Measurement Issues in the Comparative Manifesto Project Data Set and Effectiveness of Representative Democracy

by

Vyacheslav Mikhaylov

Dissertation

Presented to the

University of Dublin, Trinity College

in fulfillment

of the requirements

for the Degree of

Doctor of Philosophy

University of Dublin, Trinity College

February 2009

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this or any other University. It includes the unpublished work of others, duly acknowledged in the text wherever included. I agree that the Library may lend or copy this thesis upon request.

Vyacheslav Mikhaylov February 2009

Measurement Issues in the Comparative Manifesto Project Data Set and Effectiveness of Representative Democracy

Vyacheslav Mikhaylov

University of Dublin, Trinity College, 2009

In this dissertation I focus on the very widely used Comparative Manifesto Project (CMP) as the source of measurements of the policy positions of political parties. The CMP data confuse the separable notions of party's position on an issue and the importance of the issue to that party. Furthermore, the CMP data are provided without a basic feature considered essential to any estimate: a measure of the uncertainty surrounding the estimated quantity.

This dissertation comprises of three papers. In the first paper, drawing on results from linguistic and behavioural research, I show that party's position on an issue and the importance of the issue to that party are conceptually and empirically distinguishable. I show how to differentiate between position and importance in the CMP data, and contrast this to the saliency-based scaling models currently used by CMP consumers. I evaluate these alternative scales in several replication studies, and propose the use of the existing CMP data that is consistent with the standard spatial models of party competition. The second paper focuses on the analysis of the two main stochastic processes that are involved in the creation of the CMP data: manifesto writing and manifesto coding. Decomposition of the possible stochastic elements in the manifesto generation process that leads to the CMP estimates allows the effects of these to be simulated. Based on these simulation studies, I show how to calculate standard errors for each estimate in the CMP data set. Analysing these error estimates, I show that many CMP quantities should be associated with substantial uncertainty. Next I focus on measurement error arising from stochastic variation in the coding of a given observed text by human coders. I develop a more systematic characterisation of the problems of reliability and bias in the data than has hitherto been attempted. I set out a framework for reliability and misclassification in categorical content analysis, and ap-

ply this framework to the CMP coding scheme. In the third paper, I apply the results of the first two papers to address the question of the effectiveness of democratic representation process. In the paper I focus on one linkage element in the chain of democratic representation: between policy positions of political parties and policy output of governments observed in public spending. Using positional scaling models and correcting for measurement error in the CMP data, I show that most of the positive results previously reported in the literature can be explained by measurement issues (scaling and uncertainty) in the CMP data. Moreover, I show that spending on social security is influenced not only by parties that are elected to government, but also by parties in the opposition, thereby undermining the logical consistency of the responsible party model.

This dissertation makes several contributions. Using statistical and experimental methods, in the first two papers I address the reliability and scaling problems with the CMP data as the result making the data useful for scholars who use these in applied empirical research. This not only makes a valuable contribution to the literature, but also has a practical implication for any user of the time-series cross-section data on policy positions of political parties. The corrections to measurement issues in the CMP data proposed in the first two papers are applied to a concrete political science question in my third paper, where using statistical methods I assess the effectiveness of democratic representation in West European parliamentary democracies. I show that previous results are explained by measurement issues of the CMP data, concluding that the responsible party model cannot be viewed as a valid model of democratic representation in West European parliamentary democracies.

Acknowledgments

This thesis would not have existed without the invaluable support of my supervisor, colleagues, friends and family. I particularly wish to thank my wife Anna for her dedication and support.

Vyacheslav Mikhaylov

University of Dublin, Trinity College

February 2009

Contents

List of Figures

List of Tables

Measurement Issues in the Comparative Manifesto Project Data Set and Effectiveness of Representative Democracy

Any scholar concerned with understanding party competition is interested in measuring the policy positions of political parties. There are many different ways to do this, including but not limited to the analysis of: legislative roll calls; survey data on preferences and perceptions of political elites; survey data on preferences and perceptions of voters; surveys of experts familiar with the political system under investigation; the analysis of political texts generated by political agents of interest. Benoit & Laver (2006) review and evaluate these different approaches.

In this dissertation I focus on one source of such measurements, the long time series of estimated party policy positions generated by the Comparative Manifestos Project (CMP) and first reported in 1987. Over the years since then, the CMP has steadily built up a huge and important dataset on party policy in a large number of countries over the entire post-war period, based on the content analysis of party manifestos. This was reported in the project's core publication, *Mapping Policy Preferences* (Budge, Klingemann, Volkens, Bara & Tanenbaum 2001, hereafter *MPP*), to have covered thousands of policy programs, issued by 288 parties, in 25 countries over the course of 364 elections during the period 1945-1998. The dataset has recently been extended, as reported in the project's most recent publication *Mapping Policy Preferences II* (Klingemann, Volkens, Bara, Budge & McDonald 2006, hereafter *MPP2*) to incorporate 1314 cases generated by 651 parties in 51 countries in the OECD and central and eastern Europe (CEE) over the periods 1990-2003. These data are, commendably, freely available from the CMP and have been very widely used by the profession, as can be seen from over 800 Google Scholar citations by third-party researchers of the core CMP publications.[1]

The range of applications using the CMP data is vast, encompassing four major areas of political science: descriptive analyses of party systems (e.g. Bartolini & Mair 1990, Evans & Norris 1999, Mair 1987, Strom & Lejpart 1989, Webb 2000); empirically grounded analyses of party competition (e.g. Adams 2001, Janda, Harmel, Edens & Goff 1995, Meguid 2005, van der Brug, Fennema & Tillie 2005); models of coalition building and government formation (e.g. Baron 1991, Schofield 1993); and measuring responsiveness of representative democracy in linkages between government programs and governmental policy implementation (e.g. Petry 1991, Petry 1988, Petry 1995). Additional applications of the CMP data include the analysis of American political behavior (e.g. Erikson, Mackuen & Stimson 2002); evaluation of partisan effects on government expenditure (e.g. Bräuninger 2005); identification of structure and dimensionality of the political space of European Parliament, European Commission and Council of Ministers (e.g. Thomson, Boerefijn & Stokman 2004); evaluation of the issue convergence in US presidential campaigns (e.g. Sigelman & Emmett 2004); analysis of the relationship between budgetary cycles and political polarization and transparency (e.g. Alt & Lassen 2006); evaluation of partisan effects on trade policy (e.g. Milner & Judkins 2004); and establishing the effect of endogenous deregulation on productivity

[1] The precise number of third-party citations is hard to calculate because third-party users are likely to cite several CMP sources in the same paper.

in OECD (e.g. Duso & Röller 2003). The CMP data has also been used as a means to validate other measures of parties' policy positions, e.g. expert surveys (Laver, Benoit & Garry 2003, Ray 1999).

Notwithstanding the fact that the CMP project is ostensibly grounded in a "saliency theory" of party competition that assumes "all party programmes endorse the same position, with only minor exceptions" (*MPP*, 82), third-party scholars have overwhelmingly used these data to estimate different party *positions*. Indeed, the CMP itself has used changes in party positions over time, especially on its left-right scale, to validate its own estimates. To a very large extent, the CMP's estimated time series of parties' left-right positions has been the overwhelming attraction of the data set for third-party researchers. For scholars seeking long time series of party policy positions in many different countries, the CMP dataset is effectively the only show in town. Many significant publications have depended on these estimates.

Despite the wide range of researchers who have depended on CMP estimates of party policy positions for their key empirical results, however, these data are based on the core assumption of the CMP that the relative mention of an issue in a manifesto provides a measure of a party's position on that issue. This assumption is derived from the "saliency theory" propagated by the CMP. Saliency theory and the way that the CMP data is typically used, however, confuse the separable notions of party's *position* on an issue and the *importance* of the issue to that party. Furthermore, the CMP data are provided without a basic feature considered essential to any estimate: a measure of the uncertainty surrounding the estimated quantity. For reliable and valid use of the CMP data, such measures of uncertainty are fundamental. Without them, users of the data cannot distinguish between "signal" and "noise", making it impossible to tell the difference between measurement error and "real" movements in party policy positions from one election to another. If we cannot tell whether two CMP estimates differ because of a change in the underlying signal, or because of error in the data—whether from measurement or from fundamental variability—then this drastically undermines

CMP data in terms of their primary value for third-party research: as a rich time series of party policy positions. When covariates measured with error are used in applied empirical research, coefficients on error contaminated variables are biased and inefficient, with the bias usually towards zero ("attenuation bias") (Fuller 1987). Thus, measurement error in the CMP data could potentially prevent findings in areas of research where there are strong theoretical expectations of findings but none could be shown in empirical analyses. One such area is the evaluation of the effect of policy preferences of governments on public spending within the framework of the responsible party model of democratic representation. This research area has been described to contain many publications but still no evidence (King & Laver 1999).

This dissertation comprises of three papers. In the first paper, drawing on results from linguistic and behavioural research, I show that party's position on an issue and the importance of the issue to that party are conceptually and empirically distinguishable. I show how to differentiate between position and importance in the CMP data, and contrast this to the saliency-based scaling models currently used by CMP consumers. I evaluate these alternative scales in several replication studies, and propose the use of the existing CMP data that is consistent with the standard spatial models of party competition. I also suggest how to improve both future coding schemes and the scaling of positions from those schemes.

The second paper focuses on the analysis of the two main stochastic processes that are involved in the creation of the CMP data: manifesto writing and manifesto coding. Decomposition of the possible stochastic elements in the manifesto generation process that leads to the CMP estimates allows the effects of these to be simulated. Based on these simulation studies, I show how to calculate standard errors for each estimate in the CMP data set. Analysing these error estimates, I show that many CMP quantities should be associated with substantial uncertainty. Effects of measurement error in the data are shown in several replication studies. Next I focus on measurement error arising from stochastic variation in the coding of a given observed text by human coders. I develop a more systematic characterisation of the problems of reliability and bias in the data than has hitherto been attempted. I set out a framework for reliability and misclassification in categorical content analysis, and apply this framework to the CMP coding scheme. To come to concrete terms with reliability and misclassification in the context of the CMP, a series of coding experiments on texts for which the CMP has supplied a "correct" coding were designed and carried out. I report on these experimental results, and show that uncertainty due to systematic misclassification has a much more detrimental effect for reliability of the CMP data than measurement error due to stochastic text generation.

In the third paper, I bring the results of the first two papers to address the question of the effectiveness of democratic representation process. The question whether representative democracy actually works is a fundamental question in political science. Defining democracy as a form of government conducted in accordance with people's preferences (Dahl 1971) means that the democratic political system is effective when preferences of voters are translated into specific policy outputs (Hyland 1995). The responsible party model is usually accepted as a valid model of democratic representation in West European parliamentary democracies (Thomassen 1994, 250). The model postulates that popular will is translated into policy via the intermediation of political parties. In the paper I focus on one linkage element in the chain of democratic representation: between policy positions of political parties and policy output of governments observed in public spending. Previous empirical studies produced significant evidence for the linkage effects between policy positions of parties in government estimated using the CMP data and public spending (e.g. Klingemann, Hofferbert & Budge 1994). However, these earlier findings used scaling models that confused position and importance, and did not correct for measurement error in the CMP data. Using positional scaling models and correcting for measurement error in the CMP data, I show that most of the positive results previously reported in the literature can be explained by measurement issues (scaling and uncertainty) in the CMP data. Furthermore, positive findings that remain after correcting for measurement issues in the CMP data raise additional questions about the suitability of the responsible party model for parliamentary democracies. Thus, I show that spending on social security is influenced not only by parties that are elected to government, but also by parties in the opposition, thereby undermining the logical consistency of the responsible party model. One possible explanations is that the popular will does not affect public spending. This raises questions over the effectiveness of democratic representation. Another explanation is that the responsible party model is not a valid reflection of the democratic representation process in West European parliamentary democracies.

This dissertation makes several contributions. In the first two papers I address the reliability and scaling problems with the CMP data as the result making the data useful for scholars who use these in applied empirical research. This not only makes a valuable contribution to the literature, but also has a practical implication for any user of the time-series cross-section data on policy positions of political parties. The corrections to measurement issues in the CMP data proposed in the first two papers are applied to a concrete political science question in my third paper. By using positional scaling models and measurement corrections for the CMP data I assess the effectiveness of democratic representation in West European parliamentary democracies. I show that previously identified linkage between policy positions of governments and public spending can be explained by measurement error effects in the CMP data. Furthermore, contrary to earlier results in the literature, I show that the responsible party model cannot be viewed as a valid model of democratic representation in West European parliamentary democracies.

Substantive issues raised in this dissertation will be further explored in my future research. Thus I plan to work on creating uncertainty estimates for the CMP data that combine *both* measurement error from the stochastic process of text generation and uncertainty from human misclassification. Bringing these two error processes into one probabilistic framework has the potential to produce comprehensive standard error estimates for the CMP data. New standard error estimates would allow revisiting the assessment of the responsible party model and democratic representation.

Overall, this dissertation suggests how to more effectively use the existing rich data resource that is the CMP: clearly differentiating between positional and importance scales, and not confusing them in “saliency” scales currently used by consumers of the data. This dissertation also suggests how to improve a manifesto coding scheme in the future: adopting a clear hierarchical coding structure with each text unit coded as positive, negative or neutral reference to a policy. Finally, this dissertation suggests a new way to look at error variance component in textual data generally.

Chapter 1

Position and Importance in the CMP Data

Abstract

A huge amount of effort in political science has gone into estimating the positions of political parties, taken as the distance of their policy preferences relative to two extremes. More contentious is a related issue concerning the importance or salience of political issues, and how this salience is manifest in party election platforms. The Comparative Manifesto Project (CMP) is based on the assumption that the relative mention of an issue provides a measure of a party's position on that issue, because the "saliency theory" in which it is grounded. Saliency theory and the way that the CMP data is typically used, however, confuse the separable notions of party's *position* on an issue and the *importance* of the issue to that party. In this paper, I argue that these two features are conceptually and empirically distinguishable, drawing on results from linguistic and behavioural research. I also show how to differentiate between position and importance based on the current version of the CMP data, and contrast this to the saliency-based scaling models currently used by CMP consumers, evaluating these alternative scales in a number of replication studies. Based on the comparison of these results to classical CMP models, I offer suggestions for better implementing future manifesto-based coding and scaling schemes.

Key Words: Comparative Manifesto Project, policy position, issue salience, saliency theory, scaling models.

1.1 Positions and salience in the CMP

Any scholar concerned with empirically understanding party competition is interested in measuring the policy positions of political parties. There are many different ways to do this, including but not limited to the analysis of: legislative roll calls; survey data on preferences and perceptions of political elites; survey data on preferences and perceptions of voters; surveys of experts familiar with the political system under investigation; the analysis of political texts generated by political agents of interest.[1] This paper focuses on the estimation of policy positions from content analysis of party manifestos produced by the Comparative Manifesto Project (CMP). The CMP is one of the most extensive data collection exercises in political science. Manual content analysis of more than 3000 manifestos produced a vast data resource for political scientists. The data is readily available in Budge et al. (2001) (hereafter *MPP*) and Klingemann et al. (2006) (hereafter *MPP2*) and cited in hundreds of third-party publications. Some major empirical exercises in the field used the CMP data.

Policy positions represent points on the mathematical construct of dimension. A line segment is an example of a set of points of dimension 1, where the boundary of the interval is a *pair* of points (Courant, Robbins & Stewart 1996, 250). Points on a "dimension 1" are then characterised by the notion of distance relative to the pair of boundary points. Thus, for example, positions of political parties on the one-dimensional "taxes versus spending" scale can be distinctively described by their relative balance of the two extremes: "taxes" and "spending." However, political parties may attach different degrees of importance (or salience) to the "taxes versus spending" dimension: some parties may find it extremely salient for their political platforms while other parties will rank it much lower than, say, environment or immigration dimensions. In other words, parties attach different importance weights to different dimensions. The distances between political parties on a dimension will then be weighted by the importance attributed by the parties to that dimension.[2]

The theoretical basis of the CMP data is set in a "saliency theory" of party competition

[1] Benoit & Laver (2006) review and evaluate these different approaches.

[2] For a discussion of the concepts of position and salience see Benoit & Laver (2006, Ch.1). Differentiation between these two concepts in political competition can be found, for example, in Grofman (2004, 31) or Riker (1996, 101)

(*MPP*, 76). The central idea of saliency theory is that party leaders tend to endorse the majority point of view on each issue, and that as published in their manifestos, "party programmes endorse the same position, with only minor exceptions" (*MPP*, 82). Parties differentiate themselves through emphasising the particular issues on which they have enough reputation to deliver on their promises (*MPP*, 7). The "taking up of positions is done through emphasising the importance of certain policy areas compared to others" (Budge 1994, 455).[3] In other words, because of the assumption that parties occupy the same position on a dimension, the positions of parties on the dimension are characterised not by the notion of distance relative to the pair of extremes, but by differences in salience parties attribute to that dimension over all others.

The CMP approach can be best illustrated by the environmental protection dimension. Setting out the observable implications of parties' policy positions on the environmental dimension as translated in the CMP, the more a party mentions environmental protection, the more pro-environment it is. Conversely, a party that does not mention the environment at all (zero times) is the most possible anti-environment. Figure 1.1 plots references to the environment dimension by main parties in the UK from 1945 to 2001.

[FIGURE 1.1 ABOUT HERE]

According to the saliency theory one can infer from Figure 1.1 that in the postwar period the Liberal Democratic party has been on several occasions potentially the most anti-environment party in the UK. In fact, the Liberal Democrats did not mention environment (zero mentions) in their manifestos on more occasions than the other two main parties combined. At the same time, parties may refer to the environment but make statements that cannot be accepted as entirely pro-environmental. Below is the example from the 1988 electoral manifesto of the Danish Liberal Party.

> Miljøpolitikken måikke stille danske virksomheder dårligere, end virksomhederne i de lande vi konkurrerer med (Venstre 1988).
>
> *The environmental policy should not result in Danish companies being worse off than the companies in the countries with which we compete (Danish Liberal Party manifesto 1988).*[4]

[3] In the saliency theory approach policy dimensions are assumed to comprise of issue areas or clusters of issues (Robertson 1976, 61).

[4] I thank Martin Hansen for drawing attention to this example and for help with the translation.

While this statement relates to industry, it clearly says “Miljøpolitikken” which is translated as “The environmental policy.” As measured by this statement, the Danish Liberal Party is clearly not pro-environment, preferring instead to let the natural environment suffer in exchange for the economic benefits that presumably come from easing restrictive environmental regulations on commercial firms. At the moment this cannot be verified conclusively, since based on the CMP assumption that no party will publicly express an anti-environmental position, there is only a pro-environment category (PER501) included in the CMP coding scheme. This precludes the estimation of parties’ positions on this dimension, because the positions of political parties on a dimension can be distinctively described only by their relative distances to two extremes. Any future development of the CMP coding scheme must include a possibility for the parties to take positive and negative positions on each issue.

1.2 Estimating party positions from the existing CMP data

Once the notion that political parties can take up positions only on one side of an issue is abandoned, positions of political parties on the dimensions of interest can still be easily estimated using the existing CMP data. This can be done as the relative balance of the positive and negative stances on an issue. However, the CMP suggests that the balance of positive and negative issues should be compared with the whole manifesto length.

The mechanics of the process can be best illustrated again using the simplest example of the environment dimension. Following the CMP assumption that there are no anti-environment references, the position of a party on the environment dimension is the ratio of the number of positive references to environmental protection (PER501) and the total manifesto length. Figure 1.2 tracks change in the position of the German Green party on the environment dimension, also indicating the total number of text units in each manifesto.

[FIGURE 1.2 ABOUT HERE]

Figure 1.2 shows that in order to keep the position on the dimension constant over time in the face of changing manifesto lengths, the Greens would have to proportionally change the number of references to all other issues.

The situation is even more complicated with combined scales created over a number of issues. This can be illustrated with the widely used `rile` left-right scale created by the CMP and supplied with the data distribution. The scale is created by subtracting the number of text units referring to 13 "left" issues from the number of text units referring to 13 "right" issues relative to the total number of text units in a manifesto:[5]

$$rile_{saliency} = \frac{R}{N} - \frac{L}{N} \tag{1.1}$$

By making the scale dependent on the total number of text units in a manifesto, `rile` scale implies that a position on the left-right dimension depends on all other dimensions in a manifesto. Such scaling method is defended as being "consistent with saliency theory" (*MPP*, 23). The salience and the position of the party on an issue is thus measured as "the relative saliency given to them in the manifestos" (*MPP*, 82). In turn, "relative saliency" is operationalised as the frequency of text units allocated to an issue relative to the total number of text units in the manifesto.

1.2.1 Repetition and importance

The assumption that relative frequency of references to an issue signifies relative salience of that issue is, in turn, based on the assumption that repetition increases the strength of a message. Repetition is said to be the hallmark of party manifestos: "making policy points involves highlighting them, repeating them in slightly varied form and coming back to them in a variety of contexts" (Budge 2001, 211).[6]

The stress by the CMP on the function of repetition and the frequencies as a measure of importance is directly related to Skinner's (1957) verbal behaviour research in psychology. He stipulated that in communication the strength of a transmitted message is based on response speed, pitch level, immediate repetition, and overall frequency. Evidence for contribution of

[5] The scales are usually multiplied by 100 to present as percentages. For details on the issues that constitute "left" and "right" see Budge et al. (2001, Ch.1).

[6] See Thomson (1999, 88–91) for a discussion of repetition as an indicator of importance in the analysis of manifesto pledges.

each individual input into the strength of the message "is based on observation of frequencies alone" (Skinner 1957, 28). Skinner gives a well known example of a person exclaiming *Beautiful!* when observing a famous work of art: "the speed and energy of the response will not be lost on the owner" (Skinner 1957, 27). However, the importance of frequency has been challenged by several scholars. Chomsky (1959) took up the example of the painting and suggested that following Skinner "to shriek *Beautiful* in a loud, high-pitched voice, repeatedly, and with no delay" (Chomsky 1959, 35) would result in a strong message, with the increase in the importance of the message achieved by training "machine guns on large crowds of people who have been instructed to shout it" (Chomsky 1959, 35). This may not be the best way to convey the importance of the painting to the owner. In fact, an equally effective strategy may be to stare at the painting for a long time in silence, and softly murmur *Beautiful* (Chomsky 1959, 35).

The assumed effect of repetition (and its realisation in the CMP saliency-based scales) has also been challenged in learning theories. There the effect of increasing the importance of a message as a function of repetition is known as the semantic generation (Jakobovits 1967, Bäuml 2002). However, repetition is also known to induce semantic satiation. This is a loss of meaning of a word as a function of repetition (Black 2003, 63). Thus, both semantic generation and semantic satiation can transpire in the same text as functions of repetition. In such a situation the relationship between two effects may be governed by a "frequency law" (Jakobovits 1967). The law states that the relationship between the intensity of response and the frequency of exposure follows an inverted U-shaped distribution. An increase in meaning through repetition happens at the semantic generation stage, while the meaning is lost through continued repetition in the semantic satiation stage (Jakobovits & Lambert 1963, Jakobovits & Hogenraad 1967). The exact shape of the frequency curve and identification of the inflection point depend on individual circumstances (Jakobovits 1967), and can be identified in experimental settings.[7]

In content analysis of party manifestos, the CMP assumes that the saliency-based scales

[7] Semantic satiation is notoriously difficult to measure (Black 2003, Esposito & Pelton 1971). Recent experimental evidence, however, attests to the existence of the effect (Black 2001, Kounios, Kotz & Holcomb 2000, Kounios 2007, Pynte 1991).

imply only the semantic generation stage, while the semantic satiation stage is not even considered in the creation of the scales. The semantic satiation here would mean that if parties repeat references to an issue this may (un)intentionally result in the effect opposite to the one assumed by the CMP. Instead of raising the importance of the issue, parties will effectively reduce its importance when it becomes less meaningful as a function of frequency. Even if repetition does not reduce the effectiveness of the message, it may not add additional salience to the position. In other words, there is only one way to state that a party rejects the Euro.

This can be illustrated with the example of the UK Conservative party and its stance on the adoption of the Euro. There are 17 sentences devoted to the issue of joining the single currency in the 1997 Conservative manifesto (1.57% of the manifesto). The manifesto for the 2001 general elections contains only 6 sentences devoted to this issue (0.83% of the manifesto). The size of the Conservative manifesto shrunk from 1084 text units in 1997 to 724 text units in 2001. Following saliency theory the importance of single currency issue decreased from 1997 to 2001. Moreover, invoking the Skinnerian approach and semantic generation effect, the adjustment in the importance of the issue through the drop in repetitions of Euro related statements leads to the weakening of the overall Conservative message on single currency. Table 1.1 presents the actual sections of these two manifestos related to the Euro.

[TABLE 1.1 ABOUT HERE]

From Table 1.1 it appears that the Conservatives in 1997 talked relatively much about single currency without saying anything. The electorate took it as an ambivalent position on the Euro (Evans 2002). A position that left the Conservative party some leeway in policy making were they to win elections. By the time of the 2001 general election, the mood of the electorate was pointedly more eurosceptic (Evans 2002). Opening up a new dimension of political contestation with the single currency issue, the Conservatives forcefully and unequivocally stated their opposition to the adoption of the Euro. This example shows that, an increase in relative frequency of the message does not necessarily result in corresponding increase in the strength of the message. Furthermore, a change in relative frequency may not signal change in the position on the single currency issue, but it may still reflect change in the importance of that issue for the Conservative party in 2001 compared to 1997. Thus it is important to clearly disambiguate position on an issue and the salience of this issue for a party.

1.2.2 Positional scales

Position on an issue can be easily distinguished from the importance of this issue for a party using the existing CMP data. This can be achieved by reflecting the position through the distance relative to the pair of two extremes on the dimension. At the same time, the position on the issue cannot depend on the document size, with the positional scale independent from the issues that are accidental or irrelevant to the analysed dimension (Krippendorff 2004, 181). The standard solution to this problem is a distance function proposed in the literature on content analysis of political texts in Krippendorff (1967) (see Krippendorff 2004, 176). For `rile` left-right dimension, the positional distance function can be constructed as the following scaling model:

$$rile_{position} = \frac{R-L}{R+L} \tag{1.2}$$

This positional scaling model has been proposed independently for the CMP data in Kim & Fording (1998) and general political textual data in Laver & Garry (2000). A simple example can illustrate the difference between the saliency-based `rile` scale (Equation 1.1) and positional scale in Equation 1.2. Take a manifesto of 200 text units that contains 100 references to "left" items and 40 references to "right" items. Position of the party on the saliency-based left-right scale is then $\texttt{rile}_{saliency} = (\frac{40}{200} - \frac{100}{200}) \times 100 = -30$. Using the positional scale the party can be placed at $\texttt{rile}_{position} = \frac{40-100}{40+100} \times 100 \approx -42.86$. At the next election the party decides to expand the section of its manifesto devoted to, say, the EU by additional 200 text units leaving the rest of the text unchanged. In the new manifesto there are 400 text units of which 100 refer to "left" issues and 40 refer to "right". However, 200 text units now refer to the EU that is not part of `rile`. New position of the party on the saliency-based left-right dimension is $\texttt{rile}_{saliency} = (\frac{40}{400} - \frac{100}{400}) \times 100 = -15$. Using the positional scaling model the party is still located at $\texttt{rile}_{position} = \frac{40-100}{40+100} \times 100 \approx -42.86$. Thus, the saliency-based `rile` scaling model shows that the party moved to the centre (from -30 to -15) as the result of devoting more text to issues that are not part of the `rile` scale.

Krippendorff's (1967) distance function can also be applied to estimate the importance of a dimension for a party. It was shown in previous section that the importance of an issue may relate to the frequency of repetition of the issue in a manifesto relative to the total manifesto size. Thus the importance of the `rile` left-right dimension can be expressed as:

$$rile_{importance} = \frac{R}{N} + \frac{L}{N} \tag{1.3}$$

The importance scaling model for the CMP data has been proposed in Benoit & Laver (2007*b*). Positional and importance scaling allow disambiguation of the separate concepts that are currently mangled by the CMP into one saliency-based scale. This can be illustrated reverting to the simple example of the position of the German Green party on the environment dimension. Current CMP category construction does not allow us to identify the two extremes that characterise the position of the party. However, one can find the hint in the above quoted statement from the 1988 manifesto of the Danish Liberal Party. Thus, it is easy to construct a scale capturing a more general environment policy dimension that represents the trade-off between environmental protection and economic growth.[8] The paradigm of economic growth is represented in the CMP by category "Productivity:Positive" (PER410), while categories "Anti-Growth Economy:Positive" (PER416) and "Environmental Protection:Positive" (PER501) together capture anti-growth politics, "ecologism", and "green" politics in general.[9] Thus, the importance of the environment dimension is captured by the following scaling model:

$$Environment_{importance} = \frac{PER501}{N} + \frac{PER416}{N} + \frac{PER410}{N}$$

The position of the Green party on the environment dimension is represented by the following scaling model:

$$Environment_{position} = \frac{(PER501 + PER416) - PER410}{PER501 + PER416 + PER410}$$

[8] This correlates with the definition of one of the core four dimensions in the expert survey in Benoit & Laver (2006, 129).

[9] For full category definitions see Klingemann et al. (2006, Appendix II).

This allows the measurement of the distinct concepts of party position on a dimension and the importance of that dimension for a party. Figure 1.3 tracks the position of the German Green party on the Environment Dimension, alongside tracking changes in the importance of this dimension for the party.

[FIGURE 1.3 ABOUT HERE]

Figure 1.3 shows that the position of the Greens on the Environment Dimension remained stable after the 1987 election. However, the importance of the dimension changed over time, reflecting, among other things, changes in strategic positioning of the party in the run up to the 1998 election and coalition negotiations, and subsequently facing re-election in 2002 while in government. Comparing to the results from the saliency-based scale presented earlier (Figure 1.2), it is clear that the saliency-based measure tracks importance fairly well, but doesn't capture distinct position of the Green party. Next section conducts a more general comparison of the positional and saliency-based scaling models for other standard dimensions like the EU integration, economic left-right, social liberal-conservative, and `rile` scales.

1.2.3 Comparing positional and saliency-based scales

Considering the `rile` it has been earlier proclaimed that the positional scaling model (Equation 1.2) and the saliency-based scaling model (Equation 1.1) are "nearly identical in empirical terms" and distinguishable only on philosophical grounds (Kim & Fording 2002, 200, fn 5). This conclusion was drawn from a near perfect correlation between the two scales. However, using the Pearson's product-moment correlation coefficient to measure agreement between two scales is invalid (Altman & Bland 1983, Bland & Altman 1986). Bland and Altman (1983,1986) proposed to plot the difference between two scales against their average.[10] Lack of agreement is summarised by calculating the bias, estimated by the mean difference between two scales, d, and the standard deviation of the differences, s. Most of the differences are expected to lie within 95% limits of agreement, calculated as $d \pm 1.96 \times s$.[11] Figure 1.4 presents

[10] Bland & Altman (1995) discuss the reasons why the difference should be plotted against the average, and not against one of the scales that is taken as a standard.

[11] Bland-Altman approach is widely used in medical statistics to compare two alternative measurement techniques. Combined citation count on Google Scholar for Altman & Bland (1983) and Bland & Altman (1986) is over 14000, giving some indication of the standard-like status of the approach.

the results for the `rile` scale.

[FIGURE 1.4 about here]

Figure 1.4 points to the increase in variability, shown by the increase in the scatter of the differences, as the magnitude of the measurement increases. The bias in the measurement is shown by the tendency for the mean difference to rise with the increase in value of average positional and saliency-based scales. Figure 1.4 also shows the presence of a clear trend in the bias, indicated by the positive slope of the regression line. The presence of either the bias or the trend identifies that the methods do not agree equally through the range. It appears that the two scales agree on the location of the centrist parties, but disagree on the placement of non-centrist parties. The `rile` combines 26 out of 56 issue categories in the CMP, which is more than the median number of categories used to code manifestos (Benoit, Laver & Mikhaylov 2009). Disagreement and bias between the positional and saliency-based scales should be more severe for scales combining smaller number of categories. Figure 1.5 presents the Bland-Altman approach for the EU dimension scale, that consists only of two categories: pro-EU (PER108) and anti-EU (PER110).

[FIGURE 1.5 about here]

Once again, bias and trend are shown in Figure 1.5. Contrary to the results for `rile` most of the agreement is shown for the parties that are very pro-European. There is substantial observable bias in the saliency-based scale towards parties that take more moderate position on the EU and also those that are extremely anti-EU.[12]

Agreement between the positional and saliency-based scales can also be directly measured using the chance-corrected concordance correlation, often referred to as Lin's concordance correlation (Krippendorff 1970, Lin 1989, Lin 2000). Lin's concordance correlation combines measures of precision and accuracy to determine how close the two scales are to the line of perfect concordance.[13] Precision is measured by Pearson's product-moment correlation. Accuracy is captured by the bias correction factor that measures how far the best-fit line deviates from the perfect agreement line. Bias correction factor is the ratio of concor-

[12] See also external validation results for the positional and saliency-based EU dimension scales in Ray (2007), albeit all conducted relying on invalid product-moment correlations.

[13] For details and applications of the measure see e.g. Cox (2006).

dance correlation coefficient and Pearson's product-moment correlation coefficient, with the range (0,1]. It reaches maximum value of 1 when there is no deviation from the line of perfect concordance, and further away from 1 means less accuracy (more bias). Table 1.2 presents the results of Lin's concordance assessment of the positional and saliency-based scales for four most widely used dimensional scaling models: `rile`, Economic left-right, Social liberal-conservative dimension,[14] and EU dimension.

[TABLE 1.2 ABOUT HERE]

The results in Table 1.2 clearly indicate that high product-moment correlation coefficient between positional and saliency-based scales does not translate into high agreement. The highest concordance is found for the `rile` scale. However, its narrow 95% confidence is significantly removed from the line of absolute agreement. As suggested earlier, the concordance is much worse for the scales that consist of smaller number of categories compared to `rile`. The situation is really dire for the EU dimension scale. Despite moderately high Pearson's correlation of 0.597, its concordance correlation coefficient is only 0.029 (with narrow 95% confidence interval). Moreover, the bias correction factor for the EU dimension is close to zero, thus pointing to very high bias (low accuracy).

These results suggest that positional and saliency-based scales are not interchangeable, with substantial amounts of bias present. This is especially true for scales that consist of the smaller number of categories compared to `rile`.

1.2.4 Additional scaling models

In addition to the scaling discussed in previous sectons, a powerful feature of the CMP data is the possibility to create scales for other dimensions. This "Lego"-like feature has been used to create the environment dimension in previous section. Once extremes have been identified the positional or importance scales can be applied to place parties on the dimension of interest. Unfortunately, at the inception, the CMP refused to create coding categories that reflect pairs of reference points (extremes) for all policy issues. The CMP currently includes only 12 clearly bi-polar categories.

[14] See Benoit & Laver (2007*b*, 100) for details on constructing Economic left-right and Social liberal-conservative dimensions.

Dominance of uni-polar categories has been explained as driven by the saliency theory of party competition, and evidence showing that "the negative side attracts such few endorsements that the codings overall can be effectively taken as one-dimensional" (*MPP*, 83). This conclusion is based on aggregate results for a sample of 24 developed democracies (*MPP*, 83). At the same time, saliency theory is proposed as an explanation of party competition in individual countries (Laver 2001). Hence, a more appropriate evaluation should be conducted also on the level of party systems and not on the aggregate results for a sample of countries. The results of such country level analysis significantly differ from the CMP results. The EU dimension is taken here as an illustration. Figure 1.6 shows the distribution of party positions on the EU dimension in three countries: Denmark, Ireland, and the UK.

[FIGURE 1.6 ABOUT HERE]

The results in Figure 1.6 clearly indicate that the saliency-based and positional scales tell different stories. First, the saliency-based scale shows parties in all three political systems inhabiting the centre of the political space. At the same time, the positional scales show a more fine grained picture of party competition on the EU dimension. The median position of parties on the EU dimension in Ireland and the UK are predominantly pro-European. However, both systems exhibit significant number of parties expressing anti-EU sentiment. In Denmark there appears to be a significant bi-polar distribution of party positions on the EU dimension: the result that is not captured at all on the saliency-based scales.

This exercise highlights two things. One is that there is significant variation in party positions on the bi-polar categories as exemplified by the EU dimension. However, this becomes visible only when evaluated on a country by country basis. Thus, the claim that categories can be taken as unidimensional is unwarranted. Secondly, in addition to the issues with the saliency-based scales raised in previous section, its use also greatly simplifies the picture of party competition. One of the results is the loss of much valuable information about the distribution of party positions on policy dimensions, and general polarisation of party systems. However, any user of the CMP data can still uncover this information using the positional scaling model.

Results presented in this paper so far suggest that the saliency-based scales depend on the size of the manifesto and lump together position on the dimension and importance of the dimension. Positional and saliency-based scales are not interchangeable, with substantial amounts of bias present. This is especially true for scales that consist of smaller number of categories compared to `rile`. Positional scales depict the richness of the information on political competition, information that is simply lost when using the saliency-based scales. All these suggests that empirical results in applied research will be significantly influenced by the choice between the positional scales and the saliency-based scales. The next section presents results of two replication studies using both the positional and saliency-based scales that highlight statistical and substantive differences in empirical results under the two scales.

1.3 Positional and saliency-based scales in empirical analysis

Two recent high-profile studies are replicated here. Both studies utilised saliency-based scales in empirical application: Hix, Noury & Roland (2006) and Golder (2006). In both cases original datasets (and replication code) were made available by the authors and replicated using the positional scales. Hix, Noury & Roland (2006) employ several scales derived from the CMP data: `rile`, Economic left-right, Social liberal-conservative, and the EU dimension scales. This allows direct evaluation of the effect of using positional rather than saliency-based scaling model. Golder (2006) derives key explanatory variables using the saliency-based `rile` scale. This replication study allows evaluation of indirect effects of the scale choice. The aim of these replication studies is not to overturn some of the existing results, but rather to show the existence of tangible effects in using positional versus saliency-based scaling models in empirical applications.

