Carlin and van der Linde, 2002; Gelman et al., 2004). The means and standard deviation of the convergent Gibbs samples generated from (2.13)-(2.20') under each variant of the model can be used to summarize the posterior distributions of the parameters. These marginal posterior distributions, however, are of no direct interest for the analysis. Rather, interest lies in the effect of the explanatory variables on the proportion of invalid voting and electoral absenteeism. I compute the impact of each of the district-level and country-level regressors on $P^{I}_{i,t}$ and $P^{A}_{i,t}$ using average predictive comparisons (Katz and King, 1999; King, Tomz and Wittenberg, 2000; Gelman and Hill, 2007). The algorithm implemented to estimate these causal effects is detailed in Appendix 2.A.

Some aspects of the model deserve further comment. First, while in the presentation above it has been assumed that $T_i = T \ \forall i, i = 1,...n$ in order to simplify the notation, the model can accommodate unbalanced data sets, with different number of elections per district. In fact, the capacity and flexibility to deal with nested unbalanced data sets is one additional advantage of Bayesian multilevel models versus more traditional frequentist

---

[11] Alternatively, a strategy based on Metropolis-Hastings sampling can be incorporated into the MCMC scheme to obtain draws from υ (Seltzer et al., 2002; Gelman et al., 2004).

approaches (Bryk and Raudenbush, 2002; Shor et al., 2007). Also, a more complex specification for the components of Σ could be adopted (e.g., allowing for serial correlation of the level-1 errors – see Allenby and Lenk, 1994). Nonetheless, given the relatively small number of observations available in the application of Section 2.4 (with very few elections per state in some cases) and the inclusion of district random-effects, an i.i.d. assumption for the components of Σ seems appropriate (Carlin and Louis, 1996; Bryk and Raudenbush, 2002). Finally, as mentioned above, although I focus on two particular variants of the mixed model – i.e., with Normal and Student-t level-1 errors – assuming alternative densities for $w_{i,t}$ would allow obtaining other thick-tailed distributions – e.g., slash and contaminated Normals, as in Rosa et al. (2003) - that might be appropriate to account for the presence of outliers.

2.4 Analyzing invalid voting and electoral absenteeism in Brazil's lower house elections

2.4.1 Data and methodology

Brazil provides an interesting case to analyze the determinants of abstention in countries with mandatory voting. While invalid ballots in advanced democracies under compulsory voting such as Australia and the Netherlands have averaged about 2 to 3 percent, the equivalent rates in Brazil have been substantially higher and more volatile over time, reaching almost 42 percent of the votes cast in the 1994 lower house election (Power and Roberts, 1995; IUPERJ, 2006). In addition, despite the fact that voting has been compulsory in the country for over 60 years, electoral absenteeism has averaged 19 percent in elections held over this period, varying from 5 to 34.5 percent (IUPERJ, 2006). Changes in the institutional design and the freeness and fairness of the elections experienced by Brazil in its recent history and the sharp differences in socio-demographic characteristics among its states allow examining the impact of different factors on invalid voting and absenteeism.[12] In order to illustrate the use of the model presented in Section 2.3 and to compare the results with those obtained using alternative modeling strategies, I analyze all lower house elections held in the country between 1945 and 2006. The dataset has an unbalance structure, with 388 observations for 27 states across 16 elections.[13]

The dependent variables of interest for the analysis are the proportion of invalid votes and electoral absenteeism in lower house elections. The proportion of invalid votes among the electorate is computed as the ratio of blank and null votes cast over the population eligible to vote. Electoral absenteeism is calculated as the percentage of potential voters failing to comply with their duty. Figure 2.2 presents the proportion of invalid voting and absenteeism by state for the elections held between 1945 and 2006. As can be seen, there is considerable variation in the two sources of abstention both between states and within states across elections.[14]

---

[12] A description of the institutional, socioeconomic and political context of Brazilian elections exceeds the purposes of this chapter; an overview can be found in Power and Roberts (1995).

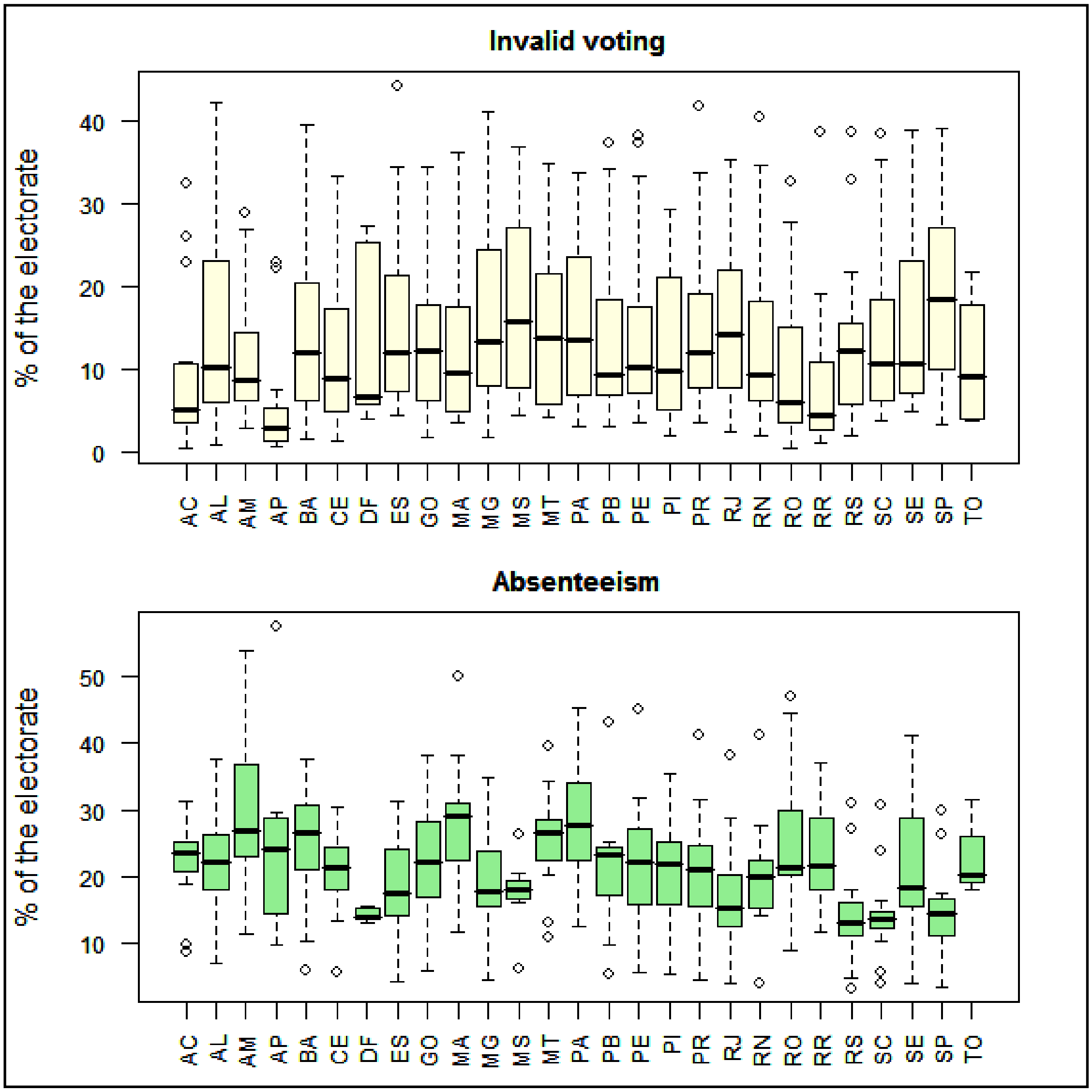
[13] The number of states in Brazil increased from twenty-two to twenty-seven during this period.

[14] The proportions are calculated based on the number of elections held in each state.

Figure 2.2

Invalid voting and absenteeism by state, as a proportion of the electorate

Lower house elections, 1945 - 2006



Note: The figure plots the proportion of invalid votes (upper panel) and electoral absenteeism (lower panel) in Brazilian lower house elections by State, in percentage points. The thick horizontal lines mark the average proportions across elections, the extremes of the colored rectangles represent the 50% intervals, and the upper and lower whiskers correspond to the 95%

intervals. Blank circles mark the outlying observations. Sources: Banco de Dados Eleitorais Do Brasil, Instituto Universitario de Pesquisas de Rio de Janeiro; Power and Roberts (1995).

In line with the different theories under consideration, socioeconomic, institutional and protest variables are included as explanatory variables in the model. The socioeconomic variables used are: *Illiteracy*, the percentage of the state's voting-age population classified as illiterate; *Urbanization*, the percentage of the state's population living in urban areas; and *FEAP*, the percentage of females in the Economically Active Population, used as a measure of women's status and the state's level of modernization. The institutional variables are: the number of *Candidates per seat*; *Franchising*, a dichotomous variable coded 1 for elections after 1985, when suffrage was extended to the illiterates, and 0 otherwise; *Electorate*, measured as the percentage of the state's total population eligible to vote; and *Ballot*, a dummy variable coded one for elections following the introduction of the single official ballot in 1962, that requires voters to write their candidate's name or registration number on a blank ballot and replaced the previous system of pre-printed ballots.[15] Finally, among the protest variables, *Manipulation* measures the degree of electoral manipulation and "political engineering", coded by Power and Roberts (1995) on

15 Prior to the introduction of the single official ballot ("cedula unica") in 1962, candidates distributed their own pre-printed ballots, which voters just had to place in the ballot box. While this required considerably less information on the part of voters, it tended to favor wealthier candidates to the detriment of less affluent ones (Power and Roberts, 1995).

a four point-scale ranging from 0 for free elections held under democratic rule to 3 for elections conducted under authoritarian tutelage; *Growth* is a two-year moving average of the percentage change in the national GDP; and *Inflation* is the natural logarithm of the country's average inflation rate in the two years preceding the election.[16] Table 2.1 provides summary statistics for the state-level and country-level predictors for the period 1945-2006.

Table 2.1

Summary statistics – Independent variables

Variable	**Mean**	**Std. Dev.**	**Min**	**percentile**	**percentile**	**Max**
State-level predictors						
Illiteracy (%)	40.0	20.8	4.7	24.6	58.9	79.8
Urbanization (%)	55.4	21.0	2.6	38.1	72.1	96.6
Females in the EAP (FEAP) (%)	23.3	16.4	3.0	10.3	40.8	58.1

[16] While the introduction of these "political protest" variables may lead to concerns about endogeneity, most previous research in this area includes either these or similar covariates, and is thus subject to the same criticism (Power and Roberts, 1995; Power and Garand, 2007). Hence, given that the focus of the chapter lies in comparing the performance of the model proposed here against other empirical approaches commonly used in the literature, I decided to keep these variables. Nonetheless, future applications of the model must explicitly take this potential problem into account.

Candidates per seat	4.4	2.8	1.0	2.3	6.0	15.4
Electorate (%)	40.4	19.2	6.9	24.3	59.0	74.4
Country-level predictors						
Franchising	0.4	0.5	0	0	1	1
Ballot	0.8	0.4	0	0.8	1	1
Electoral Manipulation	0.9	1.1	0	0	1.3	3
Growth (%)	5.3	3.6	-1.7	3.7	7.6	11.1
Inflation	3.7	1.7	1.7	2.7	4.2	7.5
Number of States	27					
Number of Elections	16					
Observations	388					

Sources: Banco de Dados Eleitorais Do Brasil, Instituto Universitario de Pesquisas de Rio de Janeiro (IUPERJ); Power and Roberts (1995).

The characterization and measurement of the independent variables closely follows Power and Roberts (1995); their data is complemented with information from IUPERJ (2006) for the 1994-2006 elections. The only difference with the respect to Power and Roberts (1995)'s work lies in the definition of *Illiteracy*: while they use the percentage of the state's electorate classified as illiterate (zero until 1985, when illiterates were enfranchised), I use the percentage of illiterates in the state's voting-age population. Although illiterates were not allowed to vote in Brazil until the 1986 election, the fact that more than sixty percent of the population had not finished the fourth grade by 1986 (Brazilian Institute of Geography and Statistics, 2003) and the difficulty of obtaining alternative reliable indicators covering the period under study led me to use illiteracy as a measure of the electorate's political skills (Power and Garand, 2007). In order to account for the effect of the enfranchisement of illiterates, I include the country-level variable *Franchising* and model the random-coefficients of *Illiteracy* as functions of it, allowing the effect of *Illiteracy* to vary across elections.

In addition, in line with Power and Roberts' (1995) argument that the country-level predictors *Ballot*, *Manipulation*, *Growth* and *Inflation* affect the proportion of invalid voting and absenteeism in each state-year, I specify the election random-intercepts $\beta_{0,t} = \left[\beta_{0,t}^{I}, \beta_{0,t}^{A}\right]'$ as functions of these variables. Given the small number of observations in the sample (Table 2.1), the coefficients of the remaining district-level variables are specified as fixed effects (i.e., their variation across elections is constrained to be 0), although the model could be written more generally to accommodate various plausible design alternatives for parametrizing these coefficients.

The following equations define the hierarchical model for district $i,\ i=1,\ldots,n$ at election $t,\ t=1,\ldots,T$:

$$\begin{aligned} Y_{i,t}^{s} &= \beta_{0,t}^{s} + \beta_{1,t}^{s} Illiteracy_{i,t} + \beta_{2,t}^{s} Urbanization_{i,t} + \beta_{3,t}^{s} FEAP_{i,t} + \\ &\quad \beta_{4,t}^{s} Candidates\ per\ Seat_{i,t} + \beta_{5,t}^{s} Electorate_{i,t} + \lambda_i^{s} + \frac{\varepsilon_{i,t}^{s}}{\sqrt{w_{i,t}}}, \qquad s = I, A \end{aligned} \tag{2.22}$$

$$\beta_{0,t}^{s} = \delta_{0,0}^{s} + \delta_{0,1}^{s} Ballot_t + \delta_{0,2}^{s} Manipulation_t + \delta_{0,3}^{s} Growth_t + \delta_{0,4}^{s} Inflation_t + \eta_{0,t}^{s}, \quad s = I, A \tag{2.23}$$

$$\beta_{1,t}^{s} = \delta_{1,0}^{s} + \delta_{1,1}^{s} Franchising_t + \eta_{1,t}^{s}, \qquad s = I, A \tag{2.24}$$

$$\beta_{k,t}^{s} = \delta_{k,0}^{s} \qquad s = I, A; k = 2,\ldots,5 \tag{2.25}$$

with

$$\left[\varepsilon_{i,t}^{I},\varepsilon_{i,t}^{A}\right]' \sim N\left(0,\Sigma\right),\ \eta_t=\left[\eta_{0,t}^{I},\eta_{0,t}^{A},\eta_{1,t}^{I},\eta_{1,t}^{A}\right]' \sim N\left(0,\Omega_{\eta}\right),\ \left[\lambda_i^{I},\lambda_i^{A}\right] \sim N\left(0,\Omega_{\lambda}\right),$$

and

$$p\left(w_{i,t}\middle|\upsilon\right)=\begin{cases}1\ \forall i,t\ \left(\text{bivariate normal prior for } Y_{i,t}\right)\ \text{or}\\ Gamma\left(\dfrac{\upsilon}{2},\dfrac{\upsilon}{2}\right)\ \forall i,t\ \left(\text{bivariate Student t prior for } Y_{i,t}\right)\end{cases}.$$

The model was fit using WinBUGS 1.4, as called from R 2.4.1.[17] All the hyperparameters in the model were assigned diffuse priors in order to let the data dominate the form of the posterior densities: the fixed effects were assigned a $N\left(\mathbf{0},\mathbf{100}I\right)$ prior, while Wishart priors with identity scale matrix and degrees of freedom equal to $rank\left(I\right)+1$ were used for the precision matrices. In order to ensure that inferences are data dependent, several alternative values for the hyperparameters were tried, yielding similar substantive results. Three parallel chains with dispersed initial values reached approximate convergence after 25,000 iterations, with a burn-in of 5,000 iterations; the results reported below are based on 1,000 samples of the pooled chains of deviates.[18]

[17] The code is available from the author on request.

[18] Approximate convergence is achieved for values of Gelman and Rubin’s (1992) estimated Potential Scale Reduction factor below 1.1.

2.4.2 Results of the compositional-hierarchical model

Table 2.2 below reports the posterior means and 90% credible intervals for the fixed effects for the two variants of the model presented in Section 2.3: assuming bivariate Normal (Model 1-a) and bivariate Student-t (Model 1-b) level-1 priors.[19] The values of the Deviance Information Criterion (DIC) for both models and the Bayes Factor for Model 1-b relative to Model 1-a are also presented.

The table shows considerable disparity in the posterior means and credible intervals of the fixed effects under both models, particularly regarding the effect of state-level predictors on the log-ratios Y^I and Y^A.[20] Comparisons between the two models based on both the DIC and Bayes Factor favor Model 1-b, indicating that the model with Student-t level-1 errors fits the data better. The evidence presented in Figures 2.3 and 2.4 further support Model 1-b. Figure 2.3 plots the mean posterior values of the standardized univariate and bivariate level-1 residuals from Model 1-a for the 388 observations in the dataset (Chaloner and Brant, 1988; Weiss, 1994). A few data points have standardized univariate residuals with absolute values larger than 5, and more than 2% of the

---

[19] In addition, I also estimated the model under the assumption of multivariate Student-t priors for the random coefficients. The main results, however, are virtually unchanged when assuming heavy tails at the higher-level of the model. Thus, I retain the assumption of multivariate normality at level-2 and focus on the effect of adopting alternative priors for the data model.

[20] It is worth noting that, when treating υ as unknown, the uncertainty regarding υ is propagated into the posterior distribution of the fixed-effects parameters (Seltzer et al., 2002).

observations are clear bivariate outliers, suggesting that a thick-tailed distribution might be better suited to the data.

Table 2.2

Estimated posterior means and 90% credible intervals for fixed effects under alternative distributional assumptions for the error terms

Parameters	**Model 1-a** Gaussian level-1 errors		**Model 1-b** Student-t level-1 errors	
	Y^I	Y^A	Y^I	Y^A
Illiteracy	-0.03 (-0.94, 0.90)	0.76 (0.19, 1.35)	0.24 (-0.52, 1.01)	0.74 (0.18, 1.24)
Urbanization	-0.88 (-1.65, -0.13)	-0.14 (-0.58, 0.34)	-0.15 (-0.79, 0.49)	-0.17 (-0.62, 0.27)
FEAP	2.30 (0.64, 3.98)	-0.18 (-1.27, 0.87)	1.00 (-0.24, 2.24)	0.48 (-0.42, 1.40)
Candidates per seat	0.03 (0.01, 0.06)	0.01 (-0.01, 0.02)	0.02 (-0.01, 0.04)	0.01 (-0.01, 0.02)
Electorate	1.50 (0.62, 2.55)	0.74 (0.17, 1.30)	1.30 (0.42, 2.17)	0.47 (-0.15, 1.10)
Franchising	1.50 (0.70, 2.30)	0.66 (0.07, 1.27)	1.50 (0.75, 2.32)	0.49 (-0.04, 1.02)
Ballot	-0.04 (-0.82, 0.68)	-0.19 (-0.97, 0.60)	0.22 (-0.45, 0.93)	-0.27 (-0.98, 0.50)

Manipulation	0.52 (0.24, 0.85)	0.33 (0.03, 0.64)	0.41 (0.15, 0.67)	0.36 (0.07, 0.66)
Growth	4.30 (-2.40, 11.20)	-5.40 (-14.1, 2.90)	5.00 (-1.70, 11.60)	-5.50 (-13.70, 2.90)
Inflation	0.38 (0.23, 0.52)	-0.01 (-0.17, 0.17)	0.34 (0.20, 0.48)	-0.01 (-0.16, 0.18)
Intercept	-5.40 (-6.50, -4.30)	-1.90 (-3.02, -0.86)	-5.30 (-6.40, -4.20)	-2.0 (-3.10, -0.90)
N (first level)	388		388	
	557.90		242.30	
Bayes	-		9.79 ×	

[1] The Deviance Information Criterion (DIC) is computed as:

$$2\left(\frac{1}{J}\sum_{j=1}^{J}-2\log p\left(y\middle|\theta^{(j)}\right)\right)+2\log p\left(y\middle|\overline{\theta}\right),$$

with $\overline{\theta}=E\left(\theta\middle|y\right)$, the posterior mean of the model's parameters. Lower values of the DIC indicate better fit to the data.

[2] The Bayes factor for model M_j relative to model M_k is given by

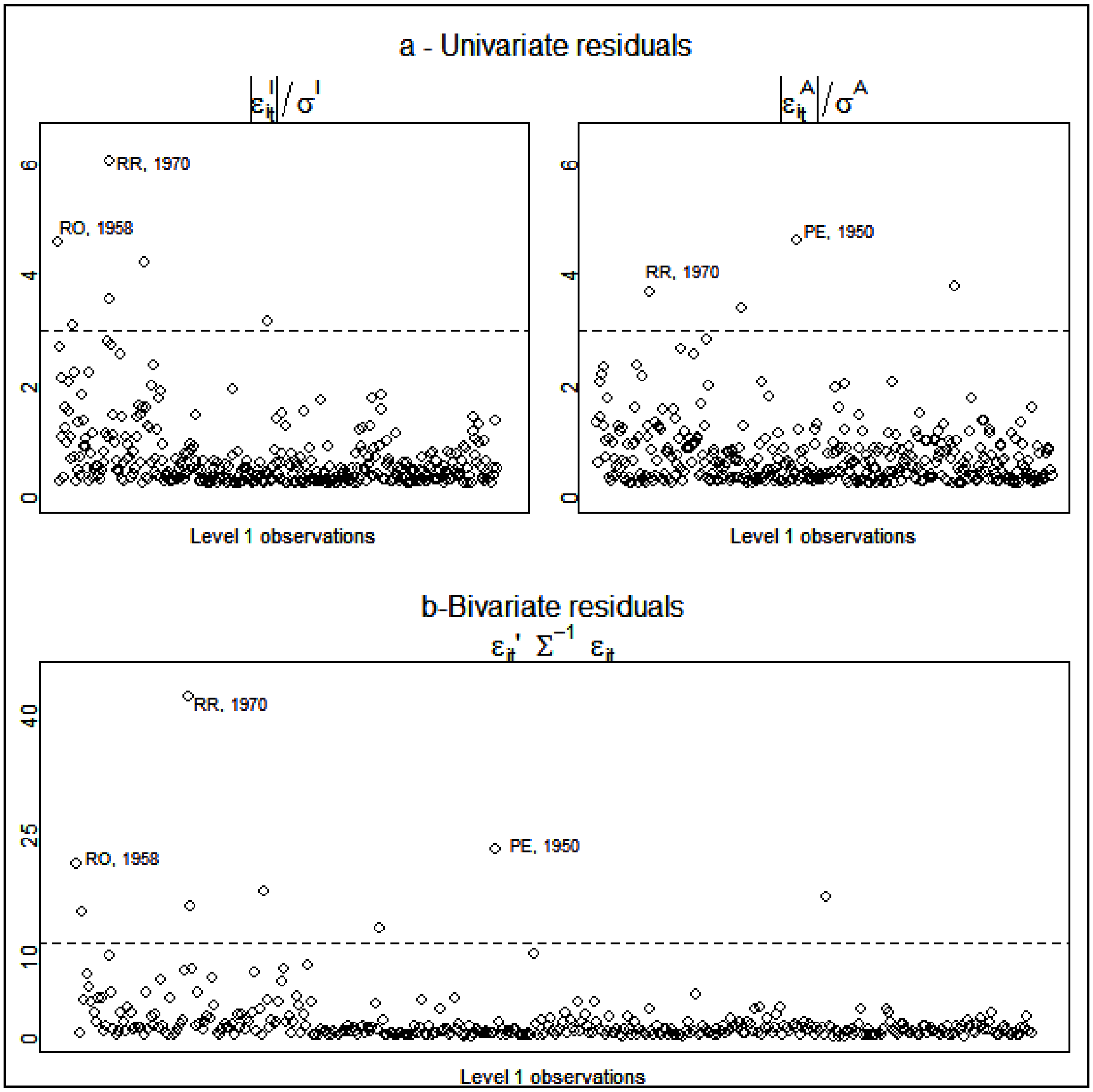
$$B_{j,k}=\frac{p\left(y\middle|M_j\right)}{p\left(y\middle|M_k\right)}=\frac{\int p\left(y\middle|M_j,\theta_j\right)p\left(\theta_j\middle|M_j\right)d\theta_j}{\int p\left(y\middle|M_k,\theta_k\right)p\left(\theta_k\middle|M_k\right)d\theta_k}.$$

I use the harmonic mean of the likelihood evaluated at the posterior draws of the parameters (Newton and Raftery, 1994; Rosa et al., 2003) as an estimate for $p\left(y\middle|M_x\right),\ \ x=j,k$:

$$\widehat{p}\left(y\mid M_x\right)=\left(\frac{1}{R}\sum_{r=1}^{R}p\left(y\mid\theta_x^{(r)}\right)^{-1}\right)^{-1}$$

Figure 2.3

Posterior means of the level-1 residuals from Model 1-a



Note: The figure plots the standardized residuals from Model 1-a, in absolute values. The univariate and bivariate residuals are computed based on the Bayesian statistics proposed by Weiss (1994):

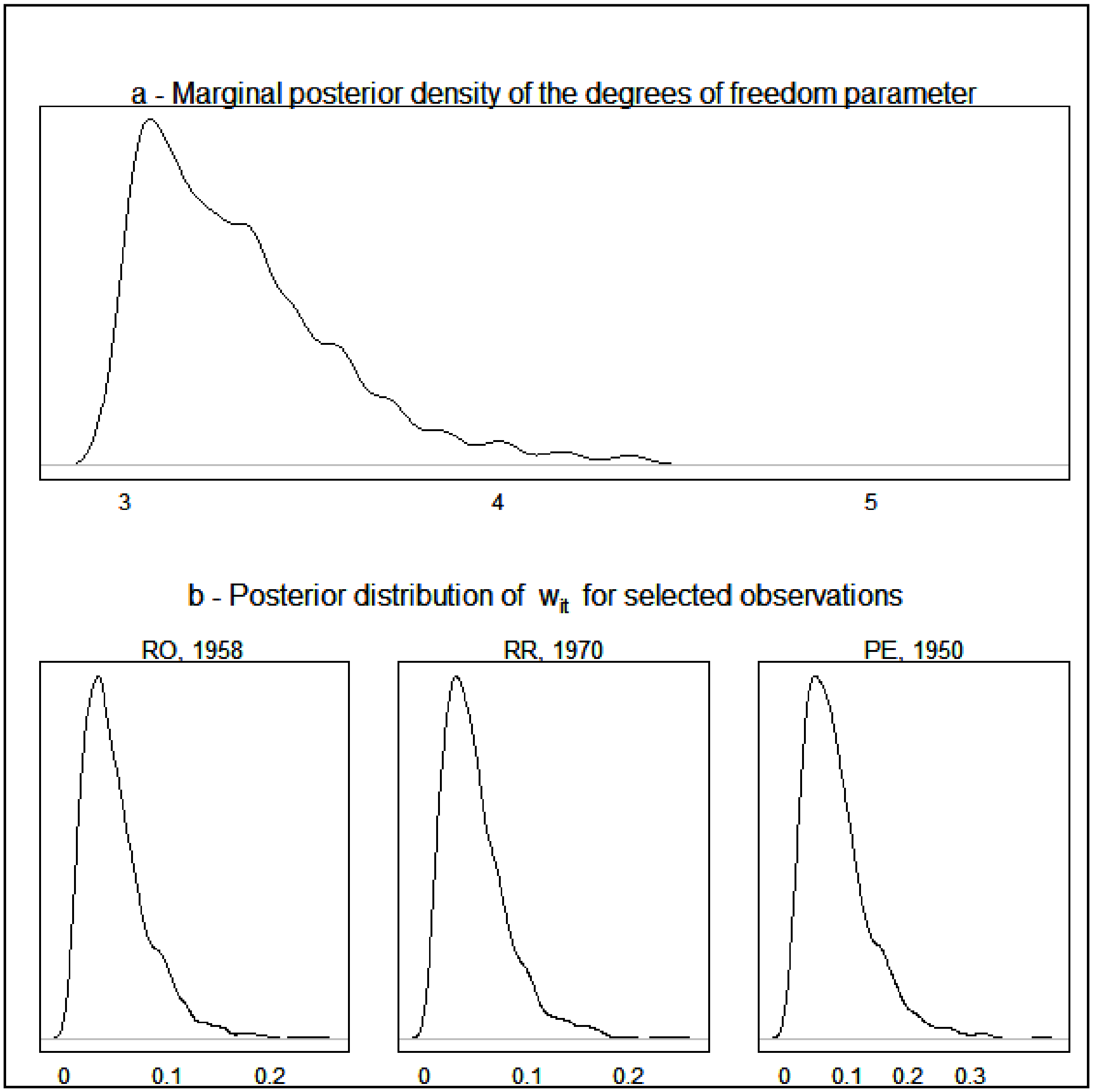
$$\frac{1}{J}\sum_{j=1}^{J}\frac{\left|Y_{i,t}^{s}-X_{i,t}\beta_{t}^{s(j)}-\lambda_{i}^{s(j)}\right|}{\sigma^{s(j)}},\ s=I,A\ \text{ and }\ \frac{1}{J}\sum_{j=1}^{J}\left(Y_{i,t}-X_{i,t}\beta_{t}^{(j)}-\lambda_{i}^{(j)}\right)'\Sigma^{-1(j)}\left(Y_{i,t}-X_{i,t}\beta_{t}^{(j)}-\lambda_{i}^{(j)}\right).$$

For the univariate residuals, the dashed horizontal lines correspond to the threshold of 3. For the bivariate residuals, the cutoff point is determined as $k=\chi^2_{2(1-\alpha)}$, $\alpha=2\times\Phi(-3)$.

In the same direction, the mean posterior estimate of υ under Model 1-b is 3.3, with its marginal posterior density concentrated around small values (Figure 2.4-a), indicating very strong departure from Normality and pointing to a heavy-tailed error distribution. As noted in Section 2.3, small values of υ determine that observations are weighted by an inverse function of the Mahalanobis distance $\left(Y_{i,t}-X_{i,t}\beta_t-\lambda_i\right)'\,\Sigma^{-1}\left(Y_{i,t}-X_{i,t}\beta_t-\lambda_i\right)$ adjusted by the degrees of freedom. Hence, for those observations identified as (bivariate) outliers in the model with Normal level-1 errors, the posterior probability that $w_{i,t}$ is equal or greater than 1 is negligible, as illustrated in Figure 2.4-b. Overall, the posterior probability that $P\left(w_{i,t}\right)\geq 1$ is less than 1% for roughly 6% of the observations in the sample, providing strong evidence of outliers (Congdon, 2003; Rosa et al, 2003). In addition, given that the "weight parameters" also reduce the influence of extreme observations on the posterior distribution of the election- and state- random coefficients, the number of level-2 bivariate outliers in Model 1-b is also halved with respect to Model 1-a, as shown in Figure 2.5.

Figure 2.4

Posterior densities of υ and $w_{i,t}$ under Model 1-b

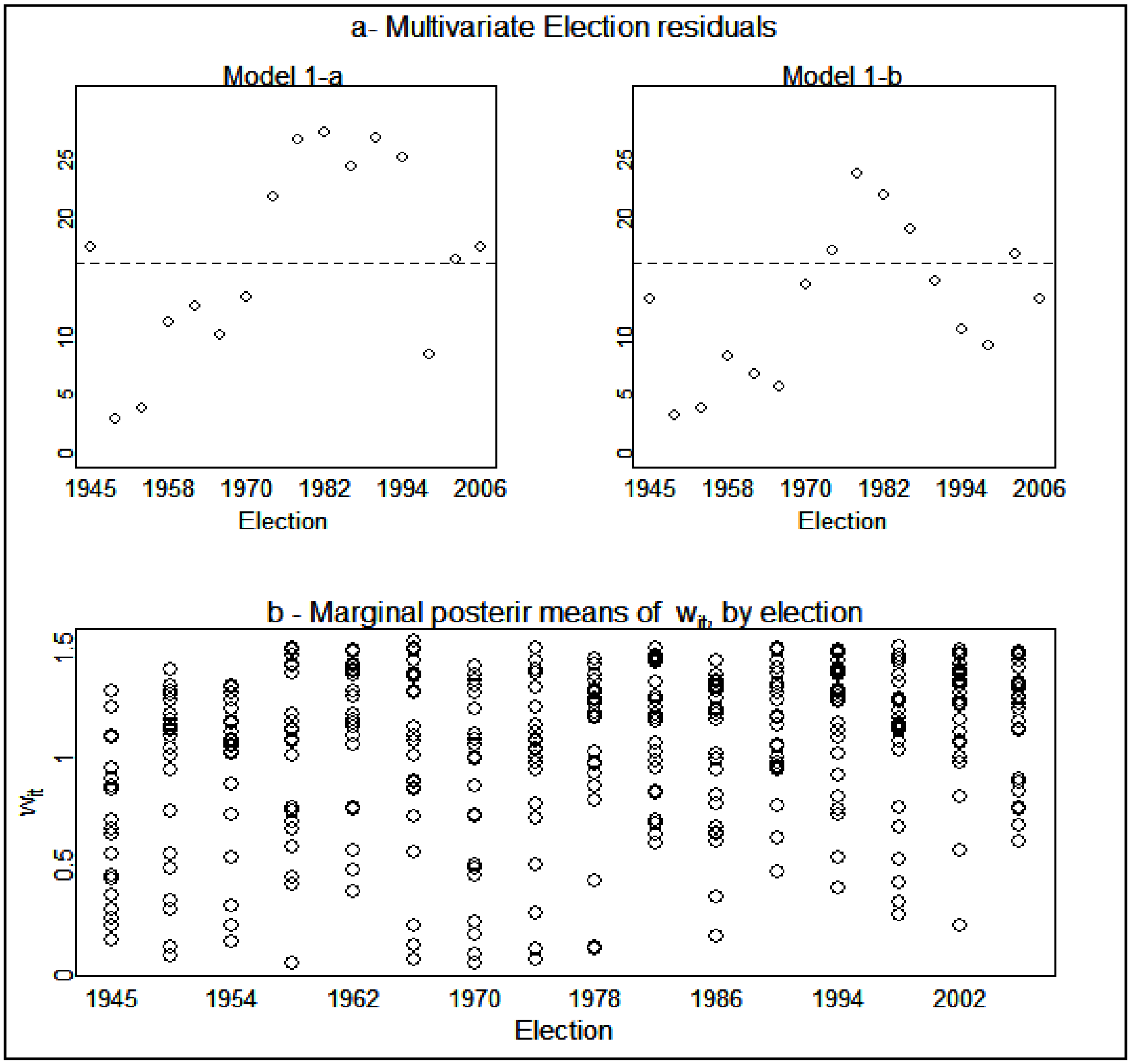


Note: The upper panel of the figure plots the posterior distribution of the degrees of freedom parameter υ of the Student-t distribution assumed for the error terms under Model 1-b. The lower panel plots the posterior distribution of the weight parameters $w_{i,t}$ for the states of Rondonia (RO), Roraima (RR) and Pernambuco (PE) in the 1958, 1970 and 1950 lower house elections.

Figure 2.5

Posterior means of the standardized election residuals and

marginal posterior means of the weight parameters



Note: Figure 5-a plots the posterior means of the standardized election residuals (in absolute values), computed as $\frac{1}{J}\sum_{j=1}^{J}\left(\beta_t^{(j)} - Z_t\delta^{(j)}\right)^{'}\Omega_{\eta}^{-1(j)}\left(\beta_t^{(j)} - Z_t\delta^{(j)}\right)$ (Weiss, 1994). The dashed horizontal lines correspond to the cutoff point $k = \chi^2_{4(1-\alpha)}$, $\alpha = 2\times\Phi(-3)$. Figure 5-b plots the marginal posterior means of the weight parameters $w_{i,t}$, by election.

Since the different comparison criteria examined above favor the model with Student-t errors, I focus on the results from Model 1-b in the remainder of the chapter. Table 2.3 reports the posterior distribution of the covariance components from the chosen model. The mean posterior correlation between the level-1 errors is moderately positive (0.24) and statistically significant at the 0.01 level, contradicting the assumption of no correlation underlying separate univariate analyses of invalid voting and absenteeism. Hence, states that experience higher relative proportions of invalid voting in an election than predicted by the model also exhibit higher relative proportions of electoral absenteeism. In addition, the bottom panel of Table 2.3 reveals that there is considerable variation in the election effects beyond that explained by the national-level variables included in the model. While the average correlation in Y^I and Y^A within states across elections are 0.28 and 0.24, respectively, the corresponding intra-election correlations between states are as large as 0.57 and 0.75, suggesting that election-specific circumstances have a substantial influence on both forms of abstention.

Table 2.3

Posterior means of variance-covariance components under Model 1-b

Level 1-errors

	Invalid Voting	**Absenteeism**
Invalid Voting	0.17 (0.03, 0.54)	
Absenteeism	0.03 (0.01, 0.09)	0.08 (0.01, 0.26)

Level 2: State random effects

	Invalid Voting	**Absenteeism**
Invalid Voting	0.16 (0.09, 0.26)	
Absenteeism	0.01 (-0.04, 0.05)	0.11 (0.07, 0.16)

Level 2: Election random effects

	Invalid voting Intercept	**Invalid voting** Illiteracy	**Absenteeism** Intercept	**Absenteeism** Illiteracy
Invalid voting Intercept	0.27 (0.12, 0.51)			
Invalid voting Illiteracy	-0.06 (-0.32, 0.14)	0.62 (0.24, 1.27)		
Absenteeism Intercept	0.05 (-0.14, 0.27)	-0.07 (-0.39, 0.23)	0.50 (0.22, 0.96)	
Absenteeism Illiteracy	-0.01 (-0.18, 0.14)	0.17 (-0.06, 0.52)	-0.12 (-0.40, 0.09)	0.29 (0.12, 0.56)

Note: 90% credible intervals reported in parenthesis.

Based on the convergent Gibbs samples of the parameters of Model 1-b, I estimate the average effect of a one-unit change in each of the state-level and national-level predictors on the proportion of invalid ballots and electoral absenteeism.[21] The results, reported in Table 2.4, reveal some interesting discrepancies regarding the determinants of the two sources of abstention. While only *Illiteracy* had a positive and significant effect on electoral absenteeism at the usual confidence levels, invalid voting in Brazil's lower house elections was strongly and positively related both to the average levels of education and skills among the electorate and to political protest. The proportion of blank and spoiled ballots rose by 0.09 percentage points for each percentage-point increase in the share of illiterates in the voting-age population, and it further rose by more than 6 points on average with the extension of suffrage to illiterates in 1985. The addition of new voters was also positively related to invalid ballots: each percent increase in the fraction of the states' population eligible to vote was associated to a 0.13 percentage-point rise in blank and null votes. Among the protest variables, higher levels of authoritarian political engineering resulted in an average increase of 3.4 percentage points in invalid voting. Although electoral manipulation also boosted illegal abstention, the impact of this predictor on absenteeism was much more variable across states and elections. The positive and significant effect of *Inflation* on invalid voting suggests that blank and null ballots might reflect not only popular dissatisfaction with inadequate representative institutions (Schwartzman, 1973; Lima, 1994), but also discontent with poor macroeconomic performance and economic

[21] In the case of the two binary variables, *Ballot* and *Franchising*, the effect is measured as a change from 0 to 1.

mismanagement by the political elites. While these results provide evidence in support of the “protest hypothesis” of invalid voting, they also suggest that less educated and newly enfranchised voters in Brazil face considerable barriers to voting (Power, 1991). The evidence is far less conclusive in the case of electoral absenteeism, underscoring the need to examine additional factors that might affect noncompliance with compulsory voting laws.

Remarkably, while all the socio-economic variables tend to affect both sources of abstention in the same direction, many of the institutional and protest variables exhibit opposite average effects on the two forms of non-voting. In particular, two relevant institutional features of the open-list PR system used in Brazil’s lower house election, namely, a large number of candidates running for office and the introduction of the single official ballot, have a positive impact on increase invalid voting but a negative average effect on illegal abstention. The opposite effect of *Ballot* and *Candidates per seat* on the two forms of non-voting suggests that there might be a certain trade-off between attracting voters to the polls and facilitating effective electoral participation. Factors that give voters more opportunities to influence electoral results ex-ante, such as the availability of more electoral options and a ballot design that gives voters more freedom to choose their preferred candidate, tend to increase turnout. However, at the moment of casting a vote, the proliferation of candidates and the requirement that voters record their preferred candidate’s name or registration number on the paper ballot tend to increase invalid voting, probably because they impose considerable informational requirements and heavy decision-making costs on the electorate, especially in the context of high illiteracy rates and massive expansion of the franchise experienced in Brazil throughout the century.

Table 2.4

Effect of a one-unit change in the predictors on invalid voting and absenteeism under Model 1-b (in percentage points)[a,b]

Predictor	Effect on Invalid voting	Effect on electoral absenteeism
Illiteracy	0.09* (0.05)	0.13** (0.05)
Urbanization	-0.02 (0.04)	-0.03 (0.04)
Females in EAP	0.10 (0.08)	0.06 (0.09)
Candidates per seat	0.11 (0.17)	-0.26 (0.17)
Electorate	0.13** (0.06)	0.07 (0.06)
Franchising	6.15*** (2.57)	1.03 (2.35)
Official Ballot	2.73 (4.17)	-5.85 (8.09)
Electoral manipulation	3.37* (2.21)	4.52 (3.44)
Growth	0.67 (0.45)	-1.00 (0.82)
Inflation	3.83*** (1.24)	-0.82 (1.72)

[a] Standard errors are reported in parenthesis.

[b] Significance levels: *** 0.01, ** 0.05, *0.1.

2.4.3 Comparison with alternative modeling approaches

In order to illustrate the differences between the model presented here and alternative approaches used to analyze abstention in compulsory voting systems, Figure 2.6 below contrasts the average causal effects of the predictors on invalid voting and absenteeism under Model 1-b with those obtained under three models that fail to account for the compositional and/or the hierarchical structure of the data. Model 2 uses separate ordinary least squares regressions for invalid voting and absenteeism, assuming independence among observations and simply pooling state-level and country-level predictors by assigning the values of the national variables to all the states in a given election. Model 3 uses separate hierarchical linear models for invalid voting and absenteeism, accounting for the temporal and geographical clustering of the data but ignoring the non-negativity and unit-sum constraints (2.1) and (2.2). Finally, Model 4 is a compositional model with random intercepts for each state but no election-random effects, again assuming a deterministic relationship between national- and state-level predictors. The specifications of Models 2, 3 and 4 are detailed in Appendix 2.B. [22]

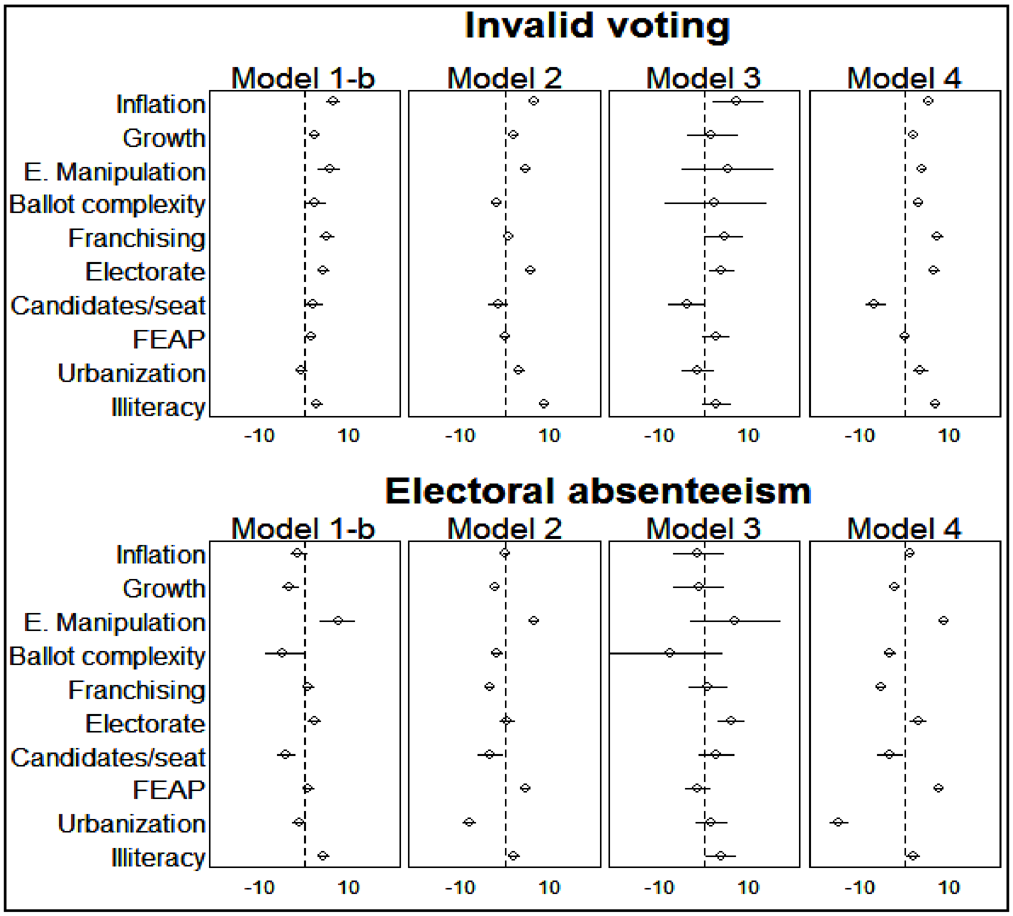
---

[22] Models 3 and 4 were fitted by MCMC simulations (Gibbs Sampling), using a normal/independent distribution for the data model, Gaussian priors for the random coefficients and diffuse conjugate priors for the hyperparameters. The substantive results remain unchanged if Gaussian level-1 errors are assumed. Details of the estimation are available from the author upon request.

Figure 2.6

Estimated marginal effects of the predictors across models

(in percentage points)



Note: The graph shows the effect of a one-unit change in each of the predictors on invalid voting and electoral absenteeism. The center dots correspond to the point estimates, the thicker lines to the 50% credible intervals, and the thinner lines to the 90% credible intervals.

The results reported in Figure 2.6 shows some noticeable differences between the four models. As seen in the upper and lower panels, the standard errors of the marginal effects of the covariates on both sources of abstention under Model 1-b tend to be considerably smaller than for Model 3 and much larger than for Models 2 and 4, particularly in the case of the country-level variables. This leads to different conclusions about the relative size and the statistical significance of the impact of the national-level predictors on invalid voting and electoral absenteeism under the different models. For instance, setting the stochastic terms in η_t to zero in Models 2 and 4 leads to significant effects of economic growth on both sources of abstention at the 0.01 level. In contrast, *Growth* has no systematic effect on either source of abstention under Models 1 and 3. At the other extreme, the large standard errors for the country-level comparisons under Model 3 determine that none of national-level variables has a significant effect on either source of abstention at the usual confidence levels.

More importantly, the four models lead to different substantive conclusions regarding the impact of some of the variables on the two sources of abstention. As seen in the lower panel of Figure 2.6, the results from Models 2 and 4 show that that the extension of voting rights to illiterates led to significantly lower levels of electoral absenteeism, suggesting that this group of new voters was more likely to show up at the polls even when, unlike for literate citizens between 18 and 70 years of age, voting is optional for illiterates. While inferences drawn from these two models tend to support the claim that “the fact that voting is…optional for illiterates seems to have little practical effect on their observance of mandatory voting” (Power and Roberts, 1995, p. 800), the average effect of *Franchising* on electoral absenteeism has the opposite sign under Model 1-b. Also, while a higher number of *Candidates per seat* has a positive average effect on invalid voting under Model 1-b, suggesting that a larger number of contestants increases the likelihood of voter error and/or makes it more difficult for voters to choose a single preferred candidate, this relationship is negative under Models 3 and 4. Finally, under Model 2, *Ballot* has a negative and statistically significant effect on invalid voting, leading to the rather implausible conclusion that the introduction of a more complex ballot system that requires considerable more information on the part of voters resulted in lower rates of blank and spoiled ballots. These examples illustrate the fact that some of the inferences drawn from the model developed in this chapter contradict the results both from the separate univariate analyses (Models 2 and 3) and from an analysis that ignores election-to-election variability in both sources of abstention beyond that explained by national-level variables (Model 4). The conflicting results from the different models lead to different conclusions about the relative validity of the alternative theories proposed to account for abstention under mandatory voting and might entail very different implications regarding, for instance, the design of electoral systems and the institutional reforms needed to promote and consolidate political participation in compulsory voting systems (McAllister and Makkai, 1993; Power and Roberts, 1995).

In order to compare the fit of the four models, I use posterior predictive simulations (Gelman et al., 2004, Gelman and Hill, 2007). Following Iyengar and Dey (2004), a plausible comparison criteria based on the discrepancy between observed and simulated data would favor the model that minimizes the predictive loss

$d\left(P^{\mathrm{Rep}}, P^{Obs}\right) = E\left(\left\|P^{\mathrm{Rep}} - P^{Obs}\right\|^2 \middle| P^{Obs}\right)$, where $P_{i,t}^{Rep} = \left(P_{i,t}^{Rep(1)}, \ldots, P_{i,t}^{Rep(J)}\right)$ denotes the replicate data sampled from the predictive distribution $p\left(P_{i,t}^{Rep} \mid P_{i,t}^{Obs}\right) = \int p\left(P_{i,t}^{Rep} \mid \theta\right) p\left(\theta \mid P_{i,t}^{Obs}\right) d\theta$ under each model.[23] The posterior predictive loss d can then be estimated as:

$$\hat{d} = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\frac{1}{J} \sum_{j=1}^{J} \left\| P_{i,t}^{Obs} - P_{i,t}^{\mathrm{Rep}(j)} \right\|^2 \right) \qquad (2.26).$$

Table 2.5 reports the estimates and 90% credible intervals for the posterior predictive loss based on $J = 1{,}000$ hypothetical replications of $P_{i,t}^{I}$ and $P_{i,t}^{A}$ for the four models. The compositional-hierarchical model exhibits the lowest discrepancy between the replicated and the actual data (at the 0.01 level). In contrast, the two models that implement separate univariate analyses for each source of abstention have the highest estimated predicted losses. In particular, Model 2, which in addition ignores the multilevel nature of the data, exhibits the worst fit. The superior performance of Model 1-b is also illustrated in Figure 2.7, which plots the actual proportions of invalid voting and absenteeism and the expected proportions under the four models, obtained by averaging $P_{i,t}^{\mathrm{Rep}(j)}$, $j = 1, \ldots, 1000$, over the simulations. As seen in the figure, Models 2 and 3 lead to negative expected proportions of invalid votes for 49% and 14% of the state-years in the sample, respectively. While both

---

[23] In the case of the compositional-hierarchical model, $P_{i,t}^{\mathrm{Rep}}$ are obtained from $Y_{i,t}^{Rep}$ using the logarithmic transformations (1.3) and (1.4).

compositional models avoid this problem, relaxing the assumption of a deterministic relationship between national- and state-level predictors and allowing for additional variability in the election effects results in a better fit for Model 1-b *vis-à-vis* Model 4. Hence, the evidence presented above indicates that the statistical model developed in this chapter provides a much improved fit over the other three modeling approaches considered, and reveals that the methodological differences between these competing empirical strategies have substantial consequences in terms of the analysis of the determinants of abstention under compulsory voting.

Table 2.5

Estimates of the Posterior Predictive Loss

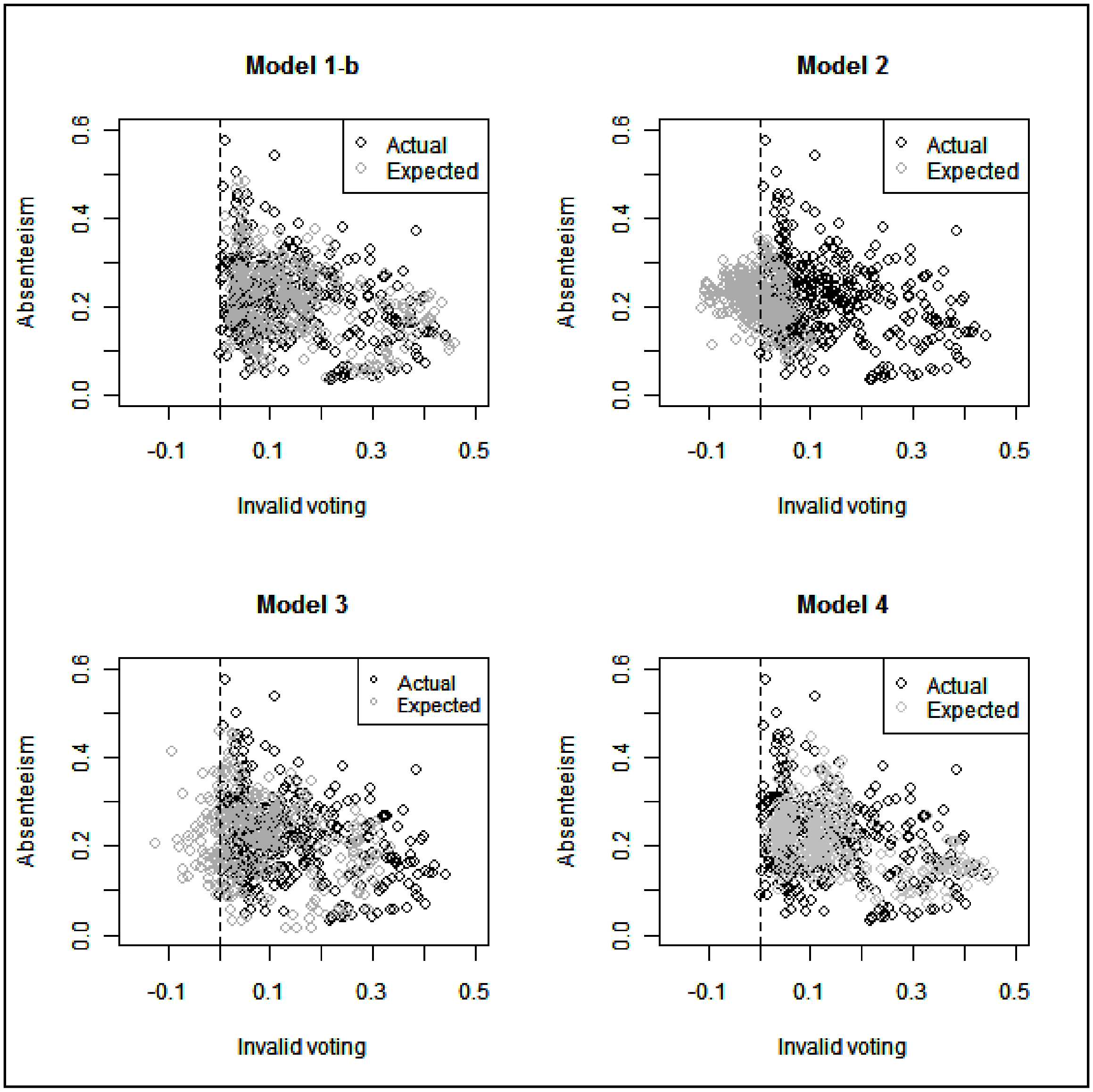
for alternative modeling

Model	$\hat{d}$
1-b	2.84 (2.33, 3.51)
2	16.71 (13.72, 19.97)
3	8.33 (7.33, 9.44)
4	6.59 (5.75, 7.55)

% credible intervals reported in parenthesis.

Figure 2.7

Actual and expected proportions of invalid voting and electoral absenteeism under alternative modeling strategies



Note: The gray circles correspond to the expected proportion of invalid voting and electoral absenteeism for each state-election of the sample for the model under consideration. The black circles correspond to the actual values.

2.5 Concluding remarks

Different theories, drawing on the literature on voter turnout in industrialized democracies, have been proposed in order to account for the phenomena of invalid voting and electoral absenteeism under mandatory voting. This chapter integrates the socioeconomic, institutional, and political-protest approaches in a statistical model aimed at analyzing the determinants of both sources of abstention in district-level elections. The model presented here accounts for the compositional and hierarchical structure of district-level electoral data and easily accommodates sensitivity analysis, encompassing a family of thick-tailed distributions that can be used for robust inference.

Results obtained from the application of the model to analyze abstention in Brazil's legislative elections allow drawing interesting substantive and methodological conclusions. The evidence presented above reveals substantial differences in the determinants of both forms of non-voting. In line with Power and Roberts (1995), I find that the proportion of blank and null ballots in Brazil's lower house elections was strongly positively related both to political protest and to the existence of important informational barriers to voting, in particular for less educated and newly enfranchised voters. The influence of these variables on illegal abstention, however, was less evident. In addition, some of the institutional characteristics of the electoral system, such as the proliferation of candidates and the introduction of a complex ballot design, seem to affect the two sources of abstention in opposite directions. Comparisons based on posterior simulations indicate that the model presented here fits the data considerably better than several alternative empirical strategies used to analyze abstention under compulsory voting. More importantly, the main conclusions and the policy implications resulting from the compositional-hierarchical model might differ significantly from those drawn using less appropriate modeling approaches prevailing in previous research in this area.

Although the model was applied to the particular case of Brazil, it provides a general tool to analyze the determinants of abstention in compulsory systems. Also, the mixed model presented in Section 2.3 can be modified in order to accommodate other possible distributions of the error terms at each level of the hierarchy (Andrew and Mallows, 1974; West, 1984; Seltzer et al., 2002; Rosa et al. 2003). An immediate extension of the chapter would be to include a larger number of countries and additional covariates in order to analyze the performance of the model and the robustness of the results from a comparative politics perspective. From a methodological standpoint, using non-parametric methods to estimate the joint density of invalid voting and absenteeism would allow examining their determinants and interactions without imposing specific parametric distributions.

Appendix 2.A

Algorithm implemented to compute the causal effects

Let $\left(\left\{\eta_t^{(j)}\right\},\left\{\lambda_i^{(j)}\right\},\delta^{(j)},\Sigma^{-1(j)},\Omega_\eta^{-1(j)},\Omega_\xi^{-1(j)},\left\{w_{i,t}^{(j)}\right\},\upsilon^{(j)}\right)$, $j=1,...,J$, denote convergent samples generated from (2.14)-(2.21). In order to compute the average effect of each of the independent variables on invalid voting and electoral absenteeism, the following algorithm is implemented (Katz and King, 1999; Bhaumik, Dey and Ravishanker, 2003; Gelman and Hill, 2007):

1. Samples of the estimated expected proportions of invalid voting and absenteeism in each district-year for given covariates are calculated using the additive logistic transformations (2.3) and (2.4):

$$\widetilde{P}_{i,t}^{I(j)}=\frac{\exp\left[\widetilde{Y}_{i,t}^{I(j)}\right]}{1+\exp\left[\widetilde{Y}_{i,t}^{I(j)}\right]+\exp\left[\widetilde{Y}_{i,t}^{A(j)}\right]},\quad j=1,...,J$$

$$\widetilde{P}_{i,t}^{A(j)}=\frac{\exp\left[\widetilde{Y}_{i,t}^{A(j)}\right]}{1+\exp\left[\widetilde{Y}_{i,t}^{I(j)}\right]+\exp\left[\widetilde{Y}_{i,t}^{A(j)}\right]},\ j=1,...,J$$

where:

$$\widetilde{Y}_{i,t}^{s(j)}=\delta_{0,0}^{s(j)}+\sum_{k=1}^{K}\delta_{k,0}^{s(j)}x_{i,t,k}+\sum_{l=1}^{L}\delta_{0,l}^{s(j)}z_{t,l}+\left[\sum_{k=1}^{K}\left(\sum_{l=1}^{L}\delta_{k,l}^{s(j)}z_{t,l}\right)x_{i,t,k}\right]+\eta_{0,t}^{s(j)}+\sum_{k=1}^{K}\eta_{k,t}^{s(j)}x_{i,t,k}+\lambda_i^{s(j)},\quad s=I,A,$$

and $x_{i,t}, z_t$ are vectors of observed district-level and country-level predictors.

2. Step 1 is repeated after changing the value of the predictor whose effect is analyzed by 1 unit, while keeping all other regressors at their observed levels, obtaining $\widehat{P}_{i,t}^{I(j)}$ and $\widehat{P}_{i,t}^{A(j)}$, $j = 1, \ldots J$.
3. The average effect of the predictor on invalid voting and absenteeism for all district-years in the sample can be estimated by averaging $\widehat{P}_{i,t}^{I(j)} - \widetilde{P}_{i,t}^{I(j)}$ and $\widehat{P}_{i,t}^{A(j)} - \widetilde{P}_{i,t}^{A(j)}$ over all $i = 1, \ldots, n;\ t = 1, \ldots, T;\ j = 1, \ldots, J$ (Bhaumik, Dey and Ravishanker, 2003; Gelman and Pardoe, 2007). Credible intervals summarizing the approximate distribution of the causal effects can also be easily constructed using standard methods from sampling theory (Gelman and Pardoe, 2007).

Appendix 2.B

Alternative strategies to modeling invalid voting and electoral absenteeism

Model 2:

$$P_{i,t}^{s} = \delta_0^s + \left(\delta_1^s + \delta_2^s Franchising_t\right) Illiteracy_{i,t} + \delta_3^s Urbanization_{i,t} + \delta_4^s FEAP_{i,t} + \\ \delta_5^s Candidates\ per\ Seat_{i,t} + \delta_6^s Electorate_{i,t} + \delta_7^s Manipulation_t + \delta_8^s Ballot_t + \\ \delta_9^s Growth_t + \delta_{10}^s Inflation_t + \varepsilon_{i,t}^s, \qquad s=I,A. \tag{2.B.1}$$

Model 3:

$$P_{i,t}^{s} = \beta_{0,t}^s + \beta_{1,t}^s Illiteracy_{i,t} + \beta_{2,t}^s Urbanization_{i,t} + \beta_{3,t}^s FEAP_{i,t} + \\ \beta_{4,t}^s Candidates\ per\ Seat_{i,t} + \beta_{5,t}^s Electorate_{i,t} + \lambda_i^s + \frac{\varepsilon_{i,t}^s}{\sqrt{w_{i,t}^s}}, \qquad s=I,A \tag{2.B.2}$$

$$\beta_{0,t}^s = \delta_{0,0}^s + \delta_{0,1}^s Ballot_t + \delta_{0,2}^s Manipulation_t + \delta_{0,3}^s Growth_t + \delta_{0,4}^s Inflation_t + \eta_{0,t}^s, \quad s=I,A \tag{2.B.3}$$

$$\beta_{1,t}^s = \delta_{1,0}^s + \delta_{1,1}^s Franchising_t + \eta_{1,t}^s \qquad s=I,A \tag{2.B.4}$$

$$\beta_{k,t}^s = \delta_{k,0}^s \qquad s=I,A; k=2,\ldots,5 \tag{2.B.5}$$

with $\varepsilon_{i,t}^s \sim N\left(0,\sigma_{\varepsilon_s}^2\right), \eta_t^s \sim N\left(0,\Omega_{\eta_s}\right), p\left(w_{i,t}^s \middle| \upsilon\right) = Gamma\left(\frac{\upsilon}{2},\frac{\upsilon}{2}\right).$

Model 4:

$$Y_{i,t}^{s} = \delta_0^s + \left(\delta_1^s + \delta_2^s Franchising_t\right) Illiteracy_{i,t} + \delta_3^s Urbanization_{i,t} + \delta_4^s FEAP_{i,t} + \\ \delta_5^s Candidates\ per\ Seat_{i,t} + \delta_6^s Electorate_{i,t} + \delta_7^s Manipulation_t + \delta_8^s Ballot_t + \\ \delta_9^s Growth_t + \delta_{10}^s Inflation_t + \lambda_i^s + \frac{\varepsilon_{i,t}^s}{\sqrt{w_{i,t}}}, \qquad s=I,A \tag{2.B.6}$$

with $\left[\varepsilon_{i,t}^{I},\varepsilon_{i,t}^{A}\right]\sim N\left(0,\Sigma\right),\left[\lambda_{i}^{I},\lambda_{i}^{A}\right]\sim N\left(0,\Omega_{\lambda}\right)$, $p\left(w_{i,t}\middle|\upsilon\right)=Gamma\left(\frac{\upsilon}{2},\frac{\upsilon}{2}\right)$,

and $P_{i,t}=\left[P_{i,t}^{I},P_{i,t}^{A}\right]$ obtained from $Y_{i,t}=\left[Y_{i,t}^{I},Y_{i,t}^{A}\right]'$ using (2.3) and (2.4).

Chapter 3

Structural Cleavages, Electoral Competition and Partisan Divide: a Bayesian Multinomial Probit Analysis of Chile's 2005 Election[24]

3.1 Introduction

Chile’s post-authoritarian party structure, dominated by two stable and solid multiparty coalitions, contrasts with the highly fragmented system existing prior to the 1973 military coup (Valenzuela and Scully, 1997; Tironi and Agüero, 1999; Alemán and Saiegh, 2007). Since the re-establishment of democracy, the center-left *Concertación* coalition, comprising the Socialist Party (PS), the Party for Democracy (PD), the Christian Democrats (CD) and the Radical Social-Democratic Party (PRSD), has been in control of the presidency and held the majority of the legislative seats. The other major coalition, the conservative *Alianza por Chile,* is made up of the Independent Democratic Union (UDI), the National Renewal Party (RN) and the Centrist Union (UCC). Although other minor parties exist outside these blocks, the two coalitions have dominated contemporary politics in Chile.

There has been considerable debate among scholars about the reshaping of the Chilean political system and about the relative influence of different factors on voters’ behavior in this new setting (Valenzuela and Scully, 1997; Tironi and Agüero, 1999; Torcal and

[24] Joint with R. Michael Alvarez. Both authors contributed equally to the following chapter, which was published in *Electoral Studies* 28(2), 177 – 189, 2009.

Mainwaring, 2003). Some authors argue that the social and cultural cleavages (in particular, class and religious divisions) that originally structured the Chilean political system still play a predominant role in defining political identities, and that the division between supporters and opponents of the authoritarian regime that marked the democratic transition was the result of a particular historical background and is likely to fade away as democracy is consolidated (Scully, 1995; Valenzuela, 1999; Bonilla, 2002). Other researchers, however, maintain that the new authoritarian-democratic cleavage has come to dominate party competition, integrating and reorganizing traditional sources of partisan divide and reflecting intense discrepancies about regime preferences and conceptions of democracy in the Chilean society that are likely to subsist (Tironi and Agüero, 1999; Torcal and Mainwaring, 2003).

The 2005 Presidential election offers an especially interesting opportunity to test these alternative explanations in an electoral setting, while at the same time exhibiting distinctive characteristics that bring about substantive and methodological implications that have received little attention in the literature on voter behavior in Chile. It was the fourth Presidential election since Chile's return to democracy, held at a time of continuing economic growth and high popularity of the incumbent *Concertación* government, and with Pinochet relegated to a marginal role in the national political scene (Bonilla, 2002; Angell and Reig, 2006). Also, for the first time in its history, the two main partners of the *Alianza por Chile*, UDI and RN, presented independent candidates who adopted relatively different electoral strategies: while Lavín (UDI) adopted an aggressive campaigning style aimed at consolidating the vote among his right-wing supporters, the candidate of the National Renewal Party, Sebastián Piñera, took a more moderate stance, distancing himself from the traditional right and the legacy of the military regime in order to capture the support of centre in view of the almost certain second-round runoff between the candidate of the *Concertación* and one of the two conservative candidates (Angell and Reig, 2006; Gamboa and Segovia, 2006). Together with the formation of the left-wing alliance *Juntos Podemos Más*, this resulted in relatively clear leftist (*Juntos Podemos Más*), center-left (*Concertación*), center-right (RN) and rightist (UDI) electoral options available for Chilean voters.

In order to analyze the relative influence of socio-demographic, ideological and political variables on voter choice at the individual level, we specify and estimate a Bayesian multinomial probit model that explicitly accounts for the multi-party character of the election by letting voters evaluate all competing candidates simultaneously and to 'group' alternatives they consider similar when choosing for which candidate to vote. Our model allows testing the relative validity of the competing theories in explaining voters' electoral behavior. In addition, it enables us to examine other factors that might have had substantial influence in this particular election, such as the presence of a second conservative contestant and its effect on voters' behavior in the view of the second-round runoff.

Therefore, the chapter offers two important contributions with respect to prior studies of the Chilean case. First, while past research analyzed citizens' party identification or vote intention (Frei, 2003; Torcal and Mainwaring, 2003), no study has so far examined actual vote choice at the individual level. Theoretical and empirical arguments indicate that party identification and vote intention are dynamic concepts influenced by election-specific circumstances and campaign effects, and that there is no linear relationship between party preferences and actual vote (Franklin and Jackson, 1983; Alvarez, 1998; Hillygus and

Jackman, 2003). In the case of Chile's 2005 election, held in a context of declining party identification among the electorate and increasing number of respondents not expressing any vote intention in opinion polls (Frei, 2003), short-term factors such as candidates' campaigning style and the impressive economic record of President Lagos' administration might have played a considerable influence on voters' decisions (Angell and Reig, 2006; Navia, 2006).

Second, all previous individual-level studies of candidate choice in Chile (Frei, 2003; Torcal and Mainwaring, 2003) employed binary choice models, restricting comparisons to pairs of parties and imposing the independence of irrelevant alternatives (IIA) property on voters. The IIA condition is a very restrictive assumption to make about voters' electoral behavior, in that it implies that the probability of a voter choosing an electoral alternative is independent of the other alternatives available and of the characteristics of these other alternatives (Alvarez, Nagler and Bowler, 2000; Train, 2003); in particular, the presence or absence of the candidate of the RN in the election would not change the relative probabilities of choosing any of the other candidates. Thus, imposing the IIA condition neglects the possibility that centrist voters who were disenchanted with the *Concertación* but were not willing to vote for a clear right-wing candidate might find a moderate conservative candidate attractive. Also, it implies that an *Alianza* supporter could not see the candidates of the UDI and the RN as substitutes, an assumption that is at odds with the view that coalition labels are meaningful for Chilean voters (Huneeus, 2006; Alemán and Saiegh, 2007) and that might have been particularly inappropriate in the context of the 2005 election, when the declining popularity of Lavín and the better prospects of Piñera in a second-round runoff against Bachelet (Gamboa and Segovia, 2006) might have driven

UDI sympathizers to vote for the RN candidate for tactical reasons. The potential for strategic voting in the 2005 election was substantially increased due to the fact that opinion polls close to the election date indicated that a *ballotage* between Bachelet and one of the conservative candidates was almost certain, and that the contest between Lavín and Piñera for the second place in the first round was very tight (Angell and Reig, 2006).

Even if relaxing the IIA condition might not necessarily improve the model fit or lead to substantially different results regarding the determinants of voter choice (Horowitz, 1980; Quinn and Martin, 1998), it allows addressing central substantive questions for the analysis of Chile's 2005 election, namely whether Piñera's entry into the race was determinant in bolstering *Alianza*'s vote support, and how it affected voters' electoral behavior. While, prior to the election, *Alianza* leaders expressed concerns that the divisions between the two conservative candidates could weaken the right-wing coalition (Gamboa and Segovia, 2006), Piñera's candidacy might in fact have contributed to its relative success in the presidential election, in which the right did considerably better than in the simultaneous legislative election and obtained more votes than the *Concertación* for the first time since Chile's return to democracy (Navia, 2006). The impact of Piñera's candidature on the election cannot be directly quantified using vote choice models that rely on the independence of irrelevant alternatives property such as the multinomial logit (Dow and Endersby, 2004). Therefore, these relevant questions have not been addressed in previous analyses of the 2005 election.

In view of the computational complexity of fitting the multinomial probit model (Train, 2003), there have been relatively few applications of this model in the political science literature (e.g., Alvarez and Nagler, 1995; Alvarez, Nagler and Bowler, 2000; Dow and

Endersby, 2004). Most applications have used maximum likelihood estimation, relying on asymptotic normality in making inferences about the error variance and covariance parameters. As shown by McCulloch and Rossi (1994), however, asymptotic approximations are quite problematic in the context of the multinomial probit model. The main advantage of the Bayesian approach based on Gibbs sampling is that it allows obtaining arbitrarily precise approximations to the posterior densities, without relying on large-sample theory (McCulloch and Rossi, 1993; Jackman, 2004). In addition, it avoids direct evaluation of the likelihood function and the resulting convergence problems exhibited by maximum likelihood optimization, and is computationally more efficient than simulation-based methods of classical estimation (Kim, Kim and Heo, 2003). Hence, the Bayesian approach overcomes some of the main criticisms that have been leveled against the use of multinomial probit in electoral studies (Dow and Endersby, 2004). Furthermore, since the Bayesian framework allows for straightforward comparisons of models that can be used to operationalize alternative sets of hypothesis (Quinn and Martin, 1998), it is particularly well suited to examine the relative validity of the different explanations proposed to account for voters' behavior in Chile.

The remainder of the chapter is organized as follows. Section 3.2 presents an initial look at voting behavior in the 2005 Presidential election using survey data. Section 3.3 presents a multinomial probit model to analyze voter-choice in multi-party elections and describes the data and methodology used to fit the model to the Chilean case. Section 3.4 presents the most salient results, and Section 3.5 concludes.

3.2 A first look at Chile's 2005 presidential election

Using data from the Comparative Study of Electoral Systems post-election survey (CSES, 2007), we provide preliminary evidence regarding the impact of different sets of variables on the support for each of the candidates running for office in the 2005 election: Michelle Bachelet, of the governing *Concertación*; Tomás Hirsch, of the left-wing coalition *Juntos Podemos Más* (JPM); and the two *Alianza* candidates, Joaquín Lavin (UDI) and Sebastián Piñera (RN). Table 3.1 presents the percentage of voters in the sample supporting each of the four candidates, based upon respondents' relevant socio-demographic traits, party identification, opinions regarding democracy and evaluation of the incumbent government.

Table 3.1

Vote choice by respondents' views and characteristics*

		Bachelet (Concertación) %	Hirsch (JPM) %	Lavín (UDI) %	Piñera (RN) %	N
Age	18 -29	45.05	12.87	8.91	33.17	202
	30-44	53.23	11.03	15.97	19.77	263
	45-64	53.36	5.83	15.25	25.56	223
	65+	55.56	6.35	23.81	14.28	63
Education	None	37.50	0.00	37.50	25.00	8
	Primary	66.41	5.47	11.72	16.41	128
	Secondary	50.00	8.55	14.74	26.70	468
	University	42.86	17.01	14.97	25.17	147
Gender	Female	56.56	8.20	13.66	21.58	366
	Male	46.23	10.91	15.32	27.53	385
Household Income	quintile	57.89	6.58	22.37	13.16	76
	quintile	55.66	8.49	13.52	22.33	318
	quintile	47.89	11.05	13.16	27.89	190
	quintile	43.38	11.76	12.50	32.35	136
	quintile	45.16	9.68	22.58	22.58	31

Religious Denomination	Yes	52.38	6.35	15.56	25.71	630
	No	45.45	26.45	9.09	19.01	121
Democracy is the best Form of government	Disagree strongly	12.50	25.00	0.00	62.50	8
	Disagree	14.55	3.64	32.73	49.09	55
	Agree	46.82	9.92	16.79	26.46	393
	Agree strongly	65.08	9.83	8.47	16.61	295
Satisfaction with democracy in Chile	Unsatisfied	13.51	29.73	16.21	40.54	37
	Not very satisfied	24.78	8.84	26.99	39.38	226
	Fairly satisfied	61.46	9.43	9.97	19.14	371
	Very satisfied	82.05	5.13	4.27	8.55	117
Government Evaluation	Very bad	0.00	16.67	16.67	66.67	12
	Bad	7.32	13.01	41.46	38.21	123
	Good	55.04	8.40	11.34	25.21	476
	Very good	81.43	10.00	1.43	7.14	140
Party identification	Concertación	83.33	5.56	3.33	7.78	90
	JPM	33.33	66.67	0.00	0.00	15
	UDI	0.00	0.00	57.14	42.86	22
	RN	0.00	0.00	20.69	79.31	29
	Others	25.00	12.50	37.50	25.00	8
	Independents	51.79	9.37	14.48	24.36	587
Sample		51.44	9.58	14.70	24.49	751

* Table entries are the percentage of each row-variable voting for the designated candidate. Percentages sum to 100 across rows.

In accord with the assumption that an authoritarian/democratic cleavage is prevalent in the restructured Chilean party system, a strong division between voters regarding their attitudes towards democracy and their regime preferences can be seen in Table 3.1. Sixty-five percent of the respondents who stated they were unsatisfied with democracy and 79% of those stating that democracy is not always the best form of government supported the RN and UDI candidates. Interestingly, those expressing more critical views towards democracy tended to support the more moderate Piñera, although dissatisfaction with democracy, however, was higher for Lavín supporters.

Socio-demographic variables also factor into the choice between the competing candidates, as seen in Table 3.1. The high support for Bachelet among women marked a clear difference with respect to previous *Concertación* candidates (Angell and Reig, 2006; Huneeus, 2006). Hirsch did twice as well among younger, better-educated voters than among the older and less educated respondents. The electoral support-base of the two conservative candidates was also quite different, with Piñera having higher support than Lavín among better educated and wealthier voters. Religion seems to have strongly affected the choice for Hirsch: agnostic, atheists and respondents with no religious affiliation were much more likely to vote for Hirsch, while those belonging to a religious denomination (Catholics and Christians, essentially) were more likely to choose one of the other three candidates.

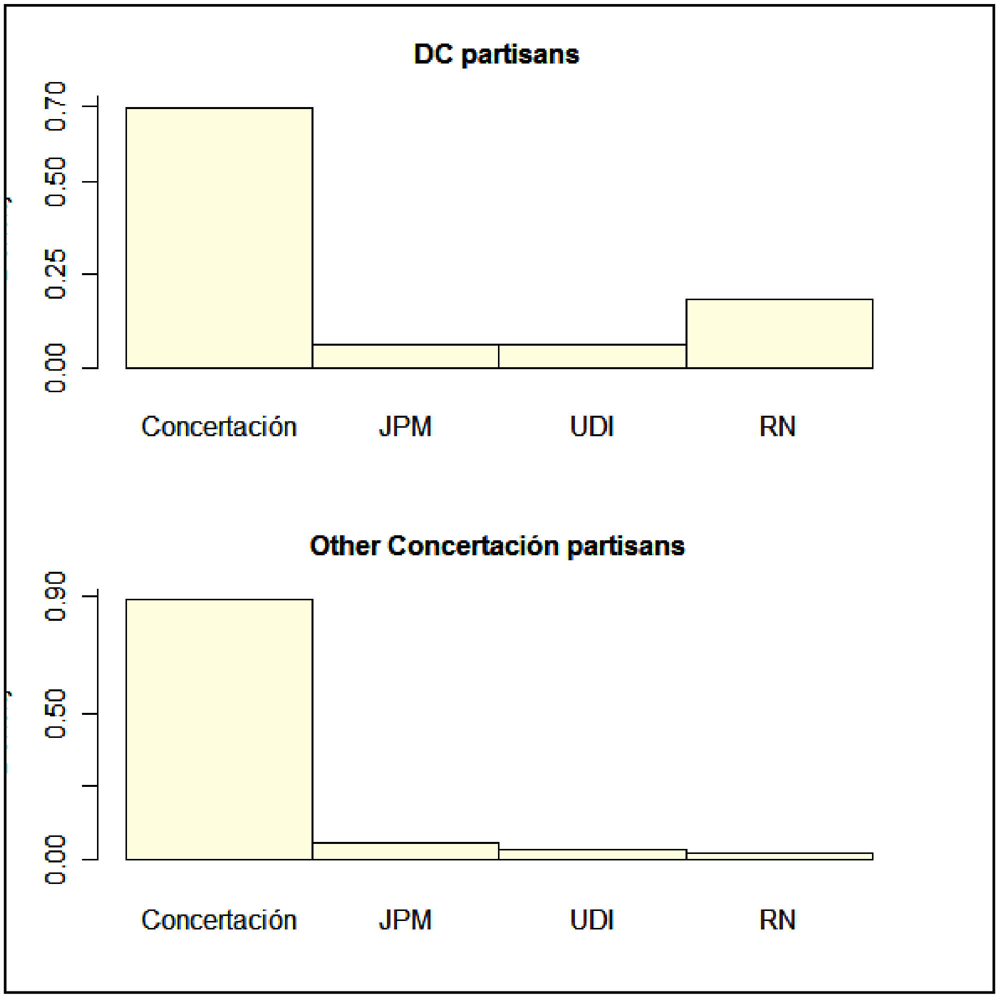
As for the effect of short-term factors, citizens' assessments of the incumbent *Concertación* government clearly influenced the choice between Bachelet and the three candidates of the opposition. Eighty percent of the respondents expressing dissatisfaction with the performance of the incumbent administration voted for the two conservative parties UDI and RN. The vote-share of *Juntos Podemos Más* was also disproportionately high among government critics, suggesting that Hirsch's vocal disapproval of the government's economic policies might have attracted the far-left voters disenchanted with the *Concertación*'s espousal of market economy and neo-liberal policies (Valenzuela and Scully, 1997; Navia, 2006). In contrast, 61% of those with favorable opinions of the government supported Bachelet. However, a majority of voters had positive evaluations of the government's performance, reflecting the unusually high popularity of President Lagos among the electorate (Angell and Reig, 2006; Navia, 2006).

Finally, another remarkable fact emerging from Table 3.1 is the relationship between partisanship and vote choice, particularly for respondents identified with the *Concertación* and UDI in the sample: 8% of the former and more than 43% of the latter voted for Piñera in the election. As mentioned in the introduction, the fact that opinion polls indicated that Bachelet would easily defeat Lavín in a two-candidate runoff while Piñera would pose a more serious challenge to the *Concertación* (Gamboa and Segovia, 2006) suggests that tactical voting might be the reason underlying the high electoral support of the RN candidate among UDI sympathizers.[25] This interpretation, however, does not account for the moderate support of Piñera among *Concertación* identifiers. Rather, the explanation in this case seems to be related to Piñera's moderate positioning and his appeal to Christian Democrats during the electoral campaign. Figure 3.1 explores this issue further by plotting the distribution of votes among *Concertación* partisans, discriminated between Christian Democrats (CD) and other *Concertación* identifiers. As shown in the figure, almost 20% of respondents in the sample expressing identification with the CD voted for the RN, a percentage 6 times higher than for other partisans of the center-left coalition.