1.3.1 Hix, Noury, and Roland (2006)

Hix, Noury & Roland (2006) are concerned with the content and character of political dimensions in the European Parliament (EP). Following an inductive scaling of roll-call votes in the EP from 1979 and 2001, Hix, Noury & Roland (2006) set out to validate their interpretation of the derived policy dimensions by regressing the mean position of each national party's delegation of MEPs on two sets of independent variables. The first set includes exogenous measures of national party positions on the `rile` left-right, Economic left-right, Social liberal-conservative, and EU dimensions. The second set relates to government-opposition dynamics and consists of categorical variables describing whether a national party was in government and whether the party had a European Commissioner, as well as dummy variables for each European party group, each EU member state, and each (session of) European Parliament. Measures of national party positions are derived from the saliency-based scaling of the CMP data, and discussed in previous section. The authors expect that national party ideal point estimates on the first dimension will be explained by the exogenous left-right policy positions, while exogenous policy positions on EU dimension explain national party ideal point estimates on the second dimension (Hix, Noury & Roland 2006, 501). The expectation then is roughly that the first dimension is predominantly about left-right and second dimension is about Europe.

[TABLE 1.3 about here]

Table 1.3 contrasts coefficients from the replications of the models using saliency-based scaling to construct variables in Hix, Noury & Roland (2006) with positional scaling introduced in previous section. (Due to space constraints replication results presented here focus only on the two models that relate to the structure of the second dimension in the European Parliament.) Model 2 aims to explain the mean positioning of political parties on the second derived EP dimension in terms of: their positions on the `rile` left-right and the European integration dimensions; categorical variables relating to whether a party was in government and had a European Commissioner; and dummy variables for each session of the EP. Model 3 extends Model 2 by replacing positions of political parties on the `rile` left-right with their positions on the Economic left-right and Social liberal-conservative dimensions. Thus, Models 2 & 3 utilise the saliency-based scaling for party positions on all widely used dimensions discussed in previous section.[15]

It is clear from Table 1.3 that the results with positional scaling are generally statistically

[15] Here I depart from original estimation using newer CMP data set made available with the publication of *MPP2*. This resulted in some very slight differences in the estimation data set. The change allowed for some results to be more pronounced in the replications than in the original.

stronger, although the coefficients are smaller in size. Differences in the magnitude of the effects seem to correspond to differences in concordance between scales across dimensions as discussed in previous section. The biggest difference in effect appears for the EU dimension, where the positional scale result is about ten times smaller than the result using the saliency-based scale. At the same time, the coefficient on EU dimension becomes statistically stronger under positional scaling in Model 2, while in Model 3 it gains statistical significance.

Substantively, the effect of using the positional instead of saliency-based scale is that the explanation of the position of a party's MEP delegation on the second dimension can be effectively done using the national party's position on the EU dimension. In turn, position on the EU dimension is more important than party's positions on either `rile` left-right or the substantive Economic left-right and Social liberal-conservative dimensions. The effects using the positional scaling are generally smaller in size, but more statistically robust.

1.3.2 Golder 2006

Amid extensive existing research on government coalitions Golder (2006) focuses on a largely ignored issue of pre-electoral coalitions in parliamentary democracies. She develops a theory of pre-electoral coalition formation and tests the theory using a data set of all potential pre-electoral coalition dyads in twenty industrialised parliamentary democracies from 1946 to 1998. In the 292 elections studied in the article, 44 per cent contained at least one pre-electoral coalition, while about a quarter of governments formed after the elections were the result of pre-electoral coalitions (Golder 2006, 194). Despite the importance of the topic, prior to Golder's article there appear to be no serious attempts to theoretically and empirically analyse factors influencing the formation of pre-electoral coalitions (195). The author shows that pre-electoral coalitions are more likely to form between ideologically compatible, similarly sized parties in party systems characterised by ideological polarisation and disproportional electoral rules.

Golder (2006, 198) argues that the utility loss associated with policy set at coalition's ideal point rather than party's ideal point is minimised when coalition partners are ideologically similar. Thus, a decision by multiple parties to co-ordinate their electoral strategies rather than contest seats alone depends on ideological distance between potential coalition partners.

Parties are also likely to form an electoral coalition if this is the best way to keep a relatively 'extreme' government from forming. More disproportionate electoral system creates an incentive for smaller parties to create pre-electoral coalitions, particularly so in more polarised party systems. Probability of forming an electoral coalition between parties that are asymmetric in size is less likely when the overall coalition size is sufficiently large.

Probability of pre-electoral coalitions between dyads of parties in a system is modelled as a function of *Ideological incompatibility*, *Polarisation* and *Electoral threshold* (plus the interaction between these two variables), *Coalition size* and *Coalition size squared*, *Asymmetry*, and an interaction between *Coalition size* and *Asymmetry*. Three variables are built from the saliency-based `rile` left-right scale: *Ideological incompatibility*, *Polarisation*, and *Polarisation* × *Electoral threshold*. This allows checking the indirect effect of positional and saliency-based scales in applied research.

Ideological incompatibility measures the ideological distance between the parties in the dyad. It is intended as a proxy for the lack of ideological compatibility of parties in a coalition. The variable is directly computed as the absolute value of the difference of the saliency-based `rile` left-right score for parties in the dyad. *Polarisation* measures ideological dispersion in a system. Parties are concerned about a potential government consisting of parties more 'extreme' relative to them. For example, centrist parties may be worried about the prospects of communists and other extreme-left parties forming a government. In such circumstances, parties are primarily concerned with the ideological positions taken by other parties. The variable is calculated as the difference of the saliency-based `rile` left-right scores for the biggest left-wing and the biggest right-wing parties in political system. Party system polarisation is hypothesised to increase the likelihood of pre-electoral coalition formation when disproportionality of the electoral system (measured as the effective *Electoral threshold*) is sufficiently high. Positive effect of electoral system disproportionality is stipulated to be even stronger in more polarised party systems, which is modelled through an interaction term *Polarisation* × *Electoral threshold*.

For replication of the analyses both CMP-derived variables are recreated using positional `rile` left-right scale. Golder (2006) tests her hypotheses using a probit model. The author estimates two specifications: random effects probit (Probit 1) and a probit model with robust standard error (Probit 2). Golder suggests that theoretical reasons and statistical tests point to random effects probit as the preferred estimation approach. The interpretation of the findings in the article is provided based on the results from Probit 1.

[TABLE 1.4 about here]

Table 1.4 contrasts results from probit model predicting the propensity of pre-electoral coalition formation based on saliency-based and positional left-right scales.[16] For three key CMP-derived variables (incompatibility, polarisation and polarisation × electoral threshold) noticeable changes occur when the models are re-estimated with variables derived from the positional scale. Coefficient on *Incompatibility* shrinks in size while remaining statistically significant and with the correct sign. As the result of using the positional scale, the coefficient for *Polarisation* changes sign from positive to negative, which brings it in line with the original theoretical expectation, albeit it remains statistically insignificant.

More importantly, *Polarisation* × *Electoral threshold* becomes statistically significant, supporting the original argument that creation of pre-electoral coalitions is more likely under higher electoral thresholds and polarisation. Substantive interpretation of the interaction term is more complex and Golder examines the marginal effect of each variable in the interaction model on probability of coalition formation graphically. Limited to variables of interest, the graphical analysis is replicated here for the two alternative scales in Figure 1.7.

[FIGURE 1.7 about here]

Figure 1.7 presents the marginal effect of 0.01 unit (corresponds to 1 unit on the original scale) increase in party system polarisation across the observed range of electoral system disproportionality (with all other variables held at their means). Solid lines indicate how the marginal effect changes with the effective threshold. Grey line represents estimation based on saliency-based scale, while black line refers to estimates based on positional scale. Dashed lines represent respective 95 per cent confidence intervals that show conditions under which polarisation has statistically significant effect on the likelihood of electoral coalition forma-

[16] This paper departs from original estimation in two ways here. First, a newer CMP data set made available with the publication of *MPP2* is used here. This resulted in some very slight differences in the estimation data set. Second, saliency-based left-right is rescaled here to range from -1 to 1. Neither change resulted in any substantive differences in the replications versus the original results.

tion.[17] The marginal effect is deemed statistically significant whenever both lower and upper bounds of the confidence interval are above (or below) the zero line.

Figure 1.7 indicates that on the saliency-based scale party system polarisation makes pre-electoral coalitions more likely when the electoral threshold becomes greater than 12. At the same time, estimates based on the positional scale put the cut-off figure for electoral threshold at 21, after which polarisation significantly affects the probability of coalition formation. Following Golder's substantive interpretation, when using the saliency-based scale 18.3 per cent of the sample have an electoral threshold greater than 12, while only 10 per cent greater than 21. In other words, an increase in party system polarisation is expected to increase the probability of pre-electoral coalition formation in just under a fifth of observed cases using saliency-based scale and in a tenth using positional scale.

The results of both replication studies show that two scales cannot be taken as equivalent in applied research. The results produced under saliency-based and positional scales are often not only statistically but also substantively different. Corresponding to the results in the previous section, the biggest difference has been found for scales comprised of small number of categories, like EU integration. Moreover, statistically and substantively different results have been found even in applications where the two scales are used to derive measures of interest (e.g., party system polarisation) and do not enter estimation models directly. Variables measuring some characteristics of political competition perform differently in empirical applications depending on the underlying scaling model.

1.4 Discussion and conclusion

The conclusions from questioning of the theoretical and empirical underpinnings of the basic design of the CMP approach to coding and scaling policy preferences from party manifestos can be summarised as follows.

First, the assumption that parties only take one side of an issue, or that "the negative

[17] As in the original article, confidence intervals are based on simulations using 10000 draws from the estimated coefficient vector and variance-covariance matrix for two probit models based on saliency-based and positional left-right scales.

side attracts such few endorsements that the codings overall can be effectively taken as one-dimensional" (*MPP*, 83), is not only wrong generally, but also demonstrably wrong given the CMP's own dataset. Consequently, policy scales should contain both positive and negative references, thus presenting two contrasting extremes of position on policy dimensions. Where issues do not appear to have a simple pro and contra side, like environmental protection, then the two extremes should be phrased in terms of relative preferences for mutually incompatible and competing policy goals, such as environmental protection versus unrestrained economic growth.

Second, the emphasis of a policy issue relative to all other manifesto statements does not measure position, but instead provides a rough indicator of the relative importance of a political issue to the party issuing the manifesto. Policy position can be considered as the difference between support for competing extremes, relative to all statements on the relative issues only. I have have termed this approach the *positional scale* and contrasted it to the CMP's *saliency scale* approach. I show not only that the two agree poorly, especially for scales with few constituent categories, but also that the saliency scale is biased because it underestimates the extremity of party positions at the extremes of the scale. The positional scaling approach is also much more in accord with standard approaches to scaling from the literature on relative frequency-based content analysis. This scale has been previously offered in the literature, but the argument made here is much stronger than has been made previously, since I advocate that it be used in every context for estimating position from manifesto content, for every issue dimension.

Third, position and importance are conceptually and practically distinct aspects of the ways that parties approach and communicate policy preferences and priorities. Accordingly, they should be measured in distinct ways. Here I have proposed an importance scale that captures the emphasis of an issue—where emphasis refers to all statements about the issue, whether positive or negative—relative to the entire manifesto.

My proposal for replacing the CMP's saliency scaling approach with a net positional one can also be viewed as a critique of the basic CMP coding scheme, since the existing scheme consists of a mixture of positional and saliency-based categories. My analysis suggests that any revision of the coding scheme would complete the step toward a fully positional coding scheme, consisting only of opposing, pro and contra categories. Of course, it would be possible to go one step further, and also to include a neutral for each confrontational policy scale, which could be ignored in the numerator of the net positional scale but counted in the denominator. This would address the concerns of McDonald & Mendes (2001*b*) about the non-reflection of neutral stances in the positional scales. In addition, the inclusion of neutral stances in the denominator of the positional scales could mitigate the bipolarity that sometimes occurs when using the net positional scale. Finally, even with a fully positional coding scheme, including one that also had neutral positions, the same information could be used to estimate the proposed "importance" measure.

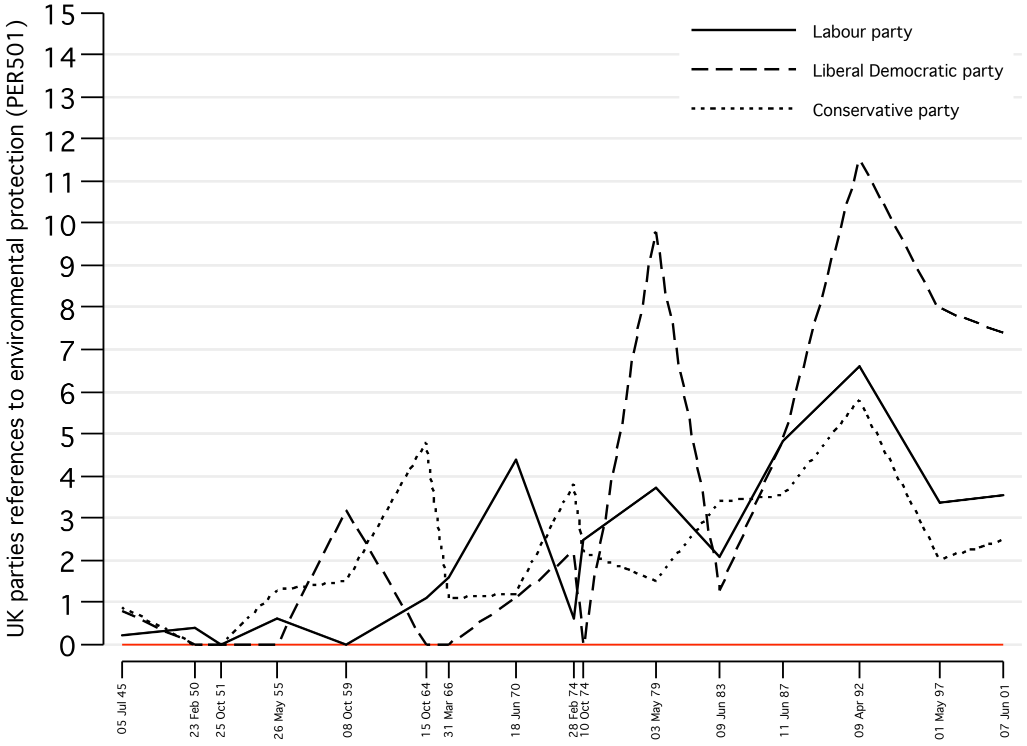


Figure 1.1: *Policy positions of main UK parties on environment dimension (PER501) over time.*

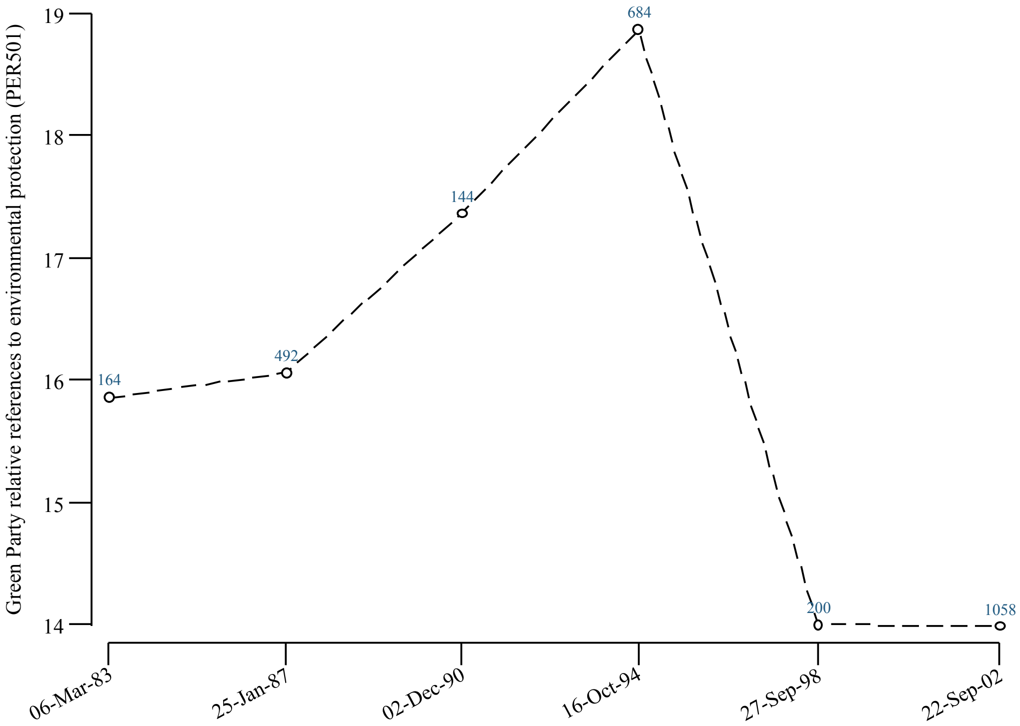


Figure 1.2: *Change in relative frequency of positive references to environmental protection (PER501) by German Green party over time.* Total number of text units in a manifesto is indicated next to the marker for relative frequency of references.

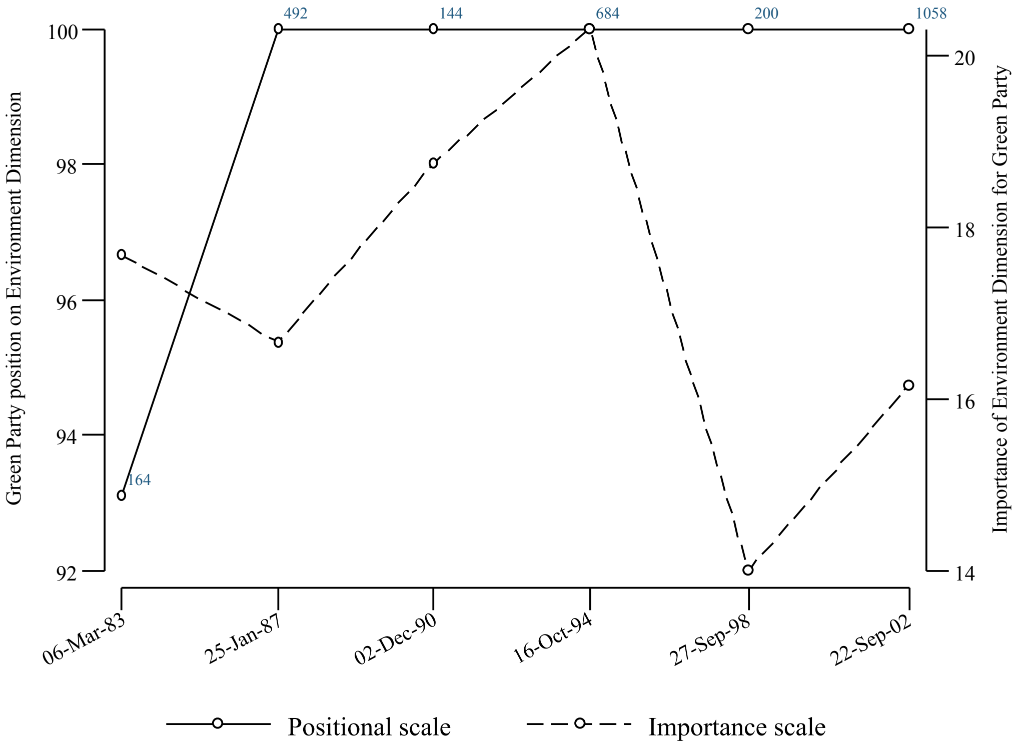


Figure 1.3: *Environmental Dimension Position and Importance scales for German Green.* Solid line tracks position of the Green party on Environment Dimension over time. Dashed line tracks the importance of Environment Dimension for the Greens, with the attributed measurements presented on the second y-axis. Total number of text units is also presented for each manifesto.

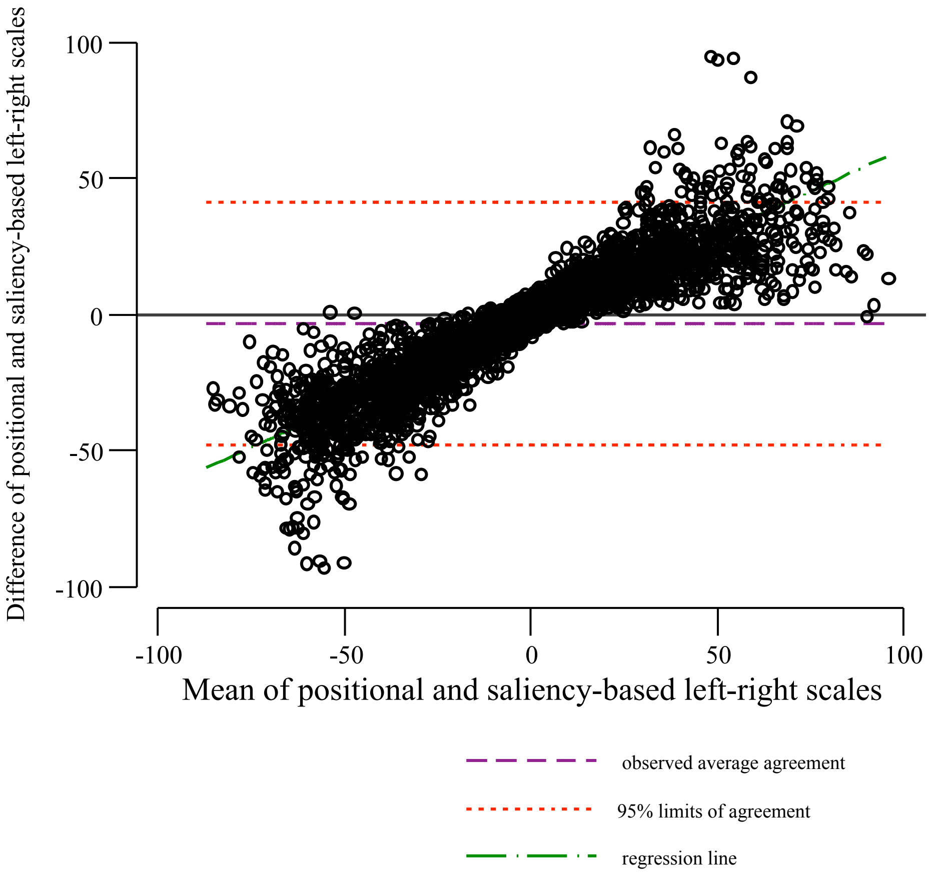


Figure 1.4: *Bland-Altman plot comparing positional and saliency-based left-right scales.* The standard Bland-Altman plot is between the difference of paired scales versus their average. It includes 95% limits of agreement around observed average agreement, and a line for perfect average agreement ($y = 0$ line). The regression line is also plotted to show the trend.

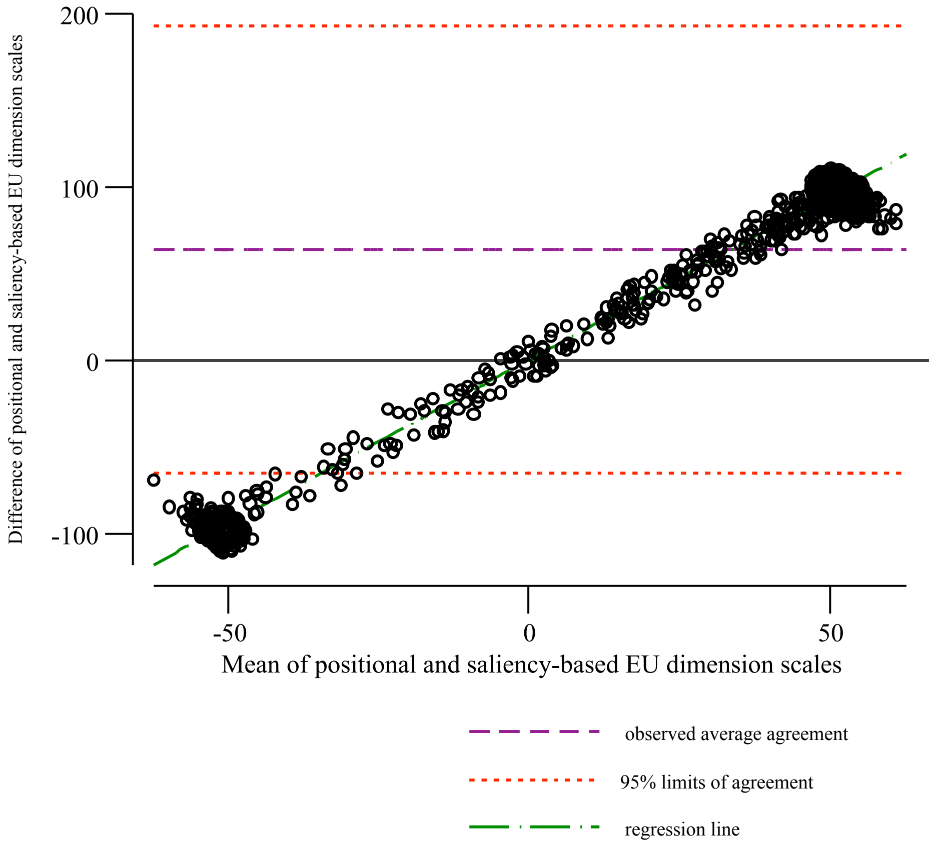
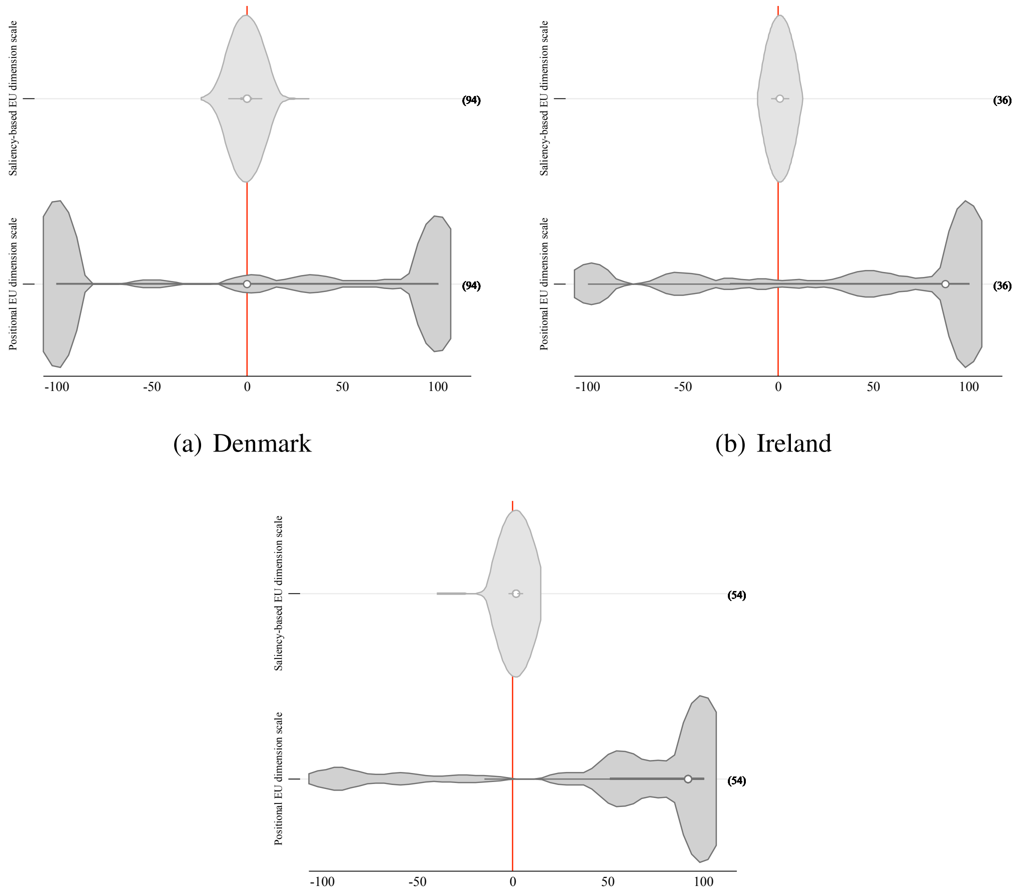


Figure 1.5: *Bland-Altman plot comparing positional and saliency-based European Dimension scales.* The standard Bland-Altman plot is between the difference of paired scales versus their average. It includes 95% limits of agreement around observed average agreement, and a line for perfect average agreement ($y = 0$ line). The regression line is also plotted to show the trend.



(a) Denmark

(b) Ireland

(c) United Kingdom

Figure 1.6: *EU dimension under saliency-based and positional scaling models in Denmark, Ireland, and the UK.* Violin plots include a marker for the median of the data, a box indicating the interquartile range, and spikes extending to the upper- and lower-adjacent values. Overlaid is the kernel density estimate.

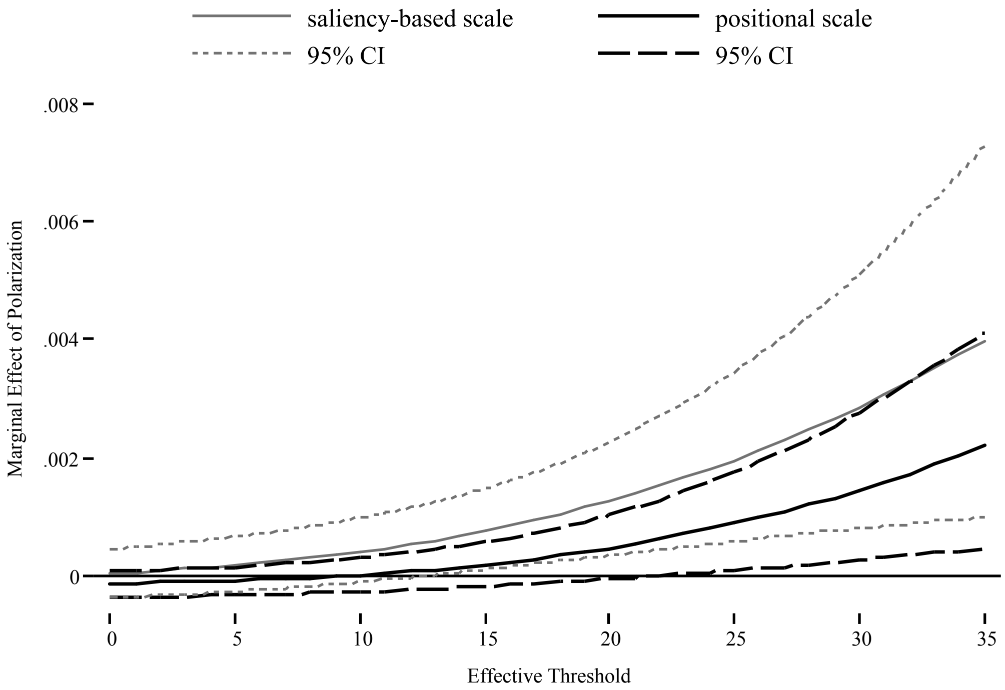


Figure 1.7: *Marginal effect of a 0.01 unit increase in party system polarisation on the probability of pre-electoral coalition formation.* Based on original presentation in Figure 2 (Golder, 2006), re-estimated for saliency-based and positional scales with a smaller incremental step (0.01 instead of 1 unit increase) following different range of scales here (from -1 to 1).

Conservative Manifesto 1997	**Conservative Manifesto 2001**
The creation of a European single currency would be of enormous significance for all European states whether they are members or not. We must take account of all the consequences for Britain of such a major development of policy. John Major secured for us at Maastricht an opt-out from the commitment to enter a single currency. It is only because of this opt-out that we have the right to negotiate and then decide whether it is in Britain's interest to join. It is in our national interest to take part in the negotiations. Not to do so would be an abdication of responsibility. A single currency would affect us whether we were in or out. We need to participate in discussions in order to ensure the rules are not fixed against our interests. The national interest is not served by exercising our option - one way or the other - before we have to. For a single currency to come into effect, European economies will have to meet crucial criteria. On the information currently available, we believe that it is very unlikely that there will be sufficient convergence of economic conditions across Europe for a single currency to proceed safely on the target date of January 1st 1999. We will not include legislation on the single currency in the first Queen's Speech. If it cannot proceed safely, we believe it would be better for Europe to delay any introduction of a single currency rather than rush ahead to meet an artificial timetable. We will argue this case in the negotiations that lie ahead. **We believe it is in our national interest to keep our options open to take a decision on a single currency when all the facts are before us. If a single currency is created, without sustainable convergence, a British Conservative government will not be part of it.** **If, during the course of the next parliament, a Conservative government were to conclude that it was in our national interest to join a single currency, we have given a guarantee that no such decision would be implemented unless the British people gave their express approval in a referendum.**	We *will* keep the pound. Labour's plan for early entry into euro is the single biggest threat to our economic stability. By keeping the pound we will keep control of our economic policy, including the ability to set interest rates to suit British economic conditions. [...] The next Conservative Government will keep the pound.

Table 1.1: *References to single currency in British Conservative party manifestos for 1997 and 2001 general elections.* Emphases in original. Text source: Richard Kimber's manifesto archive found at http://www.psr.keele.ac.uk/area/uk/man.htm

Scale	**Pearson's r**		**Concordance correlation**		**Bias correction factor**
	coefficient	95% CI	coefficient	95% CI	
`rile` dimension	0.956	[0.951, 0.960]	0.788	[0.778, 0.797]	0.824
Economic dimension	0.844	[0.831, 0.855]	0.330	[0.315, 0.345]	0.391
Social dimension	0.807	[0.794, 0.819]	0.413	[0.398, 0.427]	0.512
EU dimension	0.597	[0.557, 0.634]	0.029	[0.026, 0.033]	0.049

Table 1.2: *Comparing agreement between positional and saliency-based scales.* Comparison between the two scales is done using Lin's (1989,2000) concordance correlation. Results for Pearson's product-moment correlation are presented for comparison. Confidence intervals are based on 1000 bootstrap replications. In addition, the table includes estimates of bias correction factor. Bias correction factor is the ratio of concordance correlation coefficient and Pearson's product-moment correlation coefficient, with the range (0,1]. It reaches maximum value of 1 when there is no deviation from the line of perfect concordance, and further away from 1 means less accuracy (more bias)

Regressor	**Dimension 2: Model 2**		**Dimension 2: Model 3**	
	saliency-based scale	positional scale	saliency-based scale	positional scale
`rile` left-right	-0.0015	-0.0009		
	(0.0012)	(0.0006)		
EU integration	0.0194*	0.0018***	0.0189	0.0019***
	(0.0096)	(0.0004)	(0.0098)	(0.0004)
Social liberal-conservative			-0.0021	-0.0008
			(0.0023)	(0.0006)
Economic left-right			-0.0006	0.0003
			(0.0020)	(0.0005)
Commissioner	0.2961***	0.2989***	0.3053***	0.3122***
	(0.0658)	(0.0650)	(0.0659)	(0.0651)
In government	0.2591***	0.2177***	0.2529***	0.2055***
	(0.0571)	(0.0557)	(0.0575)	(0.0560)
Constant	-0.3027***	-0.3877***	-0.2913**	-0.3454***
	(0.0872)	(0.0951)	(0.0903)	(0.1010)
RMSE	0.4214	0.4117	0.4221	0.4122
R^2	0.2371	0.2717	0.2387	0.2739
N	304	304	302	302

* p<0.05, ** p<0.01, *** p<0.001

Table 1.3: *Replication results for second (EU) dimension in Hix, Noury & Roland (2006, Table 5).* Original results are replicated using saliency-based and positional left-right scales. Dummy variables for European Parliaments are included but not reported. Note: Robust standard errors in parentheses.

Regressor	**Probit 1**		**Probit 2**	
	saliency-based scale	positional scale	saliency-based scale	positional scale
Incompatibility	-2.008***	-1.096***	-1.487***	-0.857***
	(0.309)	(0.165)	(0.270)	(0.136)
Polarisation	0.148	-0.373	0.177	-0.212
	(0.619)	(0.323)	(0.240)	(0.112)
Threshold	0.016	0.007	0.018**	0.009
	(0.012)	(0.013)	(0.006)	(0.006)
Polarisation × Threshold	0.054	0.042*	0.026*	0.027***
	(0.033)	(0.018)	(0.013)	(0.007)
Coalition Size	0.047***	0.045***	0.041***	0.039***
	(0.012)	(0.012)	(0.008)	(0.008)
Coalition Size Squared	-0.001***	-0.0001***	-0.0001***	-0.0001***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Asymmetry	-0.056	-0.115	0.031	-0.026
	(0.319)	(0.316)	(0.230)	(0.233)
Asymmetry × Coalition Size	-0.028**	-0.026**	-0.023***	-0.022**
	(0.009)	(0.009)	(0.007)	(0.007)
Constant	-2.281***	-1.910***	-2.007***	-1.726***
	(0.337)	(0.337)	(0.185)	(0.189)
N	3383	3383	3383	3383
Log likelihood	-564.645	-561.913	-622.613	-614.593

* p<0.05, ** p<0.01, *** p<0.001

Table 1.4: *Replication results in Golder (2006, Table 1).* Original results are replicated using saliency-based and positional left-right scales.

Chapter 2

Analysis of Error Processes in Comparative Manifesto Project: Stochastic Text Generation and Human Misclassification

Abstract

Spatial models of party competition are central to modern political science. Before we can elaborate such models empirically, we need reliable and valid measurements of agents' positions on salient policy dimensions. The primary empirical times series of estimated party positions in many countries derives from the content analysis of party manifestos by the Comparative Manifesto Project (CMP). Despite widespread use of the CMP data, and despite the fact that estimates in these data arise from documents coded once, and once only, by a single human researcher, the level of error in the CMP estimates has never been estimated or even fully characterised. This greatly undermines the value of the CMP dataset as a scientific resource. It is in many ways remarkable that so much has been published in the best professional journals using data that almost certainly has substantial, but completely uncharacterised, error. This paper presents an integration of two papers that propose a remedy (Benoit, Laver & Mikhaylov 2009) and at the same time raise further questions about the reliability of the data (Mikhaylov, Laver & Benoit 2008). First, the CMP data generating processes are characterised. These inherently stochastic processes of text authorship, as well as of the parsing and coding of observed text by humans. Second, these error generating processes are simulated by bootstrapping analyses of coded quasi-sentences. This allows the estimation of precise levels of non-systematic error for every category and scale reported by the CMP for its entire set of 3,000+ manifestos. Using the estimates of these errors, we show how to correct biased inferences, in recent prominently published work, derived from statistical analyses of error-contaminated CMP data. This part is based on Benoit, Laver & Mikhaylov (2009). The focus is then shifted to error that arises during the text coding process. The paper presents the results of a coding experiment that used trained human coders to code sample manifestos provided by the CMP, allowing the estimation of the reliability of both coders and coding categories. The effect of coding misclassification on the CMP's most widely used index, its left-right scale is demonstrated. This part is based on Mikhaylov, Laver & Benoit (2008). Finally, conclusions are drawn for future use and design of the CMP data.