[25] Party identification is defined based on respondents' answer to the question "Which party do you feel closer to?" in the CSES survey.

Figure 3.1

Distribution of votes among *Concertación* identifiers



Note: The figure shows the percentage of electoral support for each of the competing parties among respondents identified with the *Concertación* in the 2005 election. The upper panel summarizes vote choices among partisans of the Christian Democrats (DC), while the lower panel reproduces the information for respondents identified with other parties of the *Concertación* coalition.

Hence, this preliminary analysis suggests that, in line with the hypothesis underscoring the prevalence of an authoritarian/democratic cleavage in Chilean politics, voters' views and attitudes towards democracy played a key role on their decision of whether to vote for the *Concertación* or the *Alianza* candidates. In contrast, while socio-demographic variables also influenced voter behavior, they did not clearly determine a division between supporters of the two main political coalitions. In addition, the evidence presented above reveals that election-specific factors such as the emergence of a moderate conservative candidate and voters' strategic considerations also had a considerable influence on electoral behavior. This indicates that the different hypothesis proposed to account for voter behavior in Chile must be considered in the light of the particular political and institutional context of the 2005 election, and that previous analyses based entirely on citizens' party identification would probably fail to provide a complete account of voting patterns in the presidential race.

These bivariate relationships, however, do not allow us to assess the relative influence of the different variables on voter choice in a controlled way. In to assess which factors were more relevant in the 2005 election and to test alternative hypothesis about the determinants of voter behavior in Chile, we specify and estimate a model of multi-candidate vote choice.

3.3 A multi-candidate model of vote choice for the 2005 election

In order to test the competing explanations and to account for possible substitution patterns between electoral choices, we specify and estimate a multinomial probit model that allows us to examine the effect of different individual characteristics on voter choice after controlling for other confounding factors, as well as to assess how changes in candidates' spatial positions affect their expected vote-share. Unlike previous models applied in individual-level analysis of Chilean elections, the multinomial probit specification assumes that the voter simultaneously considers all the electoral options when making her choice, allowing us to test for the violation of the IIA assumption and to assess whether the relative probabilities of a voter choosing between any two candidates depends on the presence of other electoral options.[26]

3.3.1 Data and research design

Our source of data is the Comparative Study of Electoral Systems post-election survey (CSES, 2007). In line with the competing theories about the determinants of electoral behavior in Chile, we examine the effect of respondents' socio-demographic characteristics, attitudes towards democracy and assessment of the incumbent Lagos' government on their vote choice. The socio-demographic variables included in the model:

[26] The IIA assumption underlying logistic models can be tested on subsets of alternatives (Hausman and McFadden, 1984) and cross-alternative variables (McFadden, 1987). However, rejection of IIA using these tests does not provide much guidance on the correct specification to use (Train, 2003).

Age; *Education*, recorded on an four point-scale ranging from no education to university degree; a dummy variable for *Female; Income*, by household quintile*;* and *Religion*, coded 1 for respondents belonging to a religious denomination (Catholicism and other Christian faiths, essentially), 0 otherwise. We also include *Regime preference*, recording respondents' agreement with the statement "Democracy is better than any other form of government"; *Satisfaction with democracy*, a variable reflecting how satisfied respondents are with the way democracy works in Chile; and *Government evaluation*, measures respondents' assessment of the performance of Lagos' government; the three variables are scored on four-point scales in ascending order. As an alternative, all variables coded on an ordered scale were discretized, with the lower category taken as baseline and dummy variables specified for the remaining categories; the main substantive findings reported in Section 3.4, based on the default parametrization, remain unchanged under this alternative specification.[27]

In addition, in line with the prevalent spatial model of voting (Hinich and Munger, 1994; Merrill and Grofman, 1999), we include *Ideological distance*, a measure of respondents' spatial perceived ideological distance from each of the candidates in the model, defined as the squared difference between the respondent's self-reported placement on an 11-point left-right scale and her placement of each of the parties on the same scale (Merrill and Grofman, 1999). The left-right ideological dimension plays a key role in terms of popular perceptions of party differences in Chile (Valenzuela and Scully, 1997; Tironi

---

[27] A complete set of results using the alternative coding scheme is available from the authors upon request.

and Agüero, 1999), where there are relatively minor differences between the main political forces regarding fundamental political and economic issues (Scully, 1995; Fuentes, 1999; Angell and Reig, 2006). In the case of *Concertación*, we use a weighted average of respondent's placements of the parties that form the coalition, while we use respondent's placements of the Communist Party to approximate the location of *Juntos Podemos Más*.[28] Although the CSES survey asks Chilean respondents only about parties' positions, we compared the ideological locations obtained from the CSES survey with candidates' perceived positions form the October-November 2005 Centro de Estudios Públicos (CEP, 2005) national survey; the ordering of the candidates on the left-right scale is the same in both surveys, and differences in the mean of respondents' placements of the candidates between the surveys are quite small.[29] While the results reported below are based on the distance measure computed from parties' perceived location in the CSES survey in order to avoid statistical complexities brought about by combining information from different sources (Lohr, 2005; Raghuanthan et al., 2006), using the candidates' placements obtained from the CEP survey yields similar results.[30]

[28] The Communist Party is the major partner of *Juntos Podemos Más* and the only party of the alliance whose location respondents were asked about in the CSES survey.

[29] Less than 0.9 points on an eleven-point scale for each of the candidates.

[30] See footnote 5.

3.3.2 Model specification and empirical strategy

Our basic model specification is grounded in the spatial voting and random utility maximization literature, and draws on Alvarez and Nagler (1995) and Alvarez, Bowler and Nagler (2000). We assume that the voter's utility for each candidate is composed of a systemic component, specified as function of characteristics of the individuals and the candidates, and a stochastic component that represents the influence of unobserved factors on voters' choice. Following Alvarez and Nagler (1995), voter i's utility for candidate j, denoted by $U_{i,j}$, is given by:

$$U_{i,j} = z_i^{'}\alpha_j + x_{i,j}^{'}\delta + \varepsilon_{i,j}, \quad j\text{=Bachelet, Hirsch, Lavín, Piñera} \qquad (3.1),$$

where z_i is a vector of characteristics of the i th voter (including a constant term), $x_{i,j}$ is a vector of characteristics of the j th candidate relative to the voter, α_j and δ are vectors of parameters to be estimated, and $\varepsilon_{i,j}$ is a disturbance term. We assume that the four error terms $\left(\varepsilon_{i,\text{Bachelet}}, \varepsilon_{i,\text{Hirsch}}, \varepsilon_{i,\text{Lavín}}, \varepsilon_{i,\text{Piñera}}\right)$ follow a multivariate normal distribution with mean vector 0 and variance-covariance matrix Σ, allowing the random components of utility to be correlated across parties. In line with random utility models, each voter is assumed to vote for the candidate that provides her with the highest utility; that is,

$$Y_i = j \quad \text{if} \quad U_{i,j} = \max\left(U_i\right) \qquad (3.2),$$

where Y_i is the observed voter choice, Given that only differences in utility matter and thus any location shift will not change the observed vote, we can solve the identification problem by taking one party as the base alternative and expressing i's utility for the other candidates relative to her utility for the base alternative. Assuming, without loss of generality, that we take Piñera (RN) as the base alternative, and defining $\tilde{U}_{i,k} = U_{i,k} - U_{i,\text{Piñera}}$, $k = \text{Bachelet, Hirsch, Lavín}$, we can express the random utility model as j:

$$\tilde{U}_i = W_i \tilde{\beta} + \tilde{\varepsilon}_i \qquad (3.3),$$

where

$$W_i = \left[z_i \otimes I_3, X_i^* \right],$$

$$X_i^* = \begin{bmatrix} x'_{i,\text{Bachelet}} - x'_{i,\text{Piñera}} \\ x'_{i,\text{Hirsch}} - x'_{i,\text{Piñera}} \\ x'_{i,\text{Lavín}} - x'_{i,\text{Piñera}} \end{bmatrix},$$

$$\tilde{\varepsilon}_i = \left(\tilde{\varepsilon}_{i,\text{Bachelet}}, \tilde{\varepsilon}_{i,\text{Hirsch}}, \tilde{\varepsilon}_{i,\text{Lavín}} \right) \sim N_3\left(0, \tilde{\Sigma}\right), \; \tilde{\varepsilon}_{i,k} = \varepsilon_{i,k} - \varepsilon_{i,\text{Piñera}},$$

and

$$Y_i\left(\tilde{U}_i\right) = \begin{cases} k & \text{if } \max\left(\tilde{U}_i\right) = \tilde{U}_{i,k} > 0, \quad k = \text{Bachelet,Hirsch,Lavín} \\ \text{Piñera} & \text{if } \max\left(\tilde{U}_i\right) < 0 \end{cases} \qquad (3.4).$$

The parameters $\theta = \left(\tilde{\beta}, \tilde{\Sigma}\right)$ are still not identified, because a scale shift will not change the observed choices.[31] We follow McCulloch and Rossi (1994) and achieve identification by normalizing the parameters with respect to $\tilde{\sigma}_{1,1}$: $\theta' = \left(\tilde{\beta}', \tilde{\Sigma}'\right) = \left(\tilde{\beta} / \sqrt{\tilde{\sigma}_{1,1}}, \tilde{\Sigma} / \sqrt{\tilde{\sigma}_{1,1}}\right)$.

The likelihood for the multinomial probit model is then given by:

$$f\left(Y \middle| W, \tilde{\beta}', \tilde{\Sigma}'\right) = \prod_{i=1}^{n} \Pr\left(Y_i \mid W, \tilde{\beta}', \tilde{\Sigma}'\right) \tag{3.5}$$,

$$P\left(Y_i \middle| W_i, \tilde{\beta}', \tilde{\Sigma}'\right) = \int_{A_j} \phi_3\left(\tilde{U}_i \middle| W_i, \tilde{\beta}', \tilde{\Sigma}'\right) d\tilde{U}_i \tag{3.6}$$,

where ϕ_3 is the trivariate normal probability density function, and

$$A_j = \begin{cases} \tilde{U}_i : \tilde{U}_{i,k} > \max\left(\tilde{U}_{i,-k}, 0\right) & \text{if } Y_i = k, k = \text{Bachelet,Hirsch,Lavín} \\ \tilde{U}_i : \tilde{U}_i < 0 & \text{if } Y_i = \text{Piñera} \end{cases}.$$

The posterior density of the parameters is given by Bayes theorem as

$$\pi\left(\tilde{\beta}', \tilde{\Sigma}' \middle| W\right) \propto f\left(Y \middle| W, \tilde{\beta}', \tilde{\Sigma}'\right) \pi\left(\tilde{\beta}'\right) \pi\left(\tilde{\Sigma}'\right) \tag{3.7}$$,

where $\pi\left(\tilde{\beta}'\right)$ and $\pi\left(\tilde{\Sigma}'\right)$ denote the prior densities of $\tilde{\beta}'$ and $\tilde{\Sigma}'$, respectively.

---

[31] That is, $Y_i\left(\tilde{U}_i\right) = Y_i\left(\alpha\tilde{U}_i\right) \;\; \forall \alpha > 0$.

The model was fit through Markov chain Monte Carlo simulations, using McCulloch and Rossi's (1994) Gibbs sampling algorithm.[32,33] As mentioned in the introduction, Bayesian procedures based on Gibbs sampling allow making exact finite sample inferences without relying on large-sample theory (McCulloch and Rossi, 1994; Kim, Kim and Heo, 2003). Because of the discrete nature of the dependent variable, a considerable sample size may be required for accurate asymptotic approximations (McCulloch and Rossi, 1994). Hence, the Bayesian approach is particularly appropriate given the relatively small dataset available to analyze the 2005 election.[34] The Bayesian approach is also better suited to deal with a large number of alternatives than the simulation-based methods of classical estimation, which require deriving the likelihood function with respect to each element of the variance-covariance matrix, thus resulting in substantial increases in computational time (Greene, 1999; Train, 2003; Kim et al., 2003).

In addition, the Bayesian model-fitting strategy allows for comparison of competing models and explanations of voter behavior in a straightforward and computationally

---

[32] See McCulloch and Rossi (1994), McCulloch, Polson and Rossi (2000) and Imai and van Dyk (2005) for a detailed presentation of the sampling algorithm. A general discussion of Gibbs sampling can be found in Gelfland and Smith (1990) and Casella and George (1992).

[33] The Gibbs sampler was implemented using the 'bayesm' package in R (Rossi, Allenby and McCulloch, 2005).

[34] McCulloch and Rossi (1994) show that non-normality of finite sampling distributions of the error variance-covariance parameters can arise with even 1,000 observations per parameter, indicating that "asymptotic theory may be of little use for the MNP model" (p. 219).

practical way using Bayes factors (Kass and Raftery, 1995). The Bayes factor for model M_j relative to model M_k is given by:

$$B_{j,k} = \frac{p\left(y|M_j\right)}{p\left(y|M_k\right)} = \frac{\int p\left(y|M_j,\theta_j\right) p\left(\theta_j|M_j\right) d\theta_j}{\int p\left(y|M_k,\theta_k\right) p\left(\theta_k|M_k\right) d\theta_k} \qquad (3.8),$$

where, in the application of Section 3.4, we used the harmonic mean of the likelihood values evaluated at the posterior draws (Newton and Raftery, 1994) as an estimate for $p\left(y|M_x\right),\ x = j,k$:

$$\hat{p}\left(y \mid M_x\right) = \left(\frac{1}{R}\sum_{r=1}^{R} p\left(y \mid \theta_x^{(r)}\right)^{-1}\right)^{-1} \qquad (3.9).$$

Diffuse proper priors were assumed for the parameters in the model, $\tilde{\beta} \sim N\left(\overline{\beta}, B^{-1}\right)$ and $\tilde{\Sigma} \sim \textit{Inverse Wishart}\left(v,V\right)$, with $\overline{\beta} = 0,\ B^{-1} = 0.0001I,\ v = 6,\ V = vI$ (McCulloch, Polson and Rossi, 2000); routine sensitivity analyses were performed to assess the robustness of the results with respect to different priors and starting values for the sampling algorithm, yielding similar results. A single Markov chain was run for 3,000,000 cycles, with the first 50,000 discarded as burn-in; while McCulloch and Rossi’s (1994) sampler is quite easy to implement, high correlation between the parameters and the latent variables introduced by the data augmentation algorithm used to form the Gibbs sampler (Tanner and Wong, 1987; McCulloch and Rossi, 1994; Imai and van Dyk, 2005), coupled with a high-dimensional parameter space, determined that the Markov chain was extremely slow in navigating the state space, and some parameters required more than 2,000,000 draws to converge.[35] The results presented in Section 3.4 are based on the last 50,000 Gibbs sample draws of the parameters.

3.4 Empirical results

3.4.1 Multinomial probit estimates

Table 3.2 reports the posterior means and 95% Bayesian credible intervals of the parameters of the multinomial probit model. The coefficients for the individual-specific variables give the effect of each the variable on the vote for Bachelet, Hirsch and Lavín relative to Piñera. The coefficient for *Ideological distance* indicates the effect of respondents' perceived ideological distance from the parties in the probability of voting: a negative coefficient indicates that a voter is more likely to vote for a party the closer the party's position is to her own. At the bottom of Table 3.2 are the estimates for the error correlations between *Concertación*, *Juntos Podemos Más* and UDI.[36] The model correctly predicts voter choice in 59.6% of the cases, while a "null model" predicting that voter choice for each respondent will take the value of the most common outcome in the sample (*Concertación*) correctly classifies 51.4% of the vote. Such a model, however, would

---

[35] Convergence was assessed using Geweke's (1992) diagnostic based on a test for equality of the means of the first 10% and last 50% of the Markov chain.

[36] In fact, we estimate the covariance matrix of the differences in utility, with RN as the base category. This does not make any substantive difference in the interpretation of the model.

provide no information about the effect of the predictors on the relative probability of voting for the different parties.

The summaries of the posterior densities shed substantial light on the relative influence of respondent's socio-demographic characteristics, attitudes towards democracy and evaluation of government performance on their electoral behavior. First, regarding the effect of socio-demographic variables, wealthier voters were more likely to vote for the Renewal Party (RN) than for *Concertación* or UDI, and younger voters were also more likely to choose Piñera over Lavín. None of these variables significantly affected the choice between the RN candidate and Hirsch. In contrast, and in line with the data presented in Table 3.1, more educated voters and those not belonging to any religious denomination were more likely to vote for JPM than for RN, but these variables did not affect the choice between Piñera and the other two candidates at the 95% level. Although these estimates indicate that socio-demographic factors did influence voters' electoral behavior, they did not necessarily affect the choice between *Concertación* and *Alianza*. Rather, the evidence indicates that some of the socio-economic variables that had a positive effect on the probability of choosing Bachelet over Piñera – e.g., *Income* - also increased the probability of voting for Lavín over the candidate of the Renewal Party.

On the other hand, respondents' regime preferences and their evaluation of the incumbent government significantly affected the choice between Bachelet and the two candidates of the *Alianza*. Respondents who stated that democracy is always the best form of government and those expressing favorable views of Lagos' administration were more likely to vote for Bachelet than for Piñera, but this variable did not affect vote choice between the UDI and the RN candidates. Voters satisfied with the way in which democracy works in Chile were more likely to vote for Bachelet and for Lavín than for Piñera, but they were less likely to choose Hirsch over the RN candidate.

Table 3.2

Posterior means and 95% credible intervals (in parenthesis)

for the parameters of the multinomial probit model

Coefficients	**Bachelet/Piñera**	**Hirsch/Piñera**	**Lavín/Piñera**
Intercept	-1.88 (-3.83, -0.73)	-0.11 (-0.57, 0.31)	-1.02 (-2.03, -0.06)
Age	0.11 (-0.00, 0.23)	0.00 (-0.06, 0.07)	0.20 (0.04, 0.38)
Education	-0.17 (-0.36, 0.02)	0.13 (0.02, 0.24)	-0.07 (-0.31, 0.17)
Female	0.11 (-0.09, 0.30)	-0.01 (-0.11, 0.13)	0.10 (-0.13, 0.35)
Income	-0.17 (-0.27, -0.06)	-0.01 (-0.08, 0.06)	-0.17 (-0.34, -0.04)
Religion	0.09 (-0.15, 0.34)	-0.34 (-0.52, -0.17)	0.21 (-0.10, 0.60)
Regime preference	0.19 (0.02, 0.36)	0.02 (-0.07, 0.11)	0.08 (-0.15, 0.28)
Satisfaction with democracy	0.33	-0.13	0.26

	(0.13, 0.51)	(-0.24, -0.04)	(0.06, 0.48)
Government evaluation	0.31 (0.07, 0.66)	0.07 (-0.02, 0.17)	-0.13 (-0.46, 0.18)
Ideological distance	-0.01 (-0.01, 0.01)		
Correlations			
$\rho_{Concertación, JPM}$	-0.77 (-1.00, 0.26)		
$\rho_{Concertación,UDI}$	0.92 (0.55, 1.00)		
$\rho_{JPM,UDI}$	-0.93 (-1.00, -0.63)		
% Correctly predicted (vs. Null Model*): 59.6% (51.44%) Number of observations: 751			

*The null model predicts that voter choice for each respondent will take the value of the most common outcome in the sample.

A remarkable result emerging from Table 3.2 is that, although the coefficient of Ideological distance has the expected negative sign, in line with the spatial voting literature, it is not statistically significant at the usual confidence levels. This finding is robust to alternative definitions of the ideological distance measure, such as using the absolute value rather than the square of the difference between the respondents' and the parties' locations on the left-right scale or approximating parties' location using the mean of respondents' placements (Rabinowitz and MacDonald, 1989; Alvarez and Nagler, 1995). Nonetheless, it must be mentioned that 24% of the respondents in the sample who placed themselves in the far-left end of the ideological scale stated that they had voted for one of the two *Alianza* candidates. This suggests that this result might stem from the methodological difficulties inherent in collecting perceptual data (Aldrich and McKelvey, 1977; King, Murray, Salomon and Tandon, 2004) or from flaws in the CSES questionnaire. In order to address this problem, we re-estimated the model using estimates of respondents' self-placement and parties' locations obtained through Aldrich and McKelvey's (1977) method of scaling, with virtually identical outcomes. Hence, although we cannot discard the hypothesis that this result is mainly driven by problems in the CSES questionnaire and well-known difficulties associated to the use of ordinal scales (King et al., 2004), a possible explanation lies in the absence of important policy differences between the three main candidates and in the fact that the first round of the election was presented as a choice between candidates' personal traits, rather than between parties or ideological positions (Gamboa and Segovia, 2006).

A different interpretation has to do with the extent of tactical voting among the Chilean electorate. Given the high probability of a *ballotage* and the highly disputed contest between Piñera and Lavín for the second place in the election, voters - in particular, *Concertación* sympathizers - might have had an incentive to cast a ballot for a candidate other than their most preferred one in order to affect the race between the two candidates of the *Alianza* and to influence who would face Bachelet in the second-round runoff (Cox, 1997). The relationship between partisanship and vote-choice reported in Table 3.1 and the high percentage of split-ticket voting between the presidential and legislative races (Navia, 2006) suggests that tactical voting might have been relatively important in the 2005 election; we explore this argument in Section 3.4.3 below.

Finally, the estimated error correlations between *Concertación* and UDI and between *Juntos Podemos Más* (JPM) and UDI are statistically significant at the usual confidence levels: we find a positive correlation between *Concertación* and UDI and a negative correlation between JPM and UDI. Although the positive correlation between *Concertación* and UDI is at odds with received knowledge about citizens' partisan identities in Chile, it is in line with Angell and Reig's (2006) observation that the RN candidate was disliked by a significant proportion of Lavín's supporters, and might help account for the fact that a considerable percentage of them voted for Bachelet in the second-round runoff against Piñera (Gamboa and Segovia, 2006; Huneeus, 2006). These results indicate that the IIA assumption is violated and that models that impose such condition might produce incorrect inferences about voter choice in Chile's 2005 election (Alvarez, Bowler and Nagler, 2000). More importantly, such models would neglect the fact that Piñera's entry into the election significantly affected citizens' probabilities of voting for the other competing candidates.

3.4.2 The effect of individual characteristics on vote choice

The coefficients reported in Table 3.2 are difficult to interpret directly due to the nonlinear functional form of the multinomial probit model and the fact that the voters' utilities are expressed with respect to a baseline alternative (Piñera). In order to assess the relative impact of the different factors proposed to account for voter behavior in Chile and to be able to make pairwise comparisons between candidates, we estimate the marginal effect of the individual-specific variables on the probability of voting for each candidate using "first differences" (King, Tomz and Wittenberg, 2000). For each respondent in the sample, we compute vectors of choice probabilities $\left[P_i\left(\text{Bachelet}\right), P_i\left(\text{Hirsch}\right), P_i\left(\text{Lavín}\right), P_i\left(\text{Piñera}\right)\right]$ based on the value of the regressors and the Gibbs sample draws of the models' parameters using the GHK algorithm (Hajivassiliou, McFadden and Ruud, 1996). Then we alter one independent variable at a time and recompute the predicted probabilities for each respondent, holding all other variables constant. Finally, we average the differences between these probabilities over all simulations and respondents, obtaining the mean value and 95% credible intervals for the causal effect of the variable under analysis. Table 3.3 summarizes the average impact on the probability of support for each party of changing the values of the predictors from one end of the scale to the other.[37]

[37] In the case of the binary variables, *Female* and *Religion*, we measure the impact of a change from 0 to 1.

Table 3.3

Marginal effect of individual-specific variables on voter choice

Variable	Bachelet (Concertación)	Hirsch (JPM)	Lavín (UDI)	Piñera (RN)
Age	0.00 (-0.02, 0.01)	-0.07 (-0.18, -0.01)	0.07 (0.01, 0.18)	0.00 (-0.01,0.01)
Education	-0.21 (-0.31, -0.01)	0.11 (0.04, 0.23)	0.04 (-0.10, 0.10)	0.07 (-0.03, 0.15)
Female	0.02 (0.00, 0.03)	0.00 (-0.01, 0.01)	0.02 (0.00, 0.04)	-0.03 (-0.04, -0.01)
Income	-0.21 (-0.27, -0.01)	0.04 (0.01, 0.12)	-0.01 (-0.17, 0.00)	0.18 (0.03, 0.24)
Religion	0.04 (-0.04, 0.10)	-0.17 (-0.28, -0.06)	0.02 (0.00, 0.07)	0.11 (0.00, 0.25)
Regime preference	0.33 (0.03, 0.47)	0.02 (0.01, 0.05)	-0.10 (-0.20, 0.10)	-0.26 (-0.33, -0.14)
Satisfaction with democracy	0.42 (0.18, 0.52)	-0.19 (-0.38, -0.08)	-0.04 (-0.11, 0.13)	-0.01 (-0.31, -0.01)
Government evaluation	0.66 (0.33, 0.85)	0.09 (0.02, 0.23)	-0.54 (-0.73, -0.29)	-0.22 (-0.35, -0.11)

95% credible intervals reported in parenthesis.

In line with the results presented in Table 3.2, the estimated first differences do not support the hypothesis that socioeconomic or religious cleavages played a key role in the choice between leftist and conservative candidates. While, *ceteris paribus*, higher education levels increased the probability of voting for the left-wing *Juntos Podemos Más* by 11 percentage points, it reduced the likelihood of voting for *Concertación* by 0.21. Respondents belonging to households in the wealthiest income quintile were 0.18 more likely to vote for Piñera than those in households at the bottom quintile, but they were also 0.04 more likely to vote for Hirsch. Also, respondents belonging a religious denomination were 0.18 less likely to cast a ballot for Hirsch than atheist or agnostic voters, but this variable had no statistically significant effect on the probability of voting for Bachelet or for either of the two candidates of the *Alianza*.

In contrast, opinions about regime preference and government performance did have substantive and opposite effects on the probability of voting for the two leftist and the two conservative candidates. Respondents would be on average 0.33 more likely to vote for Bachelet and 0.02 more likely to vote for Hirsch if they felt that democracy is always the best form of government, but they would be 0.26 less likely to vote for Piñera. Also, moving from a very negative to a very positive evaluation of the incumbent government increased the likelihood of voting for Bachelet and Hirsch by 0.66 and 0.09, respectively, while reducing the average probability of supporting Lavín and Piñera by 54 and 22 percentage points. Given the success of the President Lagos's economic and social policies and the fact that neither of the UDI nor the RN candidates proposed substantial transformations in this regard, it seems reasonable to assume that the strong positive effect of a negative evaluation of the government on the probability of supporting the *Alianza* is not necessarily reflecting retrospective voting. Rather, it might be related to a series of important democratizing reforms implemented during Lagos' term in office, such as the elimination of designated senators and the restoration of the presidential power to designate and remove the heads of the different branches of the military, as well as to the adoption of divisive "symbolic" measures like the reparations to victims of human rights violations (Angell and Reig, 2006; Navia, 2006).

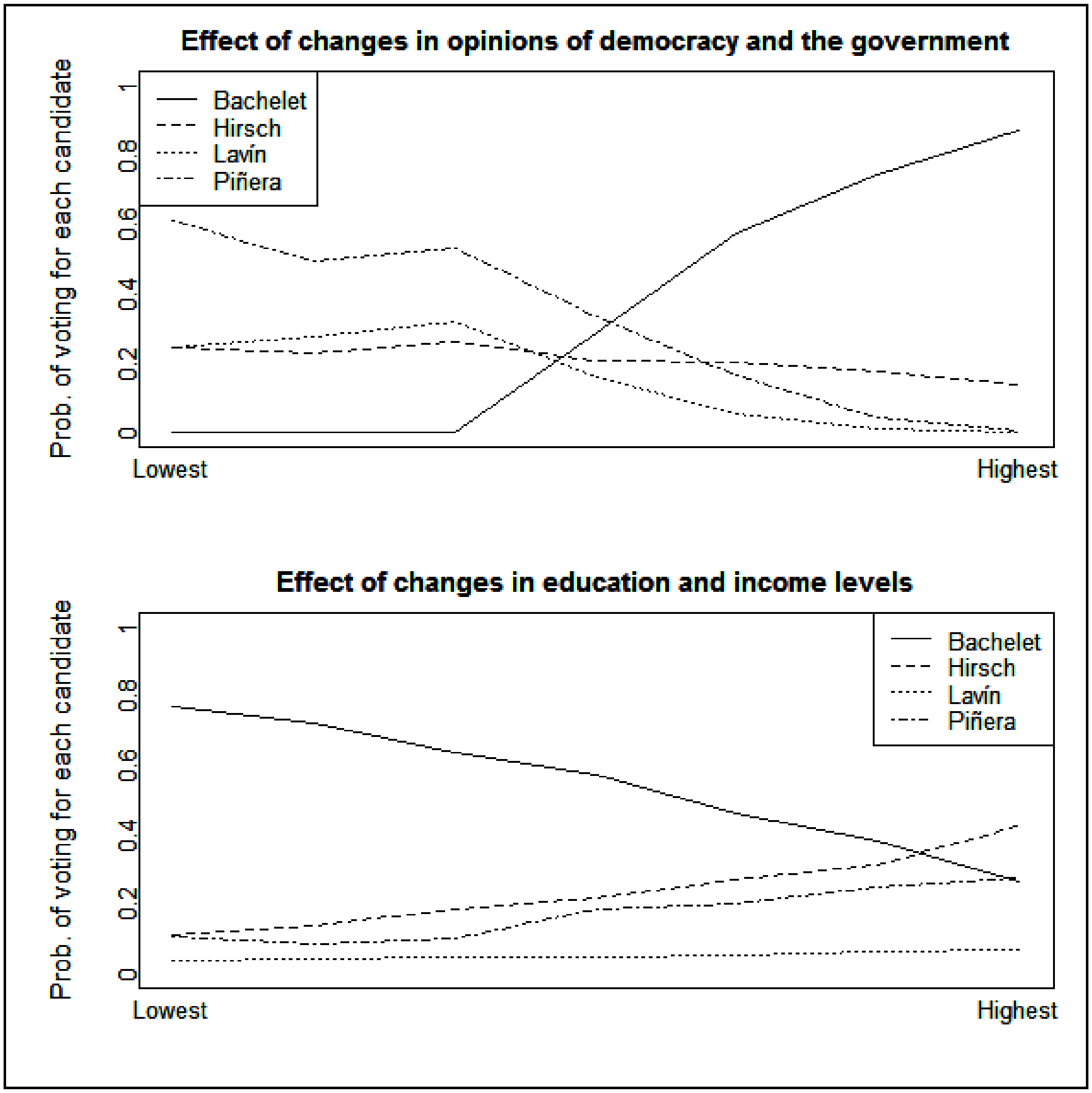
On the other hand, although *Satisfaction with democracy* also had a significant influence on voter choice, the effect of this variable does not reveal a clear left-right division. On average, moving from a very negative to a very positive opinion of the way in which democracy works in Chile increased the likelihood of voting for Bachelet by 0.42, but decreased the probability of voting for either Hirsch or Piñera by 0.19. Notice, however, that the causal effect of this variable on the probability of choosing Lavín is not statistically significant at the 0.05 level. The positive relationship between dissatisfaction with the functioning of democracy and the likelihood of voting for Hirsch and Piñera might reflect a demand for alternative electoral options among voters disenchanted with the two major blocs dominating electoral competition, rather than respondents' anti-democratic values. While Hirsch adopted a critical position towards both the *Concertación* and the conservative opposition during the campaign, Piñera emphasized the need to build a broad center-right "New Coalition" based on "Christian Humanist" principles to replace the *Concertación/Alianza* dichotomy (Angell and Reig, 2006; Gamboa and Segovia, 2006).

In order to better illustrate the relative validity of the hypotheses emphasizing the role of socio-economic and authoritarian-democratic cleavages, Figure 3.2 summarizes the effect on the choice probabilities of a hypothetical voter of shifting all the predictors used to operationalize each approach from the lower to the upper end of the scale.[38] The upper panel of Figure 3.2 plots the probabilities of voting for each candidate as a function of the voter's views on democracy and government performance, while holding the socio-demographic variables at the mean sample values. The lower panel reproduces the analysis, varying the voter's education and income levels and fixing the remaining predictors at their average values.

[38] Our hypothetical voter is male, of mean age, education and income, and belongs to a religious denomination; his opinions on democracy and the government and his ideological distance from each party are set to be at the mean sample values. Although the values of the independent variables used to construct this hypothetical voter influence the baseline probability estimates, they do not but substantively influence the effect of changes in the predictors on the voter's choice probabilities.

Figure 3.2

Effect of views on democracy and the government and of socio-economic variables on choice probabilities



Note: The upper panel plots the probabilities of voting for each candidate as a function Regime preference, Satisfaction with democracy and Government evaluation. The lower panels plots the choice probabilities as functions of Education and Income.

The comparison of the upper and lower panels in Figure 3.2 suggests that differences in the attitudes towards democracy and the evaluation of the government are the main source of divide between *Alianza* and *Concertación* supporters. Going from the lower to the upper end of the scale on *Regime Preference*, *Satisfaction with democracy* and *Government evaluation* increases the probability of voting for Bachelet from 0 to 87 percent, while decreasing the likelihood of voting for Lavín and Piñera from 0.25 and 0.65 to 0 and 0.01, respectively. Simultaneous increases in *Education* and *Income* also have a substantial effect on the likelihood of choosing Bachelet, lowering it by as much as 50 percentage points, from 0.77 to 0.27. However, the effect of such increases on the likelihood of supporting the candidates of the *Alianza* is much smaller, raising it from 0.11 to 0.27 in the case of Piñera, while having virtually no effect on the vote for Lavín. Hence, our findings support the arguments underscoring the role of the authoritarian-democratic in the choice between *Concertación* and *Alianza* (Tironi and Agüero, 1999; Torcal and Mainwaring, 2003). In contrast, the evidence presented above shows that socio-economic and cultural factors are the main determinants of the support for Hirsch.

It is worth mentioning, however, that the comparison of a model including only socio-economic variables *vis-à-vis* a model including only respondents' views on democracy and the government does not favor any of the two specifications: the Bayes factor between the second and the first model is 1.09, and remains essentially unchanged (1.11) when including the spatial distance measure in both specifications. Hence, neither the hypothesis emphasizing the role of social and cultural cleavages nor the theory underscoring the authoritarian-democratic divide provides a single best explanation of voter choice in Chile's 2005 election.

3.4.3 The role of the electoral context: candidate competition and voter calculus

The salience of the authoritarian-democratic cleavage in structuring the competition between the two main political blocs in the 2005 election suggests that Chile is still, in the words of Constable and Valenzuela (1991), a ‘nation of enemies’. In this context, it is particularly relevant to examine whether Piñera’s candidacy and his campaign strategy aimed at distancing himself from the far right and the military dictatorship, contributed not only to his victory over Lavín in the contest for the second place in the election, but also to increase the support for the *Alianza*.

In order to do so, we exploit the fact that the multinomial probit model allows us to estimate the effect of the entry of Piñera in the presidential race and determine where the RN votes had gone in his absence. For each respondent in the sample, we calculate his expected utility difference for Bachelet, Hirsch and Lavín with respect to Piñera using the Gibbs sampling draws of the coefficients. Based on these differential utilities and on the draws of elements of the variance-covariance matrix, we can simulate vectors of choice probabilities for Bachelet, Hirsch and Lavín in a three-candidate race and estimate their expected vote-shares. In order to compare the results with those obtained under a scenario in which Piñera had been the only candidate of the *Alianza*, we also computed the probability of each voter choosing between Bachelet, Hirsch and Piñera in a three-candidate race with Lavín omitted. Table 3.4 reports the simulated vote-shares of the candidates in these two hypothetical three-candidate races and contrast them with the model’s predictions for the four-candidate election.

Table 3.4

Expected vote-shares of the candidates under alternative electoral scenarios (in percentage points)

Candidate	Four-candidate race	Three-candidate races	
		without Piñera	without Lavín
Bachelet (Concertación)	50.87 (48.34, 53.31)	51.64 (49.04, 54.27)	57.10 (54.83, 59.56)
Hirsch (JPM)	8.19 (6.83, 9.79)	17.52 (15.31, 19.90)	9.57 (8.03, 11.23)
Lavín (UDI)	14.53 (12.74, 16.72)	30.84 (23.38, 33.81)	-
Piñera (RN)	26.41 (23.81, 28.17)	-	33.33 (31.60, 35.12)

95% credible intervals reported in parenthesis.

As seen in the table, in a four-candidate election, our model predicts an expected vote-share of 50.9% for Bachelet, 8.2% for Hirsch, 26.4% for Piñera and 14.5% for Lavín, close to the actual proportion of votes for each candidate in the sample (Table 3.1).[39] While the expected vote-share for the two candidates of the *Alianza* would add to almost 41% in the four-candidate election, none of the two conservative candidates running alone would have obtained more than 34% of the vote in a three-candidate race against Bachelet and Hirsch.

[39] In the CSES sub-sample of 751 respondents we use, there is a positive bias for Bachelet and Hirsch and a negative bias for Lavín, which our multinomial probit model reproduces.

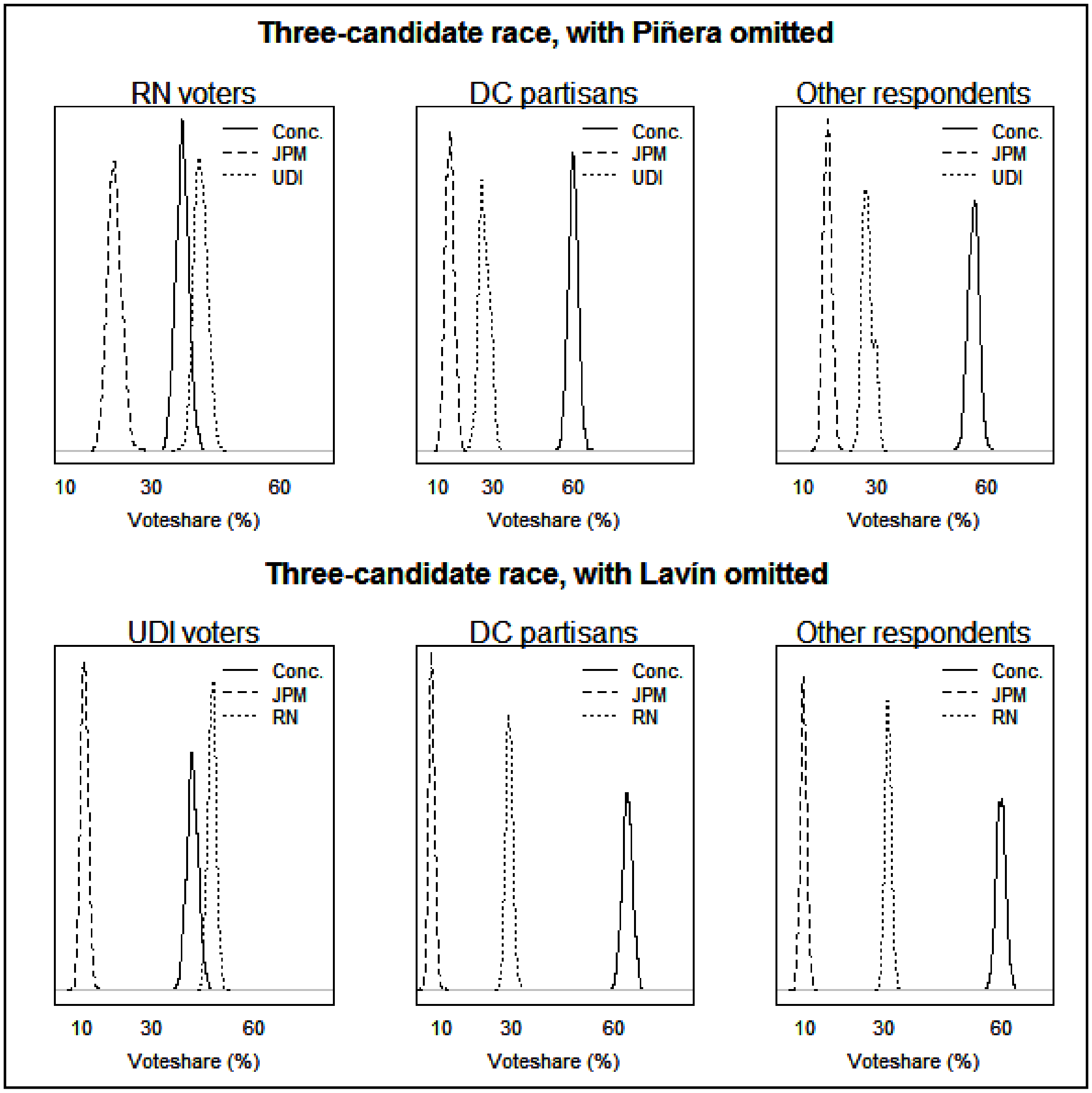
This indicates that Piñera's candidacy was an important determinant of *Alianza*'s relative success in the 2005 election, increasing the support for the center-right by more than 10 percentage points with respect to the hypothetical case in which Lavín had been the only candidate of the coalition, as originally expected. Moreover, the support for Hirsch among the respondents in the sample would more than double under this scenario when compared to the four-candidate race, suggesting that the RN candidate was backed by a segment of voters who were not willing to cast a ballot for either Bachelet or Lavín. This interpretation is in line with the results in Tables 3.2 and 3.3 showing that Hirsch and Piñera had a strong support among voters disenchanted with the workings of democracy in Chile and who might have been looking for alternatives to the two "traditional" electoral options.

Figure 3.3 explores this issue further, plotting the distribution of the candidates' vote-share among different groups of respondents in both hypothetical three-candidate elections. The upper panel of Figure 3.3 summarizes the model's predictions for a three-candidate race between Bachelet, Hirsch and Lavín, plotting the distribution of the support for the candidates among those respondents who voted for Piñera in the actual election, among Christian Democrats (DC), and among the rest of the respondents in the sample. Analogously, the bottom panel of Figure 3.3 presents the results for a three-candidate election with Lavín omitted, plotting the distribution of the support for Bachelet, Hirsch and Piñera among those who voted for Lavín in the presidential election, among DC partisans and among the remaining respondents.

The upper panel shows that, in a three-candidate race between Bachelet, Hirsch and Lavín, respondents who voted for Piñera in the actual election were more likely to vote for *Juntos Podemos Más* than the rest of the respondents in the sample. Also, comparing the two hypothetical three-candidate races between Bachelet, Hirsch and a single *Alianza* contender, the support for JPM among respondents who voted for Piñera in the presidential election would have been twice as large as among respondents who voted for Lavín. Hence, far from weakening the *Alianza*'s electoral prospects, the division between the UDI and the RN candidates seems to have actually increased the coalition's vote-share. While the *Alianza* retained its customary right-wing vote, Piñera's candidacy allowed the coalition to expand its electoral appeal to some citizens dissatisfied with the workings of democracy in Chile and demanding alternative electoral options; Lavín, *Alianza*'s "natural" candidate, would not have offered an attractive choice for this group of voters.

Figure 3.3

Predicted vote-shares in the two hypothetical three-candidate races



Note: The upper panel plots the distribution of support for Bachelet, Hirsch and Lavín among different groups of respondents in a three-candidate race with Piñera omitted. Analogously, the lower panel plots the expected-vote share of Bachelet, Hirsch and Piñera in a three-candidate race, with Lavín omitted.

Nonetheless, the evidence reported in Table 3.4 and Figure 3.3 also point out that Piñera's success was, to a large extent, due to the specific circumstances surrounding the 2005 election and, in particular, to the uncertainty about which of the two *Alianza* candidates would join Bachelet in the almost certain second-round runoff (Angell and Reig, 2006; Gamboa and Segovia, 2006). Three main results back up this claim. First, as seen in the upper panel of Figure 3.3, Piñera voters were more likely to vote for Lavín than for the other contestants if the RN candidate had not entered the race, and the expected vote-share of the UDI among them is much higher than among the other respondents in the sample. Hence, Piñera seems to have attracted many conservative voters who would have otherwise voted for Lavín and who might have seen the RN candidate as a more viable option in a second-round runoff against Bachelet. Second, as seen in the third column of Table 3.4, even though the *Alianza* would had done slightly better in a three-candidate race in which Piñera, rather than Lavín, had been the coalition's only nominee, Bachelet's expected vote-share among respondents in the sample would have peaked at more than 57% under this scenario. Moreover, Hirsch's support in this case would remain essentially unchanged in comparison to the actual four-candidate race. Hence, some voters who would support the minority candidate in an election with Bachelet and Lavín as real contenders for the presidential office would chose to cast a 'useful' vote for the *Concertación* in a less polarized election in which Piñera was the only nominee of the *Alianza*. Finally, although Christian Democrats were twice as likely to vote for Piñera than for Lavín in the actual election (Figure 3.1), Figure 3.3 shows that their propensity to vote for Piñera and for Lavín in the two hypothetical three-candidate races would not be significantly different. This result indicates that some respondents who voted for the National Renewal Party in order to prevent Lavín from advancing to the second round would have had no incentive to do so in a three-candidate election with Bachelet and Piñera as the main aspirants for office.

Therefore, our findings underscore the fact that the political and institutional context played a key role in explaining Piñera success and *Alianza*'s unprecedented support in the 2005 election. In particular, tactical voting seems to have been an important determinant of the support for the RN candidate. Interestingly, however, the strategic calculus of Chilean voters corresponds only in part to the predictions of received models of strategic vote under top-two runoff (Myerson and Weber, 1993; Cox, 1997). In line with the theoretical literature, Piñera's support among UDI sympathizers might reflect the desire of conservative voters to coordinate on the candidate of the *Alianza* that, according to public opinion polls, stood the best chance in a two-candidate runoff against Bachelet. In contrast, his support among *Concertación* partisans and potential Hirsch voters indicates that strategic voting among center-left and left-wing respondents stemmed from their desire to exclude Lavín from the *ballotage*, rather than to improve the probability of victory for their most-favored candidate (Cox, 1997).

Again, this calls attention to the persistence of the authoritarian/democratic cleavage and its influence on the election results. The prevalence of this division in the society, together with the broad consensus over economic and social issues and strong party and coalition labels, impose powerful limitations to party leaders and political elites intending to alter their long-established electoral coalitions (Valenzuela and Scully, 1997). Their success in doing so might depend on their ability to shift their bases of electoral competition by politicizing new dimensions of conflict that cross coalitions and partisan lines, such as the ones underlying the growing "demand for rights" among the electorate and the increased debate over moral and social issues (Torcal and Mainwaring, 2003; Angell and Reig, 2006).

3.5 Final remarks

The 2005 election in Chile had several unusual characteristics. Among them, the presence of two viable conservative candidates marked a clear difference with previous elections. In this chapter, we specify and fit a Bayesian multinomial probit model to study the presidential race, accounting for the multi-party character of the election and allowing estimation of substitution patterns among the candidates that enable us to assess how their expected vote-shares would change under alternative electoral scenarios. The Bayesian approach is particularly well suited for analyzing this election, given its advantages over classical estimation techniques for dealing with a relatively large number of alternatives and small sample sizes, as well as for providing a practical way of testing competing hypothesis and statistical models.

Our results shed light on the debate about the transformation of the political system in Chile and the redefinition of voters' preferences since the re-establishment of democracy. In line with Tironi and Agüero's hypothesis (1999), we find that voters' regime preferences and their attitudes towards democracy played a substantial role in the choice between the candidates of the *Concertación* and the *Alianza*. Earlier works suggested that the authoritarian-democratic cleavage would probably lose its influence over time as the memories of the dictatorship receded, as democracy was consolidated and as parties found new political issues to mobilize their supporters (Valenzuela and Scully, 1997; Torcal and Mainwaring, 2003). However, the empirical evidence from the 2005 election shows that, sixteen years after the plebiscite that marked the end of Pinochet's rule, voters' electoral behavior still reflects the durability of the division between supporters and critics of the military regime.

In addition, our analysis underscores the considerable impact of the particular electoral context of the 2005 presidential race on voter choice, an aspect that has received relatively little attention in previous analyses of Chilean elections. Specifically, we show that the entry of a second conservative candidate into the presidential race increased the vote of the right, gathering the support of some Christian Democrats and, especially, of voters who were not inclined to favor either Bachelet or Lavín. We also find that much of Piñera's support was due to strategic calculus on the part of voters in view of the almost certain second-round runoff and the tight contest between Lavín and Piñera for the second place in the first round of the election. An in-depth analysis of this argument, however, requires developing a statistical model to estimate the amount of strategic voting in multiparty elections under top-two runoff, an extension that we leave for further research.

Chapter 4

Correcting for Survey Misreports using Auxiliary Information with an Application to Estimating Turnout[40]

4.1 Introduction

Much of the empirical work in the social sciences is based on the analysis of survey data. However, as has been widely documented (Battistin 2003; Bound, Brown and Mathiowetz 2001; Poterba and Summers 1986), these data are often plagued by measurement errors. There are many possible sources for such errors. Interviewers may erroneously record answers to survey items, and respondents may provide inaccurate responses due to an honest mistake, misunderstanding or imperfect recall (Gems, Ghaosh and Hitlin 1982; Hausman, Abrevaya and Scott-Morton 1998; Molinari 2003). Also, as underscored by the social psychology literature, survey respondents tend to overreport socially desirable behaviors and underreport socially undesirable ones (Cahalan 1968; Loftus 1975). In the case of discrete or categorical variables, mismeasurement problems have been traditionally referred to as “misclassification" errors (Aigner 1973; Bollinger 1996; Bross 1954).

In the political science literature, concerns about misclassification have been particularly prevalent in the analysis of voting behavior. Empirical studies of the determinants of voter

[40] Joint with Jonathan N. Katz. Both authors contributed equally to the following chapter. Forthcoming, *American Journal of Political Science* 54(3), July 2010.

turnout focus on how the probability of an individual voting varies according to relevant observable factors, such as citizen's level of political information, registration laws, or demographic characteristics. That is, these studies are interested in estimating the conditional distribution of the turnout decision given certain characteristics of interest.[41] The decision to vote, however, is typically not observed due to the use of secret ballot in the U.S. Furthermore, even if we could observe turnout from the official ballots we would not, in general, be able to observe all the characteristics - e.g., the voter's policy preferences or information about the candidates - that presumably affect the decision. Hence, political scientist rely on the use of survey instruments, such as the American National Election Study (ANES) or the Current Population Survey (CPS), that include both measures of respondents' relevant characteristics and their self-reported voting behavior. This almost always leads to estimation of the common logit or probit models, since the turnout decision is dichotomous, although there are alternatives such as scobit (Nagler 1994) or non-parametric models (Hardle 1990) for discrete choice models.

However, it has been long established that some survey respondents misreport voting, i.e., they report that they have voted when in fact they did not do so (Burden 2000; Clausen 1968; Katosh and Traugott 1981; Miller 1952; Parry and Crossley 1950; Sigelman 1982; Silver, Anderson and Abramson 1986). The evidence that misreporting is a problem can be found in a series of validation studies that the ANES conducted in 1964, 1976, 1978, 1980, 1984, 1988 and 1990. These validation studies were possible, but expensive, because

[41] The literature is far too vast to even begin to fully cite here. See Aldrich (1993) for a review of the theoretical literature and Wolfinger and Rosenstone (1980) for an influential empirical study.

voting is a matter of public record, although for whom a voter voted is not. After administering a post-election survey to a respondent, an official from the ANES was sent to the respondent's local registrar of elections to see if in fact they were recorded as having voted in the election. This is not an easy task, since respondents often do not know where they voted, election officials differ in their ability to produce the records in a usable form, and there might be differences between the survey data and the public records due to errors in spelling or recording. This means that the validated data may also be mismeasured, but for this chapter we will assume it is correct. That said, the ANES for these years included both the respondent's self-reported vote and the validated vote. The differences between the two measures are fairly shocking. Depending on the election year, between 13.6 and 24.6 percent of the respondents claiming to have voted did in fact not according to the public records.[42] In contrast, only between 0.6% and 4.0% of the respondents in the 1964 - 1990 validated surveys who reported not having voted did vote according to the official records. Since there is no reason to believe that measurement errors should mainly be of false positives - i.e., reporting voting when the official record contradicts this claim - , this lends some credence to the social pressures argument for misreporting (Bernstein, Chadha and Montjoy 2001) and should help mitigate some of our concerns about other potential sources

[42] The Current Population Survey (CPS) also exhibits considerable turnout overreporting, although the magnitude is substantially lower than for the ANES (Highton 2004). As shown by Hausman, Abrevaya and Scott-Morton (1998) and Neuhaus (1999), however, even modest amounts of misreporting can affect parameter estimates.

of classification errors, such as inaccurate records.[43] The large differences between reported and validated turnout led to a cottage industry analyzing the causes of misreporting (Abramson and Claggett 1984, 1986a,b, 1991; Ansolabehere and Hersh 2008; Cassel 2003; Hill and Hurley 1984; Katosh and Traugott 1981; Sigelman 1982; Silver, Anderson and Abramson 1986; Weir 1975) and to a debate about how to best measure misreporting (Anderson and Silver 1986). All of these studies find that misreporting varies systematically with some characteristics of interest, but none of them provides an estimation solution to correct for possible misreporting. The open question then is what to do about the problem of respondents misreporting. One possibility would be to use only validated data. At some level this is an appealing option. If we are sure that the validated data is correct, then estimation and inference is straightforward. Unfortunately, collecting the validated turnout data is difficult and expensive, and ANES has stopped doing validation studies for these reasons. Furthermore, even if validation studies were free, some states, such as Indiana, make it impossible to validate votes. Hence, if we are going to limit ourselves to use only fully validated data, our samples will be much smaller. Moreover, would also be throwing away the useful information included in the already collected but non-validated studies.[44] On the other hand, simply ignoring misreporting and using self-

---

[43] Clearly, other reasons besides social desirability may also contribute to explain differences between self-reported and validated turnout (Abelson, Loftus and Greenwald 1992).

[44] In the case of the ANES, turnout is one of the few survey items included since the late 1940s, covering a larger period than any other continuing survey (Burden 2000). Validation studies, on the other hand, only comprise a handful of elections.

reported turnout to estimate standard probit or logit models can result in biased and inconsistent parameter estimates and inaccurate standard errors, potentially distorting the relative impact of the characteristics of interest on the response variable and leading to erroneous conclusions (Davidov, Faraggi and Reiser 2003; Hausman, Abrevaya and Scott-Morton 1998; Neuhaus 1999).[45]

In this chapter we develop a simple Bayesian approach to correct for misreporting, allowing researchers to continue to use the self-reported data while improving the accuracy of the estimates and inferences drawn in the presence of misclassified binary responses.[46] Our model draws on Hausman, Abrevaya and Scott-Morton (1998), but incorporates information on the misreporting process from auxiliary data sources, aiding in identification (Gu 2006; Molinari 2003) and making it easier to avoid the problems that limit the use of Hausman, Abrevaya and Scott-Morton (1998)'s modified maximum likelihood estimator in small samples such as those typically used in political science (Christin and Hug 2004; Gu 2006). While incorporating this information into the analysis

[45] A third strand of research focuses on procedures for reducing the frequency of overreporting, such as altering question wording or reformulating survey questions (Belli, Traugott and Rosenstone 1994; Belli et al. 1999; Bound, Brown and Mathiowetz 2001). Nonetheless, while this can improve the quality of future datasets, we would still be wasting large amounts of data collected in previous surveys.

[46] We focus on the case of misclassified responses and error-free covariates. Several methods have been proposed to adjust for measurement error in the covariates. See Carroll, Ruppert and Stefanski (1995) and Thurigen et al. (2000) for a review.

of the sample of interest using frequentist methods is far from straightforward (Prescott and Garthwaite 2005), this can be easily accomplished within the Bayesian framework via Markov Chain Monte Carlo (MCMC) simulations. Although other Bayesian approaches have been proposed to adjust for misclassification using prior information to overcome fragile or poor identifiability, they either rely exclusively on elicitation of experts' opinions (McInturff et al. 2004; Paulino, Soares and Neuhaus 2003) or assume that information on both the true and the fallible response is available for all subjects in a random subsample of the data (Viana 1994; Prescott and Garthwaite 2002, 2005). In contrast, the information on the misreport patterns incorporated into our model need not come from the sample of interest, and can be combined with elicitation of experts' beliefs if needed. In the empirical application presented in this chapter we will use earlier and small-sample validation studies to correct for misreporting. However, matched official records, administrative registers and possibly even aggregate data might be used to gain this information. Given the potential difficulties of eliciting probabilities from experts' opinions and the scarcity of internal validation designs relative to administrative data sets, external validation studies and other sources of ancillary information (Bound, Brown and Mathiowetz 2001; Garthwaite, Kadane and OHagan 2004; Hu and Riddert 2007; Wiegmann 2005), the correction developed in this chapter provides a more flexible way of incorporating prior information and can be more widely applied than existing approaches.[47] In addition, these alternative

[47] In internal validation studies, the true response is available for a subset of the main study and can be compared to the imperfect or observed response. In the case of the external validation designs, the misreport pattern is estimated using data outside the main study.

approaches focus only on the case in which the misclassification rates are independent of all covariates. As mentioned above, this assumption seems to be inappropriate in the case of the determinants of voter turnout, as well as in many other potential applications. The magnitude and direction of the biases when misreporting is covariate-dependent can be quite different than in the case of constant misclassification rates (Davidov, Faraggi and Reiser 2003; Neuhaus 1999) and, in the context of analyzing voting behavior, Bernstein, Chadha and Montjoy (2001) show that ignoring the correlation between the covariates of interest and the misreport probabilities may seriously distort multivariate explanations of the turnout decision. Finally, our approach enables us to simultaneously address another important problem with survey data, namely missing outcome and/or covariate values, using fully Bayesian model-based imputation (Ibrahim et al. 2005).

Although our model is developed in the context of estimating the conditional probability of turning out to vote, the method is general and will be applicable whenever misclassification of a binary outcome in a survey is anticipated and there is auxiliary information on the misreporting patterns. For instance, our approach could be used to analyze survey data on participation in social welfare programs (Hernanz, Malherbet and Pellizzari 2004), pension plans (Molinari 2003), energy consumption (Gu 2006), employment status (Hausman, Abrevaya and Scott-Morton 1998) and many other areas where we expect to see substantial rates of misreporting and potential correlation between some of the covariates affecting the response and the misreport probabilities. The model can also be implemented when misreporting depends on covariates other than those influencing the outcome. For example, for a substantial proportion of the CPS sample, turnout is measured by proxy, rather than self-reported (Highton 2004). In this case, the misclassification probabilities would be modeled using information on misreporting patterns among household members reporting other members' turnout decision, which could be obtained from validated CPS studies.[48] Extensions of our method to discrete choice models with more than two categories along the lines of Abrevaya and Hausman (1999) and Dustmant and van Soest (2004) are possible as well.

The chapter proceeds as follows. The next section formally lays out the estimation problem in the presence of misreporting and develops our proposed solution. Section 4.3 presents results from a Monte Carlo experiment illustrating how important misreporting can be in practice and comparing the estimates from our method with those obtained under several alternative approaches. We also evaluate the robustness of our approach to misspecification of the misreport model and assess its performances in the presence of both misclassification and missing data. In Section 4.4, we provide three applications of our methodology using data on voter turnout from the ANES. Finally, Section 4.5 concludes.

4.2 Correcting for misreporting in binary choice models

4.2.1 Defining the problem

Let y_i be a dichotomous (dummy) variable, and denote by x_i a vector of individual characteristics of interest. We want to estimate the conditional distribution of y_i given x_i, $P\left[y_i \mid x_i\right]$. However, instead of observing the "true" dependent variable y_i, assume we

[48] We thank an anonymous referee for pointing us to this potential application of our model.

observe the self-reported indicator $\tilde{y}_i$. Most studies use the observed $\tilde{y}_i$ as the dependent variable, typically running either a probit or logit model to estimate $P\left[\tilde{y}_i \mid x_i\right]$.

In order to know whether this substitution can lead to incorrect inferences, we need to know the relationship between $P\left[y_i \mid x_i\right]$ and $P\left[\tilde{y}_i \mid x_i\right]$. We can always write:

$$P\left[\tilde{y}_i = 1 \mid x_i\right] = P\left[\tilde{y}_i = 1 \mid x_i, y_i = 1\right] \bullet P\left[y_i = 1 \mid x_i\right] + \\ P\left[\tilde{y}_i = 1 \mid x_i, y_i = 0\right] \bullet P\left[y_i = 0 \mid x_i\right] \qquad (4.1)$$

by the law of total probability. All that we have done is to rewrite the probability $P\left[\tilde{y}_i = 1 \mid x_i\right]$ into two components: when the self-reported or observed variable $\tilde{y}_i$ coincides with the true response y_i, and when it does not. Also, noting that $P\left[\tilde{y}_i = 0 \mid x_i, y_i = 1\right] = 1 - P\left[\tilde{y}_i = 1 \mid x_i, y_i = 1\right]$, we can re-write the relationship as:

$$P\left[\tilde{y}_i = 1 \mid x_i\right] = \left(1 - \pi_i^{1|0} - \pi_i^{0|1}\right) P\left[y_i = 1 \mid x_i\right] + \pi_i^{1|0} \qquad (4.2),$$

where $\pi_i^{1|0} = P\left[\tilde{y}_i = 1 \mid x_i, y_i = 0\right]$ is the probability that the respondent falsely claims $\tilde{y}_i = 1$ when in fact $y_i = 0$, and $\pi_i^{0|1} = P\left[\tilde{y}_i = 0 \mid x_i, y_i = 1\right]$ is the probability the observed response takes the value 0 when the true response is $y_i = 1$. It is important to note that the probability of each type of misreporting is conditional on x_i.

Standard methods for estimating binary choice models generally assume that the conditional distribution of the dependent variable given x_i is known up to a parameter vector β. However, unless $\pi_i^{1|0} = \pi_i^{0|1} = 0 \; \forall i$, estimating the conditional probability $P\left[\tilde{y}_i = 1 | x_i\right]$ rather than $P\left[y_i = 1 | x_i\right]$ will generally lead to biased estimates of β and inaccurate standard errors, with even small probabilities of misreporting potentially leading to significant amounts of bias (Davidov, Faraggi and Reiser 2003; Hausman, Abrevaya and Scott-Morton 1998; Neuhaus 1999). In addition, the marginal effect of covariate x on the observed response $\tilde{y}_i$ and on the true response y_i will differ by:

$$\begin{aligned}\frac{\partial P\left[\tilde{y}_i = 1 | x_i\right]}{\partial x} - \frac{\partial P\left[y_i = 1 | x_i\right]}{\partial x} &= -\left(\frac{\partial \pi_i^{1|0}}{\partial x} + \frac{\partial \pi_i^{0|1}}{\partial x}\right) P\left[y_i = 1 | x_i\right] \\ &\quad - \left(\pi_i^{1|0} + \pi_i^{0|1}\right) \frac{\partial P\left[y_i = 1 | x_i\right]}{\partial x} + \frac{\partial \pi_i^{1|0}}{\partial x}\end{aligned} \qquad (4.3).$$

As a result, inferences drawn on the relationship between the covariates of interest and the response variable may change substantially when estimated based on the likelihood function defined by $P\left[\tilde{y}_i = 1 | x_i\right]$ rather than on the true model $P\left[y_i = 1 | x_i\right]$, depending on the distribution of $\beta' x_i$ and the covariate vector x_i, on the prevalence of misclassification and on the relationship between the probabilities of misreporting and the covariates in x_i (Bernstein, Chadha and Montjoy 2001; Hausman, Abrevaya and Scott-Morton 1998; Neuhaus 1999).

Different parametric models have been proposed to correct for misclassification of the dependent variable in binary choice models (Carroll, Ruppert and Stefanski 1995; Hausman, Abrevaya and Scott-Morton 1998; McInturff et al. 2004; Morrissey and Spiegelman 1999; Paulino, Soares and Neuhaus 2003; Prescott and Garthwaite 2002, 2005).[49] In particular, Hausman, Abrevaya and Scott-Morton (1998) proposed a modified maximum likelihood estimator that requires the “monotonicity" condition $\pi_i^{1|0} + \pi_i^{0|1} < 1$ to achieve identification. Using Monte Carlo simulations, they showed that their model consistently estimates the extent of misclassification and the parameter vector β, at least in large samples. More recently, however, Christin and Hug (2004) replicated the work of Hausman, Abrevaya and Scott-Morton (1998) for different sample sizes, and found that the modified maximum likelihood estimator performed consistently better than simple probit models ignoring misclassification only in samples of 5,000 or more observations. In smaller samples, standard probit estimators outperformed it in many cases, and Christin and Hug (2004) concluded that the modified maximum likelihood estimator is only advisable for large samples. As noted by Gu (2006), the failure of Hausman, Abrevaya and Scott-Morton (1998)'s estimator in small samples is likely due to the insufficiency of the monotonicity condition to ensure model identification. For such sample sizes typically available in political science, even moderate rates of misclassification may hinder model identification, so different assumptions may be required to put bounds on the misclassification rates and the regression coefficients. In addition, Hausman, Abrevaya and

[49] A comprehensive review of different methods developed to deal with misclassification and measurement errors in nonlinear models can be found in Carroll, Ruppert and Stefanski (1995).

Scott-Morton (1998) and, in fact, most empirical applications of models proposed to correct for misreporting, assume constant misclassification rates, failing to account for the potential influence of the covariates of interest on $\pi_i^{1|0}$ and $\pi_i^{0|1}$.[50] Relevant prior information on the misreport patterns is often available from auxiliary data sources, such as internal or external validation studies, small sample pilots or administrative registers, which can be used to impose restrictions on the misreport probabilities and regression coefficients to aid in identification and improve inferences on the relationship between $\boldsymbol{x}$ and y (Chen 1979; Molinari 2003).

In order to incorporate the information on the misreporting structure from auxiliary data sources, we propose a Bayesian approach based on Markov Chain Monte Carlo (MCMC) methods. This approach has three basic advantages in this setting. First, results from previous statistical studies can be easily incorporated into the model for the sample of interest within the Bayesian framework (Dunson and Tindall 2000; Ibrahim and Chen 2000; Ibrahim, Ryan and Chen 1998). Second, MCMC methods directly account for the extra uncertainty in the variances caused by using estimates of the misreport probabilities obtained from the auxiliary data instead of their true values. In contrast, in the context of frequentist estimation, this would require additional "post-estimation" steps, such as bootstrapping (Haukka 1995), applying the results of Murphy and Topel (1985) for two-

[50] Abrevaya and Hausman (1999); Hausman, Abrevaya and Scott-Morton (1998) and Paulino, Soares and Neuhaus (2003), among others, discuss extensions to deal with covariate-dependent misclassification, but they do not analyze this case in practice.

step estimators, or using numerical techniques (Kuha 1994).[51] In addition, our approach does not rely on large sample assumptions and avoids the need for complicated numerical approximations (Viana 1994) when the posterior distributions are analytically intractable. The model is a simple modification of Hausman, Abrevaya and Scott-Morton (1998)'s estimator and can be easily implemented by practitioners and applied researchers using flexible and freely available software for Bayesian analysis such as WinBUGS or JAGS (Plummer 2009; Spiegelhalter, Thomas and Best 2003).

4.2.2 A Bayesian model to correct for misreporting using auxiliary data

We are interested in accurately estimating the effect of the individual characteristics of interest on the conditional distribution of the true response. Hence, the focus of our analysis lies in the marginal posterior distribution of β, while the modelization of the conditional probabilities $\pi_i^{1|0}$ and $\pi_i^{0|1}$ can be regarded as "instrumental".

Since the observed response variable is dichotomous, we can start by assuming that, conditional on some set of relevant individual characteristics, the observations are independently and identically distributed according to a Bernoulli distribution - as in

[51] Another possible approach is to assume that misclassification rates are known and equal to those prevalent in the auxiliary data (Poterba and Summers 1995). Nonetheless, as noted by Hausman, Abrevaya and Scott-Morton (1998), not only will this lead to inconsistent parameter estimates if the assumed misclassification probabilities are not consistent estimates of the true probabilities, but the standard errors of the coefficient estimates will be understated.

Hausman, Abrevaya and Scott-Morton (1998). The probability of the sample can therefore be written as:

$$L\left(\theta \mid \tilde{\boldsymbol{y}}, \boldsymbol{x}\right) = \prod_{i=1}^{N} P\left[\tilde{y}_i \mid \boldsymbol{x}_i, \theta\right]^{\tilde{y}_i} \left(1 - P\left[\tilde{y}_i \mid \boldsymbol{x}_i, \theta\right]\right)^{1-\tilde{y}_i} \qquad (4.4),$$

with $\theta = \left\{\pi_i^{1|0}, \pi_i^{0|1}, \beta^{'}\right\}$. We will further assume that the conditional probability of the true response variable is given by $P\left[y_i = 1 \mid x_i\right] = F\left(\beta^{'} x_i\right)$, where $F\left(\bullet\right)$ is some cumulative density function. For ease of exposition, we use the probit link, so that $F\left(\bullet\right)$ is the standard normal distribution denoted by $\Phi\left(\bullet\right)$. This will lead to a probit model with a correction for misreport; the use of the logit link function would result in a logit model with a correction for misreporting. We also assume that $P\left[y_i = 1 \mid x_i\right]$is a priori independent of $\pi_i^{1|0}$ and $\pi_i^{0|1}$.[52] Substituting for $P\left[\tilde{y}_i \mid x_i\right]$in Equation 4.2 and denoting by $\mathcal{S}$ the sample data, we arrive at:

$$L\left(\pi_i^{1|0}, \pi_i^{0|1}, \beta \mid \mathcal{S}\right) = \prod_{i=1}^{N} \left[\left[\left(1 - \pi_i^{1|0} - \pi_i^{0|1}\right)\Phi\left(\beta^{'} x_i\right) + \pi_i^{1|0}\right]^{\tilde{y}_i} \times \left[\left(1 - \pi_i^{1|0} - \pi_i^{0|1}\right)\left(1 - \Phi\left(\beta^{'} x_i\right)\right) + \pi_i^{0|1}\right]^{1-\tilde{y}_i} \right] \qquad (4.5),$$

[52] This assumption simplifies the analysis considerably without entailing any obvious drawback from a practical perspective (Paulino, Soares and Neuhaus 2003).

which represents the probability of observing the sample under misreporting. The joint posterior density of $\theta = \left\{\pi_i^{1|0}, \pi_i^{0|1}, \beta^{'}\right\}$ is therefore given by:[53]

$$p\left(\pi_i^{1|0}, \pi_i^{0|1}, \beta \mid \mathcal{S}\right) \propto L\left(\pi_i^{1|0}, \pi_i^{0|1}, \beta \mid \mathcal{S}\right) \times p\left(\pi_i^{1|0}, \pi_i^{0|1}, \beta\right) \qquad (4.6).$$

Without prior substantive information, a common choice for $p\left(\pi_i^{1|0}\right)$ and $p\left(\pi_i^{0|1}\right)$ would be vague *Beta* distributions, while independent normal priors with zero means and (possible common) large variances could be assigned for the components of β (McInturff et al. 2004; Prescott and Garthwaite 2005). However, as mentioned above, using at priors for the misclassification errors will likely lead to poor identifiability (Gu 2006). In addition, specifying diffuse priors for β can also hinder convergence in some circumstances (Gu 2006; McInturff et al. 2004; Prescott and Garthwaite 2002). Incorporating prior information on $\pi_i^{1|0}, \pi_i^{0|1}$, and β from auxiliary data sources can help overcome these problems and improve the accuracy of the parameter estimates (Gu 2006; McInturff et al. 2004; Prescott and Garthwaite 2002, 2005).

Suppose that both the true and the self-reported dependent variables are recorded for all respondents in a validation study of size M. Comparing y_j to $\tilde{y}_j$ for every $j = 1, ..., M$, we can estimate the misreport probabilities for the validated sample. Let z_j^1 and z_j^2 denote sets of regressors that are useful in predicting the conditional probabilities

[53] Alternatively, a "latent variable" approach based on data augmentation (Tanner and Wong 1987) can be used to simplify the computations. See McInturff et al. (2004).

$\pi_j^{1|0}$ and $\pi_j^{0|1}$, where the notation allows for the fact we may use different regressors to predict the two types of misreporting. z_j^1 and z_j^2 may include some or all of the variables in $\boldsymbol{x}$, as well as other variables not affecting the true response. Again, for ease of exposition, we assume probit link functions and specify the conditional probabilities of misreporting as $\pi_j^{1|0} = \Phi\left(\gamma_1' z_j^1\right)$ and $\pi_j^{0|1} = \Phi\left(\gamma_2' z_j^2\right)$. Since our interest lies primarily on the distribution of β, $\gamma = \left\{\gamma_1', \gamma_2'\right\}$ could in principle be viewed as "nuisance" parameters in our setting (Ibrahim, Ryan and Chen 1998), although they help provide meaningful interpretations for the underlying misreporting process (Chen 1979).[54] Letting $\mathcal{V}$ denote the data from the validation study, the likelihood from $\mathcal{V}$ is:

$$\begin{aligned} L\left(\gamma_1, \gamma_2, \beta \mid \mathcal{V}\right) = & \prod_{j=1}^{M} \Phi\left(\beta' x_j\right)^{y_j} \left(1 - \Phi\left(\beta' x_j\right)\right)^{1-y_j} \times \\ & \prod_{y_j=1} \Phi\left(\gamma_1' z_j^1\right)^{\tilde{y}_j} \left(1 - \Phi\left(\gamma_1' z_j^1\right)\right)^{1-\tilde{y}_j} \times \\ & \prod_{y_j=0} \Phi\left(\gamma_2' z_j^2\right)^{1-\tilde{y}_j} \left(1 - \Phi\left(\gamma_2' z_j^2\right)\right)^{\tilde{y}_j} \end{aligned} \quad (4.7).$$

The posterior distributions $p\left(\gamma_1, \gamma_2, \beta \mid \mathcal{V}\right)$ or $p\left(\gamma_1, \gamma_2 \mid \mathcal{V}\right)$ could then be used to specify the priors for β, γ_1 and γ_2 in the model fit to the sample of interest by repeated application of Bayes' theorem. However, since these posteriors cannot be expressed as

[54] It is worth mentioning, however, that $\pi_j^{1|0}\left(z_j^1\right)$ and $\pi_j^{0|1}\left(z_j^2\right)$ are not necessarily identified. See Lewbel (2000).

tractable distributions, there is no straightforward way of transferring the relevant information from the validation study to the analysis of the main sample (Prescott and Garthwaite 2005). In addition, unless the validation study is a random sub-sample of the main study, heterogeneity between the two samples might in some circumstances lead to misleading conclusions if inference on β is based on the pooled datasets (Duan 2005). Hence, we consider both samples simultaneously, combining the likelihoods in 4.5 and 4.7 with vague independent priors $p(\gamma_1)$, $p(\gamma_2)$ and $p(\beta)$ and weighting the likelihood from the validated sample by a "tunning" parameter δ that controls how much influence the validated data has relative to the main sample (Chen, Ibrahim and Shao 2000; Ibrahim and Chen 2000). The joint posterior density of the unknown parameters is therefore given by:

$$p\left(\pi_i^{1|0},\pi_i^{0|1},\beta \mid \mathcal{S}\right) \propto L\left(\pi_i^{1|0},\pi_i^{0|1},\beta \mid \mathcal{S}\right) \times L\left(\gamma_1,\gamma_2,\beta \mid \mathcal{V}\right)^{\delta} \times p\left(\gamma_1\right) \times p\left(\gamma_2\right) \times p\left(\beta\right) \quad (4.8),$$

with $0 \le \delta \le 1$, where $\delta = 0$ corresponds to the case in which no auxiliary information is incorporated into the analysis for the main sample, while $\delta = 1$ gives equal weights to $L\left(\pi_i^{1|0},\pi_i^{0|1},\beta \mid \mathcal{S}\right)$ and $L\left(\gamma_1,\gamma_2,\beta \mid \mathcal{V}\right)$. δ can be assigned either a fixed value or a prior distribution - e.g., $\delta \sim Beta(c,d)$ (Chen, Ibrahim and Shao 2000; Ibrahim and Chen 2000).[55] Although Equation 4.8 is intractable analytically, inference can be performed

---

[55] In the latter case, the prior for $p(\delta)$ would be added to Equation 3.8. See the discussions in Chen, Ibrahim and Shao (2000) and Ibrahim and Chen (2000) for additional details.

using Gibbs sampling along with Metropolis steps to sample the full conditionals for β, γ_1 and γ_2 (Gelfland and Smith 1990; Casella and George 1992; Chib and Greenberg 1995). Under mild regularity conditions (Gilks, Richardson and Spiegelhalter 1996; Robert and Casella 2004), for a sufficiently large number of iterations, samples from these conditional distributions approach samples from the joint posterior. The posterior marginals obtained from these convergent samples can then summarized and used to estimate the effect of the relevant individual characteristics on the true response and the misreport probabilities. In addition, Bayes factors can be easily implemented within our modeling framework to compare alternative link functions (Paulino, Soares and Neuhaus 2003).

Thus, we only need to have validated data from a previous sample or for a sub-sample of the respondents in order to correct for misreporting in the model for the main study. In case several validation studies are available, they can be easily integrated into our analysis by adapting the method proposed in Ibrahim and Chen (2000) to incorporate historical data in binary choice models, substituting $L(\gamma_1, \gamma_2, \beta \mid \mathcal{V})$ in Equation 4.8 by:

$$\prod_{d=1}^{D} L(\gamma_1, \gamma_2, \beta \mid \mathcal{V})^{\delta_d} \qquad (4.9),$$

where $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_D\}$ denotes the data from D validation samples and $\delta = \{\delta_1, \delta_2, ..., \delta_D\}$, $0 \le \delta_d \le 1$ can be assigned I.I.D. beta priors (Ibrahim and Chen 2000; Ibrahim, Ryan and Chen 1998). Note that, while we must assume that the same error structure appears in the validated and nonvalidated samples and that the process generating misreporting is similar in both datasets, the covariates included in $\boldsymbol{x}$ and $\boldsymbol{z} = \{z_1, z_2\}$ do not have to be necessarily identical for both datasets. For instance, when estimating the determinants of the turnout decision, we could allow for election-specific factors affecting the turnout and the misreport probabilities, combining information from validation studies with experts' opinions, theoretical restrictions or even specifying diffuse priors for some of the predictors. Covariates that were not measured in previous studies can be incorporated into the analysis of the sample of interest by specifying the priors for these new covariates through the “initial” prior $p(\gamma_1, \gamma_2, \beta)$ in Equation 4.8 (Ibrahim et al. 2005).

Even if we did not have access to a validation sample, several other sources of information, such as administrative records or even aggregate data could be used to impose informative constraints on the misclassification rates and improve the parameter estimates. For example, in the analysis of voter turnout, we could observe turnout rates in small geographic areas, such as counties or congressional districts, that could be used to specify the misreport probabilities for all individuals in the sample belonging to a given area. While it will not be generally possible to specify a generalized linear model of misreporting in such circumstances, hierarchical beta priors can be used to summarize auxiliary information available on misreporting patterns by location or relevant socio-demographic characteristics (Dunson and Tindall 2000). Finally, if no relevant information to predict misreporting exists either in validation studies or other auxiliary data, constraints on the misreport probabilities could be imposed *via* elicitation of experts' opinions. Our model would then be virtually identical to McInturff et al. (2004) and Paulino, Soares and Neuhaus (2003).

Despite the advantages of our approach, it is worth mentioning that, like all parametric estimators, our model might be quite sensitive to distributional and modeling assumptions. Although semi-parametric methods have been used to estimate discrete choice models with misclassified dependent variables (Abrevaya and Hausman 1999; Hausman, Abrevaya and Scott-Morton 1998; Dustmant and van Soest 2004), they are also subject to potential misspecification (Molinari 2003). Moreover, in the case of covariate-dependent misclassification, available semi-parametric techniques require either sacrificing identification of some of the parameters in β (Abrevaya and Hausman 1999) or complex computations that are not likely to be attractive for practitioners and empirical researchers (Lewbel 2000). A different approach would be to adapt and implement non-parametric methods based on Manski (1985), Horowitz and Manski (1995) and Molinari (2003). In particular, the "direct misclassification approach" proposed by the latter allows incorporating prior information on the misreporting pattern to obtain interval identification of parameters of interest, and can be easily applied to the case in which misclassification depends on perfectly observed covariates with relatively little computational cost. However, as is well known, non-parametric methods are subject to the curse of dimensionality, which can pose a problem in applications where the misreporting probabilities might depend on a relatively large set of covariates, and is uncertain whether point identification can be achieved in this setting (Hu 2008). To the best of our knowledge, there is very little research comparing the performance of parametric versus non-parametric methods to correct for covariate-dependent misclassification and evaluating the relative weaknesses and advantages of both approaches in applied work.

4.2.3 Extending the model to account for missing data

Besides measurement errors, survey data is often plagued with large proportions of missing outcome and covariate values due to non-response or loss of data. As is well known, unless the data are missing completely at random (MCAR), using list-wise deletion and restricting the analysis only to those respondents who are completely observed can lead to biased estimates (Little and Rubin 2002; Chen et al. 2008).[56] Furthermore, even if the data are MCAR, complete-case analyses may lead to discard a large proportion of observations and can be therefore quite inefficient (Ibrahim et al. 2005). Ad-hoc approaches to dealing with missing data, such as excluding covariates subject to missingness from the analysis or using mean imputation, are easy to implement but exhibit several potential problems such as biased estimates, inefficiency and misspecification (Chen et al. 2008; Ibrahim et al. 2005; Gelman and Hill 2007).[57] On the

[56] It is worth mentioning, however, that there are situations in which inference based on a complete-case analysis might yield unbiased estimates and outperform imputation methods even when the data are not missing completely at random (Little and Wang 1996).