Key Words: Comparative Manifesto Project, content analysis, measurement error, misclassification.

2.1 Text as a source of information about policy positions

This paper integrates the results of research on measurement error processes in the Comparative Manifesto Project. It brings together analysis of uncertainty associated with stochastic generation of political text (manifestos) in Benoit, Laver & Mikhaylov (2009) and analysis of measurement error due to human misclassification of party manifestos in Mikhaylov, Laver & Benoit (2008).

Political text is a fundamental source of information about the policies, preferences and positions of political actors. This information is vital to the operationalisation of many models at the heart of modern political science.[1] Our ability to measure policy positions using political text is constrained by available methods for systematically extracting information from the vast volumes of suitable text available for analysis. Recent methods have made progress by breaking from traditional content analysis to treat text, not as an object for subjective interpretation, but as objective data from which information about the author can be estimated in a rigorous and replicable way (e.g. Slapin & Proksch 2007, Monroe & Maeda 2004, Laver, Benoit & Garry 2003, Laver & Garry 2000). Treating words as data enables the use of conventional methods of statistical analysis, allowing inferences to be drawn about unobservable underlying characteristics of a text's author, for example policy positions, from observable content of the text. This statistical approach eliminates both subjectivity and the propensity for human error, making results of text-based analysis easily replicable. A huge benefit is that it generates measures of uncertainty for resulting estimates—now recognised as a *sine qua non* for serious empirical research in the social sciences (King, Keohane & Verba 1994, 9).

A vital issue for any statistical approach to text analysis is the content validity of resulting estimates. All results, however generated, must ultimately be interpreted and judged valid by expert human analysts. This is why purely statistical techniques for text analysis can

[1] Of course there are many alternative ways to measure political positions, including but not limited to: the analysis of legislative roll calls; survey data on preferences and perceptions of political elites; survey data on preferences and perceptions of voters; surveys of experts familiar with the political system under investigation; the analysis of political texts generated by political agents of interest. Benoit & Laver (2006) review and evaluate these different approaches.

never completely replace human interpretative coding. The key advantage of computational techniques for statistical text analysis is their great potential to generate rigorous analyses of vast volumes of text, far beyond the capacity of any feasible team of human coders. Before we accept the resulting estimates as valid, however, these must be calibrated against results generated by human interpretative coders working with at least a small representative subset of the text under investigation. This means that estimates generated from human interpretative text coding must also be rigorously derived and replicable. In particular such estimates must come with associated measures of uncertainty so we can know whether they are "the same as" or "different from" other measures with which they are compared. Absent this rigour, human interpretative text coding is of no systematic value in validating results generated using other techniques. Unfortunately, results generated by human interpretative coding of a given text are often reported as point estimates with no associated measures of uncertainty. Our task here is to begin the process of addressing this issue.

While our arguments below relate to any type of text, we focus in particular on a set of political texts that has been extensively studied: party manifestos. A huge number of manifestos have been analysed, using human interpretative coders, by the Comparative Manifestos Project (CMP).[2] First reported in 1987 (Budge, Robertson & Hearl 1987), a hugely expanded version of this dataset was reported in the project's core publication, *Mapping Policy Preferences* (Budge et al. 2001, hereafter *MPP*), to have covered thousands of policy programs, issued by 288 parties, in 25 countries over the course of 364 elections during the period 1945-1998. The dataset has recently been extended, as reported in the project's most recent publication *Mapping Policy Preferences II* (Klingemann et al. 2006, hereafter *MPP2*), to incorporate 1,314 cases generated by 651 parties in 51 countries in the OECD and central and eastern Europe (CEE). Commendably, these data are freely available and have been very widely used, as can be seen from over 800 Google Scholar citations by third-party researchers of core CMP publications.[3] The CMP data are particularly attractive to scholars seeking long time series of

[2] We also note, however, that the CMP is not the only text-based measure that is based on party manifestos: Laver & Garry (2000), Laver, Benoit & Garry (2003), and Slapin & Proksch (2007) are also examples.

[3] As of August 25, 2007. The precise number of third-party citations is hard to calculate because third-party users are likely to cite several CMP sources in the same paper.

party policy positions in many different countries, for whom this dataset is effectively the only show in town. Despite their pervasive use by the profession, however, these data come with no associated measures of uncertainty. The *reliability* of many CMP scales, especially the left-right scale, has been investigated (e.g. McDonald & Mendes 2001*b*, Hearl 2001, *MPP2*, ch. 5), as has the *validity* of CMP scales in comparison with external measures (e.g McDonald & Mendes 2001*a*, Hearl 2001, *MPP2*, ch. 4). But there is no estimate of *uncertainty* that accompanies the very precise point estimates of policy emphasis that are the essential payload of the CMP and form the basis of any scales estimated from the CMP dataset.

This problem has long been noted by both the project and its critics (e.g. *MPP2*, ch. 5; Benoit & Laver 2007*a*) but we still lack a solution. Reliable and valid use of CMP data, however, mandates measurement of uncertainty in the policy estimates deployed. Without such measures, users of CMP data cannot distinguish between "signal" and "noise," between measurement error and the "real" differences in policy positions that are at the heart of so many theoretical models. As we show below, we can infer far less *actual change* in party policy from one election to the next, using *observed changes* in CMP estimates, since some of the observed change can be attributed to textual noise. Compounding this problem, CMP estimates of party policy positions are typically used as explanatory variables. Ignoring measurement error in such variables leads to biased inferences about causal relationships, and thus to flawed research findings. The unmeasured level of non-systematic error in the CMP dataset drastically undermines its primary value for the profession, as a reliable and valid set of estimates of party policy positions across a wide range of years, countries and policy dimensions. If this problem can be fixed, not only will CMP data be much more useful in themselves, they will also be much more valuable as sources of calibration for techniques of computational text analysis that can in turn be deployed in vastly more ambitious projects.

We address this problem by decomposing stochastic elements in the data generation process underlying interpretative content analysis by humans. This has two essential components: text generation and text coding. In this paper, we focus on measurement uncertainty arising from the stochastic nature of political text itself. Any observed text is but one of a huge number of *possible* texts that could have been generated by an author intent on conveying the same message. Characterising stochastic text generation allows us to systematise the blindingly obvious but hitherto neglected intuition that *longer texts tend to contain more information than shorter ones*. Thus there is huge variation in the length of texts analysed by the CMP; some coded texts are more than 200 times longer than others. Astonishing as this seems the moment we think about it, all published work using CMP data assumes all texts are equally informative.

The text coding component usually takes the approach of analysing the content of texts using a categorical scheme consisting of two steps (Krippendorff 2004, 219). First, texts are parsed into smaller units relevant to the research question, such as words, sentences, or quasi-sentences, depending on the research design. Following this first step of *unitisation*, a second step involves *coding* each unit by assigning a category from the coding scheme to each text unit. Both steps can be held to scrutiny according not just to the validity of the resulting information, but also for the reliability of the procedure, two criteria that often trade off with one another in practice.

Whenever non-deterministic instruments—such as human beings—are used to unitise and code texts, then the content analysis procedure faces potential problems with *reliability*. Depending on how unreliable the procedure is, estimates constructed from the codings may lack validity because of the noise or even bias introduced by the content analysis procedure. Reliability is no guarantee of validity, however, and in practice validity tends to suffer in the pursuit of maximising reliability. Indeed, the debate over computerised versus hand-coded content analysis largely revolves around the tradeoff between reliability and validity. Proponents of computerised schemes for estimating party positions from political manifestos (e.g. Laver, Benoit & Garry 2003, Laver & Garry 2000, Slapin & Proksch 2007) cite perfect reliability in their favour, and struggle to demonstrate validity, while hand-coded schemes such as the CMP claim validity as a central advantage and then devote huge resources to attempts to enhance reliability (see for instance Klingemann, Volkens, Bara, Budge & McDonald (2006) chs. 4–5).

As a thought experiment, suppose we want to estimate the position on a left-right scale of French president Nicolas Sarkozy, using as texts the complete set of speeches he made on the record during 2007. We could count the frequencies n_l, n_r of the letters “`l`” and “`r`” in each text and measure the Sarkozy’s position on a left right policy scale as $(n_r - n_l)/(n_r + n_l)$. This would be a superbly reliable technique, easily implemented using computers, but probably also possible using chimpanzees trained in character recognition. Anyone claiming it as a valid measure, however, would be pitied rather than published. Nor is the problem cured by taking account of the fact that the speeches are in French and redefining the measure using frequencies of the letters "`g`" and "`d`". Rather, the problem is that this overly simplistic coding scheme does not link the text units or the coding frame to valid verbal manifestations of politically left or right policy, even though either computers or trained monkeys could implement it with a perfect or at least very high degree of reliability.

We could vastly improve on validity by selecting a better coding scheme. Leaving aside more nuanced ideological differences for the moment, assume we propose a new coding scheme consisting of two categories, "left" and "right", and that the task is to tag each sentence from Sarkozy's speeches according to this binary classification. To code the texts with this scheme, we could recruit a panel of scholars accepted within the profession as the world's greatest experts on French politics, ask them to read the Sarkozy speeches and then classify each sentence as left or right. Every sane person would agree that our new measure is much more valid than the earlier letter-based approach, but we now have a new problem in that the experts will surely disagree on how at least some sentences should be classified. The experts must apply subjective judgements based on their interpretation of the each sentence's meaning—indeed this is why we chose them over the chimpanzees, whose expected agreement would have been 25% through pure chance. Subjective judgments are at the very least subject to stochastic variation, ranging from a sudden bout of acute indigestion on the part of a coder, to the fact that different coders may have listened to different news stories on different mornings before they began their coding exercise, to different toss-up judgement calls any or all of them might make at any given time. In addition, our coders might deem that many sentences in Sarkozy's text that have nothing to do with either left or right, and as retaliation for the limited choices offered by our coding scheme, may randomly assign such sentences to "left" or "right". Even worse, our coders might tend to categorise ambiguous sentences as "right" given their contextual knowledge about Sarkozy. Either way, our procedure will yield different answers each time we repeat it, with some sentences being subject to *misclassification* each time, and resulting in summary estimates whose validity is now suspect. Part of the problem has arisen from the fundamentally indeterminate nature of human judgement, but this problem has been compounded by a poor coding scheme—two interrelated aspects to which we return at length below.

Ideally, of course, we would like the policy positions we estimate from political texts to be valid and unbiased, constructed from procedures that are perfectly reliable and reproducible. A research procedure, according to Krippendorff,

> is *reliable* when it responds to the same phenomena in the same way regardless of the circumstances of its implementation...In content analysis, this means that the reading of textual data as well as of the research results is replicable elsewhere, that researchers demonstrably agree on what they are talking about. (Krippendorff 2004, 211)[4]

In any content analysis scheme using human coders to apply a coding scheme with any degree of meaning, however, perfect reliability is virtually impossible. Our first task as data analysts, therefore, is to identify and characterise problems of validity and reliability, as well as potential consequences (such as bias) in our research procedure and resulting estimates. Absent this, our estimates are worthless. Indeed they are in a real sense worse than worthless since we have no idea at all how good or bad they are, completely undermining any procedural confidence in the veracity of the results produced by the research. When it comes to interpreting data, an unreliable research procedure casts basic doubts as to the meaning of the data and what any analysis of these data would mean (Krippendorff 2004, 212). Our first priority should therefore be to characterise problems regarding validity, reliability, and bias in our research procedure, and our second task to work as hard as we can to minimise their effects.

We proceed as follows. First, we describe the CMP dataset and the processes that led to its generation. Focusing on stochastic text generation and the impact of text length on measure-

[4] Krippendorff (2004, 214) identifies three types of reliability: stability, reproducibility, and accuracy. *Stability* is concerned with possible change of coding results on repeated trials. This type of reliability has a coder reanalysing the same manifesto after a period of time in order to highlight any intra-coder disagreement. A stronger measure of reliability is *reproducibility*, also called inter-coder reliability. This measure assesses the degree of replication of coding results by two distinct coders working separately. It covers intra-coder disagreement and inter-coder differences in interpretation and application of the coding scheme. *Accuracy* tests the conformity of coding process and data generation procedure to some canonical standard, and is perceived to be the strongest test of reliability. It can be used effectively at the training stage when coder's performance can be compared to some 'true' results.

ment uncertainty, we show two different ways to calculate standard errors for each estimate in the CMP dataset; one relies on analysis, one on simulation. Analysing these error estimates we find that many CMP quantities, *even assuming perfectly reliable human coders*, should be associated with substantial uncertainty. We show how these error estimates can be used to distinguish substantive change from measurement error in both time-series and cross-sectional comparisons of party positions. We suggest ways to use our error estimates to correct analyses that use CMP data as co-variates, re-running and correcting some prominent analyses reported in recent literature.[5]

Next we focus on measurement uncertainty arising from stochastic variation in the coding of a given observed text by human coders. CMP data are widely used by third party researchers to measure policy positions of political parties on an election-by-election basis, indeed they are profession's primary source of such data. We know axiomatically that these data have problems of validity, reliability and bias, just as all data do. Here our main substantive interest lies in developing a more systematic characterisation of some of these problems than has hitherto been attempted. We set out a framework for reliability and misclassification in categorical content analysis, and apply this framework to the CMP coding scheme. To come to concrete terms with reliability and misclassification in the context of the CMP, we designed and carried out a series of coding experiments on texts for which the CMP has supplied a "correct" coding, and we report on these tests.[6]

Finally, we discuss the results of our analysis of stochastic text generation and stochastic human coding of manifestos. We discuss the implications of our results for continued use of the CMP research. Our aim in doing this is to increase the professional value of the CMP data by enhancing our ability to draw reliable, valid and unbiased statistical inferences from these.

2.2 From Policy Positions to Coded Dataset

Before we characterise error in the CMP dataset, we must understand the processes by which this error arises. These are essentially the same processes that underlie any human interpretative

[5] See Benoit, Laver & Mikhaylov (2009) for an earlier presentation.

[6] See Mikhaylov, Laver & Benoit (2008) for an earlier presentation.

coding based, wholly or partially, on text sources. They therefore apply more generally to the many social science datasets that include variables generated by humans who read some text and then record a quantitative coding conditioned on this. To aid exposition, however, we focus on the data generation processes underlying the CMP. These are summarised in Figure 2.1.

[FIGURE 2.1 ABOUT HERE]

The premise of all content analysis is that there is something to be analysed. Here, we think of this as the *true policy position*, π of the author of some text. This is fundamentally unobservable even, arguably, to the author. If the author is not a hermit, s/he may want to send signals about this position to others. These may represent "sincere" attempts to communicate π or "strategic" attempts to communicate some other position. There is a *strategic model of politics*, M, that characterises the author's incentives to signal a policy position that may or may not be the same as π - we can think of this as the *intended message*, μ. Note that μ exists only in the brain of the author and is also fundamentally unobservable.

Having formed the intention to communicate μ, the author *generates some text*, τ, to do this job. Every time the author sets out to communicate μ, s/he is likely to generate a slightly different τ. As an aid to intuition here, consider what happens when an author's hard disk crashes after a long hard day of manifesto writing. First, hair is torn out. Then an attempt is made to recreate the day's work. The recreated text is very unlikely indeed to be identical to the lost text; indeed the author may well think of "better" ways to say the same thing, when given the job of saying it all over again. Now think of different authors, with somewhat different literary styles, all trying to convey precisely the same message. In a nutshell, there are many different versions of τ that could be generated with the sincere intention of conveying the same μ. There is a *stochastic text generation process* T, that maps μ into τ.

We now have an observed text τ, which we can take as having a "certain" content, at least to the extent there are unambiguous text characters deposited on the page. The process of *reading* the text now begins. In terms of a project such as the CMP, this involves a human expert reader first breaking the text into units, "quasi-sentences" in the argot of the CMP, and then subjectively assigning these text units to categories in a predefined coding scheme. This scheme is a *measurement instrument*, I. In the CMP's case I is a 56-category scheme describing different types of policy statement the author might make, or 57 categories if the "uncoded" category is also included. The CMP scheme was defined by a particular group of scholars meeting in the mid-1980s. It is almost certain that a different group of scholars meeting at the same time, or the same group of scholars meeting at a different time, would have defined a different coding scheme. The realised CMP coding scheme I is thus one of a huge number of possible coding schemes that could have been realised.

Given an observed text τ and a realised coding scheme I, expert human readers interpret text units in τ and allocate these to coding categories in I. This coding process has both subjective and stochastic elements. The same human reader at different times, or a different human reader at the same time, may well allocate the same text unit to different coding categories. There is thus a *stochastic text coding process* C that, given I, maps τ into δ, a database of text codings. Given the stochastic processes we have outlined above, the codings in δ are associated with considerable uncertainty.

The analyst wants the database of text codings in the first place because s/he wants to estimate something about the text's author. This involves scaling the data, using some *scaling model* S. Clearly, there are many different scaling models that could be applied to the same database of text codings. The result of applying scaling model S to the database of text codings in δ will be a set of scales λ. In relation to the CMP, a very well-known scale is the left-right scale called `rile`. This is the feature of the scaled CMP dataset that is overwhelmingly the most commonly used in published work. There are, of course, many *different possible sets of scales* λ that could be developed by applying scaling model S to database δ.

Finally, the circle is closed as the analyst uses a text's measured scale positions, given λ, to make *inferences about the text's author*. These inferences may concern the author's text deposits τ, "true" position π or intended message μ. Statistical inference in these matters can rely on conventional techniques. Logically valid inferences are *increasingly dependent on underlying theoretical models* as they move back the causal chain from τ to μ to π.

We have been very explicit about all of this because it is important to focus carefully on particular features of the long process of causal inference summarised in Figure 2.1. Lack of clarity about this can, for example, lead to misplaced criticisms of the CMP data. Many of the alleged shortcomings attributed to the estimation of party positions from manifestos, for instance, concern the validity of using manifestos as unbiased, observable implications of true party positions. It is frequently argued, for example, that party manifestos are strategic documents that do not convey the "true" party position, in effect that $\mu \neq \pi$. But this is not a measurement issue. Assuming we can measure the intended message μ from the observed text τ in an unbiased way, this is a matter of specifying the correct strategic model M that maps μ into π. The claim that manifestos are strategic documents does not therefore have any bearing on CMP text codings, but rather on the logical inferences that are drawn from these about unobservable "true" policy positions π. The solution to this problem is not better text codings in δ but a better strategic model of politics, M. Similarly, it is perfectly reasonable to argue that the CMP's additive left-right scale `rile` is flawed and that other left-right scales using the same data, for example those proposed by Gabel & Huber (2000), or by Kim & Fording (1998), are more valid bases for drawing inferences about the policy positions, μ or π, of text authors. Again, this does not concern the database of CMP text codings, δ, but rather the validity of the scaling model S that maps these into a set of derived scales λ. The solution to this problem is a better scaling, not better text codings.

Figure 2.1 also helps us focus on features of the CMP dataset that are indeed intrinsic to the data collection project itself, further distinguishing between problems that can be fixed without recourse to additional data collection and those that cannot be addressed without new data on the coding of party manifestos. Thus far little attempt has been made to take account of the fact that the CMP's core measurement instrument I, its 57-category coding scheme, is but one realisation of the many possible coding schemes that could have been devised.[7] Clearly the CMP coding scheme is an utterly integral feature of the CMP dataset. Equally clearly, assessing the implications of this involves recoding the same documents using different schemes, and thus a major new data collection enterprise.

Previously very little attempt has been made, furthermore, to characterise the stochastic *coding* process, C, by estimating the extent of variation between coders in applying the same

[7] Laver & Garry (2000) recoded some party manifestos using what they felt to be a more valid, hierarchically structured, coding scheme. Schofield & Sened (2006) report results of having experts recode manifestos using national election study questionnaires coding schemes, to allow party and voter positions to be mapped into a common space.

coding scheme I to the same text τ. This cannot be investigated without conducting multiple human codings of the same document using the same coding scheme and thus also involves a major new data collection enterprise. Considerable attention has, however, been paid to the reliability and validity of scales derived from the CMP database of text codings, reflected in extensive discussion of the validity of the CMPs `rile` scale.[8] Such discussions about scaling do not hinge on the collection of a new database of new text codings, δ, but rather on how a given dataset should be scaled.[9]

We are not concerned here with building scales from the CMP data, but with another aspect of the CMP manifesto dataset that can be addressed without a major new data collection exercise. This concerns the fact that there is a stochastic text generation process, T, that maps the intended message μ into an observed text τ. We model this process below, using both analytical techniques and simulations, allowing us to formalise the intuition that longer political texts, other things being equal, convey more information about their authors. After that we turn to characterise the stochastic *coding* process, C, by estimating the extent of variation between coders in applying the same coding scheme I to the same text τ in a set of experiments.

[8] This is particularly important because the overall content validity of the CMP dataset is claimed, by the CMP itself, in terms of the extent to which time series estimates of party positions on `rile` track received wisdoms among country experts about "real" party movements over time on the left-right dimension.

[9] However, a related issue concerns the format in which the CMP data are distributed and used. Formally, the full database δ of CMP text codings comprises an ordered sequence of all coded text units for each text, each unit tagged by which coding category it was assigned to by different coders. The CMP issues, and indeed itself works with, a vastly reduced "scaled down" version of δ. (Indeed it is not clear that the full δ continues to exist for this dataset.) Thus the "semi-scaled" version of the CMP dataset familiar to most scholars involves a set λ of 57 scales, each scale measuring the relative emphasis given to each coding category as the proportion of text units coded into this category. This is, of course, only one of many possible ways of performing data reduction on the underlying dataset of text codings, δ. A scholar wanting to measure the relative importance of issues in terms of whether these were mentioned earlier rather than later in a manifesto, for example, has no way of retrieving this information from the distributed CMP dataset, even though this information did exist for all coded manifestos at some time in the history of the project.

2.3 Characterising the Stochastic Process of Text Generation

In what follows, we want to estimate the level of uncertainty in CMP estimates of party policy positions that arises from the stochastic process of text generation. Before going forward, therefore, it is important to be clear about which of the processes mapped in Figure 2.1 we are going to hold constant. Taking things from the top, we are not concerned with modeling the text authors' strategic incentives to dissemble. We thus in effect assume that $\mu = \pi$. Readers who do not believe this must specify a strategic model M of politics, mapping μ into π, that we do not consider here. The stochastic process, C, of human text coding is directly estimated below, but to make analysis feasible it is held constant here. The only assumption we make about C is that this stochastic process is unbiased. We take the CMP's 57-category coding scheme as given and do not concern ourselves with the datasets that alternative coding schemes might have produced. While the scaling model S that has been applied to the database of CMP codings clearly raises crucial issues, we take two core features of this as given in what follows. The first is the scaling assumption that measures a text's relative emphasis on a CMP coding category as the percentage of coded text units assigned to that category. The second is the precise definition of the CMP's `rile` scale. What we *do* focus on in this section is the stochastic process T that maps text authors' unobservable policy positions $\pi(=\mu)$ into observable text deposits τ.

For a given policy category j, define π_{ij} as the *true* but *unobservable* policy position of the text's author, represented as country-party-date unit i. The j categories in this case are the 56 policy categories in the CMP coding scheme, plus an additional category for "uncoded," giving a total of $k = 57$ categories. Since, according to the CMP's measurement model, true policy positions are represented by relative or "contrasting" emphases on different policy categories within the manifesto, these policy positions are relative proportions, with $\sum_{j=1}^{k} \pi_j = 1$.[10]

[10] In what follows, we refer to these quantities as policy "positions." The CMP's saliency theory of party competition is neither widely accepted nor indeed taken into any account by most third-party users of CMP data. However, inspection of the definitions of the CMP's coding categories reveals that all categories but one of the 56 are very explicitly positional in their definitions, which refer to "favourable mentions of..." , "need for...," etc. The sole exception is PER408 "Economic goals" , a category which is (quite possibly for this reason) almost never used by third-party researchers. For this reason, we do not regard it as in any way problematic that third-party users almost invariably interpret the CMP's

For example, party i's emphasis, for a given election, on the 20th issue category in the CMP coding scheme (401: Free Enterprise), is represented as π_{i20}.

We can never observe the "true" policy positions of manifesto authors, π_{ij}. It is possible, however, to have a human coder analyse party i's manifesto using the CMP's coding scheme, and thereby to measure the relative emphasis given in the manifesto to each π_{ij}. This is measured as $p_1, \ldots p_k$, where $p_j \geq 0$ for $j = 1, \ldots, k$ and $\sum_{j=1}^{k} p_j = 1$. In the absence of systematic error (bias):

$$\mathrm{E}(p_{ij}) = \pi_{ij} \tag{2.1}$$

In other words, the observed relative emphasis given to each coding category in a party's manifesto will *on average* reflect the true, fixed, and unobservable underlying position π_{ij}. The realisation of π_{ij} in any given manifesto, however, reflects the stochastic process of text authorship, yielding the observed proportions p_{ij}. Every time a manifesto is written with the intention of expressing the same underlying positions π_{ij}, we expect to observe slightly different values p_{ij}.

Given this characterisation of both observed and unobservable policy positions, which directly follows the CMP's own assumptions, we can postulate a statistical distribution for observed policy positions. If we assume each text unit's allocation to a policy category is independent of the allocation of each other text unit, then we can characterise the CMP's realised manifesto codings as corresponding to the well-known *multinomial* distribution with parameters n_i and π_{ij}, where n_i refers to the total number of quasi-sentences in manifesto i. The probability for any manifesto i of observing counts of quasi-sentences x_{ij} from given categories j is then described by the multinomial formula:

$$\Pr(X_j = x_j, \ldots, X_k = x_k) = \begin{cases} \frac{n!}{x_j! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k} & \text{when } \sum_{j=1}^{k} x_j = n \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

In the context of the CMP coding process for a given manifesto, each x_k represents the number of text units coded to a given category j, since through the multinomial expectation, $\mathrm{E}(x_{ij}) = p_{ij} n_i$. In terms of the "PER" or percentage categories reported by the CMP for each

"saliency" codings as "positional."

manifesto, what is actually reported is $x_{ij}/n_{ij}100$, or the estimate of manifesto i's "true" percentage ($\pi_{ij}100$) of the quasi-sentences from category j. We have no additional information that might lead us to conclude there is a systematic function mapping (in a biased way) the true position to a different expected observed position—already expressed by Equation 2.1. Our concern here is with non-systematic (unbiased) error, which is the extent to which $\text{Var}(p_{ij}) > 0$, even though π_{ij} is fixed at a single, unvarying point.[11]

So far we have considered only the case of a "given" manifesto, but of course the combined CMP dataset set deals with many such units—a total of 3,018 separate units representing different combinations of country, election date, and political parties for the combined (*MPP* + *MPP2*) datasets.[12] If we are to fully characterise the error from the stochastic process whereby texts are generated, then this will mean estimating $\text{Var}(p_{ij})$ for every manifesto i for all $k = 57$ categories.[13]

The lengths (n_i) of the coded manifestos underlying the CMP dataset vary significantly, although this valuable information is almost never referred to by subsequent users of CMP data. About 30 percent of all coded manifestos had less than 100 quasi-sentences, coded into one of 56 categories. Some had less than 20 quasi-sentences; some had more than 2000. Despite very wide variation in the amount of policy information in different manifestos, policy positions estimated from CMP data are almost always treated in the same way, regardless of whether they are derived from coding 20 text units, or 2000.[14] The total number of text

[11] In the language of classic reliability testing, we are concerned here with estimating the error variance σ_E^2, related to reliability classically defined as $1 - \sigma_E^2/\sigma_X^2$. When σ_E^2 is unobserved—as is always the case with manifesto coding—a variety of surrogate methods may be used to estimate the reliability of the CMP estimates, many of which have been explored previously (e.g. McDonald & Mendes 2001*b*).

[12] It is not quite accurate to state that the dataset represents 3,018 separate *manifestos*, since some of these country-election-party units share the same manifesto with other parties (`progtype`=2) or have been "estimated" from adjacent parties (`progtype`=3). See Appendix 3, *MPP*. The full CMP dataset also failed to provide figures on either total quasi-sentences or the percentage of uncoded sentences for 141 manifesto units, limiting the sample analysed here to 2,877.

[13] Note that there are reasons, however, to believe that the multinomial assumptions that the π_{ij} (and resulting X_{ij}) categories are independent and identically distributed, are almost certainly wrong, since political views of one type tend to be correlated with those of related, but separately coded types. We return to this issue below in comparing the parametric (multinomial) model to non-parametric errors estimated from bootstrapping.

[14] We also note that not all quasi-sentences can be coded, giving rise to a non-trivial category for "uncoded" content. While the median percentage of uncoded content is low, at 2.1%, the top quarter of all manifestos contained 8% or more of uncoded content, 10 percent of manifestos contained 21% or more uncoded content.

units found in a manifesto appears to be, absent systematic information or prior expectation on this matter, unrelated to any political variable of interest. Yet, while assuming that the proportions π_{ij} remain the same regardless of document length, increasing the length of a manifesto does increase confidence in our estimates of these proportions. This reflects one of the most fundamental concepts in statistical measurement: uncertainty about an estimate should decrease as we add information to that estimate.[15] Given that our characterisation of the stochastic process that produces observed text categories depends directly on the length of the text, we show next how to use this information to produce error estimates directly reflecting this basic uncertainty principle.

2.4 Estimating Error in Manifesto Generation

2.4.1 Analytical error estimation

One way to assess the error variance of estimated percentages of text units of the CMP's 56 coding categories is through the analytic calculation of variance for the multinomial distribution we have used to model category counts. The goal is to determine the variance of each of the policy ("PER") categories reported by the CMP, which in the language described above represent $\hat{\pi}_{ij}100$ for each category j and each manifesto i. Here we assume no coding bias (by Equation 2.1), where each π_{ij} represents the *true* but *unobservable* position of country-party-date unit i on issue j.

Returning to the definition of the multinomial distribution in Equation 2.2, for any multinomial count X_{ij}, the variance is defined as

$$\mathrm{Var}(X_{ij}) = n_i p_{ij}(1 - p_{ij}) \tag{2.3}$$

[15] Experience from the CMP has also found that human coders tend to unitise the texts into quasi-sentences in a less than perfectly reliable fashion, although this is an aspect of coder variance that we do not deal with here. An analysis of results from repeated codings of the training document used by the CMP to initiate new coders by Volkens (2001*b*) gives us insight into deviation by different coders from the "correct" quasi-sentence structure, as seen by the CMP. Volkens reports that average deviation from the "master" quasi-sentence length by thirty-nine coders employed in the CMP was around ten percent. In the CMP coding tests we have analysed ourselves, which involve 59 different CMP coders in the course of training, coders identified between 127 and 211 text units in the same training document, with a SD of 19.17 and an IQR of (148, 173).

Following the algebraic manipulation and dropping the manifesto index *i* for simplicity):

$$\begin{aligned}
\mathrm{E}(X_j) &= np_j \\
x_j &= np_j \\
\frac{x_j}{n} &= p_j \\
\mathrm{Var}\left(\frac{1}{n}x_j\right) &= \mathrm{Var}(p_j) \\
\frac{1}{n^2}\mathrm{Var}(x_j) &= \mathrm{Var}(p_j) \\
\frac{1}{n^2}\, np_j(1-p_j) &= \mathrm{Var}(p_j) \\
\frac{1}{n}\, p_j(1-p_j) &= \mathrm{Var}(p_j)
\end{aligned}$$

Translating into the CMP's percentage metric ($p_j * 100$):

$$\begin{aligned}
10,000\mathrm{Var}(p_j) &= \frac{10,000}{n} p_j(1-p_j) \\
\mathrm{SD}(p_j 100) &= \frac{100}{\sqrt{n}}\sqrt{p_j(1-p_j)}
\end{aligned}$$

This allows to express the variance of the proportion p_{ij}, and the rescaled percentage (used by the CMP as):

$$\mathrm{Var}(p_{ij}) = \frac{1}{n_i}\, p_{ij}(1-p_{ij}) \tag{2.4}$$

$$\mathrm{SD}(p_{ij} 100) = \frac{100}{\sqrt{n_i}}\sqrt{p_{ij}(1-p_{ij})} \tag{2.5}$$

$$\mathrm{SD}(p_{ij}) \propto \frac{1}{\sqrt{n_i}}$$

In part, then, the error will depend on the size of the true percentage of mentions $p_{ij}100$ for each "PER" category *j*. Assuming this quantity is fixed for each party-election unit *i*, however, what is variable as a result of the data generating process is the length n_i of the manifesto. This aspect of the error in the CMP estimates, therefore, is inversely proportional to the (square root of the) length of the manifesto. This should be reassuring, since it means that longer manifestos reduce the error in the estimate of any coding category *j*, irrespective of

p_j. Longer manifestos provide more information, and we can be more confident about policy positions estimated from them.

The situation is more complicated for additive measures such as the pro-/anti-EU scale (PER108 - PER110) or for the CMP's widely-used left-right scale, an additive scale obtained by summing percentages for 13 policy categories on the "right" and subtracting percentages for 13 categories on the "left." This is because, for summed multinomial counts, the covariances between categories must also be estimated, since it is a property of variance that $\text{Var}(aX+bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y)$. There are several strong reasons, including the limited observations we have of non-random ways in which different human coders code the same text unit into different categories, as well as innate substantive relationships between coding categories, to suspect that these covariances will be non-zero. For these reasons, we do not recommend using analytically derived errors for composite scales aggregated from the CMP's 56-category scheme, instead advocate a more general, non-parametric approach: simulation.

2.4.2 Estimating Error Through Simulation

Given potential analytical problems we identify at the end of the previous section, we suggest an alternative way to assess the extent of error in CMP estimates. This uses simulations to recreate the stochastic processes that led to the generation of each text, based on our belief that there are many different possible texts that could have been written to communicate the same underlying policy position. We do this by bootstrapping the analysis of each coded manifesto, based on re-sampling from the set of quasi-sentences in each manifesto reported by the CMP. Bootstrapping is a method for estimating the sampling distribution of an estimator through repeated draws with replacement from the original sample. It has three principal advantages over the analytic derivation of CMP error in the previous section. First, it does not require any assumption about the distribution of the data being bootstrapped and can be used effectively with small sample sizes ($N < 20$) (Efron 1979, Efron & Tibshirani 1994). Second, bootstrapping permits direct estimation of error for additive indexes such as the CMP "right-left" scale, without making the assumptions about the covariances of these categories required to derive an analytic variance. Since exact covariances of these categories are unknown, sample dependent, and influenced by non-random coder errors, it is highly speculative to make the assumptions needed for analytical computation of variance for additive scales. Finally, simulation allows us to mix error distributions, a key requirement in our case if we wish to incorporate additional forms of error. For instance, we might also wish to simulate coder variances such as the (possibly normally distributed) differences in text unitisation mentioned by Volkens (2001*b*), although we do not do so here. For all of these reasons, we always prefer the bootstrapped error variances over an analytic solution for additive CMP measures such as the left-right scale.

The bootstrapping procedure is straightforward. Since the CMP dataset contains percentages of total manifesto sentences coded into each category, as well as the raw total number of quasi-sentences observed, we convert percentages in each category back to raw numbers. This gives a new dataset in which each manifesto is described in terms of the number of sentences allocated to each coding category. We then bootstrap each manifesto by drawing 1,000 different random samples from the multinomial distribution, using the p_i as given from the reported PER categories. Each (re)sampled manifesto looks somewhat like the original manifesto and has the same length, except that some sentences will have been dropped and replaced with other sentences that are repeated. We feel this is a fairly realistic simulation of the stochastic text generation process. The nature of the bootstrapping method applied to texts in this way, furthermore, will strongly tend to reflect the intuition that longer (unbiased) texts contain more information than shorter ones.

One problem that is not addressed by bootstrapping the CMP manifesto codings is that, as anyone who has a close acquaintance with this dataset knows, many CMP coding categories are typically empty for any given manifesto—resulting in zero scores for the variable concerned. No matter how large the number we multiply by zero, we get zero. Thus a user of CMP data dealing with a 20-sentence manifesto that populates only 10 coding categories out of 56 must in effect assume that, had the manifesto been 20,000 sentences long, it would still have populated only 10 categories. *In extremis*, if some manifesto populated only a single CMP coding category, then every sampled manifesto would be identical. We cannot get around this problem with the CMP data by bootstrapping, unless we make some very interventionist assumptions about probability distributions for non-observed categories. We prefer to assume that zero categories—for example zero mentions of the European Union by Australian party manifestos in 1966—reflect a real intention of the text author not to refer to the matter at issue. We thus, for want of better information, take zero categories at face value.[16]

The great benefit of bootstrapping CMP estimates to simulate the stochastic process of text generation is that we can generate standard errors and confidence intervals associated with the point estimates, not only for each coding category but also for scales generated by combining these categories. Furthermore, even though we have strong reasons to believe CMP estimates follow a multinomial distribution, bootstrapping provides error estimates without needing to assume any distributional information not present in the observed quasi-sentences from the texts themselves.

[FIGURE 2.2 ABOUT HERE]

The results of this bootstrapping procedure are illustrated in Figure 2.2, which shows the relationship between manifesto length, in quasi-sentences, and the bootstrapped standard error for PER501, the "pro-environment" category (panel a), as well as for the additive CMP left-right scale (panel b).[17] As predicted in the previous section, error variances decline directly with (logged) manifesto length. The wide differences, indicated by the thick band in panel (a) reflect the very different proportions coded into $p_{i\text{PER501}}$ for different manifestos.