[57] A detailed review of different methods commonly used to handle missing data is beyond the scope of this paper. See Horton and Kleinman (2007), Ibrahim et al. (2005), Little and Rubin (2002) and Schafer and Graham (2002), among others, for a detailed discussion.

other hand, Bayesian methods such as the one presented in this paper can easily accommodate missing data. There is no distinction between missing data and parameters within the Bayesian framework, and thus inference in this setting essentially requires defining a prior for the missing values and sampling from the joint posterior distribution of the parameters and missing values, incorporating just an "extra-layer" in the Gibbs sampling algorithm compared to the complete-case analysis (Gelman et al. 2004; Ibrahim et al. 2005). In particular, our model can be immediately extended to deal with missing response and covariate values, including cases with missing responses alone, with missing covariates alone, and with missing covariates and responses. This allows us to accommodate item and unit nonresponse in both the main and the validation studies.[58]

Let $\boldsymbol{w}_i = \left(w_{i,1}, w_{i,2}, \ldots w_{i,p}\right)'$, $i = 1,..,N$ denote a $p \times 1$ vector of covariates included in $\boldsymbol{x}_i$, $\boldsymbol{z}_i^1$ and $\boldsymbol{z}_i^2$, and denote the marginal density of $\boldsymbol{w}_i$ by $p\left(\boldsymbol{w}_i \mid \alpha\right)$, where α parametrizes the joint distribution of the covariates. Adopting the notation in Chen et al. (2008), we write $\boldsymbol{w}_i = \left(\boldsymbol{w}_{i,obs}, \boldsymbol{w}_{i,mis}\right)$, where $\boldsymbol{w}_{i,mis}$ is the $q_i \times 1$ vector of missing components of $\boldsymbol{w}_i$, $0 \le q_i \le p$, and $\boldsymbol{w}_{i,obs}$ is the observed portion of $\boldsymbol{w}_i$. Similarly, we use $\tilde{y}_{i,mis}$ if the self-reported outcome $\tilde{y}_i$ is missing, and $\tilde{y}_{i,obs}$ otherwise. We assume that the missing data mechanism is *ignorable* (Rubin 1976; Little and Rubin 2002). That is, we assume that the missing data mechanism does not depend on the missing values, but may depend on the observed outcome and covariate data included in the model - i.e., the data

58 However, as seen in Equation 3.11 below, respondents with completely missing outcomes and covariates do not contribute to the likelihood function.

are missing at random (MAR) - and that the parameters governing the missing data mechanism are distinct from the parameters of the sampling model. The observed-data likelihood for the main study can then be written as:

$$
\begin{aligned}
L\left(\gamma_1,\gamma_2,\beta \mid \mathcal{S}_{obs}\right) = & \prod_{\tilde{y}_{i,obs},\mathbf{w}_i=\mathbf{w}_{i,obs}} p\left(\tilde{y}_i \mid \mathbf{w}_i,\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_i \mid \alpha\right)\times \\
& \prod_{\tilde{y}_{i,obs},\mathbf{w}_i=\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis}\right)} \int p\left(\tilde{y}_i \mid \mathbf{w}_i,\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis} \mid \alpha\right)d\mathbf{w}_{i,mis}\times \\
& \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=\mathbf{w}_{i,obs}} \int p\left(\tilde{y}_{i,mis} \mid \mathbf{w}_i,\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_i \mid \alpha\right)d\tilde{y}_{i,mis}\times \\
& \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis}\right)} \int\int p\left(\tilde{y}_{i,mis} \mid \mathbf{w}_{i,mis},\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis} \mid \alpha\right)d\tilde{y}_{i,mis}d\mathbf{w}_{i,mis}\times \\
& \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=\mathbf{w}_{i,mis}} \int\int p\left(\tilde{y}_{i,mis} \mid \mathbf{w}_i,\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis} \mid \alpha\right)d\tilde{y}_{i,mis}d\mathbf{w}_{i,mis}
\end{aligned} \tag{4.10}
$$

which, as noted by Chen et al. (2008), reduces to:

$$
\begin{aligned}
L\left(\gamma_1,\gamma_2,\beta \mid \mathcal{S}_{obs}\right) = & \prod_{\tilde{y}_{i,obs},\mathbf{w}_i=\mathbf{w}_{i,obs}} p\left(\tilde{y}_i \mid \mathbf{w}_i,\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_i \mid \alpha\right)\times \\
& \prod_{\tilde{y}_{i,obs},\mathbf{w}_i=\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis}\right)} \int p\left(\tilde{y}_i \mid \mathbf{w}_i,\gamma_1,\gamma_2,\beta\right)p\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis} \mid \alpha\right)d\mathbf{w}_{i,mis}\times \\
& \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=\mathbf{w}_{i,obs}} p\left(\mathbf{w}_i \mid \alpha\right)\times \\
& \prod_{\tilde{y}_{i,mis},\mathbf{w}_i=\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis}\right)} p\left(\mathbf{w}_{i,obs},\mathbf{w}_{i,mis} \mid \alpha\right)d\mathbf{w}_{i,mis}
\end{aligned} \qquad (4.11).
$$

As suggested by Ibrahim, Chen and Lipsitz (2002), it is often convenient to model the joint distribution $p\left(\mathbf{w}_i \mid \alpha\right)$ as a series of one-dimensional conditional distributions:

$$p\left(w_{i,1},w_{i,2},\ldots w_{i,p}\mid\alpha\right)=p\left(w_{i,p}\mid w_{i,1},w_{i,2},\ldots w_{i,p-1},\alpha_p\right)\times$$
$$p\left(w_{i,p-1}\mid w_{i,1},w_{i,2},\ldots w_{i,p-2},\alpha_{p-1}\right)\times\ldots\times p\left(w_{i,1}\mid\alpha_1\right) \quad (4.12),$$

where $\alpha_l,\ l=1,..,p,$ is a vector of parameters for the $l$lth conditional distribution, the α_l 's are distinct, and $\alpha=\left(\alpha_1,\alpha_2,...,\alpha_p\right)$. As noted by these authors, specification 4.12 has the advantages of easing the prior elicitation for α and reducing the computational burden of the Gibbs algorithm required for sampling from the observed data posterior, and is particularly well-suited for cases in which $\boldsymbol{w}$ includes categorical and continuous covariates. While the modeling of the covariate distributions depends on the order of the conditioning, Ibrahim, Chen and Lipsitz (2002) show that posterior inferences are generally quite robust to changes in the order of the conditioning. Obviously, 4.12 needs to be specified only for those covariates that have missing values. If some of the covariates in $\boldsymbol{w}$ are completely observed for all respondents in a survey, they can be conditioned on when constructing the distribution of the missing covariates.

The joint posterior density of the unknown parameters based on the observed data is then given by:

$$p\left(\gamma_1,\gamma_2,\beta,\alpha\mid\mathcal{S}_{obs}\right)\propto L\left(\gamma_1,\gamma_2,\beta,\alpha\mid\mathcal{S}_{obs}\right)\times p\left(\gamma_1,\gamma_2,\beta,\alpha\right) \quad (4.13).$$

Information on the misreport patterns and on all the parameters of interest can be incorporated from the validation study in essentially identical way as in the case with no missing data. A joint prior for $\left(\gamma_1,\gamma_2,\beta,\alpha\right)$ could be specified as:

$$p(\gamma_1,\gamma_2,\beta,\alpha) \propto L(\gamma_1,\gamma_2,\beta,\alpha \mid \mathcal{V}_{obs})^{\delta} \times p(\gamma_1) \times p(\gamma_2) \times p(\beta) \times p(\alpha) \quad (4.14),$$

where $L(\gamma_1,\gamma_2,\beta,\alpha \mid \mathcal{V}_{obs})$ is obtained from the complete-data likelihood of the validation study:

$$L(\gamma_1,\gamma_2,\beta,\alpha \mid \mathcal{V}_{obs}) = \int\int p(\tilde{\boldsymbol{y}},\boldsymbol{y} \mid \boldsymbol{w},\gamma_1,\gamma_2,\beta,\alpha)\; d\tilde{\boldsymbol{y}}_{mis} d\boldsymbol{w}_{mis} \qquad (4.15)$$

and, as mentioned above, δ is a scalar prior parameter that weights the validated data relative to the data from the main study.[59] Note that our specification allows for missing responses $\tilde{y}_i$ and covariate values in the validated sample as well, and can accommodate cases in which the missing self-reported variable depends on the true y_i. As in the case of no missing data, it is also possible to incorporate only the information from the observed probability of misreporting in the validation study to specify the priors for γ_1, γ_2 and a subset $\alpha_{\boldsymbol{z}}$ of the components of α for the main study, while using diffuse prior distributions for the remaining parameters. However, the additional information obtained from $L(\gamma_1,\gamma_2,\beta,\alpha \mid \mathcal{V}_{obs})$ can increase efficiency in many missing data problems in which certain parameters in the likelihood function are not identifiable and/or very little information is available for inference, particularly when the "gold-standard" measure y_i is observed for a large proportion of the respondents in the validation study (Ibrahim, Chen and Lipsitz 2002; Reilly and Pepe 1995; Robins, Rotnitzky and Zhao 1994).

---

[59] See Section 4 in Ibrahim, Chen and Lipsitz (2002) for details.

In principle, it is possible to extend this approach to the case of non-ignorably missing values following Huang, Chen and Ibrahim (1999), Ibrahim and Lipsitz (1996) and Ibrahim, Lipsitz and Chen (1999). However, there is usually little information on the missing data mechanism, and the parameters of the missing data model are often quite difficult to estimate (Ibrahim, Lipsitz and Horton 2001). The plausibility of the MAR assumption can be enhanced by including additional individual and contextual variables in the model specification (Gelman et al. 2004; Gelman, King and Liu 1998).

4.3 A Monte Carlo experiment

In this section, we conduct a series of simulation analyses aimed at illustrating the problems of ignoring misreporting in practice, comparing the performance of our solution *vis-à-vis* alternative parametric models proposed in the literature to account for misreporting, and assessing the sensitivity of the estimates from our model to the specification of the underlying model of misreporting.

4.3.1 Comparison of alternative approaches to dealing with misreporting

Based on the Monte Carlo design in Neuhaus (1999), we simulated 2,000 observations for two covariates: x_1 is drawn from a standard normal distribution, and x_2 is a dummy variable equal to one with probability $1/2$. The true response y_i was generated as:

$$y_i = I\left(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i \geq 0\right),$$

where $I(E)$ is the indicator function equal to one if E is true and zero otherwise, $(\beta_0, \beta_1, \beta_2) = (-1,1,1)$ and ε_i drawn from a $N(0,1)$ distribution.

The misreport probabilities $\pi_i^{1|0}$ and $\pi_i^{0|1}$ were chosen such that: i) average misclassification rates are symmetric and take values of 2%, 5%, 10% and 20%; ii) different possible relationships between $\boldsymbol{x}_i, \pi_i^{1|0}$ and $\pi_i^{0|1}$ are taken into account. This allows us to determine whether and to what extent ignoring misclassification affects the parameter estimates for different rates of misreporting and for different correlation patterns between the covariates of interest, the true response and the misreport probabilities. For reasons of space, we only present the results for the two basic scenarios considered by Neuhaus (1999), denoted as Designs A and B.[60] In Design A, $\pi_i^{1|0}$ and $\pi_i^{0|1}$ are independent of the covariates in $\boldsymbol{x}$; the observed response $\tilde{y}_i$ was generated by randomly changing y_i according to the constant misreport probabilities. Under Design B, the binary covariate x_2 is assumed to be strongly positively correlated with $\pi_i^{1|0}$ but negatively related to $\pi_i^{0|1}$; $\tilde{y}_i$ in this scenario was generated from y_i as a function of x_2, as indicated in Table 4.A.1 in Appendix 4.A. This corresponds, for instance, to the situation described in previous analysis of voter turnout that found overreporting to be clearly correlated with race (Abramson and Claggett 1984, 1986a,b, 1991; Hill and

---

[60] Simulations were also carried out allowing for $\pi_i^{1|0} \neq \pi_i^{0|1}$. The results for the entire set of simulation exercises are available from the authors upon request.

Hurley 1984; Sigelman 1982). For all simulated datasets, we impose the monotonicity condition $\pi_i^{1|0} + \pi_i^{0|1} < 1$ (Hausman, Abrevaya and Scott-Morton 1998).

In order to apply the methodology developed in Section 4.1.2, we randomly selected half of the observations in the sample and assigned them to be the validation study. The remaining 1,000 observations were assigned to be the main sample under analysis, and we ignored the true response and the information on the misclassification probabilities for these observations, using the information on conditional misreport probabilities from the validation study to fit the model in Equation 4.8 with x_1 and x_2 as regressors both in the response and the misreport models. Since the validation study is a random sub-sample of the main study, a point mass prior δ = 1 with probability 1 was used, equally weighting the validated and main samples.[61] It is worth noting that, following Equation 4.7, we adopt a probit specification for the misreport probabilities, despite the fact that using a probit link to estimate a binary choice model with non-normal error terms can yield biased and inconsistent parameter estimates (Horowitz 1993). However, as mentioned above, the main purpose of our method is to improve inferences on the conditional distribution of the true response given some covariates of interest rather than to estimate the conditional misreport probabilities. Adopting a probit specification commonly used by practitioners for the underlying misreport model allows us to assess how robust are the

[61] Changes in the values of δ have relatively little effect on the parameter estimates in our setting. See the results in Section 3.4.3.

estimates of β to common misspecification errors likely to emerge in applied work.[62] In 4.3.2 we examine the sensitivity of our method to various forms of misspecification of the misreport model in more detail.

We compared the estimates from our method with those obtained using a standard probit model ignoring misreporting, as well as from two alternative approaches proposed in the literature to correct for misclassification. Model A-1 is similar to Hausman, Abrevaya and Scott-Morton (1998)'s parametric estimator, assuming constant misclassification probabilities and ignoring the information from the auxiliary data. Model A-2 also assumes covariate-independent misreporting, but information on the posterior distribution of $\pi_i^{1|0}$ and $\pi_i^{0|1}$ from the validation study is used to define $p\left(\pi_i^{1|0}\right)$ and $p\left(\pi_i^{0|1}\right)$ for the main sample, as suggested in Prescott and Garthwaite (2002, 2005). All models were fit via MCMC methods, assigning independent $N(0,100)$ priors for the components of β. The parameters in γ_1 and γ_2 under model were assigned independent $N(0,100)$ distributions, while independent $Beta(c,d)$ priors were specified for $\pi_i^{1|0}$ and $\pi_i^{0|1}$ under Models A-1 and A-2. c and d were set equal to 1 for Model A-1 and extracted from the misclassification rates in the validated sample for Model A-2.[63]

---

[62] Clearly, researchers should also worry about misspecified response models, but this problem is common to all parametric binary choice models.

[63] See Section 2 in Prescott and Garthwaite (2002). All the models were fit in WinBUGS 1.4. Three parallel chains of length 50,000 with over-dispersed initial values and a 5,000 period

Tables 4.A.2 and 4.A.3 in Appendix 4.A report the posterior means and central 95% credible intervals for all the parameter estimates from the four different estimation approaches, and Figure 4.1 below plots the approximate posterior density for the coefficient of the binary covariate x_2. As the amount of misclassification increases, the point estimates (means and medians) for all the parameters under the standard probit specification become further away from the true values. The central 95% credible intervals from the model ignoring misclassification fail to cover the true $\left(\beta_0, \beta_1, \beta_2\right)$ for average misreport probabilities larger than 5% under Design A, as well as the true coefficients of the simulated covariates under Design B. However, even for average misclassification rates $\pi_i^{1|0}$ and $\pi_i^{0|1}$ as low as 0.05, the true coefficient of x_1 lies outside the credible intervals from the probit model under both experimental designs. In contrast, under our proposed estimation solution, the point estimates for all the parameters are much closer to the true β and the central 95% intervals cover them for all values of $\pi_i^{1|0}$ and $\pi_i^{0|1}$ under the two simulation scenarios. Similar results (not shown) are obtained for the ratios of the estimated coefficients of the simulated covariates with respect to the intercept. We also note that the standard deviations for the estimates under our model tend to be larger than for the simple probit model due to the fact that our model captures the additional uncertainty in the true latent variable y_i induced by misreporting

---

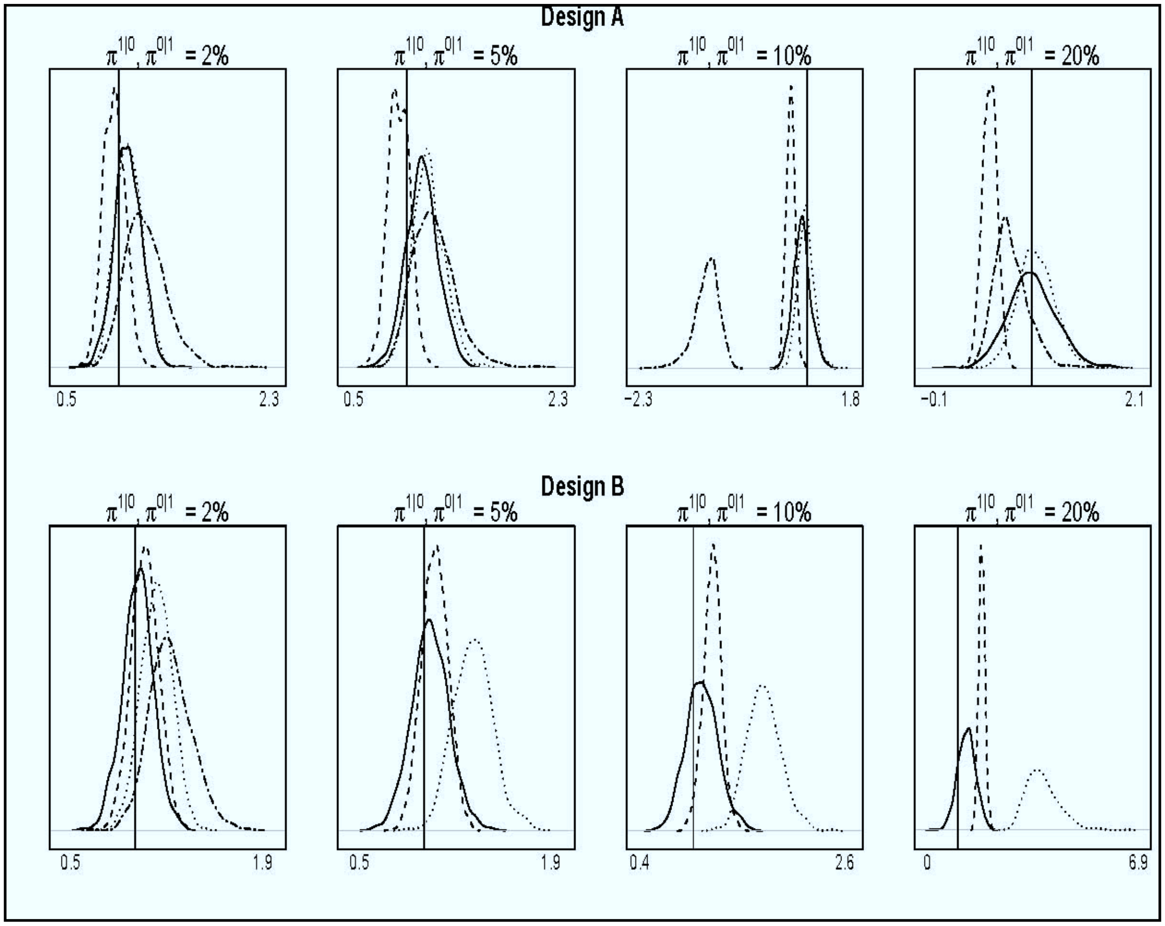
burn-in were run for each model; convergence was assessed based on Gelman and Rubin's estimated Potential Scale Reduction Factor (Gelman and Rubin 1992).

(McGlothlin, Stamey and Seaman 2008; Neuhaus 1999). For the same reason, the standard deviations from our model increase considerably with the amount of misclassification.

A comparison with the two alternative approaches to correct for misreporting shows that, when $\pi_i^{1|0}$ and $\pi_i^{0|1}$ are assumed constant, the estimates from our method do not differ substantially from those obtained under Model A-2. The widths of the central 95% credible intervals are also similar for both models, even though we might have expected that the stronger distributional assumptions about the misreport process adopted in our approach should lead to narrower intervals (Prescott and Garthwaite 2005). In the case of covariate-dependent misreporting, however, the performance of Model A-2 worsens markedly. In particular, as seen Figure 4.1, the posterior mean for β_2 becomes implausibly large as the misclassification rates and the correlation between x_2 and the misreport probabilities increase, and the point estimates for the intercept also become far away from the true β_0. Model A-1, on the other hand, fails to converge for all $\pi_i^{1|0}$ and $\pi_i^{0|1} > 0.02$ under Design B, and performs much worse than our model and Model A-2 also in the case of constant misclassification rates.

Figure 4.1

Estimated posterior densities for β_2 across models



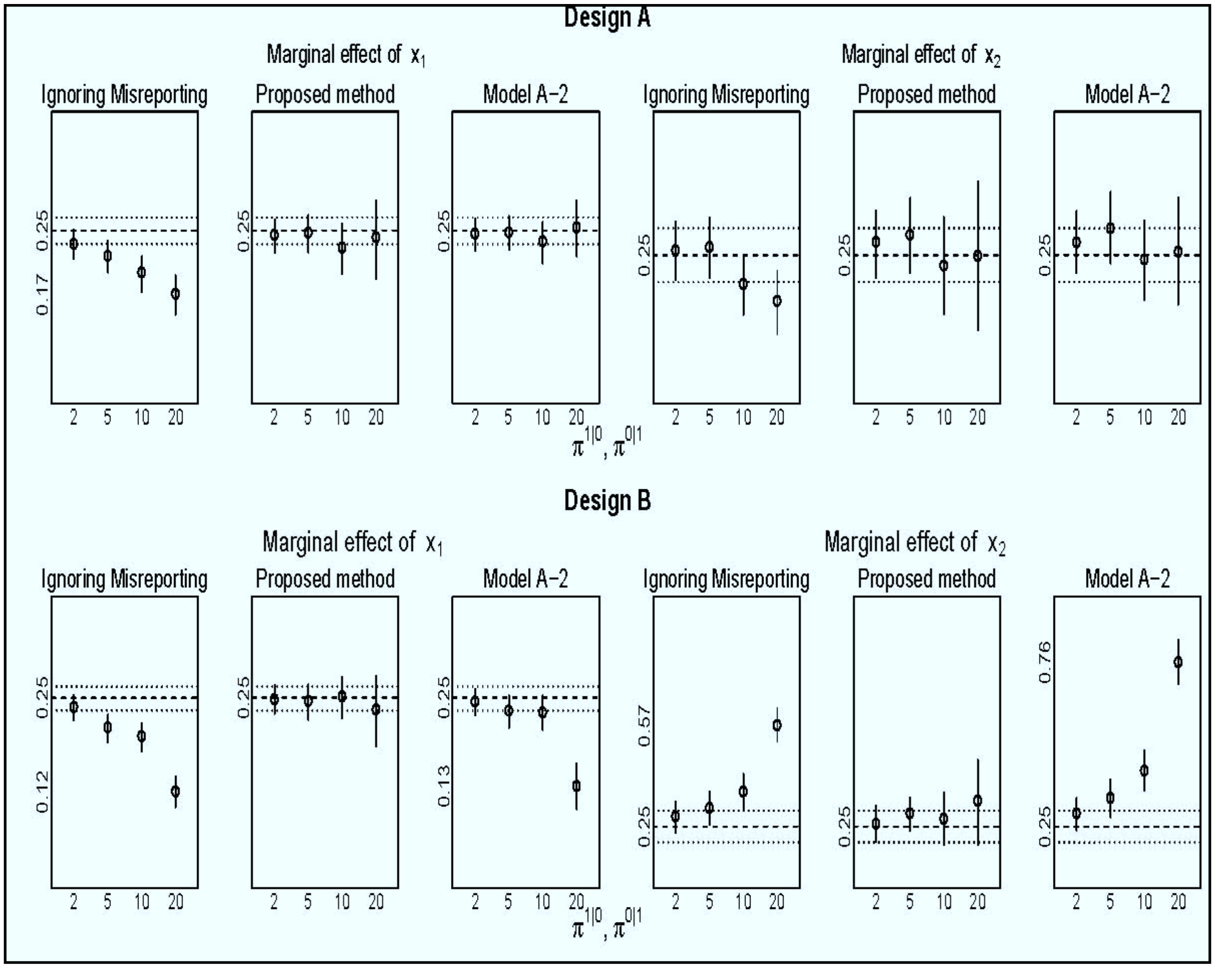
Note: The graph compares marginal posterior density for β_2 under four different estimation approaches: our proposed method (solid curve), a probit model ignoring misreporting (dashed curve), Model A-1 (double-dashed curved) and Model A-2 (dotted curve). The solid vertical line denotes the true parameter value.

Furthermore, as seen in Table 4.A.4 in Appendix 4.A, Model A-1 yields markedly biased and imprecise estimates of the average misclassification rates in the main sample, indicating that assigning diffuse distributions for $p\left(\pi_i^{1|0}\right)$ and $p\left(\pi_i^{0|1}\right)$ when the data provides very little information to estimate the misreport probabilities results in very volatile and inaccurate estimates for the parameters of interest, even when the Hausman, Abrevaya and Scott-Morton (1998)' s identification condition $\pi_i^{1|0}+\pi_i^{0|1}<1$ holds for all the cases. In contrast, the point and interval summaries of the posterior distributions under our proposed model and under Model A-2 are quite similar in most cases, although the credible intervals from the latter fail to cover the true average misreport rates in the scenario with covariate-dependent misreporting for $\pi_i^{1|0},\pi_i^{0|1}=20\%$.

In order to illustrate the differences between inferences based on the alternative estimators, Figure 4.2 plots the estimated marginal effect of x_1 and x_2 on the probability that the response takes a value of 1 under the standard probit model ignoring misreporting, our proposed method, and the approach based on Prescott and Garthwaite (2002, 2005) (Model A-2). For both simulated covariates, the true average effects estimated using y_i as the dependent variable always lie comfortably within the central 95% credible intervals from our model under both Monte Carlo designs and for all the misclassification rates considered. The maximal differences between the point estimates from our model and the true effects are at most of 2 and 8 percentage points for x_1 and x_2, respectively. In contrast, the “naïve” probit model systematically underestimates the marginal effect of x_1 in the two simulation scenarios and leads to strongly biased estimates for the effect of x_2 for $\pi_i^{1|0}, \pi_i^{0|1} = 10\%$. The differences in the performance of the two models increase with the prevalence of misreporting and are most notorious in the case in which $\pi_i^{1|0}$ and $\pi_i^{0|1}$ depend on x_2. For $\pi_i^{1|0}, \pi_i^{0|1} = 0.2$, the effect of x_1 estimated without adjusting for misclassification is less than half the true value, and the marginal effect of x_2 is overestimated by more than 30 percentage points. Also, while the marginal effects estimated using our approach and Model A-2 are quite similar under the scenario with constant misclassification rates, our model performs much better than Model A-2 when misreporting is covariate-dependent. In fact, for the binary covariate x_2, ignoring misreporting yields more accurate estimates of the marginal effects than incorporating information from the validation sample in the way suggested by Prescott and Garthwaite (2002, 2005).

Figure 4.2

Marginal covariate effects



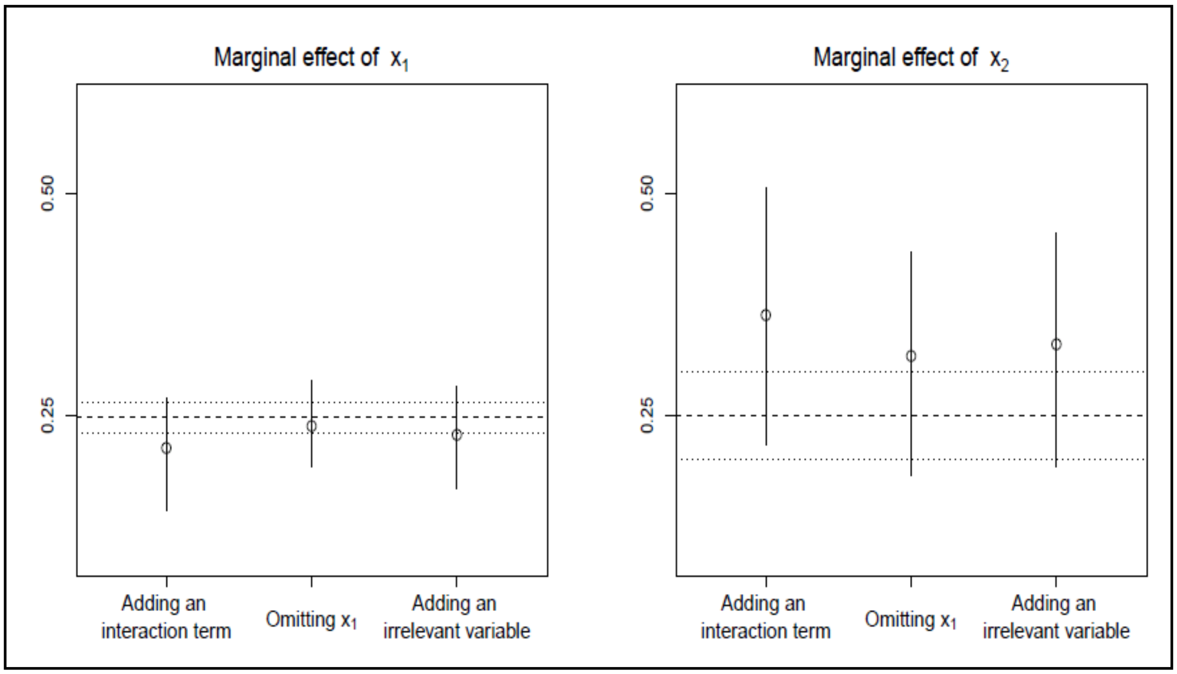
Note: The graph compares the marginal effects of the two simulated covariates x_1 and x_2 estimated under three different approaches: a standard probit model, our method correcting for misreporting, and Model A-2. The center dots correspond to the posterior means, the vertical lines to the central 95% credible intervals, and the horizontal lines represent the average effects (dashed) and 95% intervals (dotted) estimated using y_i as the response.

We also fit our model using several different specifications for the model of misreporting and found again that the point estimates of $\beta = \{\beta_0, \beta_1, \beta_2\}$ and the marginal covariate effects under our approach are closer to the true values than under any of the alternative methods considered. Figure 4.3 illustrates this, plotting the marginal effects of x_1 and x_1 for Design B and $\overline{\pi^{1|0}}, \overline{\pi^{0|1}} = 0.2$ under three alternative specifications for the linear predictor of the misreport model: adding an interaction term between x_1 and x_2, omitting x_1, and including an additional variable that is not significantly related to either the misreport probabilities or the true response. The point estimates differ slightly across specifications and are more accurate for the average effect of the normally distributed covariate. Nonetheless, a comparison with the results in Figure 4.2 shows that our method performs considerably better than the other estimators under all the specifications of the misreport model considered.

Hence, the evidence from this simulation study shows that, in the presence of misreporting, our method can considerably improve the accuracy of the parameter estimates with respect to standard binary choice models even for misclassification rates as low as 5%. When misreporting is covariate-dependent, our proposed estimation solution also performs considerably better than alternative approaches assuming constant misclassification rates, especially for non-trivial levels of misclassification. Differences in the parameter estimates obtained under alternative models may considerably affect inferences drawn from the sample under analysis, and thus ignoring misreporting or neglecting the potential correlation between the covariates of interest and the misreport probabilities could lead to quite different substantive conclusions. In addition, our results indicate that using auxiliary information may be critical to improve identifiability and convergence properties of models correcting for misclassification in relatively small samples, such as those typically used in political science.

Figure 4.3

Marginal covariate effects under different specifications of the misreport model



Note: The graph compares the marginal effects of x_1 and x_2 estimated under our proposed method, using three alternative specifications of the linear predictor in the model of misreporting. The center dots correspond to the posterior means, the vertical lines to the central 95% credible intervals, and the horizontal lines represent the average effects (dashed) and 95% intervals (dotted) estimated using y_i as the response.

4.3.2 Assessing robustness to the specification of the misreport model

The results from the simulation study reported above indicate that the model proposed in this paper can successfully adjust for misreporting under different parametric models for the misclassification mechanism. Nonetheless, the sensitivity of our method to misspecification of the model of misreporting deserves further attention, since this may lead to inconsistent estimates of β and affect inferences on the covariate of interest (Abrevaya and Hausman 1999; Hausman, Abrevaya and Scott-Morton 1998). In order to examine this issue in more detail, we draw on research analyzing a somewhat similar problem, namely, the sensitivity of the estimated treatment effects to the specification of the propensity score model (Drake 1993; Zhao 2008).

Our main goal here is to examine the influence on the estimated covariate effects of misspecifying the disturbance distribution in the model of misreporting, omitting relevant covariates from the linear predictor, including variables not related to either the true response or the misreport probabilities and adding unnecessary nonlinear terms. Specifically, using the covariates and the true response from 4.3.1, we generate a dichotomous variable d_i as:

$$d_i = \begin{cases} I\left(\gamma_{1,0} + \gamma_{1,1}x_{i,1} + \gamma_{1,2}x_{i,2} + \eta_i \geq 0\right); & \text{if } y_i = 0 \\ I\left(\gamma_{2,0} + \gamma_{2,1}x_{i,1} + \gamma_{2,2}x_{i,2} + \eta_i \geq 0\right); & \text{if } y_i = 1 \end{cases}$$

where η is an error term, and $\gamma_1 = \left\{\gamma_{1,0}, \gamma_{1,1}, \gamma_{1,2}\right\}$, $\gamma_2 = \left\{\gamma_{2,0}, \gamma_{2,1}, \gamma_{2,2}\right\}$ are chosen to obtain different levels of misclassification and different degrees of correlation between the simulated covariates and the misreport probabilities $\pi_i^{1|0}$ and $\pi_i^{0|1}$. The observed response $\tilde{y}_i$ is in then generated as:

$$\tilde{y}_i = \begin{cases} I\left(d_i = 1\right); & \text{if } y_i = 0 \\ 1 - I\left(d_i = 1\right); & \text{if } y_i = 1. \end{cases}$$

In order to analyze the sensitivity of our method to misspecification of the error disturbance in the model of misreporting, we follow Horowitz (1993); Drake (1993); Zhao (2008) and consider 4 distributions for η: a standard normal distribution, a logistic distribution, a bimodal distribution $\eta = 0.5N(3,1) + 0.5N(-3,1)$, and heteroskedastic error terms $\eta \sim N\left(1, 1 + 0.1x_1^2\right)$. We also implement 4 alternative specifications for the linear predictor of the misreport model:

Specification 1: $\alpha_{k,0} + \alpha_{k,1}x_{i,2}$;

Specification 2: $\alpha_{k,1}x_{i,1} + \alpha_{k,2}x_{i,2} + \alpha_{k,3}x_{i,1}^2$;

Specification 3: $\alpha_{k,0} + \alpha_{k,1}x_{i,1} + \alpha_{k,2}x_{i,2} + \alpha_{k,3}\left(x_{i,1} \times x_{i,2}\right)$;

Specification 4: $\alpha_{k,0} + \alpha_{k,1}x_{i,1} + \alpha_{k,2}x_{i,2} + \alpha_{k,3}x_{i,3}$;

with $k = 1, 2,$ and x_3 drawn from a log-normal distribution. We examine the effect of both forms of misspecification separately -i.e., we correctly specify the linear predictor of the misreport model when analyzing the role of misspecified error distributions and use standard normal errors when examining the influence of the functional form of the index term.[64]

Figure 4.4 reports the estimates of the marginal covariate effects when x_1 is omitted from the linear predictor of the misreport model (Specification 1) for different values of $\gamma_{1,1}, \gamma_{2,1}$ and average symmetric misreport rates of approximately 10% and 20%.[65] The estimates of the marginal effect of x_1 worsen as the average misclassification rates increase and as the correlations between the covariate and the misreport probabilities increase. However, for all values of $\gamma_{1,1}, \gamma_{2,1}$, the estimates from our model are closer to the true marginal effects obtained using the true data than the estimates from a model ignoring misreporting. The estimates for x_2, on the other hand, are not affected by the omission of x_1 from the model of misreporting and are again much more accurate than those from a standard probit model.

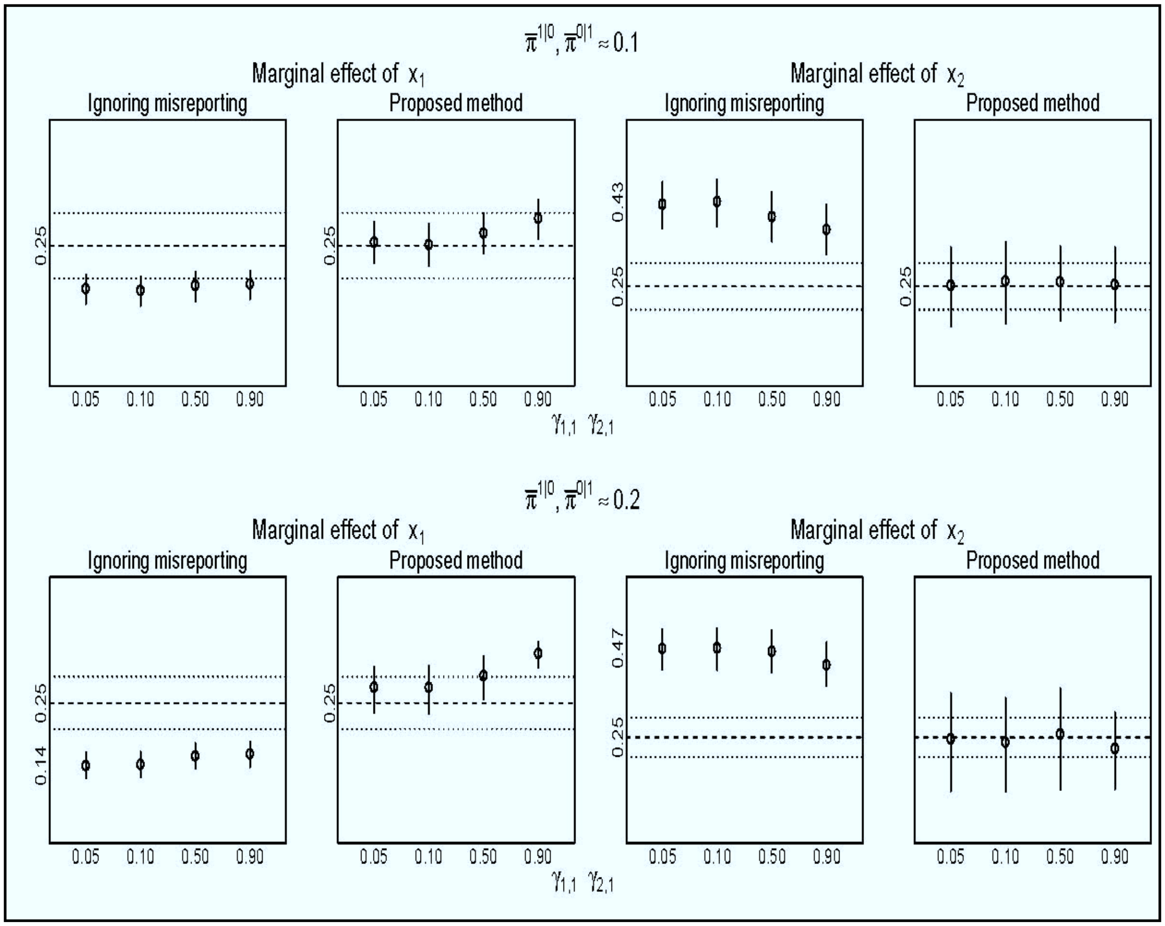
---

[64] We also let the covariates in $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ differ across specifications and consider several values of γ_1 and γ_2 with little change in the main substantive results presented in this section.

[65] In all cases, we set $\gamma_{1,2} = 1.25, \gamma_{2,2} = -1.25,$ and adjust the value of the intercept to achieve the desired average misclassification rates.

Figure 4.4

Marginal covariate effects when x_1 is omitted from the misreport model



Note: The graph plots the marginal effects of x_1 and x_2 estimated under our method when x_1 is omitted from the linear predictor of the misreport model, for different values of γ_1 and γ_2. The center dots correspond to the posterior means, the vertical lines to the central 95% credible intervals, and the horizontal lines represent the average effects (dashed) and 95% intervals (dotted) estimated using y_i as the response.

Table 4.1 complements the information from the figure, illustrating the influence of the other forms of misspecification considered for different values of $\overline{\pi^{1|0}}, \overline{\pi^{0|1}}, \gamma_1$ and γ_2. In line with the results in 4.3.1, adding irrelevant covariates and unnecessary nonlinear terms to the linear predictor of the misreport model has relatively little influence on the estimated marginal effects, and the same holds for the case of misspecified disturbance distributions. In all cases, the true average covariate effects lie within the central 95% credible intervals from our model, and the point estimates are between 4 and 18 percentage points closer to the true values than those obtained ignoring misreporting. It is worth noting that the estimates of γ_1 and γ_2 can be far away from the true coefficients when the model of misreporting is misspecified, particularly when the error terms are bimodal or heteroskedastic (Horowitz 1993; Zhao 2008). However, the estimated covariate effects seem to be quite robust to the specification of the misreport model and much more accurate than those from standard parametric models when misclassification is non-negligible. We must note, though, that these results are based on limited simulation analyses and may not be true in general.

Table 4.1

Marginal covariate effects under alternative specifications of the misreport model

	Estimator	$\partial P(y=1\mid \boldsymbol{x})/\partial x_1$	$\partial P(y=1\mid \boldsymbol{x})/\partial x_2$
	True Model	0.25 (0.24, 0.27)	0.25 (0.21, 0.30)
Linear predictor [a]			
	Specification 2	0.29 (0.24, 0.33)	0.26 (0.13, 0.38)
	Specification 3	0.29 (0.24, 0.32)	0.27 (0.15, 0.38)
	Specification 4	0.27 (0.15, 0.38)	0.26 (0.15, 0.38)
Error disturbance			
	Logistic distribution [b]	0.27 (0.21, 0.32)	0.23 (0.09, 0.38)
	Bimodal distribution [c]	0.24 (0.20, 0.28)	0.21 (0.10, 0.29)
	Heteroskedatic [c]	0.24 (0.17, 0.29)	0.28 (0.17, 0.39)
Different misreport models in both sub-samples [e]		0.24 (0.21, 0.28)	0.30 (0.21, 0.38)

[a] $\gamma_{1,0}=-1.5, \gamma_{1,1}=0.05, \gamma_{1,2}=1.25, \gamma_{2,0}=-0.2, \gamma_{2,1}=0.05, \gamma_{2,2}=-1.25, \overline{\pi^{1|0}}, \overline{\pi^{0|1}} \approx 0.2.$

[b] $\gamma_{1,0}=-1.75, \gamma_{1,1}=0.65, \gamma_{1,2}=1.3, \gamma_{2,0}=-0.75, \gamma_{2,1}=0.2, \gamma_{2,2}=-1.3, \overline{\pi^{1|0}}, \overline{\pi^{0|1}} \approx 0.2.$

[c] $\gamma_{1,0}=-1.6, \gamma_{1,1}=0.5, \gamma_{1,2}=1.3, \gamma_{2,0}=-1, \gamma_{2,1}=0.5, \gamma_{2,2}=-1.3, \overline{\pi^{1|0}}, \overline{\pi^{0|1}} \approx 0.1.$

[d] $\gamma_{1,0}=-2.05, \gamma_{1,1}=0.95, \gamma_{1,2}=0.1, \gamma_{2,0}=-1.5, \gamma_{2,1}=-2.5, \gamma_{2,2}=-0.7, \overline{\pi^{1|0}} \approx 0.1, \overline{\pi^{0|1}} \approx 0.2.$

[e] $\overline{\pi^{1|0}}, \overline{\pi^{0|1}} \approx 0.1$

Validation sample: $\gamma_{1,0}=-1.8, \gamma_{1,1}=0.52, \gamma_{1,2}=1.3, \gamma_{2,0}=-1.1, \gamma_{2,1}=0.5, \gamma_{2,2}=-1.3.$

Main sample: $\gamma_{1,0}=-2.14, \gamma_{1,1}=0.89, \gamma_{1,2}=1.74, \gamma_{2,0}=-1.22, \gamma_{2,1}=0.76, \gamma_{2,2}=-1.32.$

We also conducted additional simulations assuming a slightly different misreport processes for the validated and the main samples. Specifically, the values of γ_1 and γ_2 in the main sample were obtained by adding uniformly distributed errors to the corresponding parameters from the validation study, preserving the amount of misclassification and the direction of the relationship between the covariates and the misreport probabilities but changing the magnitude of the effect of x_1 and x_2 on $\pi_i^{1|0}$ and $\pi_i^{0|1}$. Again, as illustrated at the bottom of Table 4.1, the marginal effects estimated from our model are quite close to the true covariate effects. In contrast, the model ignoring misclassification systematically underestimates $\partial P\left(y=1|\boldsymbol{x}\right)/\partial x_1$ and overestimates $\partial P\left(y=1|\boldsymbol{x}\right)/\partial x_2$.

4.3.3 Accounting for missing response and covariate values

Finally, we compared the performance of our proposed method to other approaches in the presence of both misclassification and missing data. For this exercise, we draw x_1 from a standard normal distribution, as in 4.3.1, and simulate x_2 from a Bernoulli distribution with success probability modeled as $\Pr\left[x_{i,2}=1\right]=\Phi\left(\phi_0+\phi_1\left|x_{i,1}\right|\right)$. We assume that $x_{i,1}$ is completely observed for all subjects, and that $x_{i,2}$ and the observed response $\tilde{y}_i$ are missing at random (MAR) for some subjects. The missing mechanisms for $\tilde{y}_i$ and $x_{i,2}$ are:

$$\Pr\left[m_i^{\tilde{y}}=1\right]=\Phi\left(\alpha_{1,0}+\alpha_{1,1}x_{i,1}+\alpha_{1,2}x_{i,2}\right) \qquad \text{and}$$

$$\Pr\left[m_i^{x_2}=1\right]=\Phi\left(\alpha_{2,0}+\alpha_{2,1}x_{i,1}\right),$$

where $m_i^{\tilde{y}}=1$ or $m_i^{x_2}=1$ if $\tilde{y}_i$ or $x_{i,2}$ is observed, and 0 otherwise. Using the same

Monte Carlo designs as in 4.3.1, we generated samples of 2,000 observations with various levels of misclassification and different patterns of missing covariates and response, ignoring the true response y_i for half of the sample.

Table 4.2 illustrates the results for two combinations of misreporting and missing data patterns, contrasting the estimates of β from our method with those from a probit model ignoring misclassification and from Model A-2, based on Prescott and Garthwaite (2002, 2005).[66] For the three estimators, we use a fully Bayesian approach for inference with missing covariate and response values. In addition, we compare the estimates from our model under an all-case (AC) analysis – i.e., incorporating observations with missing values – and a complete-case (CC) analysis. (Chen et al. 2008).

Table 4.2

Posterior means and 95% credible intervals for β with missing data

Average misreport probabilities	Missing data patterns	Estimator	β_0	β_1	β_2
		True values	-1	1	1
$\overline{\pi^{1\|0}} \approx 12\%^a$ $\overline{\pi^{0\|1}} \approx 18\%$	Only $\tilde{y}_i$:14.1%[b] Only $x_{i,2}$: 25.4% $\tilde{y}_i$ and $x_{i,2}$: 8.2%	Ignoring Misreporting	-0.98 (-1.15, -0.83)	0.55 (0.44, 0.67)	1.37 (1.12, 1.63)

[66] We omit the results for Model A-1 since, as seen before, this model fails to converge for large values of $\pi_i^{1|0}$ and $\pi_i^{0|1}$.

		Proposed Method				
			AC	-0.92 (-1.43, -0.55)	0.82 (0.55, 1.24)	0.92 (0.34, 1.59)
			CC	-0.72 (-1.31, -0.05)	0.69 (0.20, 1.21)	0.67 (0.34, 1.45)
		Model A-2		-2.33 (-3.60, -1.41)	1.61 (0.94, 2.46)	3.40 (2.14, 5.13)
$\overline{\pi^{1\|0}} \approx 8\%$[c] $\overline{\pi^{0\|1}} \approx 7\%$	Only $\tilde{y}_i : 35.5\%$[d] Only $x_{i,2} : 11.8\%$ $\tilde{y}_i$ and $x_{i,2} : 13.9\%$	Ignoring Misreporting		-0.85 (-1.06, -0.62)	0.72 (0.56, 0.89)	1.05 (1.72, 1.36)
		Proposed Method				
			AC	-0.94 (-1.43, -0.52)	0.95 (0.64, 1.38)	0.97 (0.39, 1.63)
			CC	-1.29 (-2.58, -0.61)	1.57 (0.87, 2.82)	1.30 (0.54, 2.51)
		Model A-2		-1.28 (-1.79, -1.89)	1.11 (0.80, 1.56)	-1.64 (1.09, 2.35)

[a] $\pi_i^{1|0} \mid x_{i,2} = 0 : 0.13, \pi_i^{1|0} \mid x_{i,2} = 0 : 0.5, \pi_i^{0|1} \mid x_{i,2} = 0 : 0.47, \pi_i^{0|1} \mid x_{i,2} = 1 : 0.07.$

[b] $\phi_{1,0} = 1.2, \phi_{1,1} = 0.5; \phi_{1,2} = 0.9, \phi_{2,0} = 0.5, \phi_{2,1} = 0.5.$

[c] $\pi_i^{1|0} \mid x_{i,2} = 0 : 0.03, \pi_i^{1|0} \mid x_{i,2} = 0 : 0.18, \pi_i^{0|1} \mid x_{i,2} = 0 : 0.11, \pi_i^{0|1} \mid x_{i,2} = 1 : 0.02.$

[d] $\phi_{1,0} = 0.7, \phi_{1,1} = 0.45; \phi_{1,2} = 1.1, \phi_{2,0} = 0.7, \phi_{2,1} = 0.4.$

4.4 An empirical application: correcting for misreporting in the analysis of voter turnout

Next, we illustrate the potential consequences of misreporting in the context of estimating the determinants of voter turnout and provide three different applications of our methodology using data from all the validated ANES surveys between the 1978 and 1990.[67] This dataset comprises three Midterm (1978, 1986, 1990) and three Presidential elections (1980, 1984, 1988), and has the obvious advantage of allowing us to directly compare the estimates from our model to a known benchmark, i.e., the same model estimated directly on the validated vote. We assume the validated vote to be the "gold-standard" measure of turnout, although there is considerable disagreement on this point (Burden 2000; Mcdonald 2007). The concern is that the validation studies are far from perfect. As stated at the outset, vote validation is expensive and difficult. The ANES is conducted in two parts, a pre-and post-election survey. In the studies from 1978, 1980, 1984, 1986, 1988 and 1990 there were in total 11,632 completed post election surveys. Unfortunately of these completed surveys, the ANES was unable to validate 2,189

[67] We use data from the 1978–1990 validated studies in order to preserve the comparability of the survey questions regarding the conditions of the interview; we will use this information to model the conditional probability of misreporting. While we illustrate the application of our method analyzing ANES data in view of the fact that it is the most widely used survey for studying U.S. turnout (Burden 2000), the main substantive results reported in this Section hold for the Current Population Survey as well, and are available from the authors upon request.

respondents, about 19.8 percent of the usable sample.[68] The majority of these failures were caused either because no registration records were found or because the local election office refused to cooperate with the ANES. If we are willing to maintain the assumption that these errors are essentially random (in the sense of being independent of the characteristics of interest), then there is no real harm done. The measurement error will merely result in less efficient estimates of the misreporting model and a corresponding reduction in efficiency of the corrected turnout model. However, if there is systematic error, then we are just substituting one form of measurement error for another.

In Section 4.4.1, we estimate a simple model of the determinants of the turnout decision using both self-reported and validated turnout as the dependent variable in order to assess the consequences of ignoring misreporting. In 4.4.2, we re-estimate the turnout model with self-reported vote but applying our proposed solution to correct for misreporting, using a random sample of each survey as a validation sub-study. In 4.4.3, we apply our correction for misreporting under an external validation design, using information from previous ANES studies to correct for misreporting in the main sample under analysis. Both applications are based on a complete-case analysis. We deal with the problem of incomplete data in 4.4.4, where we account for item and unit non-response using the model-based, fully Bayesian imputation approach described in Section 4.2.3.

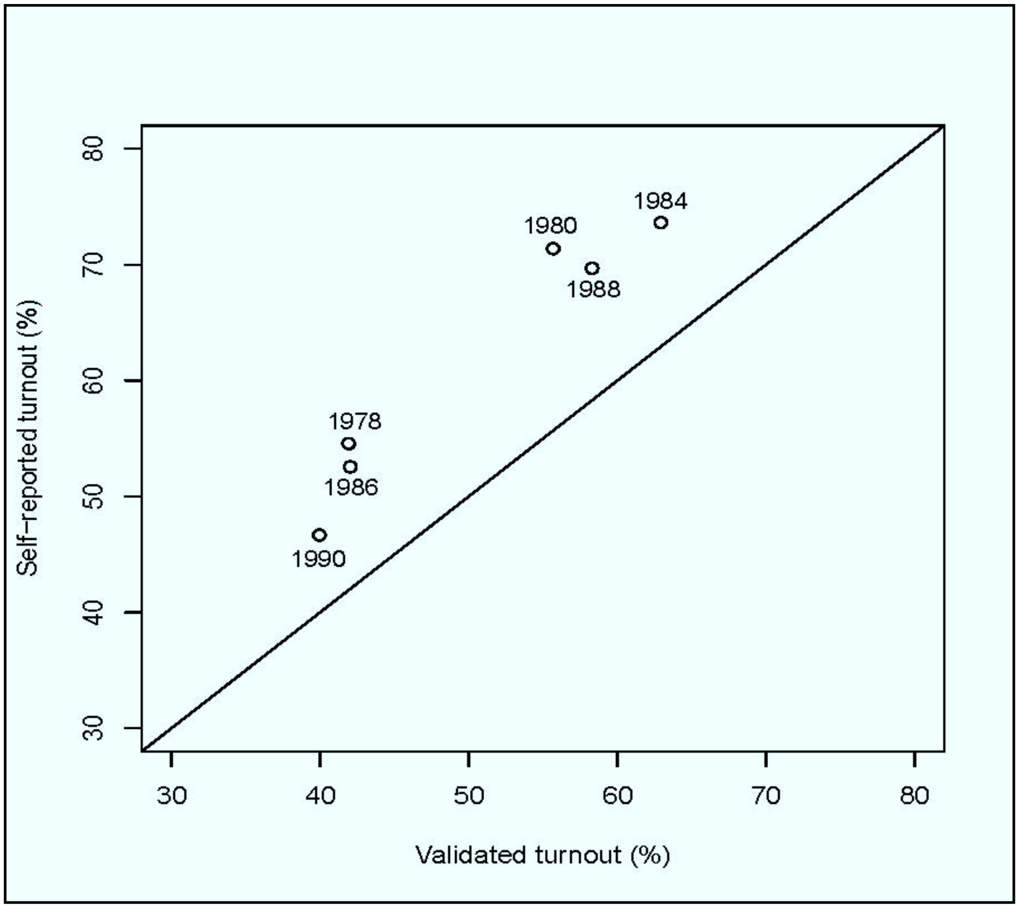
[68] The rate of non-validation varies considerably across Election Studies, from around 2% of sample in 1978 to more than 31% in 1990.

4.4.1 Turnout misreporting in the 1978 – 1990 ANES

As mentioned in the Introduction, it has long been established in the political science literature that survey respondents often report to have voted when they did not actually do so (Ansolabehere and Hersh 2008; Bernstein, Chadha and Montjoy 2001; Clausen 1968; Katosh and Traugott 1981; Miller 1952; Parry and Crossley 1950; Sigelman 1982; Silver, Anderson and Abramson 1986). Figure 4.5 illustrates the differences between turnout rates computed from self-reported and validated vote in the six ANES studies under analysis. Validated turnout is systematically lower than reported turnout, and while both rates tend to follow similar trends, differences vary considerably across years, ranging from 7 percentage points in 1990 to more than 15 percentage points in 1980. The percentage of survey respondents who claimed to have voted but did not do so according to the validated data was 17.3 percent, and more than 28% of those who did not vote according to the official records responded affirmatively to the turnout question. In contrast, only 84 respondents in the 1978-1990 ANES studies reported not voting when the official record suggested they did, representing 0.7% of the sample respondents. Additional descriptive statistics on vote misreporting in the 1978–1990 validated ANES can be found in Table 4.B.1 in Appendix 4.B.

Figure 4.5

Estimated turnout from self-reported vs. validated responses



Note: The graph shows the self-reported and validated turnout from the 1978 – 1990 ANES only in years for which there were vote validation studies. Reported turnout rates are systematically larger than the validated ones.

In order to examine whether such high rates of overreporting affect inferences on the determinants of the turnout decision, we fit two hierarchical probit models allowing for election year and regional effects with both self-reported and validated turnout as the response variable:

$$\Pr\left[\tilde{y}_i = y_i^{\text{Reported}}\right] \sim Bernoulli\left(\tilde{p}_i\right)$$
$$\tilde{p}_i = \Phi\left(\tilde{\lambda}_t + \tilde{\eta}_r + \tilde{\beta}' x_i\right)$$

and

$$\Pr\left[y_i = y_i^{\text{Validated}}\right] \sim Bernoulli\left(p_i\right)$$
$$p_i = \Phi\left(\lambda_t + \eta_r + \beta' x_i\right)$$

where the $k = 1, ..., K$ elements of β are assigned diffuse prior distributions:

$$\beta_k \sim N\left(\mu_{\beta_k}, \sigma^2_{\beta_k}\right)$$

and $\lambda_t\left(\tilde{\lambda}_t\right)$ and $\eta_r\left(\tilde{\eta}_r\right)$ are election-and region-random effects distributed:

$$\lambda_t \sim N\left(\mu_\lambda, \sigma^2_\lambda\right), \quad t = 1978, 1980, 1984, 1986, 1988, 1990;$$

$$\eta_r \sim N\left(\mu_\eta, \sigma^2_\eta\right), \quad r = \text{Northeast, North Central, South, West.}$$

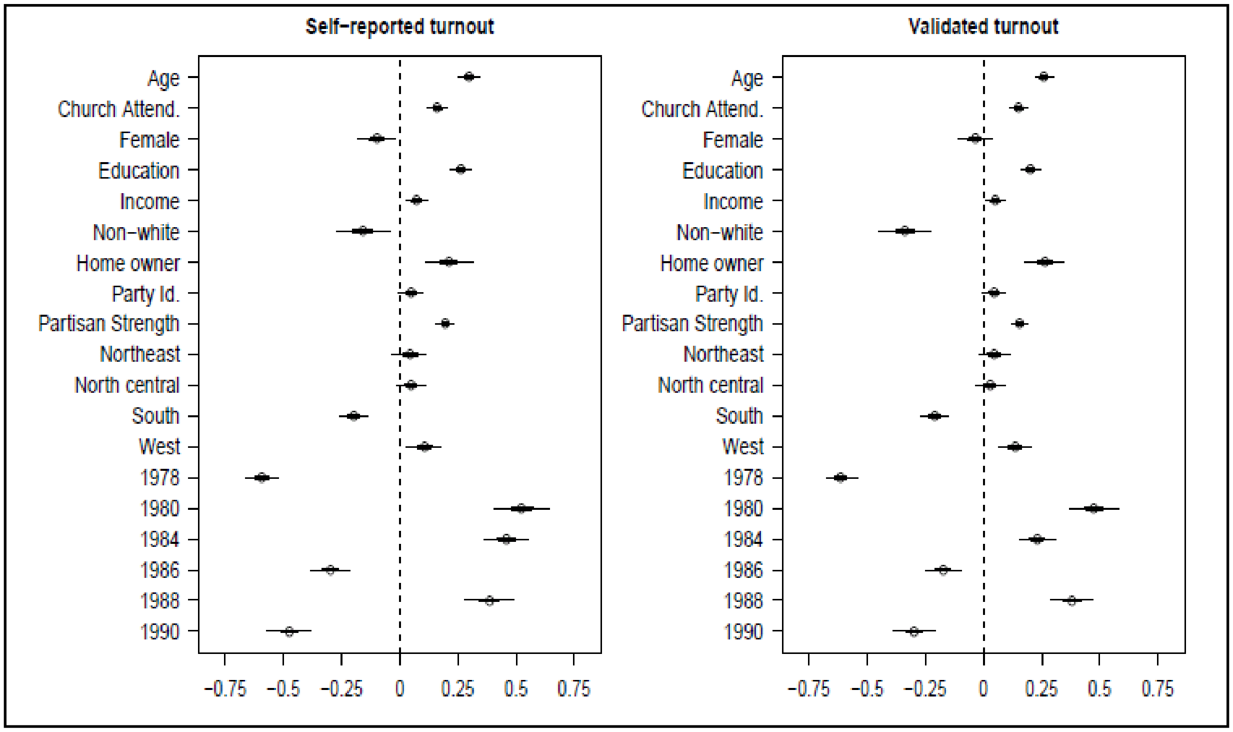
The regressors included in $\boldsymbol{x}_i$ are indicators for demographic and socio-economic conditions and political attitudes: *Age*, *Church Attendance*, *Education*, *Female*, *Home owner*, *Income*, *Non-white*, *Party Identification* and *Partisan Strength*. A description of the coding used for each of the variables may be found in Appendix 4.B. We should note that, while this specification includes some of the variables most commonly used in models of voter turnout found in the literature (Ansolabehere and Hersh 2008; Bernstein, Chadha and Montjoy 2001; Highton 2004; Leighley and Nagler 1984; Wolfinger and Rosenstone 1980), it does not examine the effect of other factors we might plausibly believe could alter turnout, such as political information (Alvarez 1998) or differences in state-level ballot laws (Wolfinger and Rosenstone 1980). The sample used in the analysis consists of 6,411 observations for the 6 elections under study and were constructed so that they are identical for both models. Only the respondents with no missing response or covariate values are included in the analysis; the remaining observations were dropped using list-wise deletion.

Figure 4.6 presents the main results from both models.[69] The left panel summarizes the posterior distribution of the model's coefficients using self-reported vote as the dependent variable, and the right panel re-does the analysis with the ANES validated vote.

[69] Three parallel chains with dispersed initial values reached approximate convergence after 50,000 iterations, with a burn-in period of 5,000 iterations. In order to ensure that inferences are data dependent, several alternative values for the hyperparameters were tried, yielding essentially similar results.

Figure 4.6

Coefficients of the probit models for self-reported vs. validated turnout



Note: The graph summaries the posterior distribution of the coefficients of the turnout model, using self-reported and validated vote as the response variable. The center dots correspond to the posterior means, the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.

Most of the parameter estimates are quite similar in both models, and inferences on the role of these predictors on the probability of voting agree with common expectations. For example, for both sets of estimates, older, wealthier and more educated respondents are more likely to turn out to vote. Also, strong partisans are on average 15 percentage points more likely to vote than independents, while respondents who attend church every week are 0.12 more likely to turn out to vote than those who never attend. Likewise, respondents are much more likely to turn out to vote in Presidential than in Midterm elections, and are less likely to vote if they live in the South. These results are essentially similar using either reported or validated vote as the dependent variable. However, there are some interesting differences between the two sets of results regarding the role of some socio-demographic variables such as gender and race. In particular, the mean posterior of the coefficient for the race indicator is more than twice as large (in absolute value) using validated vote than using self-reported vote as the dependent variable.

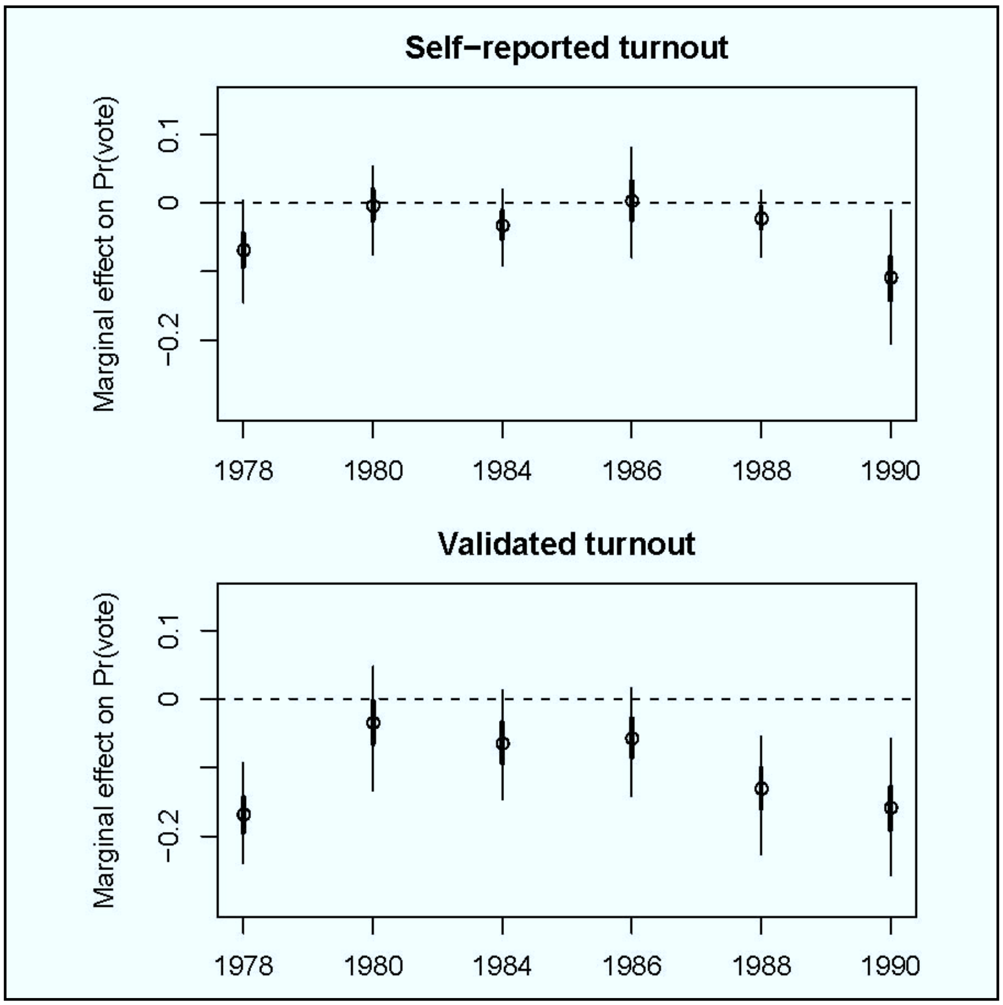
These differences in the parameter estimates can affect inferences drawn from both models regarding the impact of the covariates on the turnout decision. In order to illustrate this fact, Figure 4.7 plots the marginal effect of race on the probability of voting using reported and validated vote for each election under analysis. As seen in the figure, the negative effect of being Non-white on turnout is higher when validated vote is used as the response variable for each of the surveys considered: the average marginal effects (posterior means) are more than 6 percentage points higher than if we look only at the reported vote, with differences ranging from about 3 percentage points in the 1984 and 1986 elections to almost 11 points in the 1978 and 1988 elections. While a researcher using reported turnout would conclude that race had no significant effect on the probability of voting in the 1978 and 1988 elections at the usual confidence levels, the results obtained using validated data indicate otherwise.[70] Fitting a model of turnout

[70] In the case of the 1988 election, the marginal effect of Non-white estimated from the self-reported vote is not significant even at the 0.1 level.

using reported vote as the dependent variable will therefore tend to overpredict the probability of voting among non-white respondents and might in some cases affect substantive conclusions about the effect of race on turnout.

Figure 4.7

Marginal effects of race on turnout



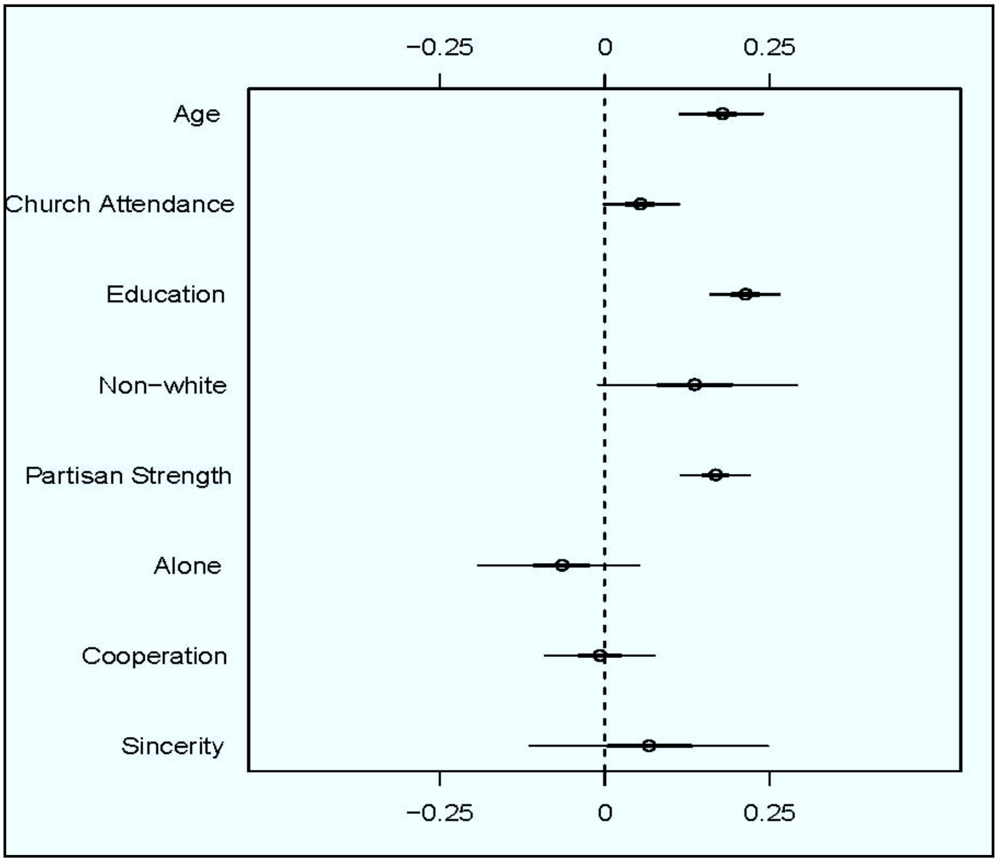
Note: The graph shows the marginal effect of the race indicator on the likelihood of voting for each election year under study, using both reported and validated vote. The center dots correspond to the point estimates (posterior means), the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.

Finally, we examine whether over-reporting varies systematically with respondents' characteristics, fitting a probit model for $\Pr\left(\tilde{y}_i = 1 \mid y_i = 0\right)$. As with the turnout model, the misreport model is fairly simple. The predictors include four variables that have been shown to be strongly correlated with overreporting in previous studies: *Age*, *Church Attendance*, *Education*, *Non-white*, and *Partisan Strength* (Ansolabehere and Hersh 2008; Belli, Traugott and Beckman 2001; Bernstein, Chadha and Montjoy 2001; Cassel 2003). In addition, we also include three additional covariates aimed at capturing some of the conditions of the interview.[71] The first is an indicator of whether the interview was conducted while the respondent was alone. According to the "social pressures" argument (Cahalan 1968; Loftus 1975), a respondent should be more likely to lie about voting if others will learn of the statement. The other two variables are the interviewers' assessments of the respondents' cooperation and sincerity during the interview. Point and interval summaries of the posterior distribution of the model's parameters are presented in Figure 4.8.

[71] All interviewers in the 1978 – 1990 ANES were asked to rate the level of cooperation and sincerity of the respondent after the completion of the survey.

Figure 4.8

Determinants of misreporting



Note: The graph shows the parameter estimates for the model of over-reporting. The center dots correspond to the point estimates (posterior means), the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.

In line with previous analyses, we find that overreporters tend to be more educated, older, more partisan, and are more likely to be regular church attendees. Also, consistent with the results reported in Figures 4.6 and 4.7, being nonwhite has a positive effect on the probability of misreporting vote status: non-whites are on average 0.05 more likely to overreport than their white counterparts, and this effect is significant at the 0.1 level. Several scholars have argued that African Americans and Latinos feel pressured to appear to have voted due to the struggles and sacrifices needed to gain voting rights for their racial or ethnic group (Abramson and Claggett 1984; Belli, Traugott and Beckman 2001; Hill and Hurley 1984), although recent research has suggested that the relationship between race and overreporting is much more complex than previously thought and depends on the demographic and geographical context (Ansolabehere and Hersh 2008; Bernstein, Chadha and Montjoy 2001; Fullerton, Dixon and Borch 2007).[72] None of the other variables has a statistically significant effect on misreporting at the usual confidence levels. In particular, the interviewers seem unable to pick up a "feeling" that

[72] It is worth mentioning that this relationship between race and vote over-reporting could also be associated to the socio-economic status of the non-white population. If it is the case that nonwhites, who are more concentrated in poorer areas, are more likely to be incorrectly validated or excluded from the validation studies because no records can be found (e.g., due to poorly staffed and maintained election offices), then this result -as well as those reported in Figures 3.6 and 3.7 -could very well be an artifact. While it is difficult to rule this claim out, addressing this concern is beyond the focus of this paper. Hence, as noted above, we proceed as if the validated data provides "gold-standard" information on turnout, or is at least not subject to systematic bias.

is not otherwise captured by the characteristics observable from the survey. This is probably caused by the fact that very few of the interviewers were willing to rank a respondent as uncooperative and/or insincere.[73]

Hence, the results from these simple models indicate that the probability of misreporting varies systematically with characteristics we might be interested in, and that failing to account for misreporting may affect parameter estimates and inferences about the determinants of voter turnout drawn from non-validated survey data. Unfortunately, as mentioned in the Introduction, the ANES has stopped conducting validation studies due to the cost and difficulty in collecting the data as well as to the fact that few researchers used the validated data. The next three sections allow us to evaluate the performance of our proposed method to correct for misreporting and improve estimates and inference obtained from self-reported turnout. Although our model accounts for the possibility of two types of misreporting, we saw before that virtually no one reports not voting when they did, and thus $\pi_i^{0|1}$ would be poorly estimated (Prescott and Garthwaite 2005). Therefore, in the applications below we will assume that $\pi_i^{0|1} = 0$, and we therefore only need to account for $\pi_i^{1|0}$.

[73] Only 1.3% of all the respondents in the sample were ranked as uncooperative by the ANES interwievers, and only 0.7% were deemed to be "often insincere".

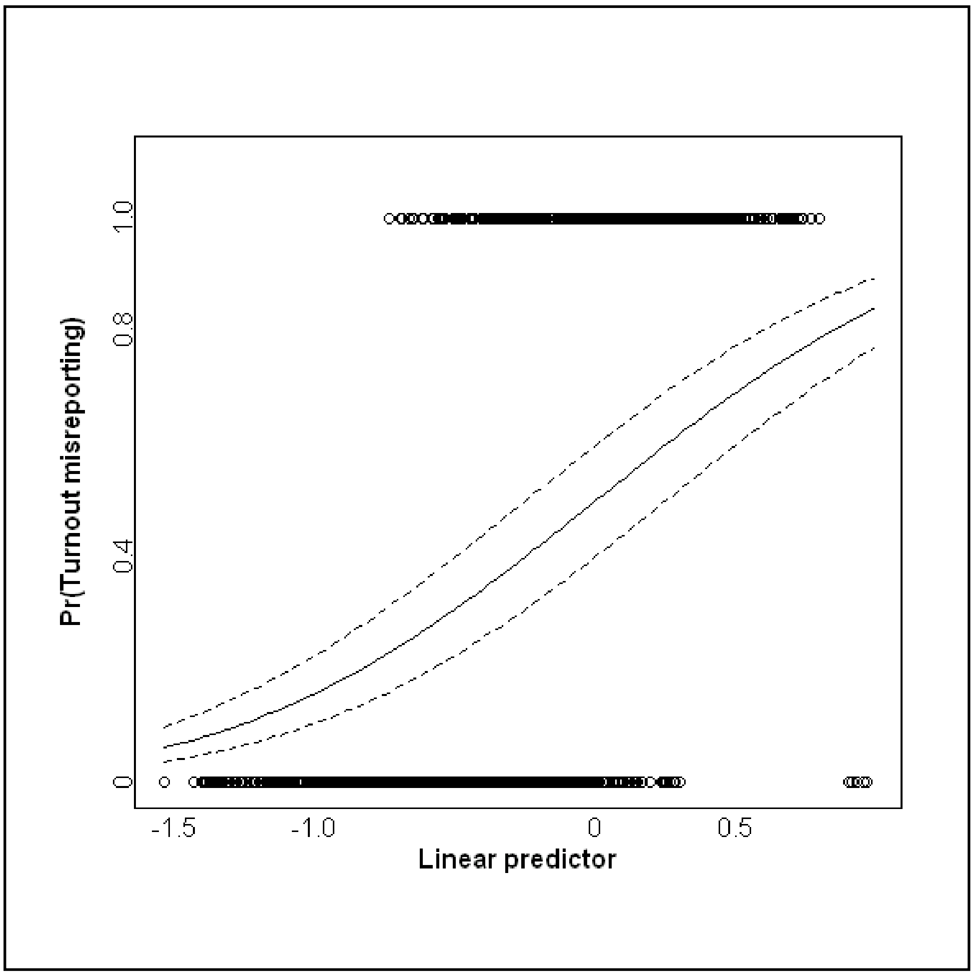
4.4.2 Correcting for misreporting using a validation sub-sample

We first apply our method assuming an internal validation design. As in the simulation exercise in Section 4.3.1, we randomly assign half of the respondents in each of the 1978–1990 surveys to be the validation sub-study and ignore the validated data for the remaining respondents. We then used the information from the validated sub-sample to correct for over-reporting in the main sample, equally weighting both datasets.

For illustrative purposes, we fit the same turnout and misreport models described in 4.4.1 for all the ANES studies considered. Nonetheless, as indicated above, the probability of voting is considerably higher in Presidential than in Midterm elections, and it is likely that different factors affect turnout in different election years. More importantly, the patterns of overreporting have also been shown to differ substantially across types of races and election years (Cassel 2003). As a result, the misreport model does not predict over-reporting very well: as seen in Figure 4.9, which shows the predicted probability of misreporting as a function of the linear predictor of the misclassification model, the covariates included in the specification do not allow clearly distinguishing overreporters from "truthful" voters. The mean error rate of the misreport model across election studies is 36%, while a null model that simply predicts that no respondent overreports has an error rate of 31%. The model correctly classifies 64% of the survey respondents in cases, and the mean predicted probability of misreporting averaged across simulations is 0.45; ideally this would be near zero or one for the entire sample (Gelman and Hill 2007). Hence, as illustrated in Figure 4.10, the predicted average misreport rates are systematically underestimated for some election-years. Therefore, while the simulation results from Section 2 suggest that our approach is quite robust to misspecification of the misreport model, we note that the performance of our proposed method would benefit from better modeling of the misreport process.

Figure 4.9

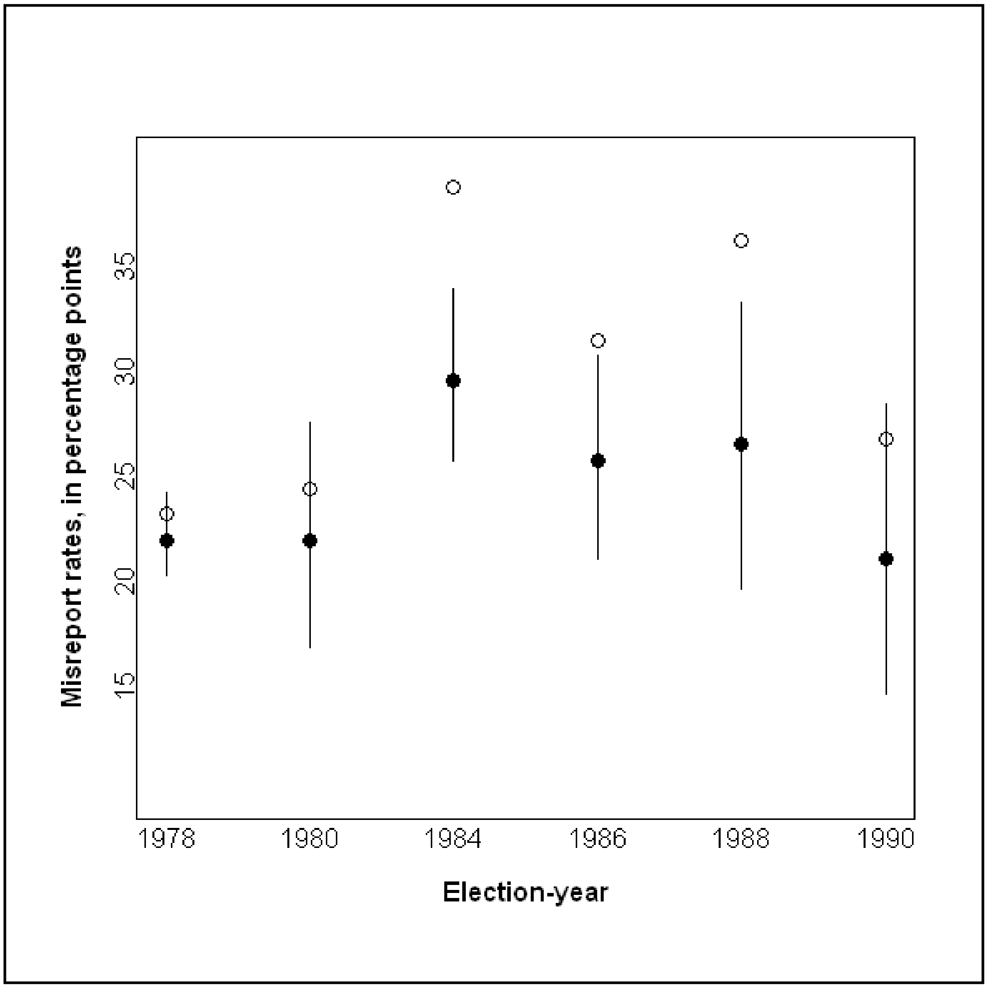
Estimated probability of misreporting for the respondents in the 1978 – 1990 ANES validated studies



Note: The graph shows the estimated probability of turnout overreporting for the respondents in the (pooled) 1978 – 1990 ANES validated studies, as a function of the linear predictor of the misreport model. The solid line represents the mean posterior probabilities of misreporting, while the dashed lines correspond to the 95% credible intervals. Dots represent the actual value of the misreport indicator.