[FIGURE 2.3 ABOUT HERE]

In Figure 2.3, we compare the bootstrapped error variance and the variance computed analytically (per Equation 2.5), for the single-category environmental policy measure (PER501).

The results of this bootstrapping provide error variances that decline as exponential functions of text length, something that holds true both for single categories and for additive scales such as the CMP "right-left". In addition, comparing bootstrapped error variance with variance computed analytically (per Equation 2.5), we get nearly identical results. The near equivalence

[16] There are several methods for dealing with empty observed categories in text analysis and natural language processing, but since these modifications systematically affect the likelihoods, they relate more to systematic than the purely non-systematic error which forms our focus here. In addition, when we tested simple methods to deal with non-zero categories—e.g. "add-one" smoothing (Jurafsky & Martin 2000, Ch. 6.3)—these changes made no noticeable differences to our results.

[17] Given the distribution of the data, both axes of Figure 2.2 uses logarithmic scales and the figure demonstrates clearly that the level of bootstrapped error in this scale error is very strongly related to manifesto length—short manifestos have much more potential for error of this type than the long ones.

of these two very different methods for estimating standard errors adds to our confidence in both the analytical derivation of CMP error variance and the method of bootstrapping text units in manifestos. When we apply our new error estimates to specific empirical research problems in the next section, we use the bootstrap-estimated error as our best approximation of overall non-systematic error in the CMP's reported estimates.

2.5 Using CMP Error Estimates in Applied Research

There are two main reasons to estimate policy positions of political actors. The first is cross-sectional: a map of some policy space is needed, based on estimates of different agent positions at the same point in time. The second is longitudinal: a time series of policy positions is needed, based on estimates of the same agent's policy positions at different points in time. Alternative techniques can estimate cross-sectional policy spaces; the signal virtue of the CMP data, and the dominant reason for its use by third-party scholars, is that it purports to offer *time series* estimates of party policy positions. However, neither cross sectional nor time series estimates of policy positions contain rigorously usable information if they do not come with associated measures of uncertainty. Absent any such measure, estimates of "different" policy positions may either be different noisy estimates of the same underlying signal, or accurate estimates of different signals.

2.5.1 Estimating valid differences

A substantial part of the discussion found in *MPP* and *MPP2* of the face validity of the CMP data comes in early chapters of each book, during which policy positions of specific parties are plotted over time. Sequences of estimated party policy movements are discussed in detail and held to be substantively plausible, with this substantive plausibility taken as evidence for the face validity of the data. But are these vaunted changes in party policy "real," or just measurement noise? We illustrate how to answer this question with a specific example related to environmental policy in Germany, a country where environmental policy is particularly salient, and also where the CMP has been based for many years. Figure 2.4 plots the time series of the estimated positions of the CDU-CSU, for a long time Germany's largest party, on PER501 (*Environment: Positive* in the CMP coding scheme). The dashed line shows CMP estimates; error bars show our bootstrapped 95 percent confidence intervals around these estimates.

[FIGURE 2.4 about here]

Error bands around CMP estimates are large in this case. Most estimated "changes" over time in CDU-CSU environmental policy could well be noise. Statistically speaking, we conclude that the CDU-CSU was more pro-environmental in the early 1990s than it was either in the early 1980s or the early 2000s; every other observed "movement" on this policy dimension can easily be attributed to noise in the textual data.

[TABLE 2.1 about here]

Table 2.1 reports the result of extending this anecdotal discussion in a much more comprehensive way. It deals with observed "changes" of party positions on the CMP's widely-used left-right scale (`rile`) and thus systematically summarises all of the information about policy movements that is used anecdotally, in the early chapters of *MPP* and *MPP2*, to justify the face validity of the CMP data. The table reports, considering all situations in the CMP data in which the same party has an estimated position for two adjacent elections, the proportion of cases in which the estimated policy "change" between one election to the next is statistically significant. These results should be of considerable interest to all third-party researchers who use the CMP data to generate a time series of party positions. They show that observed policy "changes" are statistically significant in only 38 percent of relevant cases. We do not of course conclude from this that CMP estimates are invalid. We do conclude that many policy "changes" hitherto used to justify the content validity of CMP estimates are not statistically significant, and may be noise. More generally, we argue that, if valid statistical (and hence logical) inferences are to be drawn from "changes" over time in party policy positions estimated from CMP data, it is essential that these inferences are based on valid measures of uncertainty in CMP estimates, which have not until now been available.

[FIGURE 2.5 about here]

While one of the CMP's biggest attractions is undoubtedly the time series data it appears to offer, another common CMP application involves comparing different parties at the same point in time. Considering a static spatial model of party competition, realised by estimating positions of actual political parties at some time point, many model implications depend on differences in policy positions of different parties. It is crucial, therefore, when estimating a cross-section of party policy positions, to know whether estimated positions of different parties do indeed differ from each other in a statistical sense. Figure 2.5 illustrates this problem, showing estimates of French party positions in 2002, on the CMP left-right scale. Taking account of the uncertainty of these estimates, four quite different parties—the Communists, Socialists, Greens and the Union for a Popular Movement (UMP)—have statistically indistinguishable estimated positions, even though the CMP point estimates seem to indicate differences. Only the far-right National Front had an estimated left-right position that clearly distinguishes it from other parties. On the basis of these estimates we simply cannot say, notwithstanding CMP point estimates, whether the Greens (*Verts*) were to the left or the right of the Socialists (*PS*) in 2002. The role of uncertainty in cross-sectional comparisons will differ according to context, but the French case demonstrates—for a major European multi-party democracy—that inferences of difference from CMP point estimates can be ill-informed without considering measurement error.

2.5.2 Correcting estimates in linear models

When co-variates measured with error are used in linear regression models, the result is bias and inefficiency when estimating coefficients on error-laden variables (Hausman 2001, 58). These coefficients are typically expected to suffer from "attenuation bias," meaning they are likely to be biased towards zero, underestimating the effect of relevant variables. This conclusion must however be qualified, since it depends on the relationship between the "true" predictor and the noisy proxy available to the researcher, and possibly other variables in the model. More precisely, the effect of measurement error depends on the estimation model and the joint distribution of measurement error and the other variables (Carroll, Ruppert, Stefanski & Crainiceanu 2006, 41). In the case of linear regression the effects of measurement error can range from simple attenuation bias, to masking of real effects, appearance of effects in observed data that are not present in the error-free data, and even reversal of signs of estimated coefficients compared to the case in the absence of measurement error.

By far, the most common use of policy scales derived from CMP data tends to be as explanatory variables in linear regression models. Of all the studies using CMP data as co-variates in linear regression models, however, to our knowledge not a single one has explicitly taken account of the likelihood of error in CMP estimates, or even used the length of the underlying manifesto as a crude indication of potential error. As a result, we expect many reported coefficients in studies using CMP data to be biased.

We address this issue by replicating and correcting two recent high-profile studies using CMP data: Adams, Clark, Ezrow & Glasgow (2006), and Hix, Noury & Roland (2006). In both cases we obtained datasets (and replication code) from the authors and replicated the analyses, correcting for measurement error in CMP-derived variables. We do this using a simple error correction model known as *simulation-extrapolation* (SIMEX) that allows generalised linear models to be estimated with correction for error-prone co-variates whose variances are known or assumed (Stefanski & Cook 1995, Carroll et al. 2006). While not widely used in political science, SIMEX has been applied recently by Hopkins & King (2007) as a means to correct misclassification errors in text analysis. Here, by contrast, we apply the method to correct for random measurement error in observed co-variates.

The basic idea behind SIMEX is fairly straightforward. If a coefficient is biased by measurement error, then adding more measurement error should increase the degree of this bias. By adding successive levels of measurement error in a re-sampling stage, it is possible to estimate the trend of bias due to measurement error versus the variance of the added measurement error. Once the trend has been established, it then becomes possible to extrapolate back to the case where measurement error is absent. Following Carroll et al. (2006, 98–100) the SIMEX algorithm can be succinctly described as a sequence of steps that we illustrate in Figure 2.6. The example taken is the *EU Integration* variable from Hix, Noury & Roland (2006, Model 6) replicated fully below. First, in the simulation step additional random pseudo errors are generated from normal distribution with mean 0 and variance $\zeta_m \sigma_u^2$ and added to the original data. Since m is known and chosen to satisfy $0 = \zeta_1 < \zeta_2 < \ldots < \zeta_M$ (we use typical values $\{0.0, 0.5, 1.0, 1.5, 2.0\}$), the simulation step creates m data sets with increasingly larger measurement error variances. The total measurement error variance in the m^{th} data set is $\sigma_u^2 + \zeta_m \sigma_u^2 = (1+\zeta_m)\sigma_u^2$. In the estimation step the model is fit on each of the generated error contaminated data sets. The simulation and estimation steps are repeated a large number of times (500 times in our replication example) and the average is taken for each level of contamination. These averages are plotted against the values of ζ (the filled circles in Figure 2.6), and an extrapolant function is fit to the averaged, error-contaminated estimates. In terms of ζ_m an ideal, error-free data set corresponds to $(1+\zeta_m)\sigma_u^2 = 0$, i.e. $\zeta_m = -1$.[18] Extrapolation to the ideal case ($\zeta = -1$) yields the SIMEX estimate (the hollow circle in Figure 2.6). The quadratic extrapolant function is usually preferred, since it has been shown to result in more conservative corrections for attenuation and is often more numerically stable than the alternative nonlinear function (also shown in Figure 2.6) (Carroll et al. 2006, Hardin, Schmiediche & Carroll 2003, Lederer & Küchenhoff 2006). (In our replications below, we report corrections based on the more conservative quadratic extrapolation.)

More complicated error corrections are of course possible, but here we deliberately chose a method that is simple, applicable to a wide class of generalised linear models, and for which freely available software is available that can be used with popular statistical packages.[19]

[FIGURE 2.6 about here]

Adams, Clark, Ezrow and Glasgow (2006)

Adams et al. (2006) analyse whether political parties in Western Europe adjust their ideological orientations in response to shifts in voters' policy preferences. The authors extend the "dynamic representation" model by empirically analysing whether the type of political party affects the causes and consequences of their movements on policy. In particular the article is concerned with whether "niche" parties (typically Communists, Greens or extreme-right) respond differently to public opinion shifts compared to mainstream parties (e.g. Labour, Socialist, Social

[18] More precisely, for the case of simple linear regression $\widehat{\beta}_{x,naive}$ is the naive OLS estimate of β_x, and it consistently estimates $\beta_x \sigma_x^2/(\sigma_x^2+\sigma_u^2)$ and is biased for β_x when $\sigma_u^2 > 0$. The least squares estimate of the slope from the m^{th} data set, $\widehat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2/\{\sigma_x^2 + (1+\zeta_m)\sigma_u^2\}$. The ideal case of a data set without measurement error in terms of ζ_m corresponds to $(1+\zeta_m)\sigma_u^2 = 0$, and thus $\zeta_m = -1$. See Carroll et al. (2006) for full details.

[19] For R, the `simex` package is available from CRAN. Information on SIMEX implementation in STATA can be found at `http://www.stata.com/merror/`.

Democratic, Liberal, Conservative and Christian Democratic).

The first model analysed in the original article and replicated here deals with whether mainstream and niche parties differently adjust their policies in response to public opinion shifts. Party policy shifts are operationalised as changes in a party's CMP left-right scale position in successive elections. This measure is regressed on public opinion shifts, a dummy variable for niche party status, the interaction of these two variables, lagged dependent variable, lagged vote share change, the interaction of these two terms, and a set of country dummies. The authors' expectation is that if the coefficient on *Public opinion shift* is positive and statistically significant then mainstream parties are responsive to shifts in public opinion along the lines of the dynamic representation model. They also expect to find a negative and statistically significant coefficient on the *Niche Party* × *Public opinion shift* variable, providing evidence that niche parties are less responsive to public opinion shifts than mainstream parties, thereby supporting the main "policy stability" hypothesis of the article. In our replication of Adams et al. (2006, Table 1), we focus on the effect of measurement error in both the dependent variable on the left-hand side, its lagged value on the right-hand side, and an interaction of the lagged dependent variable and lagged change in vote share. In the classical measurement error (CME) domain it is known that measurement error in the dependent variable, if uncorrelated with other co-variates, will only inflate standard error of the regression (Abrevaya & Hausman 2004), while measurement error in independent variables will bias the results.[20] We assume here and in subsequent replications that all other co-variates are measured without error. The error estimate in contaminated co-variates is derived from our bootstrapped standard error.[21]

The second model in Adams et al. (2006) tests whether policy adjustments (shifts in policy towards the centre of the voter distribution or away from it) affect parties' electoral support and whether this relationship differs between mainstream and niche parties. Key explanatory variables are constructed from the CMP and thus are expected to be error-prone: *Centrist policy*

[20] In order to remain within the CME domain we assume that measurement error in first-differences in the dependent variable is uncorrelated with error in second-differences in its lagged value. The effect of measurement error in first-difference estimation in panel data models is much higher than in level models (Arellano 2003, 50), which may somewhat explain low reported R^2s.

[21] In this and the replications that follow, our error estimates for each error-prone co-variate is the mean of the in-sample average error variance from the bootstrapping procedure (and specified in the note to each table).

shift, Noncentrist policy shift, Niche Party × Centrist policy shift, Niche Party × Noncentrist policy shift. The first variable is measured as the absolute value of the change in party's position on the CMP left-right scale when a leftist party shifts right or rightist party shifts left, and zero otherwise. The variable measuring the shift away from the centre is similarly constructed. The next two variables pick up the differences in electoral effects for niche and mainstream parties in relation to centrist and non-centrist policy shifts.[22] Adams et al. (2006) expect mainstream parties to gain votes in the centrist policy shift and lose votes in non-centrist shift, thus leading to the expectation of a positive and statistically significant coefficient on *Centrist policy shift* and a negative and statistically significant coefficient on *Noncentrist policy shift*. The authors suggest that niche parties are electorally penalised for policy adjustments regardless of the direction of this adjustment (centrist or non-centrist) in what they call "costly policy shift" hypothesis. This leads to the expectation of statistically significant and negative coefficients on both *Niche Party × Centrist policy shift* and *Niche Party × Noncentrist policy shift*. At the same time another hypothesis put forward by Adams et al. (2006) states that niche parties lose votes in comparison to mainstream parties for moderating their policy stance ("costly policy moderation" hypothesis). In turn this results in the expectation of negative and statistically significant coefficient only on the *Niche Party × Centrist policy shift* variable.

[FIGURE 2.7 about here]

Figure 2.7 presents results of our error correction for both models, taken from the two regression tables of Adams et al. (2006).[23] For each model, we compare our replication of the published results with SIMEX estimates.[24] The most profound effect of SIMEX correction of Model 1 is the expected inflation of the standard error of the regression and drop in explained

[22] Two additional control variables are based on CMP measures: *Party policy convergence and Party policy convergence × Peripheral party*. The former is operationalised as the sum of all centrist policy moves by all parties in the system. The latter is an interaction of *Party policy convergence* with a dummy variable for parties taking extreme position on left-right dimension. In addition to these six error-prone co-variates, Model 2 in Adams et al. (2006) contains dummy variables for niche parties, governing parties, coalition governments, previous change in vote share, as well as several economic control variables: changes in unemployment and GDP rates and their interaction with governing party dummy.

[23] For both replication studies, we present results in graphical form, following the suggestions (and using code from) Kastellec & Leoni (2007). Thus Figure 2.7 is presented in tabular form in Table 2.2.

[24] Our replications compare our corrected estimates to replicated rather than published estimates, since replicated and published results differ slightly due to slight errors in data preparation in each published analysis.

variance as the consequence of measurement error in the dependent variable. The effect of error-correction in the co-variates decreases the key explanatory variables in size but not to such an extent that their statistical significance is affected. The full extent of SIMEX error correction effects can be gleaned from changes in coefficients and standard errors presented in Figure 2.7 and Table 2.2. It is, however, obvious that weakness in the explanatory power of Model 1 calls for caution in suggesting that results "consistently support the Policy Stability Hypothesis" (525). There is indeed some support that niche parties' policy programs are less responsive to shifts in public opinion compared to mainstream parties (the grey row in Model 1). Evidence for this claim, however, is drawn from a somewhat weaker set of corrected estimates.

In the original article, the negative and statistically significant coefficient on *Niche Party* × *Centrist policy shift* (Model 2) is meant to support the "costly policy moderation" hypothesis that, in comparison to mainstream parties, niche parties are penalised by voters for moderating their policy positions. Results in the original article substantively mean that a one unit shift closer to the centre of the voter distribution along the 1–10 Left-Right scale, results, *ceteris paribus*, in niche parties' electoral loss of nearly 4% (i.e. approximately $-5.67 + 1.45$, see p523). This conclusion is cast into doubt as the result of the SIMEX correction, which causes the coefficient on *Niche Party* × *Centrist policy shift* to become smaller in size and statistically insignificant at the conventional 0.05 level. In turn, this forces the rethinking of some of the theoretical implications of the article. The conclusion that for niche parties "*both* vote-seeking and policy-seeking objectives motivate a stand-pat strategy" (525, emphasis in original), since moderation in policy positions is penalised by voters is not supported by empirical evidence based on the error-corrected estimates.

Moreover, Adams et al. (2006, 525) claim that their empirical results support the "costless spatial mobility" assumption typically used in spatial modeling – i.e., that political parties are not electorally penalised for shifting positions in policy space – with respect to mainstream parties. In fact, as Figure 2.7 and Table 2.2 show, the corrected coefficient for *Noncentrist policy shift* almost doubles as the result of the SIMEX correction. Indeed, if a one-tailed hypothesis test were applied to the coefficients for both *Noncentrist policy shift* and for *Niche Party* × *Centrist policy shift*, both would be considered statistically significant. In terms of the conclusions of the original article, the error-corrected results challenge its categorical conclusion that mainstream parties are not penalised for shifting policies away from the centre—suggesting that this effect occurs with at least as much confidence as the conclusion that niche parties are punished for shifting their policies to the centre.

Hix, Noury, and Roland (2006)

Hix, Noury & Roland (2006) are concerned with the content and character of political dimensions in the European Parliament (EP). Following an inductive scaling of roll-call votes in the EP from 1979 and 2001, Hix, Noury & Roland (2006) set out to validate their interpretation of the derived policy dimensions by regressing the mean position of each national party's delegation of MEPs on two sets of independent variables. The first set includes exogenous measures of national party positions on the left-right, social and economic left-right, and pro-/anti-EU dimensions. The second set relates to government-opposition dynamics and consists of categorical variables describing whether a national party was in government and whether the party had a European Commissioner, as well as dummy variables for each European party group, each EU member state, and each (session of) European Parliament. Measures of national party positions are taken directly from the CMP dataset or constructed from it. National party positions on the EU are taken as the difference between positive (category PER108) and negative mentions (category PER110) mentions of the EU. Party positions on economic and social policy are also constructed from the CMP categories (see Laver & Garry 2000, 628-629). The authors expect that national party ideal point estimates on the first dimension will be explained by the exogenous left-right policy positions, while exogenous policy positions on EU Integration dimension explain national party ideal point estimates on the second dimension. (501) The expectation then is roughly that the first dimension is predominantly about left-right and second dimension is about Europe.

[FIGURE 2.8 about here]

Figure 2.8 contrasts coefficients from our replications of the models using CMP variables in Hix, Noury & Roland (2006) with error-corrected measurements based on our bootstrapped variances.[25] We present here replications of the two models that are related to the structure of the first dimension in the European Parliament. Model 2 aims to explain the mean positioning of political parties on the first derived EP dimension in terms of: their positions on the general left-right, and European Integration dimensions; categorical variables relating to whether a party was in government and had a European Commissioner; and dummy variables for each session of the EP. Model 3 substitutes general left-right with a combination of economic left-right, and social left-right. Model 6 extends Model 3 also to include dummy variables for each European party group. Model 9 extends Model 6 including country dummy variables.

It is clear from Figure 2.8 that the SIMEX error correction has the most important effect on the "EU Integration" variable. The SIMEX estimate of *EU Integration* is about double the size of the naive estimate in Models 2 and 3 presented, and becomes statistically significant in the corrected estimates of Models 6 and 9. Substantively, the effect of noise in the CMP measure of EU policy is that, if we set out to explain the position of a party's MEP delegation, the national party's position on the EU is shown to be *more* important than its position on the substantive economic and social left-right dimensions, rather than unimportant as Hix et. al. conclude. SIMEX correction of the key *EU Integration* variable thus forces a rethinking of some of the substantive conclusions of this article. In the words of Hix, Noury & Roland (2006) interpreting their results from the naive model:

> EU policies of national parties and national party participation in government are only significant without the European party group dummies. This means that once one controls for European party group positions these variables are not relevant explanatory factors on the first dimension. (502)

In a direct challenge to this conclusion, results from the error-corrected model suggest that EU policies of national parties appear not to be relevant only because of attenuation bias caused by noise from the textually derived CMP measures of positioning on EU policy. Once this error is corrected for, the primary dimension of EP voting is shown to be influenced even more by EU policy than by general left-right positions.

[25] Core results are also presented in tabular form in Table 2.3.

2.6 Categorising Human Misclassification

In Figure 2.1 we have described the full process generating the CMP dataset; here our focus is on the coding category scheme and the way that human coders assign these categories to each text unit. That is, we concentrate here on the stochastic *coding* process, C, by estimating the extent of variation between coders in applying the same coding scheme I to the same text τ.

2.6.1 The CMP Coding Scheme and Sources of Disagreement

CMP estimates of the policy position of a particular party on a particular matter at a particular election are generated by using a trained human coder to allocate every "text unit" in the party's manifesto into one, and only one, of 57 policy coding categories (one of which is "uncoded").[26] The first CMP coding category, for example, is "101: Foreign special relationships: positive". Having counted text units allocated to each category, the CMP then uses its theoretical "saliency" model of party competition to inform a measurement model that defines the relative salience for the party of the policy area defined by each category as the percentage of all text units allocated to that category. The variable PER101 in the CMP dataset, therefore, is the percentage of all text units in a party manifesto allocated by the coder to "101: Foreign special relationships: positive". Fortunately for third-party users of the CMP data, it is not necessary to buy into CMP's distinctive "saliency" model of party competition before using the data. This is because the CMP coding scheme does not in fact comprise pure saliency categories. All but one of the 56 substantive categories (the exception is "408: Economic goals") are *positional*, in the sense that category definitions explicitly refer to a position on the policy issue concerned, not just a mention of this issue. Thus we have both "406: Protectionism positive" and "407: Protectionism negative" when a pure saliency coding scheme would imply just "protectionism: positive, neural or negative". It is, precisely, the *positional* nature of such policy codings that led to the widespread use of the CMP dataset by scholars seeking time-series estimates of

[26] In the extended coding scheme developed in *MPP2* to allow subcategories to be applied to manifestos from Central and Eastern European countries plus Mexico, an additional 54 subcategories were developed, designed to be aggregated into one of the standard 56 categories used in all countries. For the purposes of computing indices such as `rile`, however, the subcategories were *not* aggregated or used in any way. For these reasons and the general wish to keep the focus as simple as possible in this paper, our analysis here is restricted to the original 56 + uncoded standard CMP categories.

party policy *positions*, as opposed to time series estimates of the *salience* attached by parties to particular policy issues.

In what follows, we leave for future work the potential for *coding bias*, which arises because human coders are inevitably aware of the authorship of the texts they are coding, a problem especially acute for highly self-referential documents such as party manifestos. We deal above with non-systematic measurement error in CMP data that arises from *stochastic features of the text generation process*. Here, we focus on error arising in CMP data from *stochastic features of the text coding process*.[27] We refer to this coding error in general terms as *misclassification*.

2.6.2 Coding differences from human "features"

CMP data are fundamentally susceptible to coding error because, in their essence, they derive from subjective judgements made by human coders. These days, indeed, human coding is preferred to machine coding in settings where it is explicitly felt that subjective coding by human experts is more valid than objective coding by machines. Coding error arises because different human coders at the same time, or the same human coder at different times, are likely to code the same text in somewhat different ways. This process may be unbiased, in the sense that we can think of an unobservable "true and certain" value of the quantity being measured, with each human text coding being a noisy realisation of this. Assuming unbiased coding, we can take the mean of the noisy realisations as an estimate of the unobservable latent quantity, and the variation in these observations as a measure of the uncertainty of this estimate.[28]

The CMP data, however, are generated by party manifestos coded once, and once only, by a single human coder. There is no variation in noisy realisations of the unobservable underlying quantity and thus no estimate can be formed of the uncertainty of CMP estimates arising

[27] Note that another known source of random variation in the coding process is the difference in the unitisation of texts by CMP coders into "quasi-sentences," the basic text unit for the CMP scheme. Based on reports in Volkens (2001*a*, 38) preliminary analysis indicates that unitisation variance in the CMP manifesto is typically on the order of +/-10% of the total quasi-sentence units in a text, and doesn't seem to have any substantive influence.

[28] We do not deal here with a deep and interesting possibility that has largely been ignored, that the latent quantity being measured has an uncertain value—in this context that party policy on some issue is vague. In this case, it may be that variation in realisations of this latent quantity arises not just from measurement noise, but from fundamental uncertainty in the quantity being measured.

from coding errors. In a nutshell, we have no way of knowing whether subsequent codings of the same manifesto would be exactly the same as, or completely different from, the recorded coding that goes into the CMP dataset. We are very confident, however, on the basis of both anecdotal evidence and good old fashioned common sense that, if there were to be a series of independent codings of the same manifesto, then these would all differ at least somewhat from each other. Indeed, if someone reported that 1,000 highly trained coders had each coded 10,000 manifesto text units using the CMP's 57 category scheme, and that every single coder had coded every single text unit in precisely the same way, then our overwhelming suspicion would be that the data had been faked.

2.6.3 Coding differences from category ambiguities

Just as our hypothetical French experts and chimpanzees might tear out their (body) hair trying to assign the given categories to text units that do not neatly fit, CMP coders often report difficulties determining precisely which of the 56-plus-uncoded categories to assign to text units. Hence an important source of coder error are the ambiguities and overlap that exist in the way that some of the categories are defined. Consider the distinction between the following categories:

> "401: Free enterprise: Favourable mentions of free enterprise capitalism; superiority of individual enterprise over state control systems..."
>
> "402: Incentives: Need for wage and tax policies to induce enterprise..."

There is of course a difference between these category definitions but it is easy to imagine text for which the coder's decision as to which category is most appropriate would be a knife-edge judgement, one that would be made in different ways by different coders. In contrast "501: Environmental protection" is essentially the only CMP coding category making explicit reference to the environment, so there is nowhere else in the scheme to allocate text units referring to the environment (a decision that, incidentally, renders any anti-environmentalist statements uncodable by the CMP). Any text coding scheme must be viewed as a whole, taking into account overlaps and the sharpness of boundaries between categories as well as the definitions of each category on a stand-alone basis. However, we do expect some CMP coding categories to be more "reliable" (different coders tend to code the same text unit into the category in question) than others (different coders do not all use the category in question for the same text unit.) As we shall see, this is very much what we find in our coding experiments.

In practice the full 56-category coding scheme is never deployed on any one manifesto and the norm is for far fewer than the full set of categories are used in the coding of a typical manifesto. Figure 2.9 characterises the distribution of categories used across the entire set of 3,018-manifestos coded by the CMP. The typical manifesto coding uses only 25 categories, less than half of those available. Coding category usage ranges from startlingly mono-themed manifestos such as the 1951 Australian National Party manifesto which consisted of 42 text units all assigned to a single category ("703: Farmers Positive"), to a maximum of 51 different categories used to code the 365 quasi-sentences found in the 1950 British Conservative Party manifesto.

[FIGURE 2.9 ABOUT HERE]

Figure 2.10 shows the relative frequency, in *log* percentages, of text units allocated to different coding categories, from all text units coded in the consolidated CMP dataset (excluding CEE countries and Mexico that use the extended subcategories). The horizontal bars indicating frequency are grouped into three categories: those that are designated as "right" or "left" in the additive left-right index most commonly used by CMP consumers (called `rile` by the CMP), or "neither" meaning the category is not used in building the left-right index. To facilitate comparison among low-frequency categories, the percentage frequencies have been transformed using base-10 logarithms, which also serves to highlight differences in the categories used overall less than 1% of the time, shown to the left of the origin.

[FIGURE 2.10 ABOUT HERE]

2.6.4 From categories to scales

One response to overlapping or vague boundaries between text coding categories is to combine these, to produce a more reliable aggregate category. In addition, what amounts to the 56-dimensional policy space measured by the CMP manifesto codings is cumbersome to use as an operationalisation of specific models of party competition. Furthermore, as a matter of practical fact, most third-party users of CMP policy time series data are looking for something much simpler; nearly all of them, indeed, are looking for party positions on a simple left-right scale. In terms of Figure 2.1 this would be a representation of a scaling model *S* producing a set of scales λ.

In response to these interlocking demands, the CMP is best known for its left-right `rile` scale, λ, which the CMP itself calls its "crowning achievement" (Budge et al. 2001, 19). The scaling model, *S*, behind `rile` is a simple additive index based on aggregating 13 coding categories seen as being on the "left", 13 seen as being on the "right", and subtracting the percentage of aggregated left categories from those of the right. The theoretical range of this scale is thus [-100, 100], although in practice nearly all `rile` scores span the scale's middle range of [-50, 50]. In practical terms, therefore, two different types of classification are classification into the three aggregate categories of left, right, or neither. The `rile` scale is thought to be more reliable than any single coding category, since it is likely that most of the stochastic variation in text coding will result from different coders allocating the same text unit to different categories on the "left" or the "right". From the perspective of the left-right scale that most third-party users are interested in, such coding "errors" are thought to be in effect self-cancelling.[29] In our tests below, we critically examine this claim.

[29] This problem, which the CMP has termed "coding seepage" (Klingemann et al. 2006, 112), is thought to mainly take place in between categories within the same aggregate categories. Analysis of coding decisions conducted by the CMP team suggests several categories prone to systematic misclassification. Thus coding categories that have been identified as "seeping" codes (in brackets): Per101 (Per104), Per302 (Per303 and Per305), Per504 (Per503), Per601 (Per606), Per603 (Per605 and Per606), Per607 (Per705 and Per706); Per102 (Per103), Per105 (Per106 and Per107), Per505 (Per303), Per507 (Per303), Per702 (Per704), Per412 (Per403 and Per413), Per409 (Per404). (Klingemann et al. 2006, Table 6.1:114) Earlier investigation also identified per408 (per410) and per402 (per703) (Volkens 2001*a*, 38). The majority of "seepage"-prone categories belong to the same aggregate scales, however, prompting the CMP to recommend their "own preferred strategy" of using the aggregate scores to limit the effect of single category misclassifications. Because the components of the `rile` index "combine closely related categories, the coding errors created by ambiguity between these are eliminated. The overall measures are thus more stable and reliable than any one of their components" (Klingemann et al. 2006, 115). Other, lesser-used combined scale categories are "planeco," "markeco," and "welfare," representing the orientation towards a planned economy (403+404+412), a market economy (401+414), and the state provision for welfare (503+504) respectively.

2.6.5 Strategies to maximise reliability

Previous work investigating the reliability of the CMP scales has focused on different and quite distinct aspects of the issue. The CMP's approach to the coding reliability issue is to focus on procedures of coding used in data production (Klingemann et al. 2006, 107). Possible problems of coding error that we discuss below are approached by "setting and enforcing central standards on coders" by getting the coders to conform to a particular English-language standard and also by constant communication and interaction with the supervisor in Berlin (Volkens 2001*b*, 94). The focus thereby is on coder "training" (see Volkens 2001*a*, 37-40). Specifically the CMP has done this by setting out to train all CMP coders to code the same two manifestos in the same way as a CMP "gold standard" coding that is taken to reflect a "certain truth" about the policy positions expressed in those manifestos.

The CMP has invested great effort into improving the quality of its manifesto coding. Based on the first evaluation of test results, a new version of coding instructions was produced (Volkens 2007, 118).[30] The revised instructions draw particular attention to three specific ambiguities in the CMP coding scheme affecting coding reliability: when no category seems to apply to the quasi-sentence, when more than one category seems to apply, and when the statement in the quasi-sentence is unclear (Klingemann et al. 2006, 170). Several solutions to these problems are offered in the coding manual. When no category seems to apply the quasi-sentence may be marked as uncodable, with categories used seldom being the most difficult to code (Klingemann et al. 2006, 170). When more than one category seems to apply the manual offers eight decision rules ranging from re-reading the categories description, identifying connecting sentences, creating subcategories, checking section headings as cues to more explicit rules that specific categories should be chosen ahead of more general categories (e.g. per305 "political authority" and per408 "general economic goals"). When the statement seems unclear the coder is advised to seek cues from the context and/or contact the supervisor in Berlin.

Other investigations of reliability have specifically targeted possible error in the the ag-

[30] Hearl (2001) investigated possible coding differences following the structural change that happened in 1983 with the transition to the CMP from the original MRG set up. He finds no evidence of methodological error across that "fault line" with comparable analyses producing the same results in the subsample before 1983 and dataset as a whole.

gregated indexes, namely `rile`. McDonald & Mendes (2001*a*) and Klingemann et al. (2006, Chapter 5) focus on the issue of measurement error in the `rile` scale as an approach to assessing reliability. Exploiting the panel structure of the data set and using the Heise (1969) measurement model, the authors claim to be able to sift out measurement error from real change. From the results, and making some pretty strong theoretical assumptions and assumptions about the latent reliability structure, they conclude that `rile` is effectively very close to being perfect (Klingemann et al. 2006, 103). Such tests focus on very different issues from those of stability and reproducibility faced here, however, where our primary concern is whether coders can reliably implement the CMP coding instructions without serious misclassification errors.

2.7 A Framework for Stochastic Misclassification of Text Categories

Misclassification is a central concern in many fields, particularly in medicine where "coding errors" can mean the difference between avoiding an unnecessary, costly, and invasive procedure and dying from cancer. In this view, each unit (or "subject") belongs to some objectively "true" category, although our coders (or "raters") can only approximate this true category by assigning it a category according to their best judgement. The difference between the true and assigned category is misclassification, and this misclassification, to the extent that its realisation differs between coders, will reduce reliability of the coding procedure. Note that while we take the position that there is indeed a "true" category to which each sentence belongs—even if no human coders can agree on precisely what this is—reliability as we have defined above it depends only on coder *agreement*, not on coder adherence to some perfect (and possibly unknowable) standard. Because the entire foundation of the CMP approach is that each text unit can be assigned to either a given category or declared "uncoded," however, this implies the existence of a "true" coding, and all evidence so far uncovered points to coders making stochastic misclassifications roughly around these true categories. Without getting into the ultimately metaphysical questions about what proportion of gold to pure brass exists in the CMP's notion of a gold standard, therefore, we take the existence of such a standard as given and proceed on that basis.

Our discussion here follows the framework of Kuha & Skinner (1997) and Bross (1954). In formal terms, let the true categories of each text unit i be represented by A_i, whose values are well-defined and fixed, but classified with error as A_i^*. Misclassification occurs through a stochastic process

$$\Pr(A_i^* = j|A_i = k) = \theta_{jk} \tag{2.6}$$

where $j,k = 1,\ldots,m$ for m possible (nominal) classification categories. The key to this process is the parameter θ_{jk} which may be viewed as the proportion of population units in the true category k that would be represented by coders as category j. These parameters θ_{jk} form a misclassification matrix Θ of dimensions $m \times m$ whose elements are all non-negative and whose columns sum to one.

$$\begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} = \begin{pmatrix} \beta & 1-\alpha \\ 1-\beta & \alpha \end{pmatrix} \tag{2.7}$$

If our coding scheme were binary, as in the Sarkozy example, then the resulting 2×2-dimension Θ can be decomposed into two characteristics commonly known in the medical literature (see e.g. King & Lu 2008, Rogan & Gladen 1978) as *sensitivity* and *specificity*. Sensitivity is represented by α, and refers to $\Pr(A_i^* = k|A_i = k)$, or in the Sarkozy example, the probability that a sentence coded as "left" is really "left", or is coded as "right" when really "right". In the language of hypothesis testing, $(1-\alpha)$ will be familiar as the probability of a Type I error, the probability of rejecting a null hypothesis when it is really true. Specificity, on the other hand—represented by β—refers to $\Pr(A_i^* \neq k|A_i = k)$, the probability that a sentence is classified as "right" when it is really "left" or vice-versa. In the language of hypothesis testing, specificity refers to the risk of committing Type II error, or failing to reject a null hypothesis when the alternative hypothesis in fact true.

If a coding scheme could be applied to text units perfectly, then Θ would consist of an $m \times m$ identity matrix. To the extent that there are off-diagonals in Θ, however, random misclassification will reduce reliability, and when these off-diagonals are non-symmetric, the result will be systematic misclassification or bias. We can estimate this degree of bias by examining the effect of misclassification on our estimates of manifesto categories. Following Kuha & Skinner (1997), for a text, let N_j^A be the number of text units for which $A_i = j$, and let $P_j^A = N_j^A/N$, where $N = \sum N^A$ is the total number of text units. Our objective is to estimate the vector $\mathbf{P}^A = (P_1^A, \ldots, P_m^A)'$ of proportions of each category of manifesto code from the coding scheme, for our given text—in other words, the CMP's "per" variables.

With misclassification, $\mathbf{P}^A$ will be generally estimated by some vector $\mathbf{p}^{A*} = (p_1^A, \ldots, p_m^A)'$ where $p^{A*} = \sum I(A_i^* = j)$ and $I(\cdot)$ denotes the indicator function. It follows from (2.6) that

$$\mathrm{E}[I(A_i^* = j)] = \sum_{k=1}^{m} \theta_{jk} I(A_i = k) \tag{2.8}$$

so that the vector of "true" counts $\mathbf{p}^A$ in each coding category is related to what we expect to observe by

$$\mathrm{E}(\mathbf{p}^{A*}) = \Theta \mathbf{p}^A \tag{2.9}$$

The bias from misclassification will then be expressible as

$$\mathrm{Bias}(\mathbf{p}^{A*}) = (\Theta - \mathbf{I})\mathrm{P}^A \tag{2.10}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. Our task in assessing misclassification and the unreliability of the coding procedure that follows, therefore, is to obtain estimates of the misclassification matrix Θ. To the extent that this misclassification matrix differs from identity, categories are likely to be misclassified, yielding unreliable and potentially biased estimates of the content of the texts being coded.