Figure 4.10

Predicted and actual misreport rates, by election-year

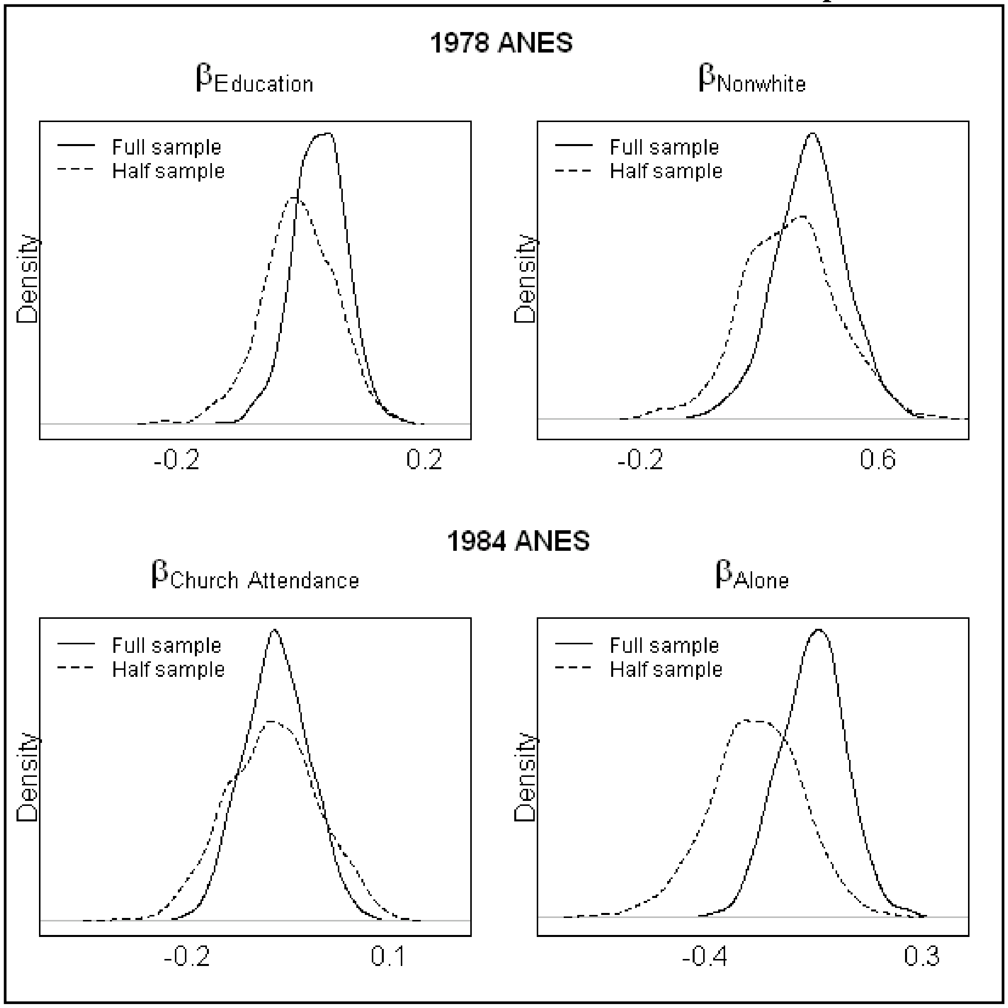


Note: The graph plots the actual and predicted rates of misreporting for each of the ANES election validated studies between 1978 and 1990. The white circles represent the actual proportion of misreporters among the respondents in each study, while the black dots correspond to the estimated proportions. Vertical lines represent the 95% credible intervals for the predicted misreport rates.

That said, the coefficients of the misreport model estimated using only half the sample do not generally differ substantially from those estimated using the whole sample, as illustrated in Figure 4.11, which plots the posterior distribution of selected parameters of the misreport model estimated for the two ANES studies with lowest (1978) and largest (1984) percentage of overreporters (See Table 4.B.1 in Appendix 4.B). More importantly, Figure 4.12 summarizes the posterior distribution of the coefficients of selected regressors estimated using validated, self-reported vote, and corrected self-reports for these two election years. Assuming that the parameters estimated using validated vote are the "correct" estimates, the point estimates (posterior means) from our model for the two elections are between 32% and 92% closer to the "true" values of each of the parameters than the estimates ignoring overreporting. In addition, like the "true" estimate, the estimate of $\beta_{Non-white}$ under our approach is significantly negative at the 0.05 level for the 1978 ANES.

Figure 4.11

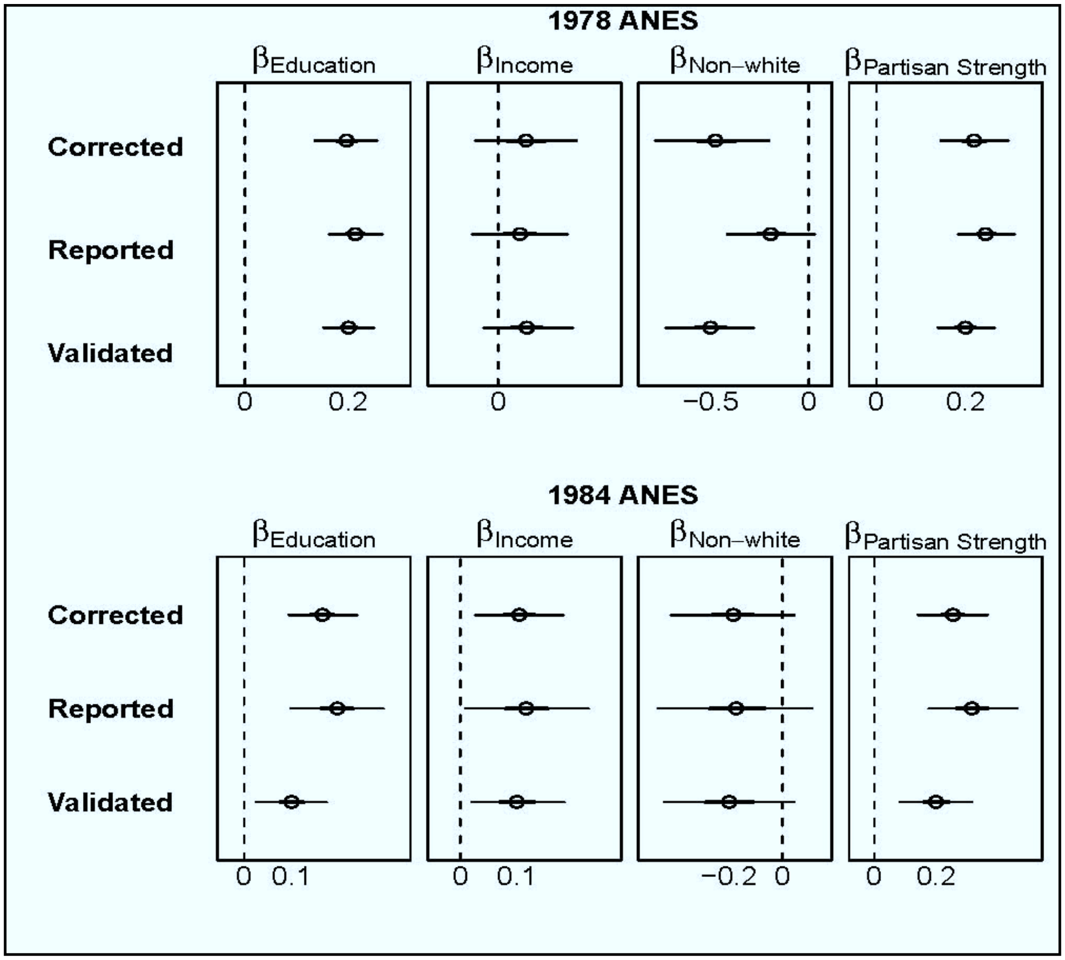
Posterior distributions of selected coefficients of the misreport model



Note: The figure compares the posterior densities of selected coefficients of the misreport model for the 1978 and 1984 ANES studies. The solid lines plot the posterior distributions of the parameters estimated using the whole sample for each survey, while the dashed lines represent the estimates obtained using only half of the sample.

Figure 4.12

Posterior summaries for selected parameters under an internal validation design

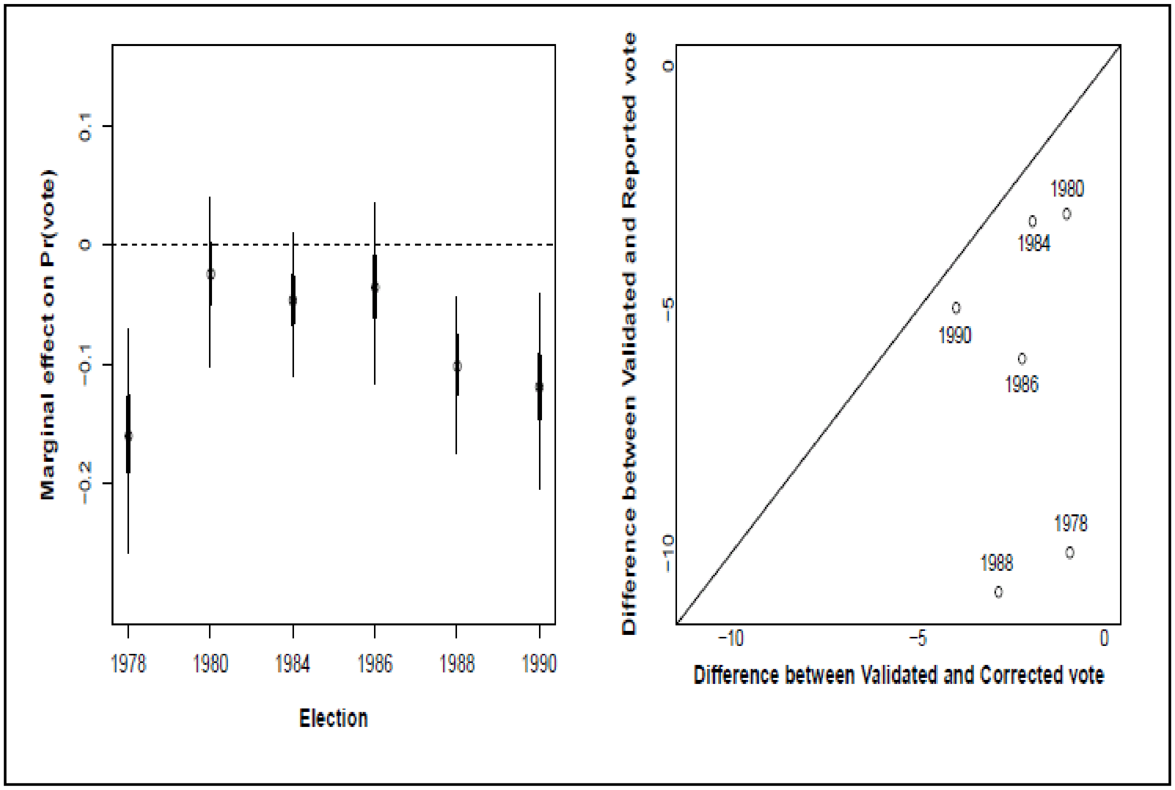


Note: The figure plots point and interval summaries of the posterior distributions of selected coefficients for the 1978 and 1984 ANES Presidential elections, using corrected, self-reported, and validated vote. The center dots correspond to the posterior means, the thick horizontal lines to the central 50% credible intervals, and the thin lines to the central 95% credible intervals from the three different models.

Figure 4.13, in turn, plots the marginal effect of race on the probability of voting estimated using our approach to correct for misreporting. A comparison of the results in the left panel of the figure with those presented in Figure 4.7 above shows that, after correcting for misreporting, the impact of race in the 1978 and 1988 elections is now statistically significant at the usual confidence levels. Moreover, as seen in the right panel, the point estimates from our model are closer to the "true" effects than those estimated from the model using self-reported vote for all the ANES studies, with differences ranging between 1 and 9 percentage points. Therefore, the evidence presented in this Section indicates that, even with the very simple model of misreporting estimated here, the improvements in the accuracy of the parameter estimates obtained using our method are important, and can eventually change the substantive conclusions drawn regarding the effect of relevant covariates on the turnout decision.

Figure 4.13

Marginal effect of race on turnout estimated under our proposed method



Note: The left panel of the graph plots the point and interval (50% and 95%) estimates of the marginal effect of race on the probability of voting estimated from our model to correct for misreporting. The right panel compares the point estimates from our model and the model ignoring misreporting with the estimates obtained using the validated data.

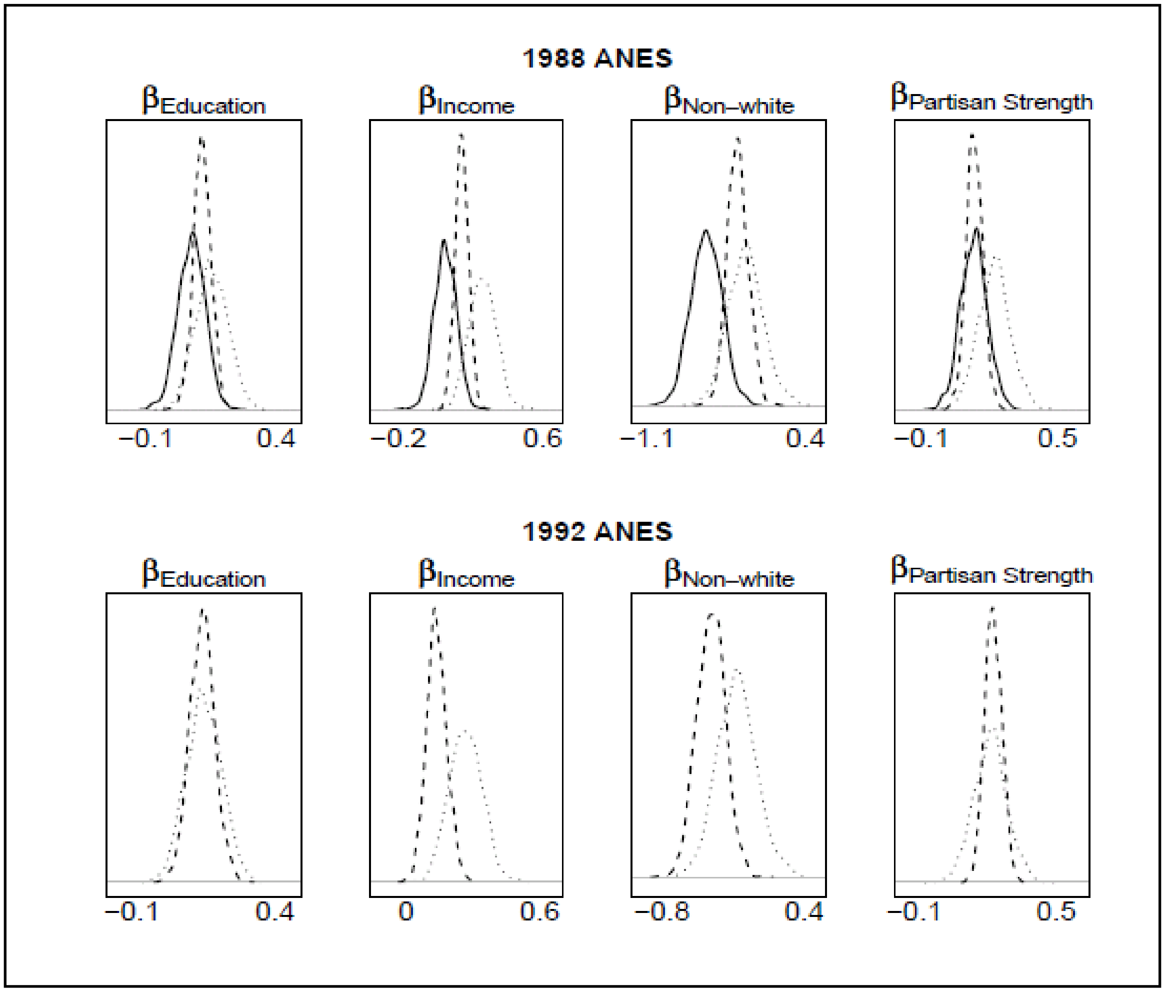
4.4.3 Correcting for misreporting using an external validation design

We also apply our correction for misreporting assuming an external validation design, ignoring the validated vote for the sample under analysis and incorporating information on the misreport probabilities and regression parameters from other ANES studies. Figure 4.14 illustrates the results of this exercise, plotting the marginal posterior distribution of selected coefficients for the 1988 and 1992 Presidential elections obtained by updating the corresponding posteriors from previous validated ANES surveys.

The upper panel the compares the posterior distributions of $\beta_{Education}$, β_{Income}, $\beta_{Non-white}$ and $\beta_{Partisan\ Strength}$ for the 1988 ANES, the last Presidential election for which vote validation is available, using validated, self-reported and corrected vote. In order to implement our correction for misreporting, we used auxiliary data from the two previous Presidential elections for which validated turnout data was collected (1980 and 1984). As seen in the figure, the marginal posterior means and modes from the model accounting for overreporting are in all cases closer to "true" values than those obtained from the unadjusted self-reports. Again, as the "correct" estimate, the estimate of $\beta_{Non-white}$ under our model is significantly negative at the 0.05 level. In the case of the 1992 ANES, for which there is no validated data, we implemented our correction for misreporting using information from the previous presidential elections for which vote validation was conducted (1980, 1984 and 1988) and compared the estimates from our model with those from a model using self-reported vote. As seen in the lower panel of Figure 4.14, the posterior distribution of some of the parameters - $\beta_{Education}$, $\beta_{Partisan\ Strength}$ - remain essentially unchanged when applying the correction for misreporting.

Figure 4.14

Posterior densities of β under an external validation design



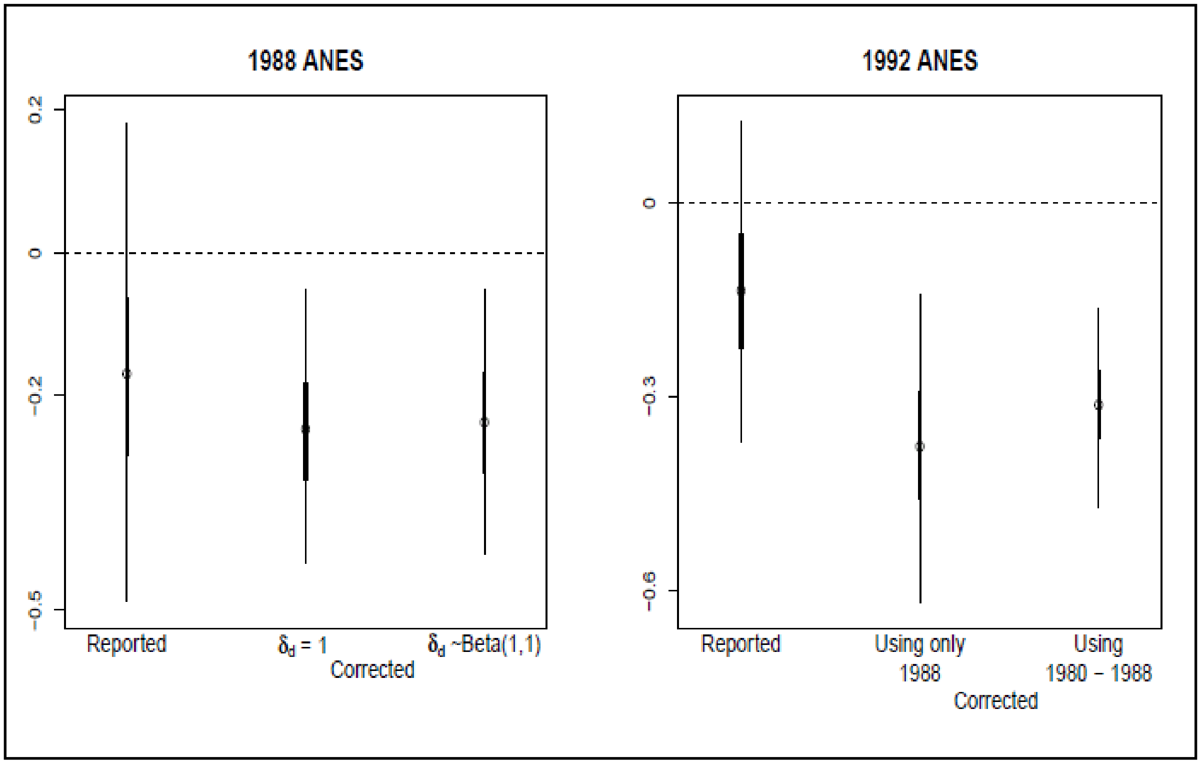
Note: The figure compares the posterior densities of selected coefficients for the 1988 and 1992 Presidential elections. The solid lines plot the posterior distributions of the parameters estimated from the validated vote, the dotted lines represent the estimates obtained using self-reported vote, and the dashed lines the ones obtained adjusting for misreporting.

However, using auxiliary information does affect the posterior distribution of the coefficients of *Income* and *Non-white*. In particular, accounting for misreporting substantially affects the marginal posterior distribution of $\beta_{Non-white}$: the mean posterior is more than twice as large (in absolute value) when using the corrected self-reports, and the effect of *Non-white* on the probability of turning out to vote is significantly negative at the 0.05 level, while it is not significant even at the 0.2 level when estimated using self-reported vote. Similar results hold when applying our model to correct for misreporting in the 1994 ANES -for which, again, vote validation was not conducted -using validated turnout data from previous Midterm elections.

We also conducted a series of sensitivity analyses aimed at assessing the robustness of the parameter estimates to changes in the composition of the auxiliary data used to correct for misreporting and in the weight assigned to the validated *vis-à-vis* the main sample. Figure 4.15 summarizes some of the results for the 1988 and 1992 ANES. The left panel plots point and interval summaries for $\beta_{Non-white}$ from our model for the 1988 ANES using two different sets of values for the weighting parameters δ_d in Equation 4.9: a point mass prior $\delta_d = 1$ with probability 1 $\forall d$, and uniform *Beta(1, 1)* priors $\forall d$, where $d = 1980,1984$. In the first case, the validated and main samples are pooled together and the estimates of β for the main sample are obtained by updating the posteriors from the previous ANES surveys via Bayes' theorem. In the second case, we allow for different a posteriori weights for each of the validated samples, thus accommodating heterogeneity between the previous ANES studies. The right panel, in turn, compares the estimates from our model for the 1992 for the cases in which only validated data from the immediate previous (1988) or from all the previous (1980, 1984, 1988) Presidential elections is used to adjust for misreporting.[74] For both election years, the estimates from our model are compared to those from the unadjusted self-reports.

Figure 4.15

Sensitivity analysis for the external validation design



Note: The graph summarizes the posterior distribution of $\beta_{Non-white}$ from our model for the 1988 and 1992 elections, using alternative strategies to incorporate information from previous validated ANES studies. The estimates are compared to those obtained using self-reported vote. The center dots correspond to the posterior means, the thicker lines to the 50% credible intervals, and the thinner lines to the 95% credible intervals.

---

[74] For the 1992 ANES, we fixed the value of δ at 1 for this sensitivity analysis.

As illustrated in the figure, the posterior standard deviations of β tend to decrease with the amount of auxiliary data used to correct for misreporting in the main sample, but the point estimates (posterior means) and the main substantive conclusions about β seem to be quite robust to changes in the values of δ and in the size and heterogeneity of the auxiliary data. In particular, correcting for overreporting using information from previous validated studies leads to stronger negative effects of being Non-white on the probability of voting than using self-reported vote, with differences of approximately 4 and 9 and percentage points for the 1988 and 1992 ANES, respectively.

4.4.4. Accounting for item and unit non-response

Both applications of our methodology in Sections 4.4.2 and 4.4.3 have been based on a complete-case analysis, including in the sample only those respondents for whom both the response to the turnout question and all the relevant covariates are completely observed. When respondents with missing covariates differ systematically from those with complete data with respect to the outcome of interest, this approach may lead to significantly biased parameters and inference (Little and Rubin 2002). In our sample from the 1978–1990 ANES studies, 14.5% of whites and 20.9% of non-whites have missing covariate values (other than race), and the percentage of missingness for the self-reported vote is almost 1.8 times larger for the latter. Since the evidence above indicates that voting patterns vary systematically with race, inferences from a complete-case analysis may be quite misleading in this setting (Ibrahim et al. 2005). In addition, list-wise deletion due to missing values in the response variable and/or the predictors leads to discard almost 45% of the respondents in the 1978–1990 ANES and more than two-thirds of the respondents in the 1994 ANES, so that complete-case analyses are extremely wasteful and potentially inefficient. Table 4.B.2 in Appendix 4.B reports the rates of item nonresponse for all the variables included in the turnout models from Sections 4.4.2 and 4.4.3.

In order to accommodate item and unit non-response, we implement the approach described in Section 4.2.3, fitting a separate model for each of the ANES studies.[75] Based on Equation 4.12, we specified probit regression models for all the dichotomous covariates in the model – *Female*, *Non-white*, *Own Home*, and *Alone* – while the remaining categorical covariates were assigned conditional normal distributions and discrete values were afterwards imputed for the missing responses (Lipsitz and Ibrahim 1996; Gelman, King and Liu 1998).[76] In all cases, we assigned vague independent normal priors for the components of α.

Figure 4.16 illustrates the results for the 1978 and 1992 ANES. For the former, 31% of the survey respondents have at least 1 missing covariate value, and 0.5% of the respondents failed to answer the turnout question, while the corresponding rates for the latter are 47% and 9%, respectively. A complete-case analysis would keep 77% of our

---

[75] See Gelman, King and Liu (1998) for an approach to multiple imputation for multiple surveys using hierarchical modeling.

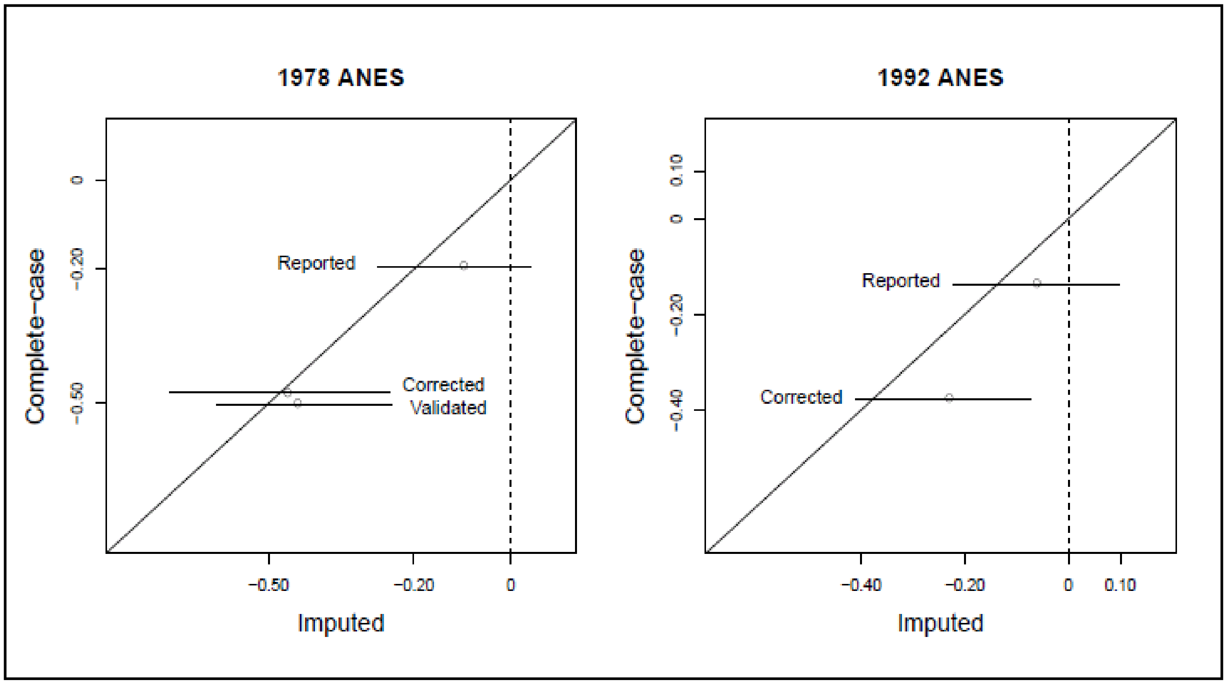
[76] The substantive results are essentially unchanged if, instead of the normal distributions, one-dimensional conditional gamma distributions are specified for these covariates, all of which are strictly positive.

sample for the 1978 ANES, and only 42% for the 1992 ANES. The left panel of the figure summarizes the marginal posterior distribution of $\beta_{Non-white}$ for the 1978 ANES using reported, validated and corrected vote. As in Section 4.4.2, our correction for misreporting was implemented based on auxiliary information from a random sub-sample of the ANES survey. The right panel plots the estimates for the 1992 ANES, for which we use validated turnout data from the previous Presidential elections, as in Section 4.4.3. In both cases, estimates obtained using Bayesian imputation are compared to those from the complete-case analyses.

Two interesting facts emerge from the figure. First, for both election-studies, the marginal posterior distribution for $\beta_{Non-white}$ estimated using our Bayesian imputation model is not statistically different from that obtained using list-wise deletion, at least at the 0.05 level. However, the standard errors tend to be lower when missing values are imputed than under list-wise deletion. This result holds in fact for most of the election-years under analysis, suggesting that by omitting the cases with missing values, much information is lost on the variables that are completely or almost completely observed, thus leading to less efficient parameter estimates (Ibrahim, Chen and Lipsitz 2002; Ibrahim et al. 2005). This is likely to be an important concern in the Election Studies examined here, given that there is substantial variation in the rates of item non-response, with most of the variables exhibiting relatively low percentage of missing values while a few others show very high rates of non-response (see Appendix 4.B).

Figure 4.16

Posterior summaries for $\beta_{Non-white}$ with list-wise deletion versus Bayesian imputation



Note: The graph plots point and interval summaries for $\beta_{Non-white}$ for the 1978 and 1992 ANES, using list-wise deletion and fully Bayesian imputation. The center dots correspond to the point estimates (posterior means), and the horizontal bars indicate the 90% and 50% credible intervals for the models with imputed missing values.

Second, imputing missing values does not change the substantive findings reported above regarding the performance of our methodology. The results for the 1978 ANES show that the estimated effects from our model correcting for misreporting are again closer to the benchmark case – using validated vote– than the effects estimated using recalled vote, and this result holds for all the ANES with validated vote. For the 1992 election, the marginal effect of race obtained from the corrected turnout model is also higher than in the uncorrected model, as was in the obtained from the corrected turnout model is also higher than in the uncorrected model, as was in the complete-case analysis. For both elections, once again, the main substantive conclusions regarding the effect of being Non-white on the probability of voting drawn from the model correcting for misreporting differ from those obtained using recalled vote.

4.5 Concluding remarks

Survey data are usually subject to measurement errors, generally referred to as classification errors when affecting discrete variables. In the political science literature, misclassification of binary dependent variables has received considerable attention in the context of estimating the determinants of voter turnout. High rates of overreporting have been documented in survey instruments commonly used to study turnout in the U.S., such as the American National Election Study (ANES) and the Current Population Survey (CPS), and most previous research has found that misreporting varies systematically with some of the relevant characteristics affecting the turnout decision.

In the presence of misreporting, standard binary choice models will generally yield biased parameter estimates and inaccurate standard errors and may lead to erroneous substantive conclusions. This paper develops a simple Bayesian method to correct for misreporting using information on the misreport mechanism from auxiliary data sources. Our model does not require full validation studies to be conducted every time a researcher is concerned about potential misreporting. As long as enough data exists to reasonably estimate the misreporting probabilities, our approach can be applied for drawing inference from the non-validated samples, improving the accuracy of the parameter estimates and inferences on the effect of covariates of interest on the true response *vis-à-vis* standard models ignoring misclassification and methods assuming constant misreport rates. This is clearly important, since obtaining "gold-standard" data is usually quite expensive and time consuming, and thus restricting the analysis only to validated studies will generally lead to discard large amounts of useful information, as in the case of the ANES.

The proposed model is fully general and modular, can be easily implemented using freely available software, and can be readily applied in the case of missing data in the response and/or covariates. While we illustrate our technique using turnout data from the American National Election Study, it could be applied in general to account for potential misclassification of a binary dependent variable in many other situations in which auxiliary data on the misreport structure is available (Bound, Brown and Mathiowetz 2001; Molinari 2003). Extensions to more general discrete choice models are also straightforward. Potential avenues for future research would be to use semi-or non-parametric methods to estimate both the misreporting and turnout models (Horowitz and Manski 1995; Molinari 2003), simultaneously account for response and covariate measurement errors within our model (McGlothlin, Stamey and Seaman 2008), and explore the possibility of incorporating semi-parametric approaches for inference with missing data (Chen and Ibrahim 2006; Robins and Rotnitzky 1995; Rotnitzky and Robins 1995).

While the primary focus of the paper has been on estimation techniques as opposed to substantive findings, the empirical application of our model to the analysis of the determinants of voter turnout has clear implications for researchers interested in race. Our results confirm that race does have a clear negative impact on turnout, and suggest that the null previous findings have been probably due to problems of misreporting, as had been argued by Abramson and Claggett (1984, 1986a, 1991). With the correction for misreporting developed in this paper, researchers could now better estimate the effect of race over the length of the ANES datasets and not just for the few years with validated turnout data. In addition, researchers might wish to revisit Wolfinger and Rosenstone (1980) findings of the effect of registration laws to see if properly correct misreporting re-enforces or diminishes their findings.

Appendix 4.A

Additional results from the Monte Carlo experiment in Section 4.3

Table 4.A.1

Misreport probabilities under Design B

Average misreport rates $\overline{\pi^{1\|0}}, \overline{\pi^{0\|1}}$	$\pi_i^{1\|0}$		$\pi_i^{0\|1}$	
	$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$
2%	0.01	0.12	0.06	0.008
5%	0.03	0.18	0.11	0.02
10%	0.05	0.35	0.24	0.04
20%	0.13	0.50	0.47	0.07

Table 4.A.2

Posterior means and 95% credible intervals - Design A

$\overline{\pi^{1\|0}}, \overline{\pi^{0\|1}}$	Estimator	β_0	β_1	β_2
-	True values	-1	1	1
0.02	Ignoring misreporting	-0.86	0.88	0.92
		(-1.01, -0.74)	(0.77, 0.99)	(0.74, 1.11)
	Proposed method	-0.99	0.99	1.04
		(-1.19, -0.82)	(0.83, 1.17)	(0.80, 1.27)
	Model A-1	-1.16	1.15	1.20
		(-1.56, -0.86)	(0.89, 1.48)	(0.89, 1.61)
	Model A-2	-1.00	1.01	1.05
		(-1.20, -0.82)	(0.85, 1.18)	(0.84, 1.30)
0.05	Ignoring misreporting	-0.82	0.78	0.90
		(-0.96, -0.68)	(0.68, 0.89)	(0.71, 1.08)
	Proposed method	-1.00	1.02	1.10
		(-1.21, -0.80)	(0.84, 1.22)	(0.84, 1.38)
	Model A-1	-1.05	1.08	1.20
		(-1.41, -0.76)	(0.82, 1.41)	(0.89, 1.62)
	Model A-2	-1.03	1.04	1.16
		(-1.26, -0.84)	(0.87, 1.24)	(0.90, 1.44)

0.10	Ignoring misreporting	-0.67	0.64	0.64
		(-0.79, -0.55)	(0.55, 0.74)	(0.48, 0.80)
	Proposed method	-0.96	0.88	0.86
		(-1.26, -0.70)	(0.69, 1.10)	(0.55, 1.21)
	Model A-1	-0.81	0.77	0.76
		(-1.32, -0.45)	(0.56, 1.15)	(0.54, 1.21)
	Model A-2	-1.03	0.96	0.94
		(-1.34, -0.78)	(0.76, 1.20)	(0.66, 1.26)
0.20	Ignoring misreporting	-0.46	0.51	0.50
		(-0.58, -0.34)	(0.42, 0.60)	(0.34, 0.66)
	Proposed method	-1.01	1.00	0.97
		(-1.47, -0.62)	(0.68, 1.42)	(0.48, 1.52)
	Model A-1	-0.71	0.75	0.74
		(-1.21, -0.34)	(0.52, 1.12)	(0.44, 1.16)
	Model A-2	-0.98	1.07	1.02
		(-1.38, -0.68)	(0.79, 1.39)	(0.64, 1.46)

Constant misclassification rates.

Sample size: N, M = 1,000.

Table 4.A.3

Posterior means and 95% credible intervals - Design B

$\overline{\pi^{1\|0}}, \overline{\pi^{0\|1}}$	Estimator	β_0	β_1	β_2
-	True values	-1	1	1
0.02	Ignoring misreporting	-0.95	0.93	1.04
		(-1.10, -0.82)	(0.81, 1.05)	(0.85, 1.23)
	Proposed method	-0.94	0.99	0.99
		(-1.12, -0.77)	(0.85, 1.16)	(0.77, 1.22)
	Model A-1	-1.18	1.12	1.23
		(-1.55, -0.89)	(0.90, 1.37)	(0.94, 1.56)
	Model A-2	-1.02	1.01	1.13
		(-1.20, -0.85)	(0.87, 1.16)	(0.91, 1.34)
0.05	Ignoring misreporting	-0.92	0.76	1.05
		(-1.06, -0.79)	(0.65, 0.86)	(0.87, 1.23)
	Proposed method	-1.01	1.01	1.02
		(-1.23, -0.80)	(0.82, 1.21)	(0.75, 1.29)
	Model A-1	-	-	-
	Model A-2	-1.20	0.99	1.33
		(-1.44, -0.97)	(0.81, 1.19)	(1.07, 1.64)

0.10	Ignoring misreporting	-0.88	0.70	1.18
		(-1.01, -0.75)	(0.59, 0.80)	(0.98, 1.37)
	Proposed method	-0.95	1.04	1.07
		(-1.24, -0.67)	(0.82, 1.28)	(0.75, 1.43)
	Model A-1	-	-	-
	Model A-2	-1.34	1.07	1.73
		(-1.66, -1.08)	(0.85, 1.32)	(1.40, 2.13)
0.20	Ignoring misreporting	-1.16	0.45	1.77
		(-1.30, -1.01)	(0.36, 0.55)	(1.58, 1.99)
	Proposed method	-1.06	0.98	1.27
		(-1.55, -0.56)	(0.66, 1.36)	(0.70, 1.82)
	Model A-1	-	-	-
	Model A-2	-2.58	1.00	3.76
		(-3.55, -1.92)	(0.70, 1.42)	(2.96, 4.92)

Covariate-dependent misclassification.

Sample size: N, M = 1,000.

Table 4.A.4

Posterior means and 95% credible for $\overline{\pi^{1|0}}$ and $\overline{\pi^{0|1}}$ [a]

Monte Carlo design	True sample values	Proposed method	Model A-1	Model A-2
A				
	$\overline{\pi^{1\|0}} = 2.75$	2.85	2.89	6.06
		(1.79, 4.29)	(1.77, 4.27)	(1.33, 11.28)
	$\overline{\pi^{0\|1}} = 1.84$	2.46	2.39	4.95
		(1.20, 4.07)	(1.10, 4.07)	(0.29, 2.12)
	$\overline{\pi^{1\|0}} = 5.17$	5.29	5.26	5.68
		(3.84, 7.02)	(3.77, 7.00)	(1.16, 11.12)
	$\overline{\pi^{0\|1}} = 4.47$	6.56	6.56	8.24
		(4.40, 9.21)	(4.34, 9.25)	(1.11, 16.49)
	$\overline{\pi^{1\|0}} = 9.53$	10.35	10.36	6.74
		(8.27, 12.71)	(8.08, 12.65)	(1.88, 11.72)
	$\overline{\pi^{0\|1}} = 9.13$	9.34	8.87	5.17
		(6.79, 12.27)	(6.36, 11.82)	(0.74, 9.60)
	$\overline{\pi^{1\|0}} = 20.68$	21.23	20.48	11.92
		(18.45, 24.28)	(17.57, 23.33)	(1.58, 21.77)

	$\overline{\pi^{0\|1}} = 20.26$	20.44	20.55	10.29
		(16.47, 24.53)	(16.81, 24.44)	(0.45, 24.85)
B				
	$\overline{\pi^{1\|0}} = 2.42$	1.62	1.52	4.55
		(0.84, 2.71)	(0.70, 2.63)	(0.64, 9.73)
	$\overline{\pi^{0\|1}} = 2.37$	2.60	2.20	2.76
		(1.23, 4.47)	(1.06, 3.75)	(0.13, 8.61)
	$\overline{\pi^{1\|0}} = 5.01$	5.14	5.36	-
		(3.72, 6.97)	(3.79, 7.10)	
	$\overline{\pi^{0\|1}} = 6.32$	5.01	4.56	-
		(3.15, 7.36)	(2.71, 6.74)	
	$\overline{\pi^{1\|0}} = 10.98$	9.94	9.96	-
		(7.92, 12.10)	(7.89, 12.38)	
	$\overline{\pi^{0\|1}} = 8.69$	11.04	10.04	-
		(8.27, 14.20)	(7.49, 12.95)	
	$\overline{\pi^{1\|0}} = 20.19$	19.59	17.37	-
		(17.03, 22.17)	(15.02, 19.65)	
	$\overline{\pi^{0\|1}} = 21.05$	21.54	17.54	-
		(17.97, 25.40)	(14.15, 21.02)	

[a] In percentage points.

Appendix 4.B

Sample description and variable coding

Table 4.B.1

Vote misreporting in the 1978-1990 ANES [a]

Election	$\Pr\left(\tilde{y}_i = 1 \mid y_i = 0\right)$	$\Pr\left(y_i = 0 \mid \tilde{y}_i = 1\right)$	$\Pr\left(\tilde{y}_i = 0 \mid y_i = 1\right)$	$\Pr\left(y_i = 1 \mid \tilde{y}_i = 0\right)$
1978	23.27	24.55	3.02	2.84
1980	24.48	16.52	0.58	1.37
1984	38.83	13.63	0.22	1.70
1986	31.55	17.70	0.66	1.40
1988	36.30	14.63	1.06	7.10
1990	26.83	16.83	3.67	6.46

[a] In percentage points.

Variables used in the turnout and misreport models

Age: 1 if Age < 30; 2 if 30 ≤ Age < 45; 3 if 45 ≤ Age < 60; 4 if Age ≥ 60.

Church Attendance: Frequency of church attendance. Coding: 1 if never; 2 if a few times a year; 3 if once or twice a month; 4 if every week or almost every week.

Education: Highest grade of school or year of college completed. Coding: 1 if 8 grades or less; 2 if 9–12 grades with no diploma or equivalency; 3 if 12 grades, diploma or equivalency; 4 if some college; 5 if college degree.

Female: 1 if the respondent is female, 0 if male.

Home owner: 1 if the respondent owns his house, 0 otherwise.

Income: Household income. Coding: 1 if 0–16th percentile; 2 if 17h–33d percentile; 3 if 34th–67th percentile; 4 if 68th–95th percentile; 5 if 96th–100th percentile.

Non-white: 0 if white, 1 otherwise.

Party Identification: -1 for Democrats, 0 for Independents, 1 for Republicans.

Partisan Strength: Coded on a four-point scale ranging from 1 for pure independents to 4 for strong partisans.

Alone: 1 if the respondent was interviewed alone, 0 otherwise.

Uncooperative: Respondent's level of cooperation in the interview, as evaluated by the interviewer. Coding: 1 if very good; 2 if good; 3 if fair; 4 if poor; 5 if very poor.

Sincerity: How sincere did the respondent seem to be in his/her answers, as evaluated by the interviewer. Coding: 1 if often seemed insincere; 2 if usually sincere; 3 if completely sincere.

In order to reduce the correlation between the parameters and to accelerate convergence and mixing of the Gibbs sampling algorithm, all variables where centered at their mean values (Gu, 2006).

Table 4.B.2

Rates of non-response for the variables included in the voter turnout models

Variable	1978 – 1990 Validated ANES	1992 ANES
Age	2.07	0.00
Church Attendance	13.20	33.72
Education	0.80	2.61
Female	4.28	0.00
Income	13.58	10.66
Non-white	4.41	1.41
Home owner	0.70	6.44
Partisan Strength	4.44	0.56
Party Identification	2.60	0.36
Alone	4.55	1.57
Cooperation	4.49	0.16
Sincerity	0.47	0.24
Reported turnout	6.12	9.30
Total sample	11,632	2,485
Complete-case sample	6,411	1,206

C h a p t e r 5

Concluding remarks

Voting behavior research has been traditionally taken to be the one areas of political science where theory can be systematically and quantitatively measured and tested and where statements of causal determinants can be more reliably formulated (Eldersveld, 1951). As a result, the importance of sophisticated data analysis methods for academic studies of electoral politics can hardly be understated. In fact, while many of the fundamental questions in the field were already defined in the 1940s and 1950s, our answers to these questions have changed considerably in the past few decades, paralleling changes in our ways of studying them. Although many of the major controversies in this area are not yet settled, the development of increasingly refined research techniques has probably contributed to improve our understanding of these controversies and to getting us closer to solving them (Niemi and Weisberg, 2001).

The increasing adoption of Bayesian methods – and, more specifically, Markov chain Monte Carlo (MCMC) algorithms – for estimation and inference in political science research provides further opportunities to advance in this direction. As shown in the previous three chapters of this dissertation, MCMC methods can be implemented to address substantive and methodological questions regarding voter participation and choice in settings in which other estimation techniques would be intractable, problematic or inefficient. Each chapter underscores the advantages of Bayesian simulation *vis-à-vis* alternative approaches for dealing with the specific problems at hand and points to possible avenues for future research. More generally, the applications included in this dissertation illustrate the power of MCMC methods to handle models long thought to be "too hard" to estimate, contributing to the development, implementation and testing of more elaborated theories of voting behavior and, thereby, to the scientific study of electoral politics, and area in which concepts, data, methods and conclusions are intimately intertwined (Converse, 2006).

Perhaps the main drawback of Bayesian simulation based on MCMC algorithms is that it is extremely computationally intensive and time-consuming. In order to ensure an accurate approximation to the posterior densities of random quantities of interest (e.g., parameters and missing data), a large number of draws from the corresponding conditional densities must be obtained, and the "quality" of the approximation is an increasing function of the number of iterations of the sampling algorithm. Still, computing resources available to scholars have become increasingly faster and cheaper during the last decade, and the release of flexible and freely available software for Bayesian inference (e.g., BUGS and JAGS) has lowered the levels of statistical and programming expertise required to implement MCMC methods (Jackman, 2000a). In addition, considerable efforts are currently being devoted by statisticians and computer scientists to developing new samplers in order to speed convergence and reduce execution times. In this direction, MCMC methods are likely to continue making strides in political science in the next years, reaching a growing number of academics interested not only in methodological but also in substantive research questions (Gill, 2000; Jackman, 2009).

B i b l i o g r a p h y

Abelson, Robert P., Loftus, Elizabeth F., and Greenwald, Anthony G. 1992. Attempts to Improve the Accuracy of Self-Reports of Voting. In J Tanur (ed.) *Questions about Questions: Inquiries into the Cognitive Bases of Surveys*, pp. 138 - 153. New York: Russel Sage.

Abramson, Paul R., and Claggett, William. 1984. Race-Related Differences in Self-Reported and Validated Turnout. *Journal of Politics* 46:719–738.

Abramson, Paul R., and Claggett, William. 1986a. Race-Related Differences in Self-Reported and Validated Turnout in 1984. *Journal of Politics* 48:412–422.

Abramson, Paul R, and Claggett, William. 1986b. Race-Related Differences in Self-Reported and Validated Turnout in 1986. *Journal of Politics* 51:397–408.

Abramson, Paul R., and Claggett, William. 1991. Race-Related Differences in Self-Reported and Validated Turnout in the 1988 Presidential Election. *Journal of Politics* 53:186–197.

Abrevaya, Jason, and Hausman, Jerry A. 1999. Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells. Mimeo, Department of Economics, Massachusetts Institute of Technology.

Aigner, Dennis J. 1973. Regression with a Binary Independent Variable Subject to Errors of Observation. *Journal of Econometrics* 1:49–60.

Aitchison, John. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, John, and Shen, Sheng M. 1980. Logistic-Normal Distribution: Some Properties and Uses. *Biometrika*, 67: 261-72.

Aldrich, John. H. 1993. Rational Choice and Turnout. *American Journal of Political Science* 37: 246 – 278.

Aldrich, John H., and McKelvey, Richard D. 1977. A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections. *American Political Science Review* 71(1): 111-130.

Alemán, Eduardo, and Saiegh, Sebastián. 2007. Legislative Preferences, Political Parties and Coalition Unity in Chile. *Comparative Politics* 39(3): 253-272.

Allenby, Greg M., and Lenk, Peter J. 1994. Modeling Household Purchase Behavior with Logistic Normal Regression. *Journal of the American Statistical Association* 89 (428): 1218 – 1231.

Alvarez, R. Michael. 1998. *Information and Elections*. Ann Arbor: The University of Michigan Press.

Alvarez, R. Michael, and Nagler, Jonathan. 1995. Economic, Issues and the Perot Candidacy: Voter Choice in the 1992 Presidential Election. *American Journal of Political Science* 30(3): 714-744.

Alvarez, R. Michael, Nagler, Jonathan, and Bowler, Shaun. 2000. Issues, Economics, and the Dynamics of Multiparty Elections: The British 1987 General Election. *American Political Science Review* 94(1): 131-149.

Ames, Barry. 1995. Electoral Strategy under Open-List Proportional Representation. *American Journal of Political Science* 39(2): 406-33.

Anderson, Brady A., and Silver, Brian D. 1986. Measurement and Mismeasurement of the Validity of the Self-Reported Vote. *American Journal of Political Science* 30:771–785.

Andrews, D. F., and Mallows, C. L. 1974. Scale mixtures of normality. *Journal of the Royal Statistical Society*, Series B, 36: 99-102.

Angell, Alan, and Reig, Cristobal. 2006. Change or Continuity? The Chilean Elections of 2005/2006. *Bulletin of Latin American Research* 25(4): 481-502.

Ansolabehere, Stephen and Eitan Hersh. 2008. Vote Validation in the 2006 CCES. Mimeo.

Antweiler, Werner. 2001. Nested random effects estimation in unbalanced panel data. *Journal of Econometrics* 101: 295 – 313.

Australian Joint Standing Committee on Electoral Matters. 2000. *The 1998 Election: Report of the Inquiry into the conduct of the 1998 Federal Election and matters related thereto.* Canberra: Parliament of the Commonwealth of Australia.

Barndorff-Nielsen, Ole E., and Jørgensen, Bent. 1991. Some Parametric Models on the Simplex. *Journal of Multivariate Analysis* 39: 106 – 116.

Battistin, Erich. 2003. Errors in Survey Reports of Consumption Expenditures. London: Working Paper 0307, Institute for Fiscal Studies.

Belli, Robert F., Traugott, Michael W., Young, Margaret, and McGonable, Katherine A. 1999. Reducing Vote Overreporting in Surveys: Social Desirability, Memory Failure, and Source Monitoring. *Public Opinion Quarterly* 63(1): 90–108.

Belli, Robert F., Traugott Michael W., and Beckman, Matthew N. 2001. What Leads to Voting Overreports? Contrasts of Overreporters to Validated Voters and Admitted

Nonvoters in the American National Election Studies. *Journal of Official Statistics* 17(4): 479–498.

Belli, Robert, Traugott, Santa, and Rosenstone, Steven. 1994. Reducing Over-Reporting of Voter Turnout: An Experiment Using a ‘Source Monitoring’ Framework. ANES Technical Report Series (010153).

Berger, James O. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

Bernstein, Robert, Chadha, Anita, and Montjoy, Robert. 2001. Overreporting Voting. Why it Happens and Why it Matters. *Public Opinion Quarterly* 65(1):22–44.

Bhaumik, Amitabha, Dey, Dipak K., and Ravishanker, Nalini. 2003. A dynamic Linear Model Approach for Compositional Time Series Analysis. Technical Report, University of Connecticut.

Billheimer, Dean, Guttorp, Peter, and Fagan, William F. 2001. Statistical interpretation of Species Composition. *Journal of the American Statistical Association* 96(456): 1205 – 1214.

Blais, André, and Dobrzynska, Agnieszka. 1998. Turnout in Electoral Democracies. *European Journal of Political Research* 33: 239-261.

Bollinger, Chris R. 1996. Bounding Mean Regressions When a Binary Regressor is Mismeasured. *Journal of Econometrics* 73: 387–399.

Bonilla, Claudio 2002. A Micro Application of the Spatial Theory of Voting. *Revista de Ciencia Política* 21(2): 3-16.

Bound, John, Brown, Charles, and Mathiowetz, Nancy. 2001. Measurement Error in Survey Data. In J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, Vol. 5, pp. 3705-3843. North Holland: Elsevier.

Brazilian Institute of Geography and Statistics. 2003. *Estatísticas do Século XX*, Rio de Janeiro: IBGE.

Bross, Irwin. 1954. Misclassification in 2 X 2 Tables. *Biometrics* 10:478–486.

Browne, William J., and Draper, David. 2001. Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models. *Computational Statistics* 15(3): 391-420.

Burden, Barry C. 2000. Voter Turnout and the National Election Studies. *Political Analysis* 8(4): 389–398.

Bryk, Anthony S., and Raudenbush, Stephen W. 2002. *Hierarchical linear models: Applications and data analysis methods*. 2nd ed. Newbury Park, CA: Sage Publications.

Cahalan, Don. 1968. Correlates of Respondent Accuracy in the Denver Validity Survey. *Public Opinion Quarterly* 32(3): 607–621.

Carlin, Bradley P., and Louis, Thomas A. 1996. *Bayesian and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Carroll, Raymond J., Ruppert, David, and Stefanski, Leonard A. 1995. *Measurement error in nonlinear model*s. London: Chapman and Hall.

Casella, George, and George, Edward L. 1992. Explaining the Gibbs Sampler. *The American Statistician* 46(3): 167-174.

Cassel, Carol A. 2003. Overreporting and Electoral Participation Research. *American Politics Research* 31(1): 81–92.

Centro de Estudios Públicos (CEP). 2005. Estudio Nacional de Opinión Pública Nº 23 – Tercera Serie. Octubre – Noviembre 2005.

Chaloner, Kathryn, and Brant, Rollin. 1988. A Bayesian Approach to Outlier Detection and Residual Analysis. *Biometrika* 75: 651-660.

Chen, Qingxia, and Ibrahim, Joseph G. 2006. Semiparametric Models for Missing Covariate and Response Data in Regression Models. *Biometrics* 62: 177–184.

Chen, Qingxia, Ibrahim, Joseph G., Chen, Ming-Hui, and Senchaudhuri, Pralay. 2008. Theory and inference for regression models with missing responses and covariates. *Journal of Multivariate Analysis* 99: 1302–1331.

Chen, Ming-Hui, Ibrahim, Joseph G., and Shao, Qi-Man. 2000. Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference* 84: 121–137.

Chen, T. Timothy. 1979. Log-Linear Models for Categorical Data With Misclassification and Double Sampling. *Journal of the American Statistical Association* 74(366): 481–488.

Chib, Siddharta, and Greenberg, Edward. 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician* 49(4): 327–335.

Christin, Thomas and Simon Hug. 2004. Methodological Issues in Studies of Conflict Processes: Misclassifications and Endogenous Institutions. Paper presented at the Annual Meeting of the American Political Science Association, Chicago, IL, September.

Clausen, Aage. 1968. Response Validity: Vote Report. *Public Opinion Quarterly* 32: 588–606.

Comparative Study of Electoral Systems (CSES). 2007. CSES Module 2 Election Study

Archive.

Congdon, Peter. 2003. *Applied Bayesian Modelling*. London: Wiley & Sons.

Constable, Pamela, and Valenzuela, Arturo. 1991. *A Nation of Enemies: Chile under Pinochet*. New York: Norton.

Converse, Philip E. 2006. Researching Electoral Politics. *American Political Science Review* 100(4): 605 – 612.

Cox, Gary. 1997. *Making Votes Count*. Cambridge: Cambridge University Press.

Davidov, Ori, Faraggi, David, and Reiser, Benjamin. 2003. Misclassification in Logistic Regression with Discrete Covariates. *Biometrical Journal* 5: 541–553.

Derks, Anton, and Deschouwer, Kris. 1998. Vrijzinigen, ongelovigen en protest. In M. Swyngedouw, J. Billiet, A. Carton and R. Beerten (eds.), *Kiezen is verliezen. Onderzoek naar de politieke opvattingen van Vlamingen*, pp. 85-112. Acco: Leuven.

Dow, Jay K., and Endersby, James W., 2004. Multinomial probit and multinomial logit: a comparison of choice models of voting research. *Electoral Studies* 23(1): 107-122.

Drake, Christiana. 1993. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biometrics* 49:1231–1236.

Draper, David. 2001. *Bayesian hierarchical modeling*. New York: Springer.

Duan, Yuyan. 2005. A Modified Bayesian Power Prior Approach with Applications in Water Quality Evaluation. Blacksburg, VA: Doctoral Dissertation, Department of Statistics, Virginia Polytechnic Institute and State University.

Dunson, David B., and Tindall, Kenneth R. 2000. Bayesian Analysis of Mutational Spectra. *Genetics* 156:1411–1418.

Dustmant, Christian, and van Soest, Arthur. 2004. An analysis of speaking fluency of immigrants using ordered response models with classification errors. *Journal of Business and Economic Statistics* 22: 312–321.

Eldersveld, Samuel J. 1951. Theory and Method in Voting Behavior Research. *Journal of Politics* 13(1): 70 – 87.

Escobar, Marcelo, Calvo, Ernesto, Calcagno, Natalia, and Minvielle, Sandra. 2002. Ultimas Imágenes Antes del Naufragio: Las Elecciones del 2001 en Argentina. *Desarrollo Economico* 42: 25-44.

Fornos, Carolina. 1996. *Explaining Voter Turnout in Latin America*. Master's Thesis, Louisiana State University.

Fornos, Carolina, Power, Timothy, and Garand, James. 2004. Explaining Voter Turnout in Latin America, 1980 to 2000. *Comparative Political Studies* 37: 909 – 940.

Franklin, Charles H., and Jackson, John E. 1983. The Dynamics of Party Identification. *Political Science Review* 77(4): 957-973.

Francese, Robert J. 2005. Empirical Strategies for Various Manifestations of Multilevel Data. *Political Analysis* 13: 430 – 446.

Franklin, Mark, and Hirczy de Mino, Wolfgang. 1998. Separated powers, divided government, and turnout in U.S. presidential elections. *American Journal of Political Science* 42: 316-326.

Frei, Eugenio. 2003. Los Partidos Políticos Chilenos Cambio y Estabilidad en el Comportamiento Electoral, 1990 -2000. *Revista de Ciencia Política* 23(2), 109-147.

Fullerton, Andrew S., Dixon, Jeffrey C., and Borch, Casey A. 2007. Bringing Registration into Models of Vote. *Public Opinion Quarterly* 71(4):649–660.

Gamboa, Ricardo, and Segovia, Carolina. 2006. Las Elecciones Presidenciales y Parlamentarias en Chile, Diciembre 2005 – Enero 2006. *Revista de Ciencia Política* 26(1): 84-113.

Garthwaite, Paul H., Kadane, Joseph B., and OHagan, Anthony. 2004. Elicitation. Pittsburgh, PA: Technical Report 808, Department of Statistics, Carnegie Mellon University, Department of Statistics.

Gelfland, Alan E., Hills, Susan E., Racine-Poon, Amy, and Smith, Adrian F. 1990. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association* 85(412): 972 – 985.

Gelfland, Alan E., and Smith, Adrian F. 1990. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85(410): 398 – 409.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. 2004. *Bayesian Data Analysis*. Boca Raton: Chapman & Hall.

Gelman, Andrew, and Hill, Jennifer. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Gelman, Andrew, King, Gary, and Liu, Chuanhai. 1998. Not Asked and Not Answered: Multiple Imputation for Multiple Surveys. *Journal of the American Statistical Association* 93(443): 846– 857.

Gelman, Andrew, and Pardoe, Iain. 2007. Average predictive comparisons for models with nonlinearity, interactions and variance components. *Sociological Methodology* 37(1): 23-51.

Gelman, Andrew and Rubin, Donald B. 1992. Inference for iterative simulation using multiple sequences. *Statistical Science* 7: 457-472.

Geman, Stuart, and Geman, Donald. 1984. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721-741.

Gems, Barbara, Ghaosh, Dhirendra, and Hitlin, Robert. 1982. A Recall Experiment: Impact of Time on Recall of Recreational Fishing Trips. *Proceedings of the Section on Survey Research Methods* 7: 168–173.

Geweke, John. 1992. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. Bernardo, J. Berger, A. Dawid and A. Smith (eds.), *Bayesian Statistics*, Vol. 4, pp. 169-193. Oxford: Oxford University Press.

Gilks, Walter R., Richardson, Sylvia, and Spiegelhalter, David J. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gill, Jeff. 2000. *Bayesian Methods: A Social and Behavioral Sciences Approach*. New York: Chapman & Hall.

Goldstein, Harvey. 1995. *Multilevel Statistical Models*. London: Arnold.

Greene, William. 1999. *Econometric Analysis*, 5th edition. New York: Prentice Hall.

Gu, Yuanyuan. 2006. Misclassification of the Dependent Variable in Binary Choice Models. Australia: Masters' Dissertation, School of Economics, University of the New South Wales.

Hajivassiliou, Vassilis A., McFadden, Daniel, and Ruud, Paul A. 1996. Simulation estimation methods for limited dependent variable models. In G. Maddala, C. Rao, and H. Vinod (eds.), *Handbook of Statistics*, Vol. 11, pp. 519-543. Amsterdam: North Holland.

Hardle, Wolfgang. 1990. *Applied nonparametric regression*. New York: Cambridge University Pres.

Haukka, Jari K. 1995. Correction for covariate measurement error in generalized linear model -a bootstrap approach. *Biometrics* 51: 1127–1132.

Hausman, Jerry A., Abrevaya, Jason, and Scott-Morton, Fiona M. 1998. Misclassification of the dependent variable in a discrete response setting. *Journal of Econometrics* 87(2): 239–269.

Hausman, Jerry, and McFadden, Daniel. 1984. Specification tests for the multinomial logit Model. *Econometrica* 52(5): 1219–1240.

Hernanz, Virginia, Malherbet, Franck, and Pellizzari, Michelle. 2004. Take-up of Welfare Benefits in OECD Countries: A Review of the Evidence. OECD Social, Employment and Migration Working Papers 17.

Highton, Benjamin. 2004. Self-reported versus Proxy-reported Voter Turnout in the Current Population Survey. *Public Opinion Quarterly* 69(1): 113–123.

Hill, Lisa. 2002. On the Reasonableness of Compelling Citizens to ‘Vote’: The Australian Case. *Political Studies* 50: 80 -101.

Hill, Kim Q., and Hurley, Patricia A. 1984. Nonvoters in Voters’ Clothing: The Impact of Voting Behavior Misreporting on Voting Behavior Research. *Social Science Quarterly* 65:199–206.

Hillygus, D. Sunshine, and Jackman, Simon. 2003. Voter Decision Making in Election 2000: Effects, Partisan Activation and the Clinton Legacy. *American Journal of Political Science* 47(4): 583-596.

Hinich, Melvin J., and Munger, Michael. 1994. *Ideology and the Theory of Political Choice*. Ann Arbor: The University of Michigan Press.

Hirczy de Mino, Wolfgang. 1994. The impact of mandatory voting laws on turnout: A quasi-experimental approach. *Electoral Studies* 13: 64-76.

Horowitz, Joel. 1980. The accuracy of the multinomial logit model as an approximation to the multinomial probit model of travel demand. *Transportation Research* B 14: 331-341.

Horowitz, Joel L. 1993. Semiparametric and Nonparametric Estimation of Quantal Response Models. In G. Maddala, C. Rao, and H. Vinod (eds.), *Handbook of Statistics*, Vol. 11, pp. 45–72. Amsterdam: Elsevier.

Horowitz, Joel L., and Manski, Charles F. 1995. Identification and Robustness with Contaminated and Corrupted Data. *Econometrica* 63(2): 281–302.

Horton, Nicholas J., and Kleinman, Ken P. 2007. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* 61(1): 79–90.

Huneeus, Carlos. 2006. Las elecciones presidenciales y parlamentarias del 2005 en Chile. *Publicaciones de Coyuntura*, Centro de Estudios de la Realidad Contemporanea.

Hu, Yingyao. 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics* 144: 27–61.

Hu, Yingyao, and Riddert, Geert. 2007. Estimation of Nonlinear Models with Mismeasured Regressors Using Marginal Information. Mimeo.

Huang, Lan, Chen, Ming-Hui, and Ibrahim, Joseph G. 1999. Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates. *Biometrics* 61: 767–780.

Ibrahim, Joseph G., and Chen, Min-Hui. 2000. Power Prior Distributions for Regression Models. *Statistical Science* 15(1): 46–60.

Ibrahim, Joseph G., Chen, Ming-Hui, and Lipsitz, Stuart R. 2002. Bayesian Methods for Generalized Linear Models With Covariates Missing at Random. *Canadian Journal of Statistics* 30: 55–78.

Ibrahim, Joseph G., Chen, Ming-Hui, Lipsitz, Stuart R., and Herring, Amy H. 2005. Missing-Data Methods for Generalized Linear Models: A Comparative Review. *Journal of the American Statistical Association* 100(469): 332–346.

Ibrahim, Joseph G., Lipsitz, Stuart R., and Chen, Min-Hui. 1999. Missing Covariates in Generalized Linear Models when the Missing Data Mechanism Is Non-ignorable. *Journal of the Royal Statistical Society* B 61(1): 173–190.

Ibrahim, Joseph G., Lipsitz, Stuart R., and Horton, Nick. 2001. Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Applied statistics* 50(3): 361–373.

Ibrahim, Joseph G., Ryan, Luise M., and Chen, Ming-Hui. 1998. Using Historical Controls to Adjust for Covariates in Trend Tests for Binary Data. *Journal of the American Statistical Association* 93(444): 481–488.

Instituto Universitario de Pesquisas de Rio de Janeiro (IUPERJ). 2006. Banco de Dados Eleitorais Do Brasil (accessed December 2006).

Imai, Kosuke, van Dyk, David A. 2005. A Bayesian analysis for the multinomial probit model using marginal data augmentation. *Journal of Econometrics* 124: 311-334.

International Institute for Democracy and Electoral Assistance, Voter Turnout Database, Compulsory Voting: ://www.idea.int/vt/index. (accessed December 2007).

Iyengar, Malini, and Dey, Dipak K. 2004. Bayesian analysis of compositional data. In D. Dey, S. Ghosh and B. Mallick (eds.), *Generalized linear Models: A Bayesian Perspective*, pp. 349-364. New York: Marcel Dekker.

Jackman, Robert W. 1987. Political institutions and voter turnout in the industrial democracies. *American Political Science Review* 81: 405 – 423.

Jackman, Simon. 2000a. Estimation and Inference via Bayesian Simulation: An Introduction to Markov Chain Monte Carlo. *American Journal of Political Science* 44(2): 369 – 398.

Jackman, Simon. 2000b. Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation. *Political Analysis* 8(4): 307 – 332.

Jackman, Simon. 2001. Voting: Compulsory. In N. Smelser and P. Baltes (eds.), *International Encyclopedia of the Social & Behavioral Sciences*. Oxford: Elsevier Science.

Jackman, Simon. 2004. Bayesian Analysis for Political Research. *Annual Review of Political Science* 7: 483-505.

Jackman, Simon. 2009. *Bayesian Analysis for the Social Sciences*. West Sussex: Wiley & Sons.

Jackman, Robert W., and Miller, Ross A. 1995. Voter Turnout in the Industrial Democracies during the 1980s. *Comparative Political Studies* 27: 467-492.

Jocelyn-Holt, Alfredo. 1998. *El Chile Perplejo*. Santiago: Planeta/Ariel.

Johnson, Norman L., and Kotz, Samuel. 1972. *Distributions in Statistics: Continuous Multivariate distributions*. New York: Wiley.

Kass, Robert E., and Raftery, Adrian E. 1995. Bayes Factors. *Journal of the American Statistical* Association 90: 773-795.

Katosh, John P., and Traugott, Michael W. 1981. The Consequences of Validated and Self-Reported Voting Measures. *Public Opinion Quarterly* 45: 519–535.

Katz, Jonathan, and King, Gary. 1999. A Statistical Model for Multiparty Electoral Data. *American Political Science Review* 93: 15 – 32.

Kim, Yeonbae, Kim, Tai-Yoo, and Heo, Eunnyeong. 2003. Bayesian estimation of multinomial probit models of work trip choice. *Transportation* 30: 51-365.

King, Gary. 1997. *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

King, Gary, Murray, Christopher J., Salomon, Joshua A., and Tandon, Ajay. 2004. Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. *American Political Science Review* 98(1): 191-207.

King, Gary, Tomz, Michael, and Wittenberg, Jason. 2000. Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *Journal of Political Science* 44: 341-355.

Kitschelt, Herbert. 1995. *The Radical Right in Western Europe. A Comparative Analysis*. Ann Arbor: The University of Michigan Press.

Kostadinova, Tatiana. 2003. Voter Turnout Dynamics in Post-Communist Europe. *European Journal of Political Research* 42: 741-759.

Kuha, Jouni. 1994. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine* 13: 1135–1148.

Leighley, Jan E., and Nagler, Jonathan. 1984. Individual and Systemic Influences on Turnout: Who Votes? *Journal of Politics* 54: 718–740.

Lewbel, Arthur. 2000. Identification of the Binary Choice Model with Misclassification. *Econometric Theory* 16:603–609.

Lijphart, Arend. 1997. Unequal Participation: Democracy's Unresolved Dilemma. *American Political Science Review* 91: 1-14.

Lindley, Dennis V., and Smith, Adrian F. 1972. Bayes Estimates for the Linear Model. J*ournal of the Royal Statistical Society*, Series B, 34: 1-41.

Lipsitz, Stuart R., and Ibrahim, Joseph G. 1996. A Conditional Model for Incomplete Covariates in Parametric Regression Models. *Biometrika* 83: 916–92.

Little, Roderick J., and Rubin, Donald B. 2002. *Statistical analysis with missing data*. New York: Wiley.

Little, Roderick J., and Wang, Yongxiao. 1996. Pattern-Mixture Models for Multivariate Incomplete Data With Covariates. *Biometrics* 52: 98–111.

Liu, Chuanhai. 1996. Bayesian Robust Multivariate Linear Regression With Incomplete Data. *Journal of the American Statistical Association* 91(435): 1219-1227.

Loftus, Elizabeth F. 1975. Leading Questions and the Eyewitness Report. *Cognitive Psychology* 7:145–177.

Lohr, Sharon L. 2005. Inference from Multiple Frame Surveys. Statistics and Biostatistics Colloquium Series, Department of Statistics, Ohio State University.

Lubbers, Marcel, and Scheepers, Peer. 2000. Individual and contextual characteristics of the vlaams blok vote. *Acta Politica* 35: 363 – 398.

Mainwaring, Scott. 1991. Politicians, Parties and Electoral Systems: Brazil in Comparative Perspective. *Comparative Politics* 24: 21-43.

Maas, Cora J., and Hox, Joop J. 2004. Robustness issues in multilevel regression analysis. *Statistica Neerlandica* 58(2): 127–137.

Manski, Charles F. 1985. Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator. *Journal of Econometrics* 82:46–51.

McAllister, Ian, and Makkai, Toni. 1993. Institutions, Society of Protest? Explaining Invalid Votes in Australian Elections. *Electoral Studies* 12: 23-40.

McCulloch, Robert E., Polson, Nicholas G., and Rossi, Peter E. 2000. A Bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics* 99: 173-193.

McCulloch, Robert E., and Rossi, Peter E. 1994. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics* 99: 173-193.

Mcdonald, Michael P. 2007. The True Electorate: A Cross-Validation of Voter Registration Files and Election Survey Demographics. *Public Opinion Quarterly* 71(4): 588–602.

McFadden, Daniel. 1987, Regression-based specification tests for the multinomial logit model. *Journal of Econometrics* 34: 63–82.

McGlothlin, Anna, Stamey, James D., and Seaman, John W. 2008. Binary Regression with Misclassified Response and Covariate Subject to Measurement Error: a Bayesian Approach. *Biometrical Journal* 50:123–134.

McInturff, Pat, Johnon, Wesley O., Cowling David, and Gardner, Ian A. 2004. Modelling risk when binary outcomes are subject to error. *Statistics in Medicine* 23(7):1095–1109.

Merrill, Samuel III, and Grofman, Bernard. 1999. *A Unified Theory of Voting – Directional and Proximity Spatial Models*. Cambridge: Cambridge University Press.

Miller, Mungo. 1952. The Waukegan Study of Voter Turnout Prediction. *Public Opinion Quarterly* 32:588–606.

Moisés, Alvaro. 1993: Elections, Political Parties and Political Culture in Brazil: Changes and Continuities. *Journal of Latin American Studies* 25: 575-611.

Molinari, Francesca. 2003. Contaminated, Corrupted and Missing Data. Evanston, IL: Doctoral Dissertation, Department of Economics, Northwestern University.

Morrissey, Mary J., and Spiegelman, Donna. 1999. Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons. *Biometrics* 55(398): 338–344.

Murphy, Kevin M., and Topel, Richard H. 1985. Estimation and Inference in Two Step Econometric Models. *Journal of Business and Economic Statistics* 3: 370–379.

Myerson, Roger B., and Weber, Robert J. 1993. A Theory of Voting Equilibria. *American Political Science Review* 87(1): 102-114.

Nagler, Jonathan. 1994. Scobit: An Alternative Estimator to Logit and Probit. *American Journal of Political Science* 38(1): 230–255.

Navia, Patricio. 2006. La elección presidencial de 2005 en Chile. El Debate Político. *Revista Iberoamericana de Análisis* Político 3(4/5): 215-228.

Niemi, Richard G., and Weisberg, Herbert F. 2001. The Study of Voting and Elections. In R. Niemi and H. Weisberg (eds.), *Controversies in Voting Behavior*, Fourth Edition, pp. 1 – 21. Washington: CQ Press.

Neuhaus, John M. 1999. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 86(4):843–855.

Newton, Michael A., and Raftery, Adrian E. 1994. Approximate Bayesian Inference by the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society*, Series B, 56: 3-48.

Parry, Hugh, and Crossley, Helen. 1950. Validity of Responses to Survey Questions. *Public Opinion Quarterly* 14: 61–80.

Paulino, Carlos D., Soares, Paulo, and Neuhaus, John. 2003. Binomial Regression with Misclassification. *Biometrics* 59: 670–675.

Pinheiro, Jose C., Liu, Chuanhai H., and Wu, Ying N. 2001. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 10: 249 – 276.

Powell, Bingham G. 1986. American voter turnout in comparative perspective. *American Political Science Review* 80: 17-43.

Power, Timothy, and Garand, James. 2007. Determinants of Invalid Voting in Latin America. *Electoral Studies* 26(2): 432-444.

Power, Timothy, and Roberts, Timmons. 1995. Compulsory Voting, Invalid Ballots, and Abstention in Brazil. *Political Research Quarterly* 48: 795-826.

Plummer, Martyn. 2009. JAGS version 1.03 manual. www-ice.iarc.fr/~/software/jags/.

Poterba, James M., and Summers, Lawrence H. 1986. Reporting Errors and Labor Market Dynamics. *Econometrica* 54: 1319–1338.

Poterba, James M., and Summers, Lawrence H. 1995. Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics* 77(2): 207–216.

Prescott, Gordon J., and Garthwaite, Paul H. 2002. Simple Bayesian Analysis of Misclassified Binary Data with a Validation Substudy. *Biometrics* 58: 454–458.

Prescott, Gordon J., and Garthwaite, Paul H. 2005. Bayesian analysis of misclassified binary data from a matched case-control study with a validation sub-study. *Statistics in Medicine* 24(3).

Quinn, Kevin M., and Martin, Andrew D. 1998. Operationalizing and Testing Spatial Theories of Voting. Paper presented at the annual meeting of the Midwest Political Science Association, Chicago, IL, April 1998.

Raghuanthan, Trivellore E., Xie, Dawei, Schenker, Nathaniel, Parsons, Van L., Davis, William W., Dodd, Kevin W., and Feuer, Eric J. 2006. Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening. The University of Michigan Department of Biostatistics Working Paper Series, Working Paper 58.

Rabinowitz, George, and MacDonald, Stuart E. 1989. A Directional Theory of Issue Voting. *American Political Science Review* 83(1): 93-121.

Rayens, W. S., and Srinivasan, C. 1991. Box Cox Transformations in the Analysis of Compositional Data. *Journal of Chemometrics* 5: 227 – 239.

Reilly, Marie, and Pepe, Margaret S. 1995. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika* 82: 299–314.

Robert, Christian P., and Casella, George. 2004. *Monte Carlo Statistical Methods*. New York: Springer.

Robins, James M., and Rotnitzky, Andrea G. 1995. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association* 90: 122 – 129.

Robins, James M., Rotnitzky, Andrea G., and Zhao, Lue P. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89: 846–866.

Robinson, William S. 1950. Ecological Correlation and the Behavior of Individuals. *American Sociological Review* 15: 351-357.

Rosa, Guilherme J., Padovani, Carlos R., and Gianola, Daniel. 2003. "Robust Linear Mixed Models with Normal/Independent Distributions and Bayesian MCMC Implementation". *Biometrical Journal* 45(5): 573-590.

Rosenstone, Steve, and Hansen, John. 1993. *Mobilization, participation and democracy in America*. New York: Macmillan.

Rossi, Peter E., Allenby, Greg M., and McCulloch, Robert E. 2005. *Bayesian Statistics and Marketing*. West Sussex: John Wiley and Sons.

Rotnitzky, Andrea G., and Robins, James M. 1995. Semiparametric Regression Estimation in the Presence of Dependent Censoring. *Biometrika* 82: 805–820.

Rubin, Donald B. 1976. Inference and missing data. *Biometrika* 63:581–592.

Scully, Timothy R. 1992. *Rethinking the Center. Party Politics in Nineteenth and Twentieth Century Chile*. Stanford: Stanford University Press.

Scully, Timothy R. 1995. Reconstituting Party Politics in Chile. In S. Mainwaring and T. Scully (eds.), *Building Democratic Institutions. Party Systems in Latin America*, pp. 100 – 137. Stanford: Stanford University Press.

Schafer, Joseph L., and Graham, John W. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods* 7(2): 147–177.

Seltzer, Michael, and Choi, Kilchan. 2002. "Model checking and sensitivity analysis for multilevel models". In *Multilevel modeling: Methodological advances, issues and applications*, Nahiua Duan and Steven P. Reise (eds.). NJ: Lawrence Erlbaum Associates.

Seltzer, Michael, Novak, John, Choi, Kilchan, and Lim, Nelson. 2002. "Sensitivity Analysis for Hierarchical Models Employing t Level-1 Assumptions". *Journal of Educational and Behavioral Statistics* 27(2), 181-222.

Shor, Boris, Bafumi, Joseph, Keele, Luke, and Park, David. 2007. A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data. *Political Analysis* 15: 165–181.

Sigelman, Lee. 1982. The Nonvoting Voter in Voting Research. *American Journal of Political Science* 26:47–56.

Silver, Brian D., Anderson, Barbara A., and Abramson, Paul R. 1986. Who Overreports Voting? *American Political Science Review* 80:613–624.

Spiegelhalter, David J., Best, Nicola G., Carlin, Bradley P., and van der Linde, Angelika. “Bayesian measures of model complexity and fit”. *Journal of the Royal Statistical Society*, Series B, 64 (4): 583-639.

Spiegelhalter, David J., Thomas, Andrew, and Best, Nicky G. 2003. WinBUGS, Version 1.4. User Manual. Cambridge, UK: Medical Research Council Biostatistics Unit, University of Cambridge.

Tanner, Martin A., and Wong, Wing H. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* 83: 528-540.

Thum, Yeow M. 1997. Hierarchical linear Models for Multivariate Outcomes. *Journal of Educational and Behavioral Statistics* 22(1): 77-108.

Thum, Yeow M. 2003. Measuring Progress Toward a Goal. Estimating Teacher Productivity Using a Multivariate Multilevel Model for Value-Added Analysis. *Sociological Methods & Research* 32(2): 153-2007.

Thurigen, Dorothee, Spiegelman, Donna, Blettner, Maria, Heuer, Cartsten, and Brenner, Hermann. 2000. Measurement error correction using validation data: a review of methods and their applicability in case-control studies. *Statistical Methods in Medical Research* 9: 447–474. Tironi, Eugenio, and Agüero, Felipe. 1999. ¿Sobrevivirá el Actual Paisaje Político Chileno? *Estudios Públicos* 74: 151-168.

Torcal, Mariano, and Mainwaring, Scott. 2003. The Political Recrafting of Social Bases of Party Competition: Chile, 1973–95. *British Journal of Political Science* 33: 55-84.

Train, Kenneth. 2003. *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Valenzuela, Julio S. 1995. Orígenes y Transformaciones del Sistema de Partidos en Chile. *Estudios Públicos* 58: 6-77.

Valenzuela, Julio S. 1999. Respuesta a Eugenio Tironi y Felipe Agüero: Reflexiones sobre el presente y el futuro del paisaje político chileno a la luz de su pasado. *Estudios Públicos* 75: 273-290.

Valenzuela, Julio S., and Scully, Timothy R. 1997. Review: Electoral Choices and the Party System in Chile: Continuity and Changes at the Recovery of Democracy. *Comparative Politics* 29(4): 511-527.

Van der Brug, Wouter, and Fenemma, Meindert. 2003. "Protest or mainstream? How the European anti-immigrant parties have developed into two separate groups by 1999". European Journal of Political Research, 42: 55 – 76.

Viana, M. A. 1994. Bayesian small-sample estimation of misclassified multinomial data. *Biometrics* 50(1): 237–243.

Weir, Blair T. 1975. The Distortion of Voter Recall. *American Journal of Political Science* 19:53–62.

Weiss, Robert E. 1994. Residuals and Outliers in Repeated Measures Random Effects Models. Technical Report 161, Department of Biostatistics, University of California at Los Angeles.

West, Mike. 1984. Outlier Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society*, Series B, 46: 431-439.

Wiegmann, Douglas A. 2005. Developing a Methodology for Eliciting Subjective Probability Estimates During Expert Evaluations of Safety Interventions: Application for

Bayesian Belief Networks. Hampton, VA: Final Technical Report AHFD-05-13/NASA-05-4, NASA Langley Research Center.

Wolfinger, Raymond E., and Rosenstone, Steven J. 1980. *Who Votes?* New Haven: Yale University Press.

Verba, Sydney, Norman Nie, and Jae-On Kim. 1978. *Participation and Political Equality: A Seven Nation Comparison*. New York: Cambridge Univ. Press.

Zellner, Arnold. 1917. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

Zhao, Zong. 2008. Sensitivity of Propensity Score Methods to the Specifications. *Economics Letters* 98(3):309–319.