2.8 An Experiment to Assess Coder Agreement

2.8.1 Methods and Data

Our method for evaluating misclassification and reliability in the CMP coding procedure was to perform a simple experiment: to see how much agreement could be obtained by multiple coders applying the CMP scheme to the same texts. Our experiment employed two texts, both taken from the "Manifesto Coding Instructions" provided in Appendix II to Klingemann et al. (2006). Apart from detailed instructions for coders, Appendix II also contains two fully coded sample texts designed to serve as examples. Using these two texts held several key advantages. First, each text had already been "officially" parsed into quasi-sentences by the CMP, meaning that we could take the unitisation step as given, and focus in the experiment only on the assignment of codes to each quasi-sentence. Second, because each text was also officially coded by the CMP, the CMP codings serve as a "gold standard" for comparing to tester codings. Finally, since these two texts had been chosen for their clarity and codeability to be instructional examples, they also made good texts for comparing tester agreement in our experiments.

The first sample text is an extract from the UK The Liberal/SDP Alliance 1983 manifesto. The text consists of 107 text units coded by the CMP into 19 categories. The second sample text is an extract from New Zealand National Party 1972 manifesto, containing 72 text units coded by the CMP into 11 categories. The National Party manifesto text contains only one unique code not present in The Liberal/SDP Alliance manifesto text. Overall, therefore, our reliability experiment could effectively estimate coder bias and misclassification in relation only to 20 out of 57 available categories, although these categories were among the most common of those found in most manifestos (see Figure 2.10).

Our test was set up on a dedicated web page containing digitised versions of sample texts, already divided into quasi-sentences. Each page also contained detailed instructions adapted directly from from "Manifesto Coding Instructions" in Appendix II to Klingemann et al. (2006). Coders were asked to select for each text unit an appropriate category from a scroll-down menu. We also collected some minimal information on coder identifiers and previous experience in coding manifestos. Only completed manifestos could be submitted into the system. Going for a mix of experience and youth we sent out invitations to participate in our experiment to the majority of trained CMP coders[31] and a selection of usual suspects: staff and

[31] Andrea Volkens has kindly provided us with a list of names of 84 CMP coders of which 60% were matched with email addresses. We also used publicly available e-mail addresses of coders trained by the CMP for a separate *Euromanifestos Project*. (See Wüst, A. and A. Volkens, *Euromanifesto Coding Instructions*, MZES.)

postgraduates at several European and North American universities. We ended up with a list of 172 names with active emails who were randomly assigned to one of the two test documents.

Our response set consisted of 39 coders, but some of these results were discarded. To be as fair as possible to the CMP, we discarded the bottom fourth of test coders in terms of their reliability, while dropping none from the top. Overall, the New Zealand manifesto was completed by 12 coders and the UK manifesto by 17. The coders whose results are reported here had a mixture of prior experience with coding manifestos using the CMP scheme, without any discernible pattern of more experienced coders being near the top, in terms of agreement with the gold standard.

2.8.2 Methods of Assessing Agreement

Previous analysis of inter-coder variation, coder bias, and misclassification can only be characterised as limited. The extent to which coder training was successful was measured on the aggregate level in relation to only one test manifesto. Reliability was calculated for each trainee as a correlation between the percentage of quasi-sentences coded into each category and the CMP "gold standard". Depending on which test we are talking about, these correlations range from 0.70 to 0.80. For 23 coders that were trained from the the second version of coding manual, their average correlation with the "gold standard" was reported to be 0.83. Of these coders fourteen were new hires taking the test for the first time. Their average correlation with the master copy is 0.82. Nine coders on the second contract took the test again with results for this group going up from 0.70 in the first round to 0.85 in the second round (Volkens 2007, 118). Klingemann et al. (2006, 107) report that coders on another contract retaking the test showed an average correlation coefficient of 0.88. These reported results are collected in Table 2.4.

[TABLE 2.4 ABOUT HERE]

Several serious issues with these reported results become immediately apparent to anyone who has ever used the CMP data. The key issue in reliability tests taken by the CMP coders is whether they agree on unitisation and categorisation of text units with the "gold standard". There is a clear distinction, however, between measuring *agreement* and measuring *association*. Strong association is required for strong agreement, but the reverse is not true

(Agresti 1996, 243).

The association measure reported by the CMP is the Pearson product-moment correlation aggregate for all categories, thus precluding disambiguation of category specific reliability. Furthermore, product-moment correlation only measures the degree of *linear trend* between two (at least) ordinal variables. That is, the degree to which values of one variable linearly predict values of the other variable. On the other hand, measures of *agreement* assess the equality of two variables and not linear prediction. If a coder *consistently* miscategorises quasi-sentence of a particular type, then association with the "gold standard" will be strong even though the strength of agreement is poor. Moreover, the Pearson product-moment correlations are not applicable for nominal-level data, which is the case in the analysis of (mis)coding of individual text units. For these reasons, states Krippendorff (2004, 245), correlations should be avoided since "in content analysis their use is seriously misleading".

Another problem with the CMP's coder reliability data concerns the issue of zero-category inflation. As discussed earlier and clearly shown in Figures 2.9 and 2.10, for any given manifesto only a small subset of the available categories tend to be used. The test manifesto used by the CMP to assess reliability is no exception, and since the correlation vectors from the CMP's reliability are indexed by category, this means a majority of the elements in the correlation vectors will have zeroes. The effect is to register high correlations based not on how well coders agree on applicable categories, but how well they agree on categories that clearly do not apply (such clear agreement on the absence of any EU-category quasi-sentences in the 1966 New Zealand training document).

Beyond the measures of association there are standard measures of agreement that are used extensively in the literature on content analysis. One standard measure is Krippendorff's α, which is "the most general agreement measure with appropriate reliability interpretations in content analysis" (Krippendorff 2004, 221). Outside the content analysis literature by far the most widely used method of statistical analysis of agreement for categorical variables is the κ measure (Roberts 2008, 811). Hayes and Krippendorff (2007) compare Krippendorff's α and Fleiss' κ and suggest that they are very similar. We also find that in most practical contexts both measures produce essentially identical coefficients. Both α and κ coefficients have a range from zero (perfect disagreement) to one (perfect agreement). Both measures also take into account the fact that some agreement may occur purely by chance.

It should be noted that there are two major issues with applying any measure of agreement and association to the CMP reliability results. First, since unitisation differs between coders, as it does +/-10% in the tests reported in Table 2.4, it is not clear on what, if anything, the coders are supposed to agree on. Second, coders report only aggregate percentages for each category leaving open the question whether coders actually agreed on codes applied to individual text units. Only by fixing the units and analysing agreement at the category level as in our experiment can true reliability be assessed, something which our test controls for.

The CMP group prefers to focus on reliability of composite indicators on the basis of their performance within the data set (Klingemann et al. 2006, 107). Reliability results for individual estimates are viewed of limited importance with the emphasis placed on general tendencies and patterns (2006, 108). Although it has been declared that "the data-set as a whole is reliable" (2006, 108), we believe that reliability can only be assessed by data that is additional to the data whose reliability is in question (Krippendorff 2004, 212). In the case of the CMP, this means analysing reliability data obtained through duplication of coding exercise by several independent coders.

2.9 Results of the Coding Experiment

2.9.1 Inter-coder Agreement

Reliability, state Hayes & Krippendorff (2007, 78), "amounts to evaluating whether a coding instrument, serving as common instructions to different observers of the same set of phenomena, yields the same data within a tolerable margin of error. The key to reliability is the agreement among independent observers." Applied to the CMP, reliability refers to the extent that different coders, coding the same manifesto independently, are able to agree on the categories to which each quasi-sentence belongs. In what follows we therefore report the simplest and easiest to test indicator of the CMP coding's reliability: how well different test coders agreed with one another when assigning categories to each quasi-sentence. Note that in assessing this form of reliability, we need make no reference to the master or “true” coding at all. If we find significant coder disagreement, then we can directly conclude that misclassification is occurring, since by necessity not every disagreeing coder can be correctly classifying each text unit.

Perfect reliability is never to be expected, but there are some widely agreed guidelines for interpreting our primary reliability measure κ. It is usually accepted that $\kappa = 0.80$ is the threshold above which a research procedure is considered to have an acceptable reliability. In the context of content analysis Krippendorff (2004, 241) suggests not to rely on variables with reliabilities below 0.80, and to consider variables with reliabilities between 0.667 and 0.80 only for drawing tentative conclusions.[32]

[TABLE 2.5 ABOUT HERE]

The results of our reliability scores from test coder results are summarised in Table 2.5. The table reports results for the British manifesto, the New Zealand manifesto, and the two combined. The first column reports κ for all coders by category. In theory, each quasi-sentence could have been rated by each coder as belonging to any one of the 56 policy categories or classified as “uncoded”, although in practice many categories were never used by any coder (for instance there was never any question that any of the categories might have dealt with the European Union, categories 108 and 110).

Because each category also plays a role in the definition of the CMP’s centrally important `rile` index—being one of the 13 left or 13 right categories, or one of the 31 categories that is neither—we also compared the “`rile` category” assigned by each coder to the quasi-sentences, reported in the second column (“By RILE”) of Table 2.5. This allowed us to test whether reliability could be improved—as expected by the CMP—when only this reduced set of three categories was used. By this view, two coders assigning “403” and “404” to the same quasi-sentence would be viewed in perfect agreement, since both of these categories are classified as “left” in the `rile` scale.

Finally, for the categories that the CMP’s master coding identified as being present in

[32] In a slightly more lenient set of guidelines, Fleiss (2003, 604) following Landis and Koch (1977) proposed guidelines for interpreting the kappa statistic with values above 0.75 may be taken to represent excellent agreement beyond chance, values below 0.40 show poor agreement beyond chance, and intermediate values represent fair to good agreement beyond chance.

the test manifestos, we are also able to report individual κ statistics for the reliability of each category. These figures are shown in the bottom part of Table 2.5, indicating how well different coders could agree on quasi-sentences being designated as specific categories, by category. The results are broadly consistent with the summary results, although several exceptionally unreliable categories stand out. From the left side of the `rile` scale, “202: Democracy Positive” is extremely poor, with $\kappa = 0.18$, as are “701: Labour Groups: Positive” and “Economic Planning: Positive”. On the Right, “605: Law and Order: Positive” and especially “305: Political Authority: Positive” are flagged by our experiment as being extremely unreliable. In general, categories identifying broad policy objectives such as “economic goals” seem to be very highly prone to inter-coder disagreement when it comes to assigning them to specific text units.

Overall, these results show that regardless of whether coders are compared in the full category tests or on the reduced three-fold `rile` classification, rater agreement is very low. In the test using the British manifesto the agreement is estimated between 0.35 and 0.36. In the test using the New Zealand the agreement is higher between 0.40 and 0.47. The test on the reduced three-fold `rile` classification showed no differences for the British text, but was slightly higher in the New Zealand test. When both sets of results were combined, the results were even lower, at 0.31-0.32. These figures are undeniable evidence that even after receiving detailed instructions, and even when at least one-third of our test coders have previous experience with coding manifestos for the CMP, reliability for the CMP scheme is significantly below conventionally acceptable standards.

2.9.2 Coder Agreement with the Master

Another way to assess reliability is by comparing the agreement of each coder with the CMP’s master coding, taking the master coding as a “gold standard” representing the correct set of categories. Indeed, this is the standard benchmark applied by the CMP in previous tests of reliability (e.g. Volkens 2007, Klingemann et al. 2006, 107). If the training process has succeeded and coders are successfully able to apply the coding scheme to actual text units, then their agreement with the master coding should be high. Agreement with the master coding can also be taken as a measure of the errors introduced by the difficulty of the coding scheme.

[FIGURE 2.11 ABOUT HERE]

The results of our tests were not encouraging. For the British manifesto test, the New Zealand manifesto test, and combined tests respectively, the median κ test coders' agreement with the master were 0.43, 0.54, and 0.46 respectively. The best coder agreed 0.74 with the master, and the worst 0.22. The full results are portrayed in Figure 2.11, separated by test. This histogram shows the frequency of different levels of κ for coder-master agreement from the 17 and 12 coders for the British and New Zealand texts respectively. The solid black line indicates the median results (0.42 and 0.54) from each test. For comparison with the conventional minimum level of acceptable reliability, we have also plotted a dashed red line indicating the conventional 0.80 cutoff for acceptable reliability. As can be clearly seen, the main density of the distribution of results for individual coders was well below standard levels of reliability, on both test documents.

2.9.3 Misclassification

Comparing the different coders' categorisations of the same text units not only allows us to estimate reliability, but also allows us to characterise precisely the nature of this misclassification. Using the master codings as an external validation sample, we are able to determine for each "true" category, what the probabilities were that test coders would assign a text unit to the correct categories versus incorrect categories. In the earlier language of or framework for misclassification, we are able to use the empirical 57×57 matrix of true versus observed codings to estimate the misclassification matrix Θ. By Equation (2.10), we know that the size of the off diagonals (or $\Theta - \mathbf{I}$) will determine the difference between the true categories A_i and the observed categories A_i^*.

In order to make the misclassification matrix manageable, we have reduced the focus to the probability that individual categories will be misclassified in terms of the three-fold `rile` classification. Looking at misclassification in this way tests the CMP assertion that errors in classification will be "self-cancelling," and also focuses attention on important errors, such as whether a category that is really "left" will be classified as one which is considered "right" in the CMP's rile scale, and vice-versa. Because the rile index—as are all other quantities in the CMP dataset—are considered as proportions of all text units, we also consider misclassifications into the "Other" category that is neither left nor right.

Full misclassification probabilities are reported in Table 2.7, for each CMP coding category. Categories are sorted so that the 13 "rile-left" categories are listed first, the 13 "rile-right" categories second, followed by the "rile-other" categories. The probability that an individual policy category will be classified as belonging to its own rile classification are highlighted in boldface. For quasi-sentences that really belong to "202: Democracy: Positive" for instance—a relative high-frequency category at 3.55% of all CMP quasi-sentences in the combined dataset—the probability is only 0.50 from our tests that it will be assigned a CMP code that is one of the 13 "rile-left" categories. The probability is almost even (0.47) that it will be coded as a category that is not part of the rile index, and just 0.03 that it will coded as a "rile-right" category. Similar interpretations can be made for each of the other CMP coding categories listed in Table 2.7. (The limited set of categories in our test documents meant that we could only report misclassification probabilities for the 20 categories identified by the CMP's Master coding.)

[TABLE 2.6 ABOUT HERE]

Table 2.6 provides the most reduced summary of this misclassification, according to a 3×3 table. The coders from our two tests provided a total of 1,668 text unit classifications, which we could identify from the CMP's master coding as belonging to a left, right, or neither rile category. Comparing these to the rile categories of the coding category that our testers identified, we see significant frequencies in the off-diagonal cells. "Left" text units in particular were prone to misclassification, as 0.35 or 35% of the time these were assigned a category that was not in the rile scale. Conversely, about 19% of the text units that were not in a category found in the rile scheme were classifies instead as "left". Overall, the highest diagonal proportion was just .70, indicating that 30% or more of the quasi-sentences were classified into a wrong rile category.

[FIGURE 2.12 ABOUT HERE]

A graphical summary of the misclassification probabilities from Table 2.7 is given in

Figure 2.12, which shows the misclassification for each coding category in our tests, by the proportions assigned by coders. The dashed lines partition the categories into those that are really "`rile`-left", "`rile`-right", and "`rile`-other"; each vertical bar is sized with its width proportional to the frequency of that category from our tests. Actual classification into left is medium grey, right is dark grey, and none is light grey. In the absence of misclassification, all bars in each division (true left, right, or other) would be the same shade, which they are clearly not.

[FIGURE 2.13 ABOUT HERE]

Another way to characterise the misclassification, as well as to single out visually the worst categories from the standpoint of misclassification, is to use a ternary plot. Figure 2.13 plots each category according to its probability of (mis)classification into the three-fold `rile` set: left, right, and other. The categories that are truly right are in black, left in medium gray, and other in light grey. In addition, the mean misclassification probabilities for each of the three categories are shown as points with a circle (these correspond to the proportions in Table 2.6.) If no misclassification existed, then all categories of the same colour would be clustered in the corners of the triangle, which as can be clearly seen does not happen. The worst categories from a misclassification standpoint are located the furthest from the corners. Categories 305 and 404 in the middle, for instance, are severely misclassified—confirming the probabilities reported in Table 2.7. As discussed previously, category 202 is also shown as being almost evenly split between left and other, while the probability that a coder assigns 202 a "right" category is almost nil.[33]

2.10 Demonstrating the Effects of Misclassification

We know from just the reduced 3×3 `rile` misclassification matrix (estimated in Table 2.6) that the probability is 30% or more that a text unit will be assigned the wrong left-right category. The question for practical purposes is: just how badly will this affect our resulting estimates?

To answer this question we use simulation of the type of misclassification identified in

[33] Briefly mentioning how to read these plots: you go from 0 to 1 on each axis and branch to the left on the grid lines. So 202 is .5 for left, .47 for Other, and .03 for right.

our results above. By simulating the effect of stochastic misclassification on a range of `rile` values at different levels of reliability, we can assess the degree of error, both systematic and non-systematic, that are likely to be present in the CMP's reported `rile` estimates. From the combined CMP dataset, we know that the population proportions of the `rile` left, right, and neither text units are roughly 0.25, 0.25, and 0.50 respectively. Our range of `rile` therefore fixes the other category at 0.50 and lets the other frequencies vary so that we can observe `rile` values from -50 to +50, once again a range taken from the empirical range in the combined CMP dataset.[34]

[FIGURE 2.14 ABOUT HERE]

The results of simulated misclassification are shown in Figure 2.14. Here we have manually manipulated the misclassification matrix to be symmetric and to produce reliabilities of (reading from top left to right) $\kappa = 0.90$, 0.80, 0.70, 0.60, and 0.50. The last panel (lower left) shows the effect of simulating error using the misclassification probabilities from Table 2.6, and having a median reliability of 0.47. A faint cross-hair indicates the origin, and a dashed line shows the identity point at which $A_i^* = A_i$. Finally, the red line shows the least-squares slope.

Two patterns clearly emerge from our simulation of misclassification. First, even at relatively high levels of reliability, misclassification adds significant noise to the resulting `rile` estimates, meaning that any individual realisation of the `rile` index is likely to contain a significant degree of random error. Because `rile` is most commonly used as an explanatory variable in the political science models—in fact this is the single most common usage of the CMP dataset by far—this means that such models are likely to have biased estimates. Second, simulation results show the tilt in the observed values away from the identity line, making observed values positioned flatter along the zero line, and causing a centrist bias in the estimated `rile` values even when the misclassification matrix is strictly symmetric. The reason is quite general: the more the true value consists of any single category, the greater the tendency of misclassification to dilute this category. (At the extreme of being, for instance, pure left, any

[34] Simulations here were performed 8 times each for even-valued "true" `rile` values ranging from -50 to 50. Misclassification was generated using the `misclass()` function from the R `simex` 1.2 package. A tiny amount of jitter has been added to the x-axis values in the plots.

misclassification can only move the estimate away from this extreme.) At the levels of reliability indicated by our tests—call it 0.50—this bias is quite severe, cutting the estimate of a "true" `rile` value of -50 or 50 almost in half. The effect on estimates when `rile` is used as an explanatory variables is to compact the range of the variable, further afflicting regression coefficients with attenuation bias.

2.11 Concluding Remarks and Recommendations

Bodies of text are data. We can analyse these data using well-known statistical tools. The implications of this are deep and general. Our discussions in this paper apply to the analysis of most bodies of text, and in particular to analyses of text based on interpretative coding by human experts. While we focus in this paper on text observed in party manifestos and analysed by the CMP, the problems we identify and set out to correct apply to any dataset based on human interpretative coding. Our focus on the CMP reflects the very widespread use of this dataset within the profession, generating a large number of publications in the best professional journals. These publications never take account of the fact that the data analysed clearly contain measurement error, and that this measurement error can clearly bias research findings.

We approach this problem by considering ways in which manifestos provide systematic information about the policy positions of their authors, in the form of text units deposited as random variables in a process of authorship that is inherently stochastic, even when the author's underlying position is fixed. We simulate this process, thereby computing error estimates for the entire CMP dataset and show how such errors affect descriptive and causal inferences based on CMP measures. Building on this method, we offer a "corrected" version of the CMP dataset with bootstrapped standard errors for all key estimates.

The substantive consequences of our new estimates of error in CMP data are far from trivial. Many apparent "differences" in CMP estimates of party policy positions—differences over time in the position of one party or differences between parties at one point in time—are probably attributable to stochastic noise in textual data rather than real differences in policy positions. Only about one quarter of all CMP-estimated "movements" in parties' left-right policy positions over time were assessed on the basis of our simulations to be statistically significant.

Replicating two recently-published articles in which error-prone CMP variables are used as co-variates, we show how to correct these using a SIMEX error correction model, based on bootstrapped estimates of likely error. The probable systematic effect of error contaminated variables is the inflation of standard error of the regression in the case of measurement error in dependent variable and bias with measurement error in co-variates. While error in co-variates typically causes attenuation bias in linear models, as our replication of the Adams et. al. results have shown, this is not always true for more complicated models. Some error-corrected effects are stronger, and more significant, than those estimated in models taking no account of error in the co-variates. Other times the effect of error correction is the opposite: making co-variates statistically insignificant. Measurement error correction can cause substantively important reinterpretation of results. A good example is what emerges as the potentially flawed inference that national party policy positions on the EU have no influence on their EP roll-call voting behaviour, an inference that is reversed once account is taken of error-contamination in the CMP dataset's sparsely-populated variables measuring EU policy. Similarly, a conclusion that in comparison to mainstream parties niche parties are penalised by voters for moderating their policy positions has also been shown to be flawed once the effects of measurement error are corrected.

Turning to another component of error in CMP – human misclassification – we know with absolute certainty, from information published by the CMP itself and summarised in Table 2.4, that CMP coders disagree with CMP master codings when assigning text units to CMP coding categories. Since different coders all have different correlations with the CMP master codings, we also know with absolute certainty that different CMP coders disagree with each other when coding the master documents. In this paper, we characterise this disagreement as stochastic coding error and set out to derive estimates of the scale of this. This is crucially important since each point in the widely used CMP time series is based on a single coding by a human coder and comes with no estimate of associated error. Before we can draw statistically valid inferences from these data, however, we need estimates of the error associated with their generation.

Table 2.5 summarises our findings on the broad scope of the stochastic error arising from multiple independent human interpretative codings of the master documents. Bearing in mind that the minimum standard conventionally deemed acceptable for the reliability coefficients reported in Table 2 is 0.8, the coefficients we find are worryingly low, almost all in the range [0.3, 0.5]. From this we infer that, had multiple independent human coders indeed been used to code every document in the CMP dataset, then the inter-reliability of these codings would be unacceptably low. While this has previously been suspected on common sense grounds, it has not previously been demonstrated in a systematic way by analysing multiple codings of the same document using the CMP coding scheme.

We also found that some categories in the CMP scheme are much more susceptible to coding error than others. Findings on this are summarised in Figure 2.12 and given in more detail in Table 2.7. We see for example that CMP coding categories “305: Political authority” and “404: economic planning: positive” generate coding errors on a very frequent basis. More worryingly for users of the CMP left right scale, they often generate coding errors that assign text units “master coded” as right (305) or left (404) to a coding category on the “wrong” side of the left right scale. This in turn means that problems arising from coding error are not solved by using the CMPs aggregate “left” and “right” categories, or the additive scale constructed from these.

The results we report above imply that the effect of coder misclassification is to add considerable noise to existing CMP estimates, substantially more than estimated to arise from the text generation process. In addition, the coder misclassification, by coding as “left” what should be “right” and vice versa, causes a centrist bias as a result of which extreme positions tend to be coded as more centrist than they “really” are. The additional noise, plus the bias caused by misclassifications towards the middle, are likely to cause additional attenuation bias of estimated causal effects when CMP quantities, especially `rile`, are used as co-variates in regression models.

The coding experiments we report above reinforce in a fairly precise way the conclusion that the CMP data, based on human interpretative coding of party manifestos, are very noisy, and give us some sense of the scale of this noise. These results strongly indicate further systematic work on this important matter, since our tentative conclusions in this paper are based on limited multiple codings of only two English-language manifestos. We used the master documents coded by the CMP in this limited exercise because we wanted to have some sense of how the multiple codings we generated compare with the CMP's own view of the "true and certain" position of each document. What is clearly now indicated, however, is a project that would procure multiple independent codings of a much larger sample of CMP documents, for which no master coding exists, to allow more confident conclusions to be drawn about the extent of unsystematic inter-coder (un)reliability and the biasing effects of systematic coder misclassification. The work we report above establishes a strong *prima facie* case that this is a problem to be taken very seriously indeed by third-party users of the CMPs estimates of time-series of party policy positions.

The importance of estimating and making use of measurement error and general uncertainty in political science data, of course, is not limited to manifesto coding and the CMP dataset. Many commonly used measurements, such as survey data, roll call votes, expert surveys of party policy (Benoit & Laver 2006), categories of legislation (e.g. Mayhew 1991), the democraticness of regime type (see Treier & Jackman 2008), and a myriad of other commonly used variables are measured with levels of error. Even when estimates of measurement error are provided—as is the case with surveys, expert surveys, and more recently, roll call votes (e.g. Clinton, Jackman & Rivers 2004)—political scientists rarely, if ever, make use of these estimates in the ways we encourage here.

While we have taken an important first step towards providing a practical and theoretically supported means to estimate non-systematic measurement error in CMP estimates, the solution we provide here is hardly the last word on the topic. Analyses of coder differences and/or coder error, for example, could uncover systematic error leading to bias, something not considered in this paper but certainly meriting investigation. Experiments with multiple independent codings of the same texts would provide important information on stochastic coding, a type of error we identified but did not focus on here. Finally, other means of implementing error correction models are certainly possible, including Bayesian-MCMC methods that can take into account the unit-specific nature of error in our error estimates. Indeed, we hope our focus on error in the widely used CMP estimates will stimulate a broader dialogue on measurement error in many of the most commonly used measures in political science, such as opinion survey results, expert survey measures, or other computed quantities. Given our knowledge of measurement error and the wide availability of techniques for dealing with this, there is no longer any excuse for scholars to use error-prone measures as if these were error free.

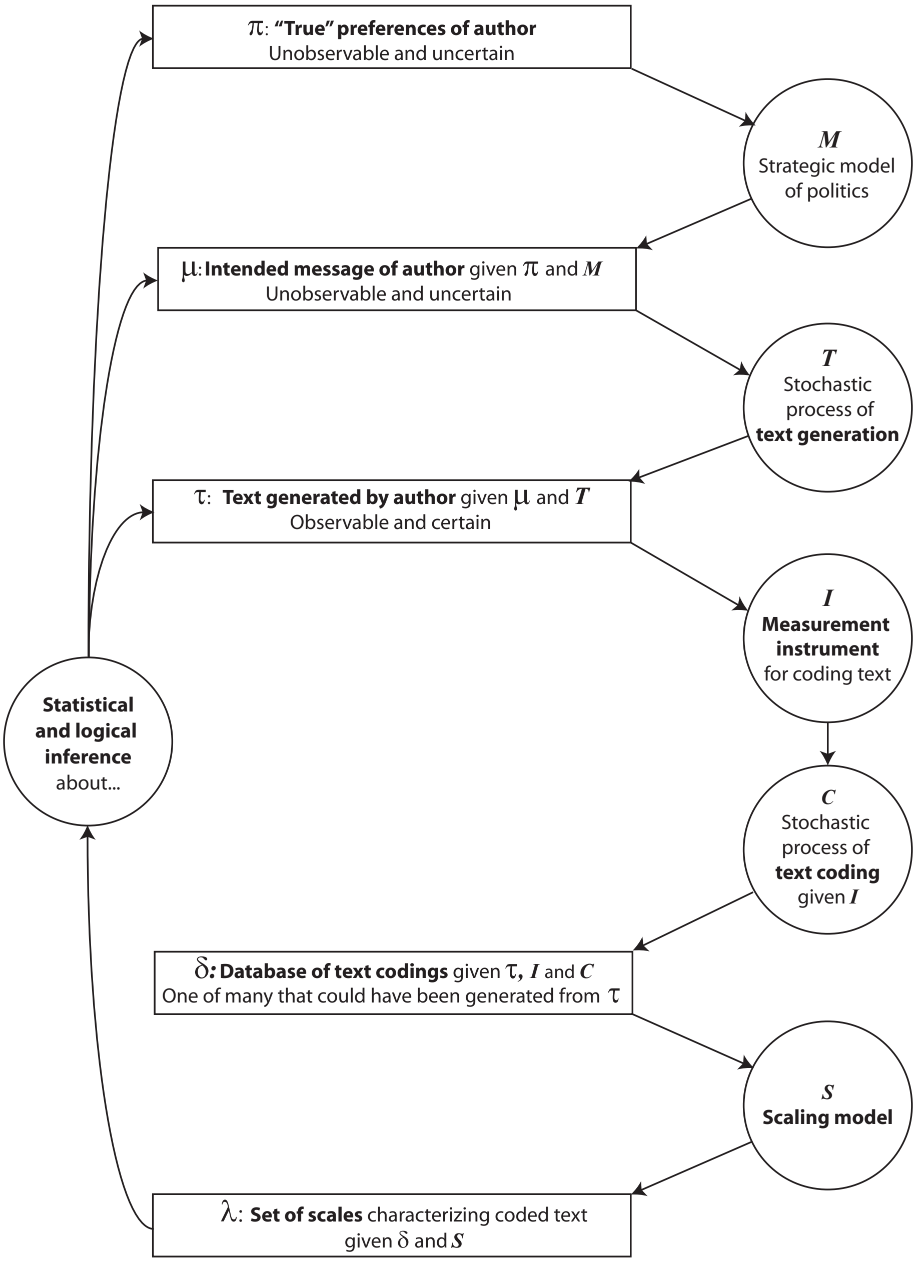

π: **"True" preferences of author**
Unobservable and uncertain
M
Strategic model of politics
μ: **Intended message of author** given π and *M*
Unobservable and uncertain
T
Stochastic process of **text generation**
τ: **Text generated by author** given μ and *T*
Observable and certain
I
Measurement instrument for coding text
Statistical and logical inference about...
C
Stochastic process of **text coding** given *I*
δ: **Database of text codings** given τ, *I* and *C*
One of many that could have been generated from τ
S
Scaling model
λ: **Set of scales** characterizing coded text given δ and *S*

Figure 2.1: *Overview of the Positions to Text to Coded Data Process.*

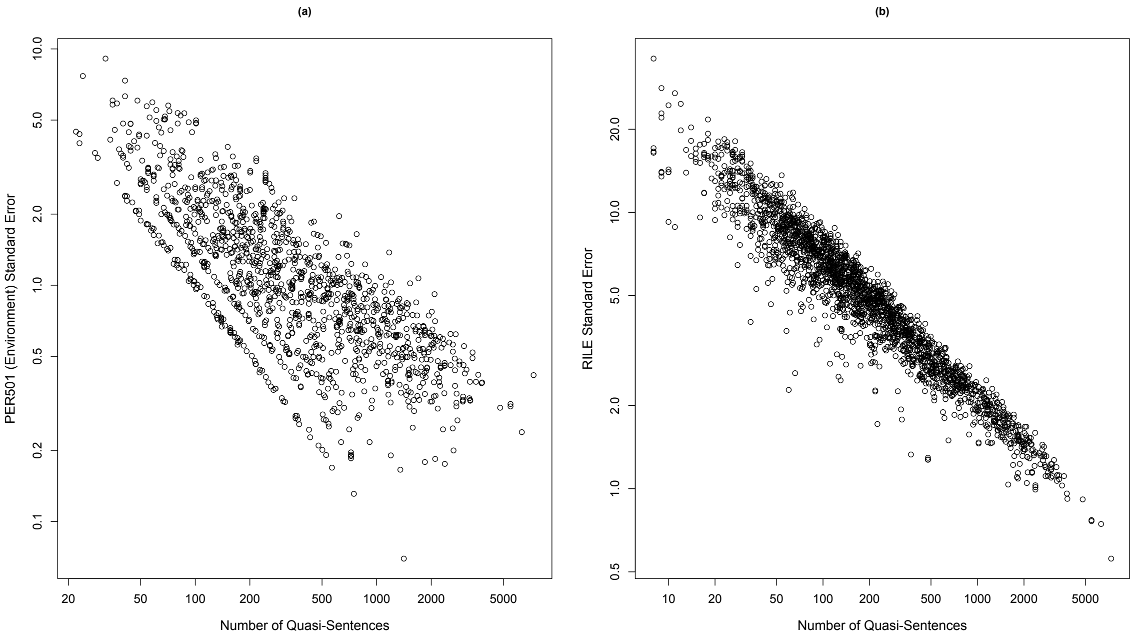


Figure 2.2: *Bootstrapped standard errors Environment (PER501) and the CMP Left-Right scale (`rile`).* Axes use log scales.

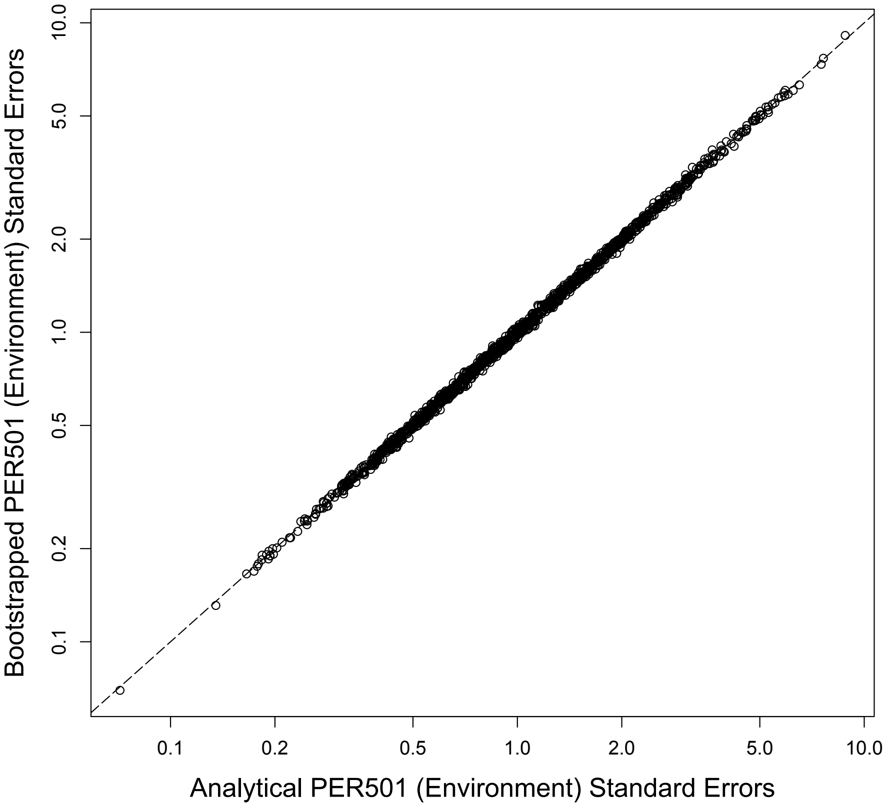


Figure 2.3: *Comparing analytical to bootstrapped standard errors for Environment (PER501).* Axes use log scales.

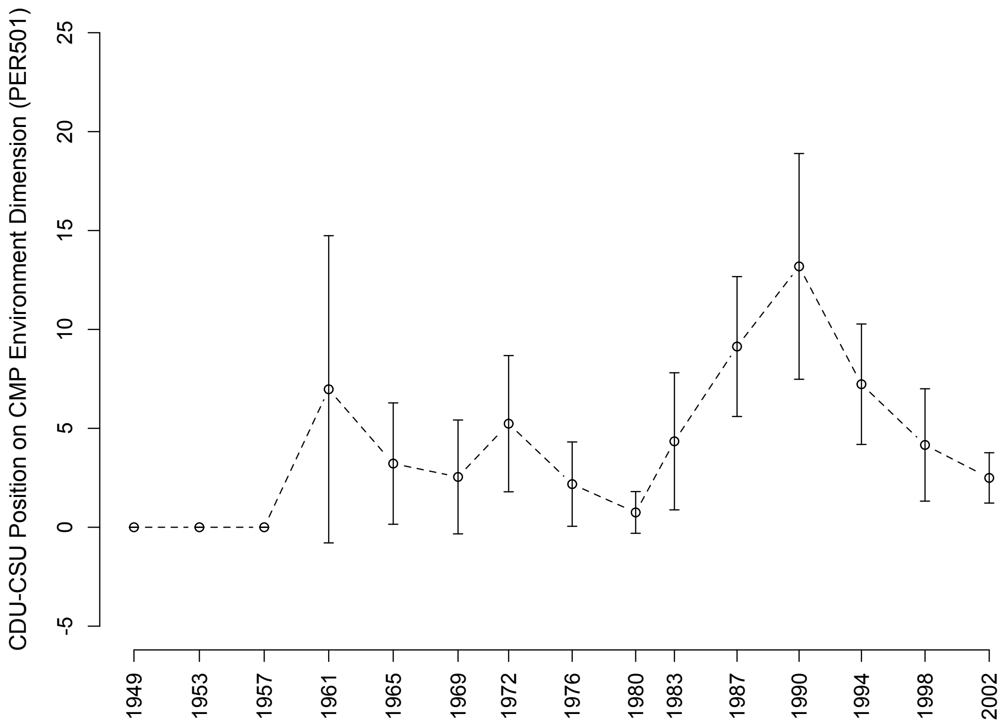


Figure 2.4: *Movement on environmental policy of German CDU-CSU over time.* Movement of Dashed line is % environment with 95% CI; dotted line is the number of quasi-sentences per manifesto coded PER501.

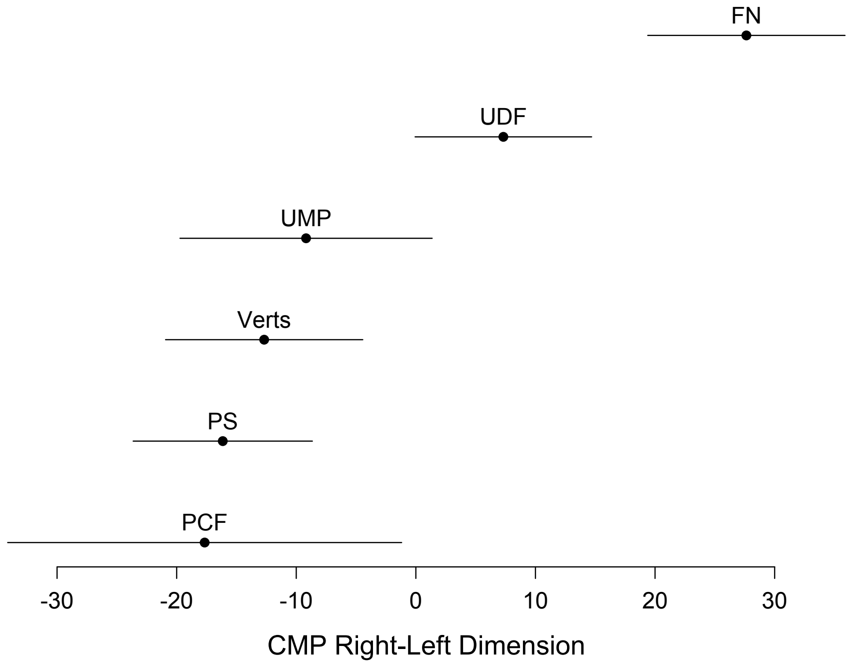


Figure 2.5: *Left-Right placement of the major French parties in 2002.* Bars indicate 95% confidence intervals.

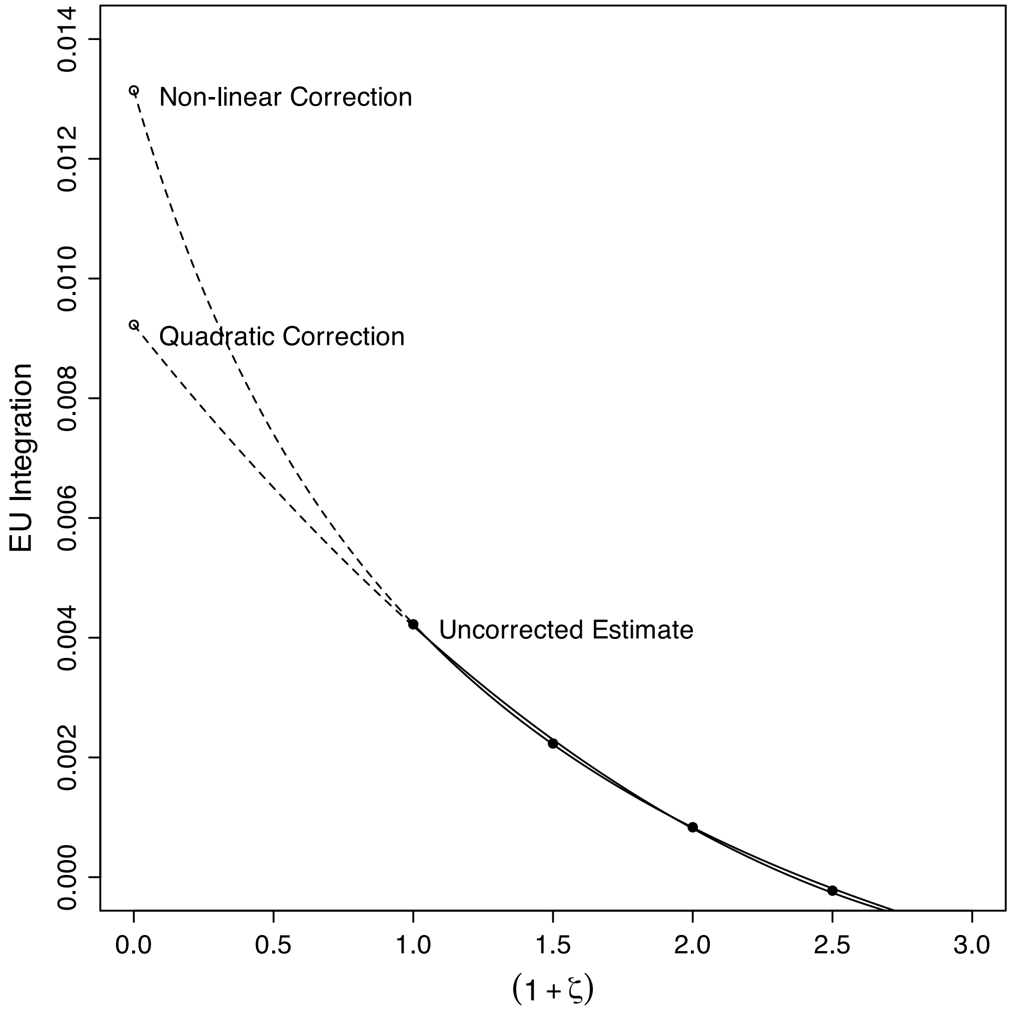


Figure 2.6: *SIMEX error correction in* EU Integration *with quadratic and nonlinear extrapolant functions, from Hix, Noury, and Roland (2006).*

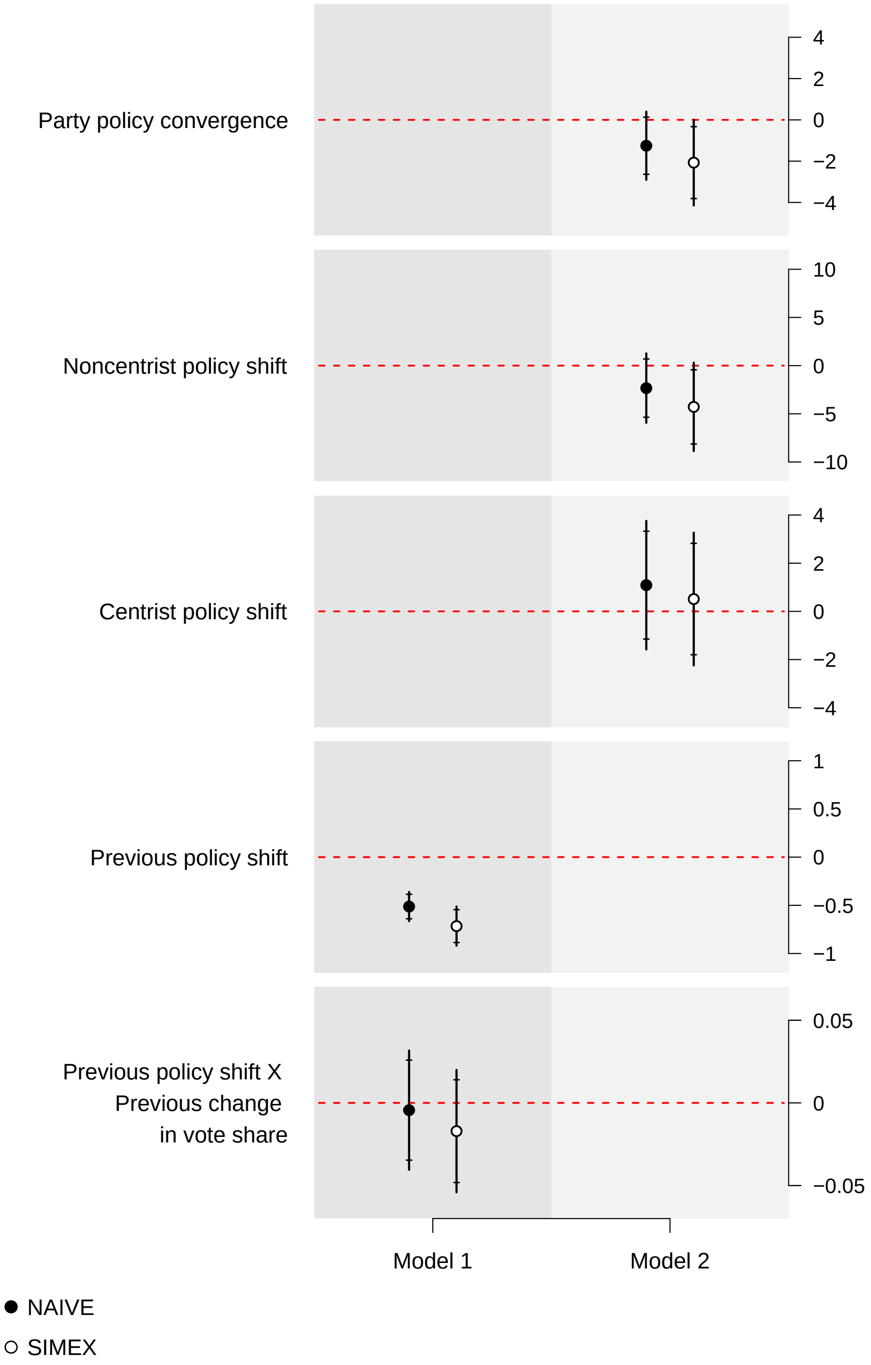


Figure 2.7: *Results of SIMEX error correction in Adams et. al.(2006).* Bars indicate 95% confidence intervals, with whiskers on each bar indicating the 90% confidence interval. Full results for all variables are available from tables below.

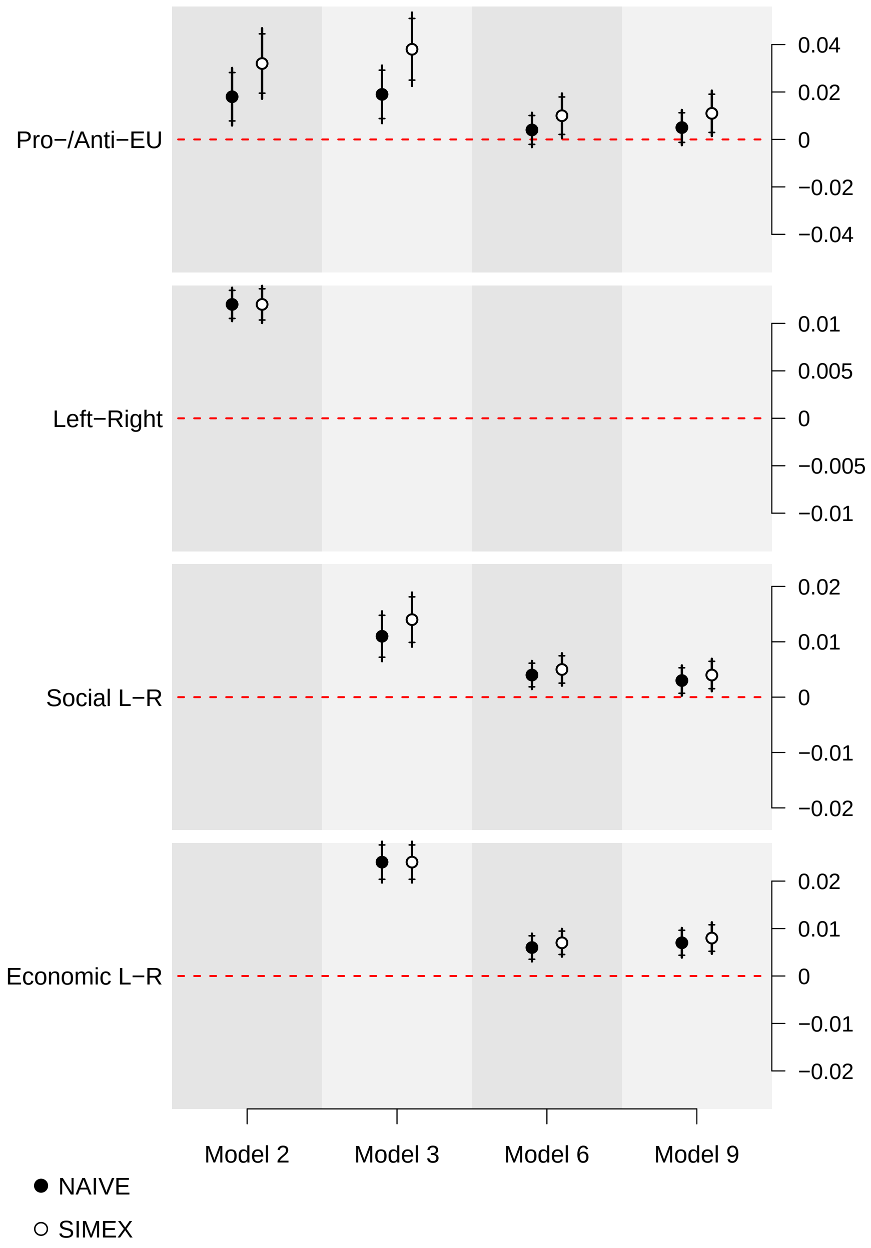


Figure 2.8: *Results of SIMEX error correction in Hix, Noury & Roland (2006, 503–504, Table 4).*

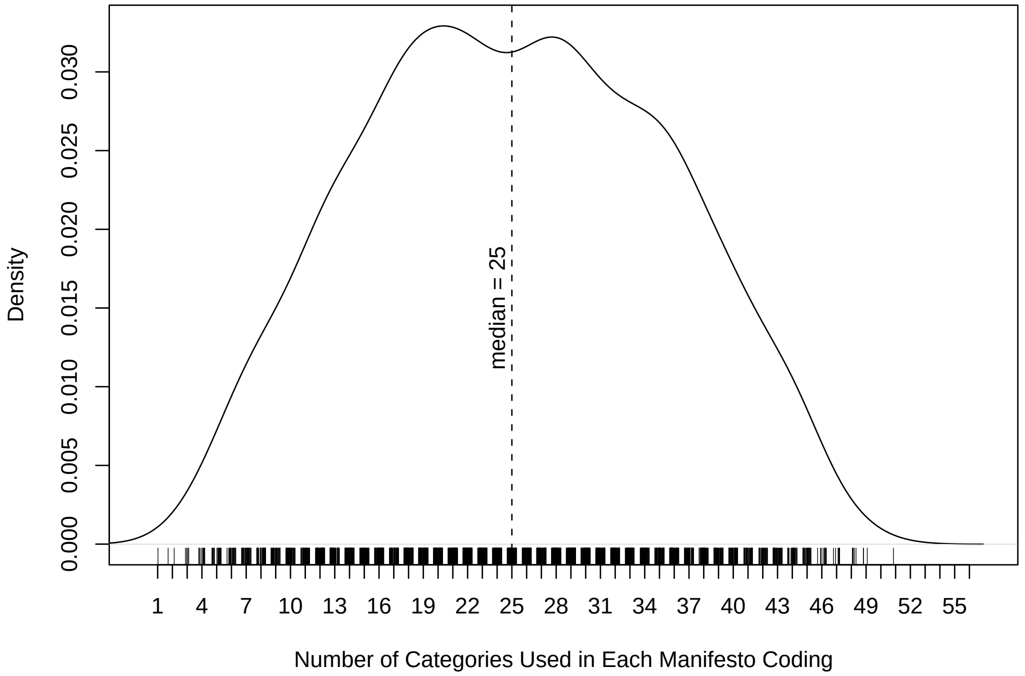


Figure 2.9: *Category Usage Profile for the Combined CMP Dataset, n=3,018 manifestos.*

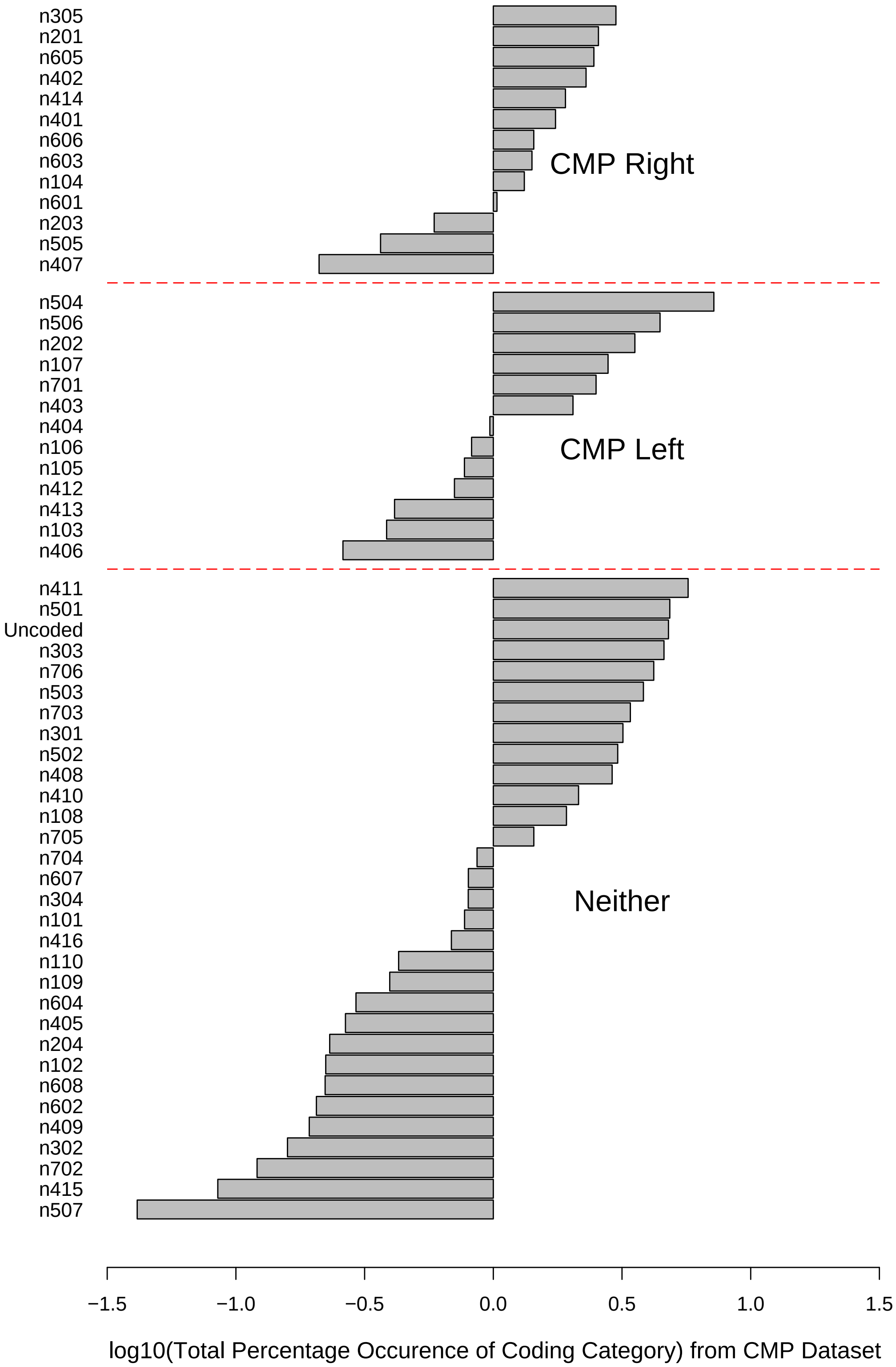


Figure 2.10: *Category Frequency Across All manifestos, as* $\log_{10}$ *percentage.*

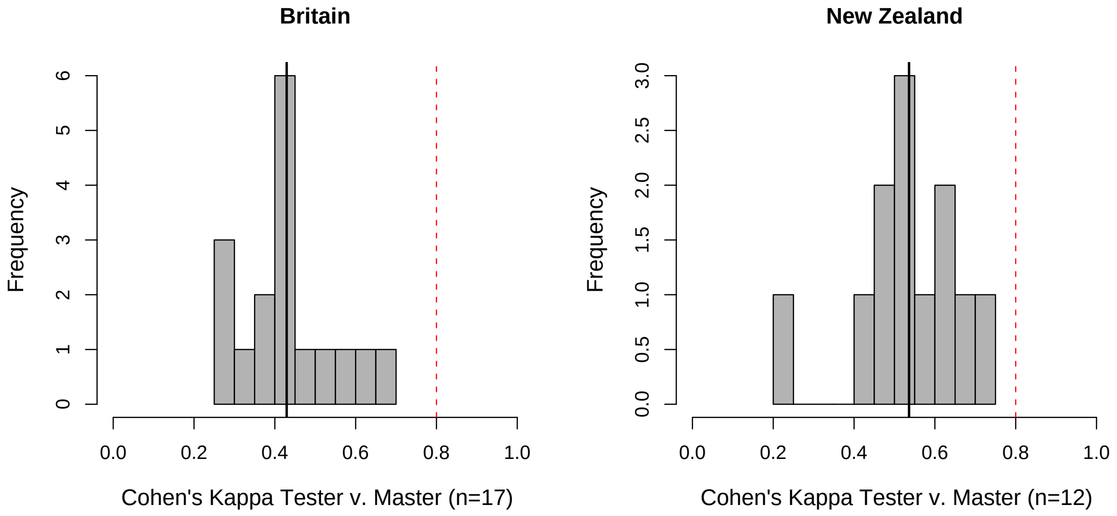


Figure 2.11: *Summary of Coder Reliabilities Compared to Master, Cohen's* κ. The dashed red line indicates the conventional lower bound as to what is considered "reliable" in interpretations of Cohen's κ. The solid black lines are the median value of κ from all the coders completing the tests.

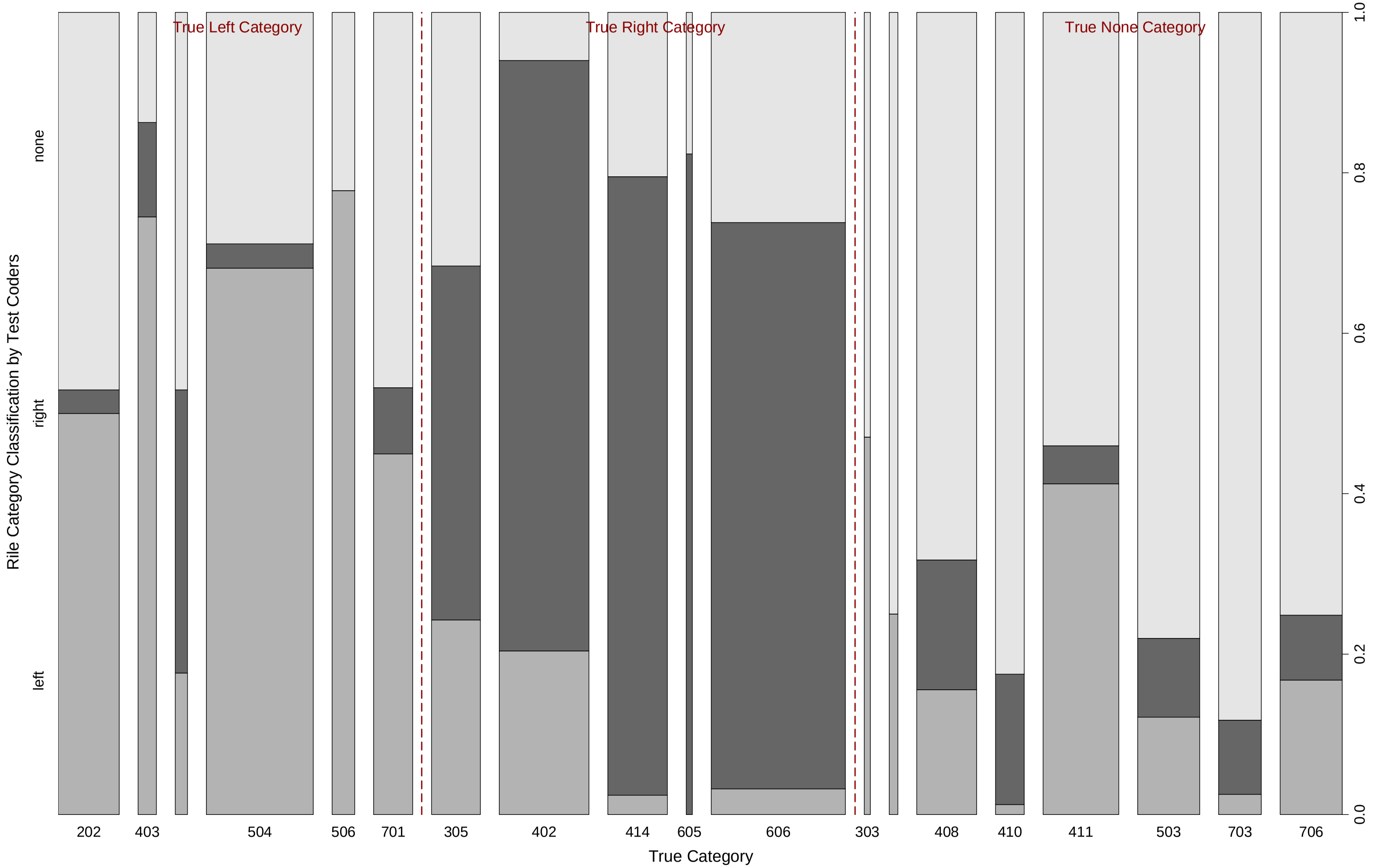


Figure 2.12: *Empirically computed misclassification matrix, True Categories by Observed* `rile` *Categories of Left, Right, or Other.*

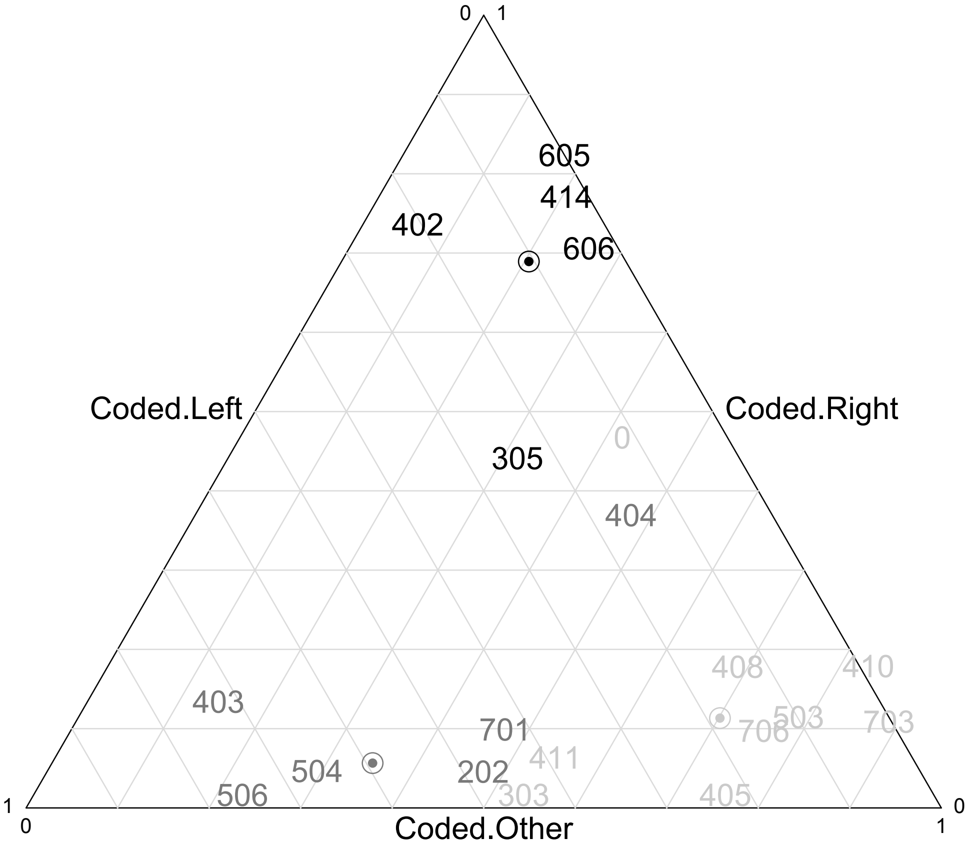


Figure 2.13: *Misclassification into Left, Right, or Other by coding category, from experiments.* The dark circles in hollow points represent the misclassification for the 3×3 left-right-other misclassification matrix. Each number plotted identifies the probability of this category being coded as a left, right, or other category, where the black numbers are really right, the medium gray numbers really left, and the light gray numbers really other categories. If no misclassification existed, all numbers would cluster together into their respective corners, which clearly does not happen.

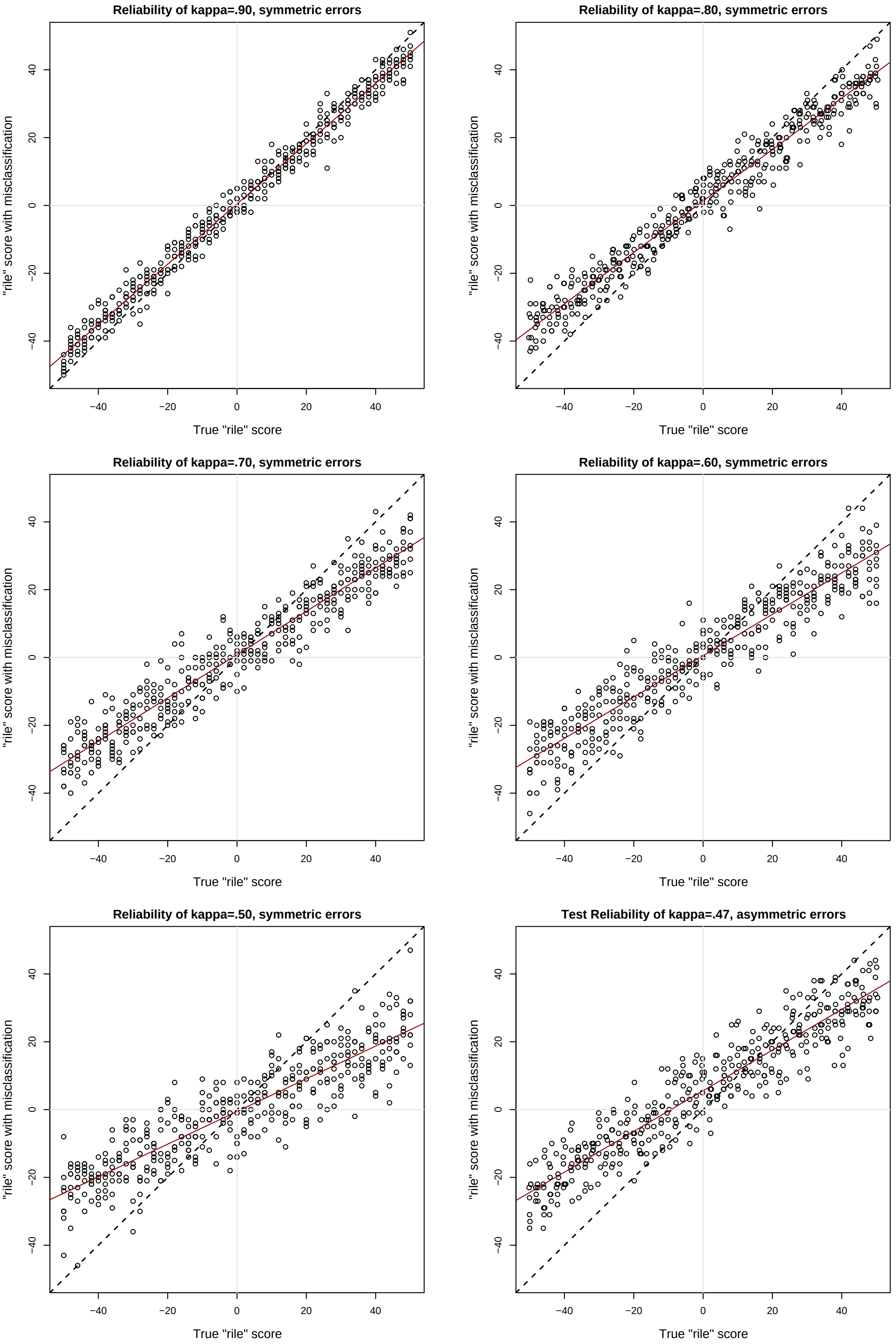


Figure 2.14: *Simulated Misclassification at Different Levels of* κ. The misclassification matrix Θ_{ji} is simulated from a manifesto with 50% uncoded content, for different levels of κ, except for the last panel, which uses $\hat{\Theta}_{ji}$ estimated from the coding experiments. Misclassification is simulated 8 times for each even-numbered true `rile` score from -50 to 50.

Statistically Significant Change?	**Elections**	**% of Total**
No	1,308	62.3%
Yes	791	37.7%
Non-adjacent	778	–
Total	2,877	100.0%

Table 2.1: *Comparative over-time mapping of policy movement on Left-Right measure, taking into account statistical significance of shifts.*

	Model 1		Model 2	
	OLS		OLS	
Variable	**Replication**	**SIMEX**	**Replication**	**SIMEX**
Public opinion shift	**0.90075**	**0.76750**		
	(0.23380)	(0.23057)		
Niche party	0.10632	0.11277		
	(0.13302)	(0.13512)		
Public opinion shift × Niche party	**-1.56752**	**-1.50785**		
	(0.44883)	(0.39125)		
Previous policy shift	**-0.51274**	**-0.98379**		
	(0.07722)	(0.09119)		
Previous change in vote share	0.01651	0.01947		
	(0.01017)	(0.01012)		
Previous policy shift ×	-0.00442	**-0.03926**		
Previous change in vote share	(0.01843)	(0.01649)		
Centrist policy shift			1.44645	1.08413
			(1.39517)	(1.41363)
Noncentrist policy shift			-2.01063	-3.70891
			(1.80682)	(2.22539)
Niche party			-1.26891	-1.65960
			(1.84939)	(1.86287)
Niche party × Centrist policy shift			**-5.66693**	-5.43235
			(2.83277)	(2.81981)
Niche party × Noncentrist policy shift			2.22222	3.12453
			(2.66182)	(2.74454)
Public opinion shift			**5.27443**	**5.74876**
			(1.76163)	(1.79644)
Party policy convergence			-1.49079	**-2.38873**
			(0.87608)	(1.09570)
Peripheral party			-0.12058	0.30199
			(1.80851)	(1.83141)
Party policy convergence × Peripheral party			1.29769	1.03848
			(1.15784)	(1.16804)
Governing party			-2.87585	-2.93974
			(1.69239)	(1.67980)
Governing in coalition			1.47135	2.01117
			(1.28295)	(1.32331)
Change in unemployment rate			-0.79846	-0.90825
			(0.72688)	(0.72734)
Change in GDP			-0.37888	-0.48947
			(0.39690)	(0.40312)
Governing party × Change in unemployment rate			0.10271	0.02515
			(1.06397)	(1.06045)
Governing party × Change in GDP			-0.09629	-0.12300
			(0.50160)	(0.49895)
Previous change in vote share			-0.15090	-0.15618
			(0.09344)	(0.09302)
RMSE	0.59067	0.66800	4.39424	4.45137
R^2	0.35780	0.17871	0.26220	0.24292
N	154	154	122	122

Table 2.2: *Results of SIMEX error correction in Adams, Clark, Ezrow & Glasgow (2006).* Country dummies are included in the estimations but not reported here. Italicised variables are error-corrected as follows: (Model 1) Policy shift (dependent variable)=.36606, Previous policy shift=.36988, interaction of Previous policy shift and Previous change in vote share=.36988; (Model 2) Centrist policy shift=0.19421, Noncentrist policy shift=0.15219, Party policy convergence=0.73376, interaction of Niche party and Centrist policy shift=0.0603, interaction of Niche party and Noncentrist policy shift=0.03977, interaction of Party policy convergence and Peripheral party=0.33419. Coefficients in bold are statistically significant at the $p \leq .05$ level; SIMEX standard errors are based on jackknife estimation.

	Model 3		**Model 6**	
	OLS		**OLS**	
Variable	**Replication**	**SIMEX**	**Replication**	**SIMEX**
EU Integration	**0.01875**	**0.03464**	0.00422	**0.00923**
	(0.00623)	(0.00774)	(0.00369)	(0.00455)
Social L-R	**0.01051**	**0.01343**	**0.00405**	**0.00493**
	(0.00227)	(0.00241)	(0.00135)	(0.00143)
Economic L-R	**0.02352**	**0.02413**	**0.00622**	**0.00683**
	(0.00217)	(0.00215)	(0.00149)	(0.00152)
Commissioner	0.07879	0.07054	0.02175	0.02173
	(0.04947)	(0.04942)	(0.03044)	(0.03037)
In government	**0.10265**	0.07942	**0.06087**	**0.05700**
	(0.04336)	(0.04365)	(0.02589)	(0.02589)
Socialists			**-0.55953**	**-0.54927**
			(0.03508)	(0.03542)
Italian Communists and allies			**-0.64108**	**-0.62645**
			(0.21150)	(0.21119)
Liberals			**-0.16767**	**-0.16405**
			(0.03590)	(0.03587)
Greens			**-1.00344**	**-0.98107**
			(0.05104)	(0.05233)
British Conservatives and allies			0.07714	0.07317
			(0.09930)	(0.09926)
Radical left			**-0.82003**	**-0.79043**
			(0.04959)	(0.05191)
French Gaullists and allies			0.09861	0.10952
			(0.06228)	(0.06241)
Non-attached members			**-0.23046**	**-0.22548**
			(0.05390)	(0.05389)
Regionalists			**-0.78486**	**-0.76795**
			(0.05675)	(0.05732)
Radical right			**0.44665**	**0.4529**
			(0.12441)	(0.12420)
Constant	**-0.14899**	**-0.17493**	**0.36410**	**0.3433**
	(0.05961)	(0.05981)	(0.04405)	(0.04537)
RMSE	0.35782	0.36254	0.49224	0.20203
R^2	0.41120	0.39561	0.81360	0.81232
N	349	349	349	349

Table 2.3: *Results of SIMEX error correction in Hix, Noury & Roland (2006, 503–504, Table 4).* All models include dummies for parliament but these are not shown. Italicised variables are error-corrected as follows: Social L-R=1.9907, Economic L-R=1.88742, EU Integration=1.69393. Coefficients in bold are statistically significant at the $p \leq .05$ level; SIMEX standard errors are based on jackknife estimation.

Test description	Mean Correlation	N	Reference
Training coders' solutions with master	0.72	39	Volkens (2001a, 39)
Training coders' second attempt with master	0.88	9	MPP2 (2006, 107)
All pairs of coders	0.71	39	Volkens (2001a, 39)
Coders trained on 2nd edition of manual	0.83	23	Volkens (2007, 118)
First time coders	0.82	14	Volkens (2007, 118)
First test of coders taking second contract	0.70	9	Volkens (2007, 118)
Second test of coders taking second contract	0.85	9	Volkens (2007, 118)

Table 2.4: *Coder reliability test results reported by CMP.* Sources are (Klingemann et al. 2006; Volkens 2001a, 2007); figures reported are Pearson's *R* for the aggregate percentage measured across 56 coding categories for the test document found in *MPP2*, pp181–186.

Reliability Test	Fleiss's κ By Category	 By RILE
British Manifesto Test (107 text units, 17 coders)	0.35	0.36
New Zealand Manifesto Test (72 text units, 12 coders)	0.40	0.47
Combined Test Results (144 text units, 24 coders)	0.31	0.32
Combined Results by Category:		
504: Welfare State Expansion: Positive (L)	0.50	
506: Education Expansion: Positive (L)	0.46	
403: Market Regulation: Positive (L)	0.29	
202: Democracy: Positive (L)	0.18	
701: Labour Groups: Positive (L)	0.14	
404: Economic Planning: Positive (L)	0.05	
402: Incentives: Positive (R)	0.46	
414: Economic Orthodoxy: Positive (R)	0.46	
606: Social Harmony: Positive (R)	0.44	
605: Law and Order: Positive (R)	0.13	
305: Political Authority: Positive (R)	0.10	
703: Farmers: Positive	0.82	
503: Social Justice: Positive	0.35	
411: Technology and Infrastructure: Positive	0.34	
706: Non-economic Demographic Groups: Positive	0.29	
405: Corporatism: Positive	0.21	
410: Productivity: Positive	0.17	
408: Economic Goals	0.13	
000: Uncoded	0.11	
303: Govt'l and Admin. Efficiency: Positive	0.02	

Table 2.5: *Reliability Results from Coder Tests.* The (L) or (R) designates whether a CMP category was part of the `rile` left or right definition, respectively.

		Coded `rile`			
		Left	Right	None	
True `rile`	Left	430	41	254	725
		0.59	0.06	0.35	
	Right	100	650	193	943
		0.11	**0.69**	0.20	
	None	188	115	712	1,015
		0.19	0.11	**0.70**	
		718	806	1,159	1,668

Table 2.6: *Misclassification matrix for true versus observed* `rile`. The top figure in each cell is the raw count; the bottom figure is the row proportion. The figures are empirically computed from combined British and New Zealand manifesto tests.

		Overall		**Fleiss κ**		**$\mathrm{Pr}(\mathrm{A}^*\vert\mathrm{A}) =$**		
Code	**Description**	**%**	**RILE**	**All**	**RILE**	**Left**	**Right**	**Other**
103	Anti-Imperialism: Positive	0.38	L					
105	Military: Negative	0.77	L					
106	Peace: Positive	0.82	L					
107	Internationalism: Positive	2.79	L					
202	Democracy: Positive	3.55	L	0.18	0.07	**0.50**	0.03	0.47
403	Market Regulation: Positive	2.04	L	0.29	-0.03	**0.75**	0.12	0.14
404	Economic Planning: Positive	0.97	L	0.05	-0.05	**0.18**	0.35	0.47
406	Protectionism: Positive	0.26	L					
412	Controlled Economy: Positive	0.71	L					
413	Nationalisation: Positive	0.41	L					
504	Welfare State Expansion: Positive	7.19	L	0.50	0.10	**0.68**	0.03	0.29
506	Education Expansion: Positive	4.44	L	0.46	n/a	**0.78**	0.00	0.22
701	Labour Groups: Positive	2.51	L	0.14	0.05	**0.45**	0.08	0.47
104	Military: Positive	1.32	R					
201	Freedom and Human Rights: Positive	2.56	R					
203	Constitutionalism: Positve	0.59	R					
305	Political Authority: Positive	3.00	R	0.10	0.14	0.24	**0.44**	0.32
401	Free Enterprise: Positive	1.74	R					
402	Incentives: Positive	2.29	R	0.46	0.03	0.20	**0.74**	0.06
407	Protectionism: Negative	0.21	R					
414	Economic Orthodoxy: Positive	1.91	R	0.46	0.16	0.02	**0.77**	0.20
505	Welfare State Limitation: Positive	0.36	R					
601	National Way of Life: Positive	1.03	R					
603	Traditional Morality: Positive	1.41	R					
605	Law and Order: Positive	2.46	R	0.13	n/a	0.00	**0.82**	0.18
606	Social Harmony: Positive	1.44	R	0.44	0.24	0.03	**0.71**	0.26
101	Foreign Special relationships: Positive	0.77	-					
102	Foreign Special relationships: Negative	0.22	-					
108	European Integration: Positive	1.92	-					
109	Internationalism: Negative	0.40	-					
110	European Integration: Negative	0.43	-					
204	Constitutionalism: Negative	0.23	-					
301	Decentralisation: Positive	3.19	-					
302	Centralisation: Positive	0.16	-					
303	Governmental and Administrative Efficiency: Positive	4.60	-	0.02	n/a	0.47	0.00	**0.53**
304	Political Corruption: Negative	0.80	-					
405	Corporatism: Positive	0.27	-	0.21	n/a	0.25	0.00	**0.75**
408	Economic Goals	2.90	-	0.13	0.02	0.16	0.16	**0.68**
409	Keynesian Demand Management: Positive	0.19	-					
410	Productivity: Positive	2.14	-	0.17	0.12	0.01	0.16	**0.83**
411	Technology and Infrastructure: Positive	5.71	-	0.34	0.29	0.41	0.05	**0.54**
415	Marxist Analysis: Positive	0.09	-					
416	Anti-Growth Economy: Positive	0.69	-					
501	Environmental Protection: Positive	4.85	-					
502	Culture: Positive	3.04	-					
503	Social Justice: Positive	3.83	-	0.35	0.24	0.12	0.10	**0.78**
507	Education Limitation: Positive	0.04	-					
602	National Way of Life: Negative	0.21	-					
604	Traditional Morality: Negative	0.29	-					
607	Multiculturalism: Positive	0.80	-					
608	Multiculturalism: Negative	0.22	-					
702	Labour Groups: Negative	0.12	-					
703	Farmers: Positive	3.41	-	0.82	0.04	0.03	0.09	**0.88**
704	Middle Class and Professional Groups: Positive	0.86	-					
705	Underprivileged Minority Groups: Positive	1.44	-					
706	Non-economic Demographic Groups: Positive	4.20	-	0.29	0.11	0.17	0.08	**0.75**
000	Uncoded	4.79	-	0.11	0.10	0.41	0.14	**0.45**

Table 2.7: *Complete category listing of misclassification estimates.* Detailed information on categories, reliability, and classification probabilities from tests.

Chapter 3

The Effectiveness of Democratic Representation

Abstract

This paper raises the question of the effectiveness of democratic representation in West European parliamentary democracies. It is usually accepted that a valid model of political representation in these countries is the responsible party model. The model is based on the assumption that the popular will is translated into government policy with the intermediation of political parties. Strong empirical support for the model has been presented in the literature earlier. This study finds that most of previous results can be explained by the effects of measurement error in variables measuring policy preferences of political parties. *Social security & welfare* is the only policy area where the representational effect is robust to measurement error. However, the results show that spending on social security is affected by the mandate of both parties that won elections and parties that lost elections. This contravenes the logical consistency of the responsible party model. The results presented in this paper can indicate that the popular will does not affect public spending, thus undermining the effectiveness of democratic representation. Another possible explanation of the results is that contrary to established opinion in the field the responsible party model is not a valid reflection of the democratic representation process in West European parliamentary democracies

Key Words: democratic representation, responsible party model, public spending, measurement error.

3.1 Introduction

A fundamental question in political science is whether representative democracy actually works. Defining democracy as a form of government conducted in accordance with people's preferences (Dahl 1971) means that the democratic political system is effective when preferences of voters are translated into specific policy outputs (Hyland 1995). In the context of West European parliamentary democracies, characterised by high degree of party discipline essential for the survival of the government, it is usually accepted that a valid model of political representation is the responsible party model or party government model (Thomassen 1994, 250).[1] In short, this model holds that popular will is translated into policy via the intermediation of political parties.

The representational chain within this model has been the subject of much research over the last six decades. It has been argued that the logical consistency of the model collapses when one of the requirements of the model is not met, with the chain being as strong as its weakest link (Converse & Pierce 1986, 698). It has long been held that the weakest link in the party government model is the requirement that voters vote according to their policy preferences (Thomassen 1994). This view has been recently challenged by Ansolabehere, Rodden & Snyder (2008) who show that previous empirical evidence is largely driven by measurement error associated with analysis of individual survey items. Implementing a simple measurement error correction method, Ansolabehere, Rodden & Snyder (2008) show that, contrary to much of survey research over the last six decades, issue preferences have a large effect on voting.

Evaluating the logical consistency of the model requires us to identify another candidate for the "weakest link" in the political representation chain. Some of the most puzzling and controversial results in this field have been produced when evaluating the translation of policy preferences of governing parties into policy output in the form of public spending. The link has been found to exist in countries where the responsible party model is generally considered to be inapplicable (e.g. the US in Budge & Hofferbert 1990). Furthermore, public spending has been found to depend not only on the policy preferences of governing parties, but also on those of parties in opposition (e.g. Klingemann, Hofferbert & Budge 1994). When the electoral choice

[1] See also Powell (2004).

of voters has no consequence for policy output of the government, this clearly constitutes a violation of the logical requirements of party government model.

Earlier results have been repeatedly questioned on methodological grounds. At the same time, later empirical work in this field does not estimate possible effect of opposition policy preferences. The approach taken here re-evaluates the model of translation of policy preferences of governments into public spending, explicitly controlling for potential influence of the opposition. Such empirical evaluation requires two basic things: data on public spending and data on policy preferences of political parties both in government and opposition. The latter data have been usually drawn from the Comparative Manifesto Project (CMP) data set. However, it has been shown that this data set contains large amount of measurement error (e.g., Benoit, Laver & Mikhaylov 2009, Mikhaylov, Laver & Benoit 2008). Empirical evaluation in this paper implements a SIMEX procedure to correct for the effects of measurement error in a co-variate.

This paper identifies that models estimated ignoring measurement error (termed here "naive" estimation) show encouraging signs of support for responsible party model. However, when corrected for measurement error much of the effect disappears. Only one policy area shows large and robust effects across both naive and error-corrected estimations. *Social security & welfare* spending appears to be very strongly influenced by policy positions of governing parties.[2] However, the results presented in this paper highlight that this policy area is also affected by policy preferences of the opposition.

The findings undermine the evidential support for the model since all but one spending area are unaffected by policy preferences of political parties. This raises doubts over the effectiveness of democratic representation. It is widely accepted on normative grounds that in West European parliamentary democracies the interests of the population are served by political parties. Voters evaluate policy packages presented by parties at the elections, and vote for parties closest to their own preferences. Thus, parties at elections receive a mandate from the voters to implement certain policies. The results presented in this paper suggest that most policy areas are unaffected by the mandate. At the same time, the only spending area where

[2] Miller & Stokes's (1963, 56) analysis shows that the domain social welfare most closely conforms to the responsible party model.

the mandate function seems to exists is being redistributed according to the mandate of both parties that won elections and parties that lost elections. Such an outcome would correspond to effectively functioning representative democracy only if there were general agreement in the society about policies relating to *Social security and welfare*. This seems unlikely, since as the largest budgetary item it also has the largest redistributive potential, and hence the potential for societal conflict. In effect this could mean that the popular will does not affect a significant proportion of public spending. Alternatively, it means that for one specific area of public policy democratic representation is being conducted outside the responsible party model. Taken together with negative findings for other spending areas, this suggests that the normative concept of democratic representation via responsible party model needs to be re-evaluated.

This paper briefly reviews responsible party model and its logical requirements in the next section. Third section evaluates existing empirical evidence for the link between policy preferences of governments and public spending. The linkage is re-evaluated while correcting for measurement error in co-variates and controlling for possible effects of opposition parties. Conclusions are drawn in the final section after presenting the results of the estimations and robustness studies.

3.2 Responsible party model and effectiveness of representation

Before proceeding with the analysis it is important to identify the base line model of representation used here, its basic components and requirements for implementation. It is also important to identify the elements that will be investigated here, and the elements that are being held constant for the purposes of current analysis.

3.2.1 Responsible party model

A populist view on representative democracy conceptualises it as a form of government conducted in accordance with people's preferences (Dahl 1971).[3] This political system is effective when preferences of voters are translated into specific policy outputs (Hyland 1995). In the context of West European parliamentary democracies, characterised by high degree of party discipline essential for the survival of the government, it is usually accepted that a valid model of political representation is the responsible party model or party government model (Thomassen 1994, 250). Popular sovereignty in a representative democracy can only be achieved through party democracy or party government (Thomassen forthcoming, Ch.1). Drawing on Schumpeter's (2000[1943]) theory of democracy the party government model was systematically presented in Schattschneider (1950). The model characterises the process of political representation as containing voters and political parties as the only relevant actors; with political parties assumed to be unitary actors with strong party discipline (Thomassen 1994, 251).

Thomassen (1994, 252) identifies the core assumption of the responsible party model: the popular will must translate into government policy. From that he further deduces more specific requirements of the model:[4]

> 1. Political parties must present different policy alternatives to the voters. Elections are meaningful only when a serious alternative is present.
> 2. Internal party discipline must be sufficient to enable the implementation of policy programmes. Parties can put forward credible electoral promises and keep them only if they are united and well disciplined.
> 3. Voters vote according to their policy preferences, i.e. they vote for the party whose programme is closest to their own policy preferences. This implies that voters have actual policy preferences, and they can differentiate policy programmes across parties. (252)

Thomassen (forthcoming, Ch.1) adds two additional requirements:

> 4. The party or coalition of parties that won the elections subsequently takes over the government. Elections can be deemed democratic only if the for-

[3] In liberal theory of democracy a less rigid link is stipulated through a basic division of labour between the electorate and their representatives (see e.g Riker 1982, Thomassen 1994).

[4] A similar set of conditions has been recently discussed in McDonald & Budge (2005, 21).

mation and control of the government is related to the outcome of the electoral process.

5. Policy programmes of political parties and policy preferences of the electorate are constrained by a single ideological dimension.[5] In most West European countries the left-right dimension has been shown to structure the behaviour of both parties and the electorate (e.g., Converse & Pierce 1986, Fuchs & Klingemann 1990, van der Eijk & Franklin 1996, van der Eijk, Franklin & van der Brug 1999, van der Brug & van der Eijk 2007).

These assumptions, when taken together, form a "chain," presented in Figure 3.1 below. The effectiveness of the model can be assessed by the amount of distortion to the preferences of voters, sent as a signal through the chain of representation, at the policy output stage.

[FIGURE 3.1 ABOUT HERE]

The logical consistency of the model collapses when one of the requirements of the model is not met, with the chain being as strong as its weakest link (Converse & Pierce 1986, 698). The requirement that voters vote according to their policy preferences (link **A** in Figure 3.1) has been identified as the weakest link (Thomassen 1994). Empirical results for several countries showed that voters do not meet the criteria of the model. In the United States (e.g., Stokes & Miller 1962) and France (e.g., Converse & Pierce 1986) voters were found not to hold coherent issue preferences, leading to a general conclusion that empirical evidence in favour of the party government model has been "devastating" (Kirkpatrick 1971).

This conclusion has been recently challenged by Ansolabehere, Rodden & Snyder (2008). Using panel data from American National Election Study, the authors show that empirical evidence for the responsible party model is largely driven by measurement error associated with analysis of individual survey items. Implementing a simple measurement error correction method, Ansolabehere, Rodden & Snyder (2008) show that, contrary to most survey research over the last six decades, issue preferences have a large effect on voting. This conclusion is

[5] Thomassen suggests that political parties offer voters a package deal, hence an elector is forced to vote for the whole package. Aggregating individual preferences at the level of the political system results in the Ostrogorski paradox (Rae & Daudt 1976): absence of logical relationship between electoral majority and policy majority on any specific issue. The election reveals the first preferences of voters among candidates, and the majority of first preferences among candidates may not be equivalent to the revelation of first preferences for a specific policy (Dahl 1956). The solution to the paradox is the assumption that both parties drafting programmes and voters choosing a party at the elections are constrained by the same unidimensional ideology, thus conforming to the basic Downsian model (Thomassen 1994, 254).

also supported by recent experimental data in Fowler & Smirnov (2007, Ch.6) and results in Lee, Moretti & Butler (2004). Overall, it appears that the previously identified "weakest link" may be not so weak afterall.

A new candidate for the "weakest link" can be identified by going through stages in Figure 3.1. Much of empirical research in political science suggests that the translation of voters' preferences through electoral systems and formation of governments in parliamentary democracies (links **B** and **C** in Figure 3.1) are generally uncontroversial in terms of stipulated mechanisms (e.g., Cox 1997, Laver & Shepsle 1996, Laver & Schofield 1998). Indeed the link from government to policy appears to be the weakest link in the representational chain of the party government model.

Analysis of the effect of government on policy output produced conflicting results. We can identify two distinct approaches in evaluation of this link in the literature. One focuses on the concrete pledges made in party programmes by parties forming a government and evaluation of the fulfilment of these pledges in government policy after elections. This strand of research produced some evidence for the translation of policy preferences into government policy in Ireland, Greece, Netherlands, UK, Canada and the US (see overview of the approach in Mansergh & Thomson 2007, Thomson 1999).

Another approach in the literature focuses on public spending as a direct manifestation of government policy. In this approach, the policy stances of parties forming governments are expected to be translated into specific spending programmes. "Money is not all there is to policy, but there is precious little policy without it" (Klingemann, Hofferbert & Budge 1994, 41), hence this approach remains the most prominent research strand. As shown below, it is also the research agenda that has produced so far largely inconclusive and disputed results, which, invoking Converse & Pierce (1986), could make or break the responsible party model. Existing evidence is reviewed next with sources of disagreement in the literature identified.

3.2.2 Government policy preferences and public spending

The link between policy preferences of parties in government and public spending in the US has been evaluated in Budge & Hofferbert (1990). Budge & Hofferbert (1990) find evidence for the existence of party mandate in the US. Petry (1991) takes this analysis to multiparty settings in France, and Petry (1995) to Canada. Klingemann, Hofferbert & Budge (1994) further extend the analysis to a set of OECD countries. These studies produced evidence linking policy preferences and public spending of both governing parties and opposition parties. Moreover, in some spending areas the policy positions of opposition parties are a better predictor of government expenditure than the policy preferences of government parties.

This finding was explained as an illustration of the accommodating nature of representative democracy, where electoral losers are not excluded from the political process and preferences of voters they represent are translated into policy outcomes. However, as suggested by Petry (1991) the evidence of opposition influence does not fit the requirements of the party government model. The representational chain in Figure 3.1 becomes meaningless if policy output does not depend on the outcome of the electoral contest. Furthermore, the policy preferences of the losing minority, which possibly does not coincide with preferences of the majority, may translate into policy output. Thus, the core assumption of implementation of popular will in government policy is being fulfilled but without the intermediate steps of representative democracy.

The results in Budge & Hofferbert (1990) have been also questioned on substantive grounds. The political representation process in the US does not fit the requirements of party government model, and evidence provided in the study means "that virtually every observer of the American party system in this century has been wrong" (King & Laver 1999, 597). King & Laver (1993) question the findings on methodological grounds. They show that after correcting for an algebraic error in calculation of standard errors, and introducing to the model dynamics of partial adjustment of budgetary process, there remains no evidence of causal relationship. The response by Hofferbert, Budge & McDonald (1993) claims that their initial *APSR* (1990) study aimed only at finding "association" and not causation. There is no reason, however, to study non-causal, chance, relationships within the framework of a causal theory like the responsible party model (King & Laver 1999, 597–598). Addressing this criticism and further methodological commentary in Thome (1999), McDonald, Budge & Hofferbert (1999) estimate a dynamic model with partial adjustment of the linkage between policy preferences of governing parties and spending on social security and welfare. They show that in a panel of OECD countries from 1972 to 1991, government preferences on social welfare exert a small and marginally statistically insignificant effect on government outlays in this spending area.

Unfortunately, the model did not test the effect of policy preferences of opposition parties – an outcome of previous studies that was theoretically contravening the requirements of the responsible party model. Gibbons (2004) re-evaluates Klingemann, Hofferbert & Budge's (1994) model addressing some of the subsequent criticism, albeit only for one country – the UK. He also finds an effect of policy preferences of the opposition on government public spending. It remains an open question whether this result holds in a panel of West European parliamentary democracies, most compatible with the party government model (Thomassen 1994).

Another outstanding issue concerns the data used to evaluate the linkage between policy preferences of parties and public spending. All studies in this research approach derive policy preferences of the parties from the Comparative Manifesto Project (CMP) data set. Positions of political parties in CMP are derived from the content analysis of electoral manifestos of political parties. However, we know that these data contain large associated uncertainty that has not been adequately acknowledged. Benoit, Laver & Mikhaylov (2009) show that part of measurement error is derived from the stochastic nature of political text generation; while Mikhaylov, Laver & Benoit (2008) highlight an additional component of measurement error due to human misclassification in the process of manifesto coding. Consequently, taking the cue from analysis in Ansolabehere, Rodden & Snyder (2008), it is likely that the remaining "weakest link" is also be affected by measurement error. This could then explain puzzling and conflicting results discussed here.

Next this paper evaluates the linkage between the policy preferences of governing parties and public spending in a panel of West European parliamentary democracies. This paper extends the results in McDonald, Budge & Hofferbert (1999), using an original, new data set of government expenditure, explicitly evaluating the effect of opposition policy preferences on public spending. Furthermore, the particular focus of this paper is on the potential effects of measurement error evaluated in a simple error-correction method. The results indicate that the responsible party model has some empirical support in naive estimation, but this disappears when error-correction is implemented. Strong supporting evidence is also identified for the simultaneous effect of policy preferences of governments and opposition, thus raising questions over the effectiveness of representative democracy in the EU 15.

3.3 Empirical analysis

3.3.1 Model and data

Extending McDonald, Budge & Hofferbert (1999), this paper investigates the effects of the policy preferences of governments and opposition parties on the widest possible cross-nationally comparable set of government expenditure functions. Thus, in addition to spending on social welfare (Bräuninger 2005, McDonald, Budge & Hofferbert 1999) responsible party linkage is evaluated for government outlays on public services, defence, education, health care, housing, culture, and general economic affairs. Each spending area is modelled as a function of policy preferences of both government and opposition, as well as a set of co-variates adopted from models of public spending in Persson & Tabellini (2003) and Brender & Drazen (2005).

Dynamic specification of the model is based on the advice in De Boef & Keele (2008). Previous studies assumed either static specification or a partial adjustment mechanism. While the former assumes immediate achievement of desired policy output upon taking office (e.g., Budge & Hofferbert 1990, Klingemann, Hofferbert & Budge 1994); the latter is characterised by incremental adjustment and a long "memory" of public spending (e.g., King & Laver 1993). Preference for a dynamic mechanism was usually justified on the theoretical grounds. Indeed there is evidence that even in tight budgetary timetables governments have some room for manoeuvre. Upon assuming office they usually find ways to quickly 'fine tune' the structure of the outlays inherited from the previous administration (Hallerberg, Strauch & von Hagen 2001). Budgetary process is often characterised as following the incremental adjustment process (Wildavsky 1964) interspersed by sharp rises (or falls) (Peacock & Wiseman 1961). This approach to budgetary process is naturally related to recent work on punctuated equilibrium theory, which suggests that periods of stability in budgetary policy can be interrupted by periods of rapid adjustment and change (Robinson, Caver, Meier & O'Toole 2007, Jones

& Baumgartner 2005, True, Jones & Baumgartner 1999). Evidence for budget punctuations has been demonstrated for Denmark, Germany, the UK (Breunig 2006), France (Baumgartner, Foucault & François 2006), the US (Jones, Sulkin & Larsen 2003, Jones & Breunig 2007), and the US state level (Breunig & Koski 2006).

De Boef & Keele (2008, 187) suggest that broad theoretical justification is a necessary, but not a sufficient condition. They suggest that a more general dynamic model should be estimated first, with any additional restrictions statistically tested. Following general guidelines in De Boef & Keele a general error-correction model (ECM) is specified here. In this approach a change in public spending is a function of its lagged level, and differences and lagged levels of each co-variate, with specific lag length tested and evaluated. In the context of public spending a variation of this model has been previously implemented by Franzese (2002) and Bräuninger (2005). One of the advantages of estimating a general ECM model is that both short- and long-run movements in budgetary composition can be interpreted as the result of short- and long-term changes in policy preferences of political parties.

Basic specification of the model takes the form:

$$\Delta Y_t = Y_{t-1} + \Delta GOV_t + GOV_{t-1} + \Delta OPP_t + OPP_{t-1} + \Delta \mathbf{X}_t + \mathbf{X}_{t-1} + \varepsilon_t \tag{3.1}$$

where Y is expenditure on one of eight spending functions[6] (% GDP), GOV is policy preferences of government and OPP is policy preferences of opposition in parliament, and $\mathbf{X}$ is a set of controls adopted from models of public spending in Persson & Tabellini (2003) and Brender & Drazen (2005).

Following McDonald, Budge & Hofferbert (1999) the data on government spending composition for fifteen West European parliamentary democracies (EU 15) is taken from the IMF Government Finance Statistics.[7] It provides annual data (1972-2000) on government ex-

[6] There are in fact ten categories in general classification. However, “Public order & safety” is a relatively new category, and “Other expenditure” is a catch-all remainder category with no substantively interesting interpretation. Hence this study focuses on eight remaining categories that have substantive policy interpretation.

[7] Use of these data has been criticised as confusing appropriations and outlays in the US context (Wlezien & Soroka 2003), and expenditure and policy in the UK context (Soroka & Wlezien 2005) (see also methodological analysis of expenditure data compilation in the UK in Soroka, Wlezien & McLean (2006)). I feel, however, that in the dynamic setting of responsible party model with rational voters the focus should be on actual expenditure as policy output rather than simple policy statements.

penditure consistently aggregated into eight functional categories. Definition of the spending categories is presented in Table 3.1.

[TABLE 3.1 ABOUT HERE]

The data cover all expenditures for consolidated central government including budgetary, social security, and extra-budgetary accounts. Figures exclude transactions between central government units but take account of transfers between central governments and other state or regional levels of government. Each functional category is expressed in percentages of GDP. A summary comparison across categories can be seen in Figure 3.2.

[FIGURE 3.2 ABOUT HERE]

Social security and welfare expenditure is by far the largest public spending item. It also shows the biggest variation across countries and over time. Spending functions in Figure 3.2 are rank ordered in terms of average size. The plot is a combination of the summary statistics displayed by box plots and overlaid kernel density plot. For example, *Health care* appears to be bi-modally distributed, with median just under 5% of GDP and upper-adjacent value at just over 10%.

Model 3.1 will be estimated for all spending categories for the full sample of EU 15 countries from 1972 to 2000. However, to save space interpretation of the results will concentrate on social security and welfare spending as the largest budgetary item. Figure 3.3 gives a feel for the dynamics in social security spending.

[FIGURE 3.3 ABOUT HERE]

Average spending for this category increased in the EU 15 until early 1980s, then remained stable for a decade, and experienced decline from early 1990s. Although as indicated by the increase in the shaded area (95% confidence interval) from early 1990s, the average decrease in this category of public spending corresponds to its divergence across EU 15 countries.

The data for policy preferences of parties in government and opposition are derived from the Comparative Manifesto Project (CMP) data set (Budge et al. 2001, Klingemann et al. 2006). The data set contains estimates of policy positions of almost all political parties in the Western democratic tradition in the postwar period. The data set is produced by hand-coding party manifestos (3018 in current data set) on a range of policy issues. As discussed above, one of the requirements of responsible party model is the constraint of both parties and voters to a single ideological dimension, frequently identified as left-right. The CMP also provides a left-right scale widely used in empirical research as a difference between elements identified as belonging to the right (R) and those belonging to the left (L), $(R-L)$. Following Kim & Fording (1998) and Laver & Garry (2000), the measure is rescaled here, $\frac{(R-L)}{(R+L)}$, to reflect only issues relevant to the left-right dimension.While original scale ranges from -100 to 100, the rescaled version ranges from -1 to 1.

The CMP data set does not differentiate between governing and opposition parties. An original data set was created for all parties in the CMP based on available results in Woldendorp, Keman & Budge (2000) and electronic resources (Wikipedia) for some of the most recent elections, supplementing for earlier elections from McDonald, Budge & Hofferbert (1999) and Cusack & Engelhardt (2003). Then, the government's policy position on the left-right dimension is created as an average of policy positions of coalition partners weighted by the contribution of the party to the parliamentary majority of governing coalition (Browne & Franklin 1973).[8] Results in the literature on government formation suggest that coalitions tend to adopt positions that are relatively close to the centre of a policy dimension of interest (Gallagher, Laver & Mair 2006, Martin & Stevenson 2001, Volden & Carrubba 2004, Carruba & Volden 2000).[9] For opposition parties, their policy position is an average of policy positions of parties comprising opposition in parliament, weighted by their relative seat share in the opposition.

In the end, one score was produced for each government and opposition in parliament following each election. The CMP data is recorded only at the election time with an issue of a new manifesto. Therefore, time in office or opposition uninterrupted by elections contains the same information on party policy position. This may be viewed as problematic[10], however, Klingemann, Hofferbert & Budge (1994, 43) rightly note that in order to relate policy positions to annually recorded events (expenditure outlays) it is necessary to statistically account

[8] For minority governments this means that policy position of the government is the weighted position of the single party in government.

[9] In the analyses of partisan effects on public spending this approach has been adopted, for example, in Cusack (1997), Franzese (2002), and Bräuninger (2005).

[10] This issue is revisited in the discussion of the robustness studies below.

for annual nature of expenditure recording. Following their guidelines, party policy position is entered for the year of the election if election takes place before July 1st, and the year immediately after the election if it takes place on or after July 1st. Left-right score is then carried over to all subsequent years until the next election. Summary comparison of left-right positions of government and opposition is presented in Figure 3.4.

[FIGURE 3.4 ABOUT HERE]

Average dynamics of two policy position variables in the sample of the EU 15 countries are presented in Figure 3.5.

[FIGURE 3.5 ABOUT HERE]

Following Persson & Tabellini (2003) and Brender & Drazen (2005), control variables, **X**, in estimation Model 3.1 represent a set of predictors that have been shown in the literature to correlate systematically with the size of government. One of the most often cited indicators reflects Wagner's law that government expenditure increases with income. This variable is included as (the log of) real GDP per capita. Another set of determinants characterises the demographic structure of a society that may influence functional composition of government spending. Two variables are used to reflect this in Model 3.1: percentage of the population between 15 and 64 years of age and over 65. Total size of population has also been linked to increases in government spending, and is included in the model (in log form). More open economies have also been shown to have larger share of public spending in GDP, reflecting demand for social insurance in more globalised, riskier economies. Thus a measure of country's openness (sum of exports and imports as a share of GDP) is included in the model. The data for these variables are taken from the World Bank's *World Development Indicators* (2008).

Another variable measures country-specific business cycles thus capturing idiosyncratic economic shocks. It is created for each country as the (log) difference between real GDP and its trend derived from a Hodrick-Prescott filter.[11] The resulting measure (*Output gap*) is interpreted as the deviation of aggregate output from its trend value in percent. The model also includes country fixed effects centred on Germany as the largest economy in the EU 15.[12]

[11] See Persson & Tabellini (2003, 48) for details

[12] Interpretation of centred effects will be in contrasting to the grand mean. Reported constant term is the grand mean, and individual coefficients are contrasts with that mean.

3.3.2 Measurement error correction

Recent analyses in Benoit, Laver & Mikhaylov (2009) and Mikhaylov, Laver & Benoit (2008)[13] decidedly point to the existence of substantial measurement error in the CMP data. When covariates measured with error are used in linear regression models, the result is bias and inefficiency when estimating coefficients on error-laden variables (Hausman 2001, 58). These coefficients are typically expected to suffer from "attenuation bias," meaning they are likely to be biased towards zero, underestimating the effect of relevant variables. This conclusion must be qualified, however, since it depends on the relationship between the "true" predictor and the noisy proxy available to the researcher, and possibly other variables in the model. More precisely, the effect of measurement error depends on the estimation model and the joint distribution of measurement error and the other variables (Carroll et al. 2006, 41). In the case of linear regression the effects of measurement error can range from simple attenuation bias to masking of real effects, appearance of effects in observed data that are not present in the error-free data, and even reversal of signs of estimated coefficients compared to the case in the absence of measurement error.

In terms of Model 3.1, this means that variables measuring policy positions of government and opposition are contaminated with measurement error, in turn biasing coefficients in studies of the responsible party model using CMP data. This aspect has not been acknowledged previously in any of the empirical evaluations of the party government model using the CMP data. In fact, Ansolabehere, Rodden & Snyder's analysis showed that correction for measurement error can overturn existing results in the field. Measurement error in one variable necessarily affects coefficient estimates in other co-variates through variances and co-variances of all variables (Ansolabehere, Rodden & Snyder 2008, 226).

This issue is addressed by, first, estimating Model 3.1 holding measurement error at zero. This corresponds to all previous analyses in the literature, albeit with an improved estimation approach following De Boef & Keele (2008). Next I replicate the analyses, correcting for measurement error in CMP-derived variables as suggested in Benoit, Laver & Mikhaylov (2009).

[13] See companion paper in this volume that integrates both approaches and shows a generalised view of error processes in the CMP data.

Correction is done using a simple error correction model known as *simulation-extrapolation* (SIMEX) that allows generalized linear models to be estimated with correction for error-prone covariates whose variances are known or assumed (Stefanski & Cook 1995, Carroll et al. 2006).

The basic idea behind SIMEX is fairly straightforward. If a coefficient is biased by measurement error, then adding more measurement error should increase the degree of this bias. By adding successive levels of measurement error in a resampling stage, it is possible to estimate the trend of bias due to measurement error versus the variance of the added measurement error. Once the trend has been established, it then becomes possible to extrapolate back to the case where measurement error is absent. Following Carroll et al. (2006, 98–100) the SIMEX algorithm can be succinctly described as a sequence of steps that I illustrate in Figure 3.6. The example taken is from the model of *Social security & welfare* spending category, and discussed in detail in the section below. Following Model 3.1 error-correction is implemented for four variables measuring for both government and opposition the change in position on left-right dimension and position on this dimension in previous time period. First, in the simulation step additional random pseudo errors are generated from normal distribution with mean 0 and variance $\zeta_m \sigma_u^2$ and added to the original data. Since m is known and chosen to satisfy $0 = \zeta_1 < \zeta_2 < \ldots < \zeta_M$ (I use typical values {0.0, 0.5, 1.0, 1.5, 2.0}), the simulation step creates m data sets with increasingly larger measurement error variances. The total measurement error variance in the m^{th} data set is $\sigma_u^2 + \zeta_m \sigma_u^2 = (1 + \zeta_m)\sigma_u^2$. In the estimation step the model is fit on each of the generated error contaminated data sets. The simulation and estimation steps are repeated a large number of times (1000 times in all the error-correction studies) and the average is taken for each level of contamination. These averages are plotted against the values of ζ (hollow circles in Figure 3.6), and an extrapolant function is fit to the averaged, error-contaminated estimates. In terms of ζ_m an ideal, error-free data set corresponds to $(1 + \zeta_m)\sigma_u^2 = 0$, i.e. $\zeta_m = -1$.[14] Extrapolation to the ideal case ($\zeta = -1$) yields the SIMEX estimate (hollow diamond in Figure 3.6). The quadratic extrapolant function is usually pre-

[14] More precisely, for the case of simple linear regression $\widehat{\beta}_{x,naive}$ is the naive OLS estimate of β_x, and it consistently estimates $\beta_x \sigma_x^2 / (\sigma_x^2 + \sigma_u^2)$ and is biased for β_x when $\sigma_u^2 > 0$. The least squares estimate of the slope from the m^{th} data set, $\widehat{\beta}_{x,m}$, consistently estimates $\beta_x \sigma_x^2 / \{\sigma_x^2 + (1 + \zeta_m)\sigma_u^2\}$. The ideal case of a data set without measurement error in terms of ζ_m corresponds to $(1 + \zeta_m)\sigma_u^2 = 0$, and thus $\zeta_m = -1$. See Carroll et al. (2006) for full details.

ferred, since it has been shown to result in more conservative corrections for attenuation and is often more numerically stable than an alternative nonlinear function, and preferable to linear extrapolant (Carroll et al. 2006, Hardin, Schmiediche & Carroll 2003). In the error-correction studies, I implement corrections based on the more conservative quadratic extrapolation, using STATA realisation of SIMEX.[15]

[FIGURE 3.6 ABOUT HERE]

Estimate of the error variance in the CMP data is taken directly from Benoit, Laver & Mikhaylov (2009).[16] Error variance for $\frac{(R-L)}{(R+L)}$ is calculated using the formula for the approximation of the variances of a ratio of coefficients with known variances (De Boef & Keele 2008, 192):

$$Var\left(\frac{a}{b}\right) = \frac{1}{b^2}Var(a) + \frac{a^2}{b^4}Var(b) - 2\frac{a}{b^3}Cov(a,b) \tag{3.2}$$

For tractability, the covariance here is assumed to be zero.[17].

3.3.3 Results

As discussed above Model 3.1 is estimated twice for each spending category: naive estimation (assuming measurement error in CMP variables to be zero) and SIMEX error correction. Naive estimation is implemented using fixed-effects regression with panel corrected standard errors (PCSE) (Beck & Katz 1995).[18] Detailed estimation results are presented in Tables 3.2 & 3.3.

[TABLE 3.2 ABOUT HERE]

[TABLE 3.3 ABOUT HERE]

Results in Tables 3.2 & 3.3 provide an illustration of the fact that measurement error in a variable can affect other co-variates in the model. Thus, comparing naive estimates with the results of the SIMEX correction, the effects on seemingly unrelated economic control variables

[15] Information on SIMEX implementation in STATA can be found at `http://www.stata.com/merror/`.

[16] For SIMEX error-correction in sample mean of unit error variances is used.

[17] Resolving some computational issues can lead to the direct estimation of the error variance of $\frac{(R-L)}{(R+L)}$ within the simulations framework in Benoit, Laver & Mikhaylov (2009)

[18] Longer lag structure has been investigated and tested with AIC. I also find no evidence of autocorrelation in the estimated models using an LM test.

can be witnessed. For example, the short-term effect of change in the openness of the economy becomes statistically insignificant after SIMEX correction of *Education* spending. Similarly, the transitory effect of income on *Health* spending also becomes insignificant.

Political variables can have both short- and long-run effects. De Boef & Keele (2008) argue that correct interpretation of dynamic models requires estimation of both short-term effects and long-term multipliers (LRM), with the attributed standard errors around the two terms.[19] The long-run multiplier is the total effect of policy position of government (or opposition) on expenditure distributed over future time periods. Following the discussion of the responsible party model above, the LRMs are of greater substantive interest than short-run effects.[20] Table 3.4 presents naive and SIMEX estimates of the error correction rate, and short-term effects and LRM for both government and opposition.

[TABLE 3.4 ABOUT HERE]

Following suggestions in Kastellec & Leoni (2007) a more intuitive presentation of the results is also presented in Figure 3.7.

[FIGURE 3.7 ABOUT HERE]

The results indicate that the most robust short- and long-term effect of policy preferences of parties can be found in the area of *Social security & welfare*. Both naive and SIMEX estimates are large and statistically significant. However, one problem with the results is that they seem to vindicate questions asked in Petry (1991) about the effectiveness of representation process. The requirements of the responsible party model seem to be violated.

The conclusion holds, however, only for the particular policy area of social security spending. I find other spending areas to be immune to the effects of changes in opposition policy positions. Expenditure on education, economic affairs, housing, and culture all show some evidence for the effects of changes in government policy positions on left-right dimension. In fact, only health care, defence, and public services expenditure appear insulated from political effects. This is not entirely surprising since *Defence* spending has been shown to depend on the

[19] The rate of return to equilibrium or error-correction rate identifies the responsiveness of the process. To keep focus I do not substantively interpret it here by itself, but use it to calculate the LRM, and mean and median lags. I do, however, present it as summary statistic below.

[20] LRM standard errors are calculated using formula 3.2, because an alternative Bewley transformation could not be estimated for SIMEX results. For a discussion of all calculated quantities and procedures for their estimation see an exposition in De Boef & Keele (2008).

spending of both allies and enemies (e.g. Sandler & Hartley 2001). On the other hand, *General public services* spending refers to expenditure on the maintenance of government functions and together with *Health care* could be viewed as above the fray of everyday politics at least in West European democracies in the sample.

Turning to the effect of measurement error correction we can see that *Education* spending in naive estimation is affected in the long-term by changes in government policy position. However, with SIMEX error-correction the effect is significant only at 90% confidence level. Spending on *Recreational, cultural and religious affairs* is affected in the long-term by changes in government policy position. The effect is statistically significant only at 90% confidence level, but holds for both naive estimation and SIMEX correction. *Economic affairs* spending appears to be affected in the short-run by changes in government policy position, at least at 90% confidence level. This effect disappears after correcting for measurement error. Similar results are found for expenditure on *Housing*.

To interpret the results, imagine a 1-point move to the right. This would constitute a major upheaval in political terms, since our empirical distribution of positions on left-right dimension ranges within [-0.5,+0.5] bounds (see Figure 3.4). Hence, a one point movement to the right is akin to an extreme right-wing government taking over from extreme left-wing government (a more realistic example is discussed below). Thus, a one point move to the right in government left-right positions results in outlays on *Social security* (as % GDP) immediately increasing by 1.36 (0.4642) percentage points (standard error is brackets), and over the long-run spending increases by 3.8 (1.7251) percentage points. The results after SIMEX correction are much larger (and still statistically significant) with the immediate increase of 3.51 (1.5794) points and long-term effect at 7.9 (2.5704) points. The effect of opposition is much larger than government effect. A one point movement to the right for the opposition results in a 5.23 (1.689) point increase immediately, and overall increase over time of 13.87 (6.1782) percentage points.

For a more realistic example we can consider government policy position moving to the left by 0.2 points. A shift of this magnitude, for instance, occurred when Labour party replaced the Conservatives in government after the 1997 elections in the UK. The movement to the left by 0.2 points on the left-right dimension results in the immediate decrease in *Social security & welfare* spending by 0.27 percentage points, and over the long-run that leads to the drop in spending in this area by 0.77 points. SIMEX estimates puts the numbers at 0.7 and 1.6 points respectively. For the opposition we can consider movement to the right by 0.05 points. Once again commensurate to the change in opposition policy position after Labour and Conservatives traded places in government in 1997. Such a move in the opposition results in the immediate increase in *Social security & welfare* spending by 0.26 points, and in the long-run spending is increased by 0.7 points. Correcting for measurement error puts the numbers at 0.77 and 1.63 respectively. We can put it into perspective by considering that average GDP in the sample is 384 billion (constant 2000 USD). Thus even the smallest effects that we identify are very substantial in real terms.

In addition to estimating the magnitude of the total effect of a shock depicted by the LRM, De Boef & Keele suggest that it is informative to know how many periods it takes for the effect of the shock to dissipate. Thus, the median lag identifies the first lag at which at least half of the adjustment to long-run equilibrium has taken place after a change in policy position of government (or opposition). This provides information about the speed with which majority of the change in position dissipates. Mean lags show the average amount of time for the change in policy position to dissipate. Calculation results for both mean and median lag lengths are presented in Figure 3.8.

[FIGURE 3.8 ABOUT HERE]

Naive estimation and SIMEX error-correction diverge in their estimates of the median lag length for *Social security & welfare* and *Culture*. In the former naive estimation suggests it would take a year longer, while in the latter SIMEX results show a longer decay period. The average amount of time for the change in policy position to dissipate is estimated very similarly with and without error-correction. The only exception being that naive estimation for *Social security & welfare* consistently suggests that it would take about a year longer compared to SIMEX results.

Figure 3.9 shows the distributions of long-term effects of changes in policy positions for government and opposition.

[FIGURE 3.9 ABOUT HERE]

The dynamics of the effects dissipation are shown for three spending functions. Much of the effect for *Social security & welfare* spending comes through immediately, both for government and opposition. Smaller incremental changes are then spread out over several subsequent years. On the other hand, for *Education* the immediate effect of government policy shifts is extremely small. A year into the process the effect stands at a moderate rate and then slowly dissipates thereafter. For *Culture* the immediate effect is almost equivalent to the effect in the subsequent period, with a fast decay thereafter.

When government moves on left-right dimension, 35% of the decrease in *Social security & welfare* spending happens immediately, with 13% in next year, and 10% subsequent year, and so on until all of the effect is worked through. Under SIMEX error-correction 49% of total effect would come through immediately, 12% next year, and 9% year after, slowly decaying in subsequent years until the effect is fulfilled.

3.3.4 Robustness

As part of the robustness studies Model 3.1 was estimated with two-way fixed effects (adding year dummy variables). The results were not substantively different. Similarly, controlling for the election year did not change substantive results.

One of the requirements of the responsible party model is that both voters and parties are constrained by one ideological dimension, usually understood as left-right. Such a left-right scale is constructed by the CMP and widely used in empirical research. Indeed, this scale has also been used in current study with minor modifications. Among several issues raised with the scale, Benoit & Laver (2006) provide evidence that the meaning of left-right is not identical across countries. Benoit & Laver (2007*b*, 100) suggest that two substantive policy dimensions — economic left-right and social liberal-conservative — should be more applicable for cross-country time-series analysis. Both substantive scales are constructed from the CMP left-right scale. Model 3.1 was re-estimated with the positions of government and opposition evaluated on two substantive scales. With one minor exception the results are not substantively different from those presented earlier. The only difference is that government position on economic dimension appears to have a marginal effect on the expenditure on general public services (third smallest public expenditure area).

As discussed earlier, measures of policy positions on the left-right dimension by construction of the CMP data set register change only at the election time. This results in identical left-right scores being awarded to governments and opposition between elections. Policy positions of parties may also experience change between elections. One of the approaches implemented here was to apply a 3-year moving average filter (tapping retrospective and forward looking parties) and re-estimate the base model using new (smoothed) data on policy positions of governments and opposition. The results were not substantively different from the estimates presented above.

Following Franzese (2002) I also re-estimated the results controlling for spatial dependence in spending between the EU 15 countries.[21] The spatial multiplier was built using average functional expenditure in other countries in the sample for a particular year, with countries given equal weight. Once again, the results for political effects were not substantively different from the base model estimates.

I also estimated a more complex model of economic controls of public spending. I estimated models used in several recent studies on budgetary composition (Tridimas & Winer 2005, Tridimas 2001, Sanz & Velázquez 2003, Kneller, Bleaney & Gemmell 1999). A core component of these models is based on the premise that government competes with the private sector in a large number of markets. Population in a country will demand more public goods only if government is capable of producing them efficiently. Thus, the key term is the relative efficiency of the public sector compared to the private sector. The relation is approximated as the ratio of the public sector deflator to the GDP deflator. In turn, the public sector deflator is calculated as weighted mean of the government investment deflator, the public consumption deflator and public transfers. Public transfers are proxied by the consumer price index. In addition models included measures of income, total government expenditure on public services as percentage of GDP, total size of population, share of population below 15 and above 65. Estimating this more complex model resulted in essentially identical output for political variables

[21] See also Franzese & Hays (2007).

of interest. The only differences found were strong short-term effects of changes in government policy positions on *Health care* expenditure. In the end, it was judged that the complexity of new model far outweighs potential benefits of its implementation compared to the base line model used in current study.

Overall the results presented in previous section appear very robust, raising confidence in the interpretation of the outcomes.

3.4 Conclusions

The results in this paper indicate that on the face of it, the responsible party model enjoys some modest empirical support. I found that in five out of eight spending areas there is statistically significant, albeit mostly at the 90% confidence level, evidence for the translation of policy preferences of government into policy outcomes. Correcting for measurement error in the CMP derived variables further reduces the effect of policy preferences of parties. The most significant and robust effect has been found for spending on *Social security & welfare*. However, spending decisions in this policy area have also been found to be affected by policy preferences of government together with policy preferences of the opposition.

The findings undermine the empirical support base of the responsible party model. Positive findings previously reported in the literature appear to be affected by measurement error. Correcting for measurement error reduces significant effects of political variables to only one policy area (*Social security & welfare*). However, positive effects identified in the analysis of *Social security & welfare* spending appear to undermine the logical consistency of the responsible party model. Normative concepts of democratic representation in West European parliamentary democracies stipulate the key role of political parties in serving the interests of the population. Voters evaluate policy packages presented by parties at the elections, and vote for parties closest to their own preferences. Thus, parties at the elections receive a mandate from the voters to implement certain policies. The results presented in this paper suggest that most of government expenditure is being redistributed outside any popular mandate. While the largest budgetary spending area (*Social security & welfare*) is being affected by popular mandate, it is being redistributed according to the mandate of both parties in government and in parliamentary opposition. Such an outcome would correspond to effectively functioning representative democracy only if there were general concensus in the society about policies relating to *Social security and welfare*. This seems unlikely, since this spending area is the largest budgetary item and as such it has the largest redistributive potential, and hence the potential for societal conflict. One possible explanation is that the popular will does not affect public spending, while political parties represented in parliament may still affect spending. This raises questions over the effectiveness of democratic representation. Another explanation is that the responsible party model is not a valid reflection of the democratic representation process in West European parliamentary democracies. Evidence suggests that the representative process in Western Europe may more inclusive than suggested by the responsible party model. Decision making process in social security spending appears to incorporate policy preferences of both government and opposition parties. In fact this results are in line with earlier evidence presented in Klingemann, Hofferbert & Budge (1994), supporting their argument for deliberative and more inclusive nature of democracy in Western Europe.

There is also more information available about the CMP data and its inherent measurement error. Some of the latest advances have not yet translated into tangible error estimates for the CMP scores. Error estimates used in this study are based on the uncertainty due to stochastic nature of political text. We know now that there is a substantial component of measurement error in the CMP data that is derived from human misclassification in the text coding process (see companion paper in this volume and Mikhaylov, Laver & Benoit (2008) for details). The component of human error and human bias in coding manifestos has not yet been introduced into error estimates of CMP scores.

There is currently a resurgence of interest in political science in the effects of measurement error. Researchers are returning to and re-evaluating some of the established “wisdoms” in the field. For example, Treier & Jackman (2008) estimate the amount of measurement error in the widely used Polity IV indicator of democracy and show that error correction revises some of the established outcomes in the literature on “democratic peace.” The advancement of techniques dealing with measurement error in political science contexts in addition to more robust estimates of measurement error in the CMP data can lead to a similar revision of the results presented here. The author intends to continue work in this area and revisit empirical evaluation of the party government model in future work.

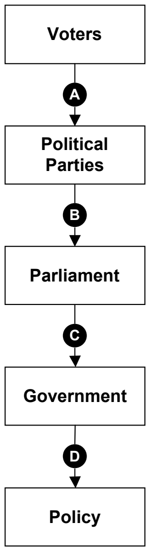


Figure 3.1: *Summary of elements in party government model.*

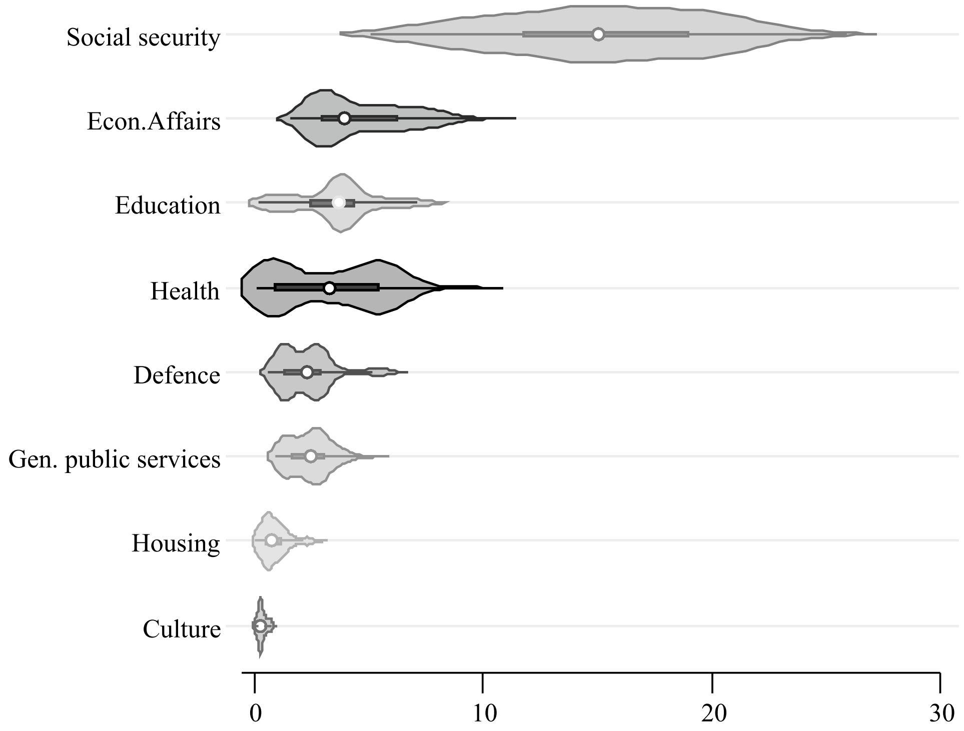


Figure 3.2: *Summary information for spending areas in the data set.* Violin plots showing kernel density with inlaid boxplot for each spending area in the data set.

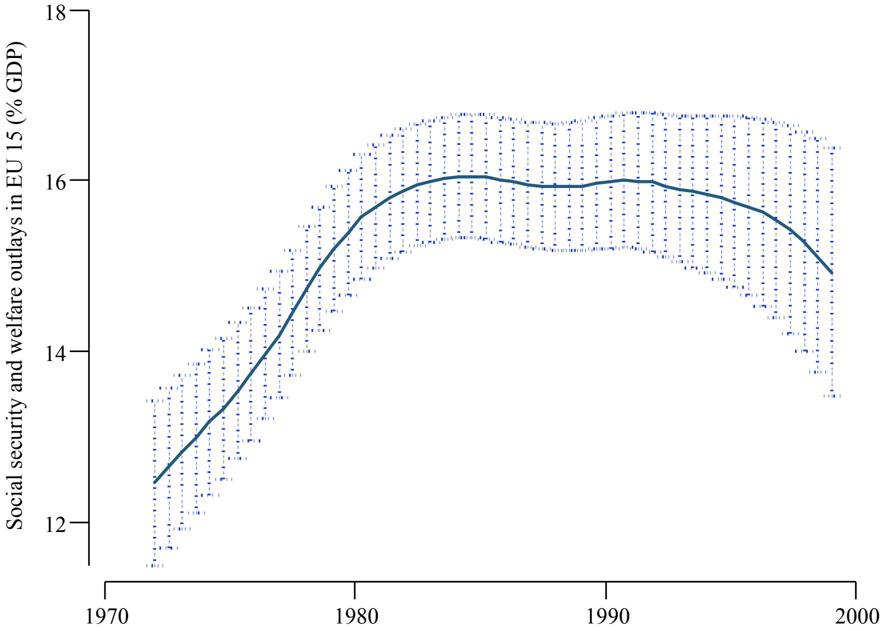


Figure 3.3: *Average spending on social security and welfare in EU 15.* Local polynomial of average central government outlays on social security and welfare in EU 15 countries over time. Shaded area represent respective 95 % confidence interval.

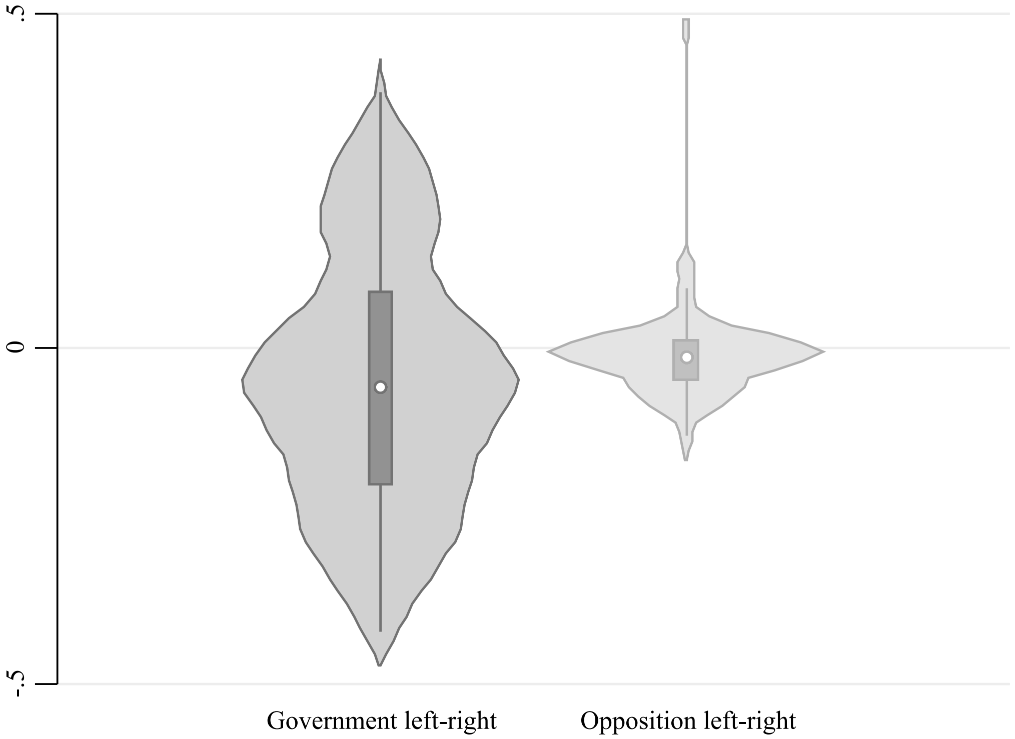


Figure 3.4: *Summary of government and opposition left-right positions in the data set.*
Violin plots showing kernel density with inlaid boxplot for left-right positions of government and opposition in the data set.

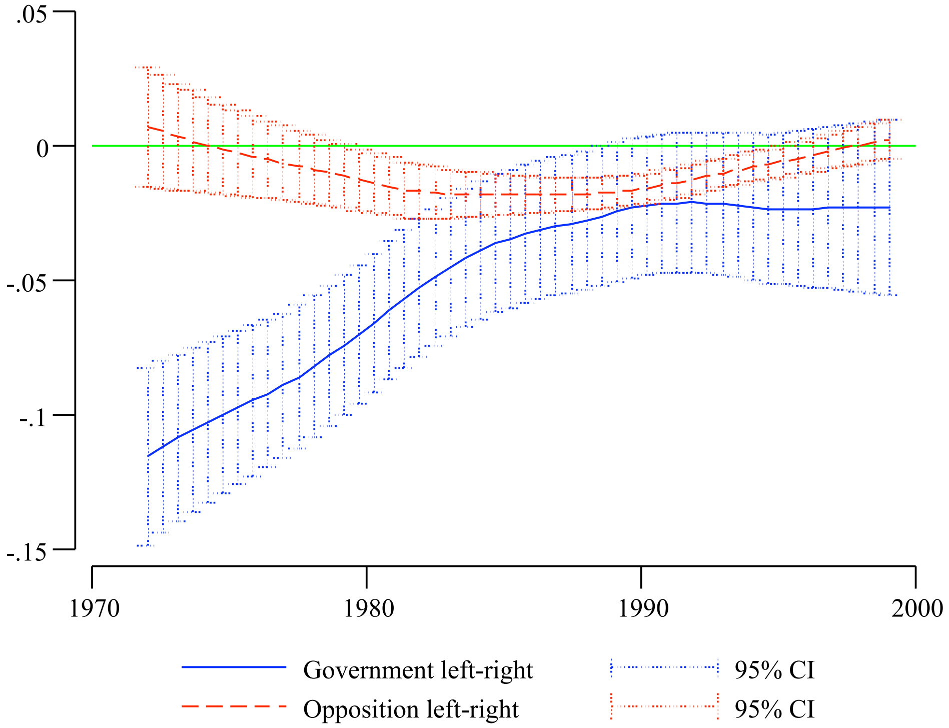


Figure 3.5: *Government and opposition average positions on left-right dimension over time in EU 15.* Local polynomial of average positions of governments and opposition in EU 15 countries over time. Shaded areas represent respective 95 % confidence intervals.

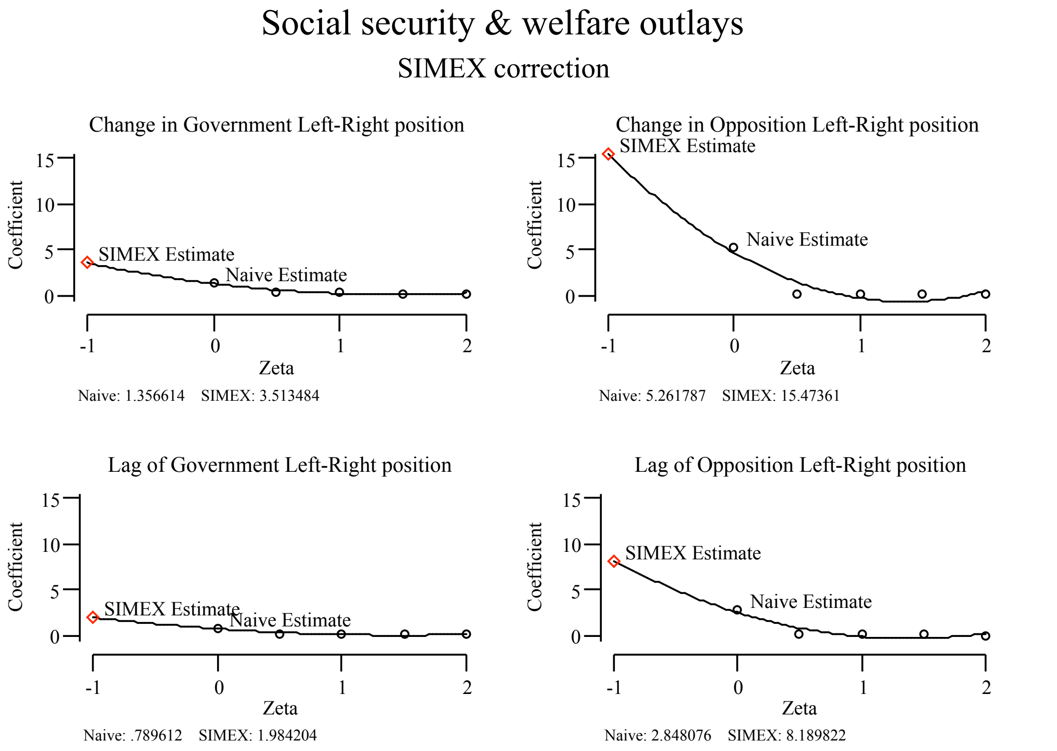


Figure 3.6: *SIMEX correction.* SIMEX correction of policy positions of government and opposition in the model of “Social security and welfare” outlays.

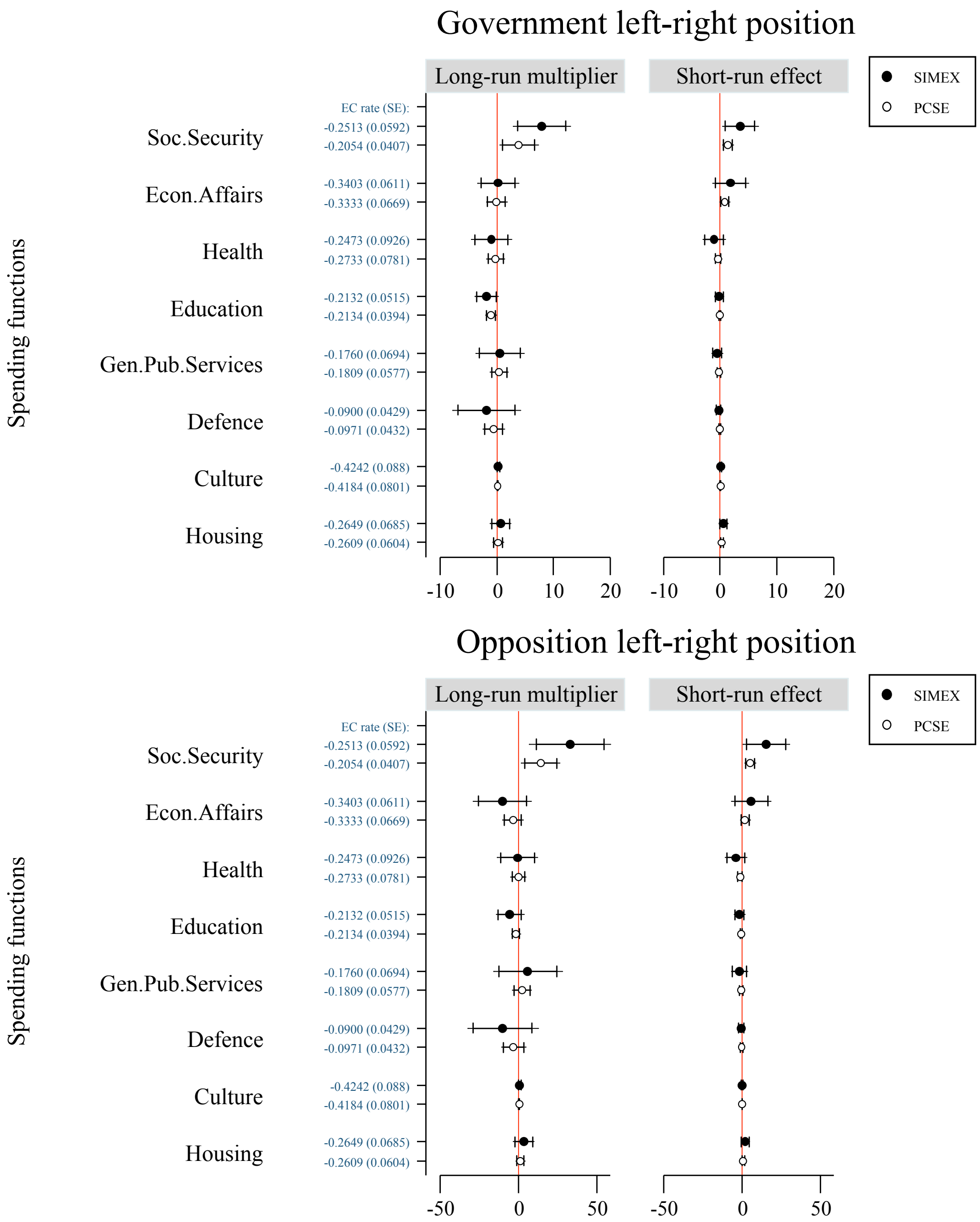


Figure 3.7: *Core estimation results.* Estimation results for government and opposition left-run position by spending area. SIMEX and PCSE estimates are compared for each outlays function for both short-run effects of change in left-run position and its long-run multiplier. Rates of return to equilibrium (error correction rates) with standard errors are shown on the margins of y-axis. Whiskers around point estimates represent 90 % and 95 % confidence intervals. SIMEX standard error estimates are based on 1000 boostrap simulations.

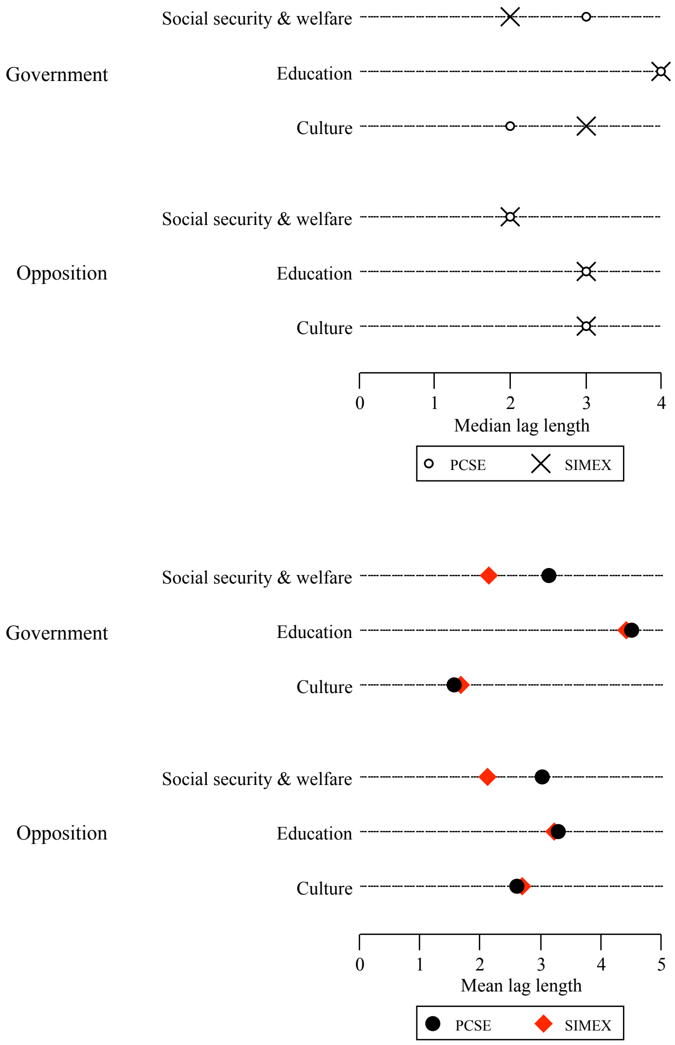


Figure 3.8: *Mean and median lag lengths.* Mean and median lag lengths calculated of the long-term effects of policy positions of government and opposition in models for *Social security and welfare*, *Education*, and *Culture* outlays.

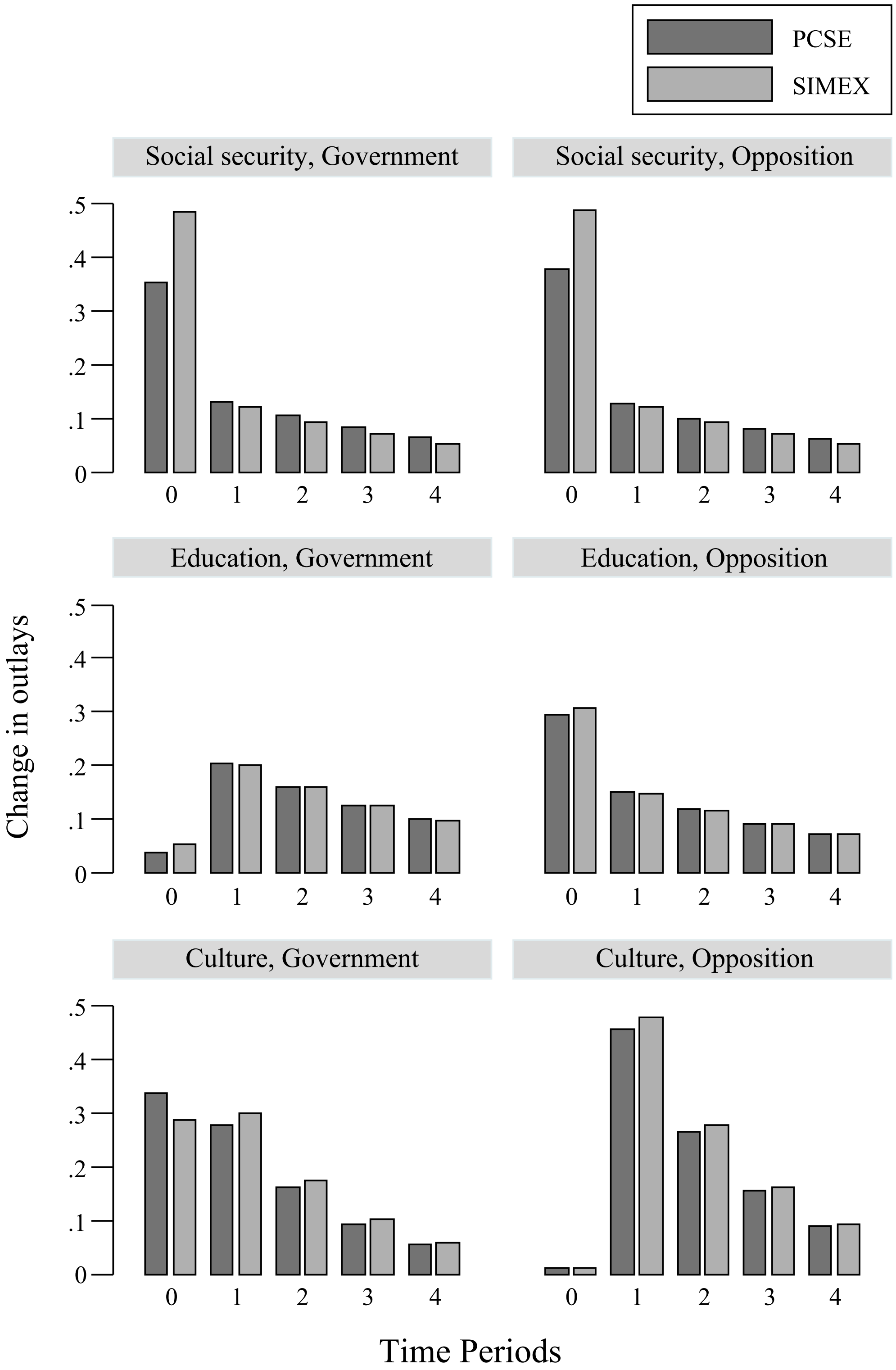


Figure 3.9: *Distribution of effects over lag lengths.* Distribution over lags of the long-term effects of policy positions of government and opposition in models of changes in *Social security and welfare*, *Education*, and *Culture* outlays.

Expenditure function	**Definition**
General public services	includes executive and legislative organs, financial and fiscal affairs, external affairs; foreign economic aid; general services; basic research; R&D general public services; public debt transactions; transfers of a general character between different levels of government
Defence	includes military defence, civil defence, foreign military aid, R&D defence
Education	includes pre-primary and primary education; secondary education; postsecondary non-tertiary education; tertiary education; education not definable by level; subsidiary services to education; R&D education
Health care	includes medical products, appliances, and equipment; outpatient services; hospital services; public health services; R&D health
Social security & welfare	includes sickness and disability; old age; survivors; family and children; unemployment; housing; social exclusion; R&D social protection
Housing	includes housing development; community development; water supply; street lighting; R&D housing and community amenities
Recreational, cultural, and religious affairs	includes recreational and sporting services; cultural services; broadcasting and publishing services; religious and other community services; R&D recreation, culture, and religion
Economic affairs	includes general economic, commercial, and labour affairs; agriculture, forestry, fishing, and hunting; fuel and energy; mining, manufacturing, and construction; transport; communication; other industries; R&D economic affairs

Table 3.1: *Definitions of government spending functions under analysis.*

Variables	Social security & welfare		Education		Economic affairs & services		Health care	
	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX
Δ Government Left-Right	1.3566**	3.5135*	-0.0431	-0.1812	0.7894+	1.7749	-0.4128	-1.1005
	(0.4642)	(1.5794)	(0.1181)	(0.4116)	(0.4586)	(1.6195)	(0.2987)	(1.0235)
Government Left-Right $_{t-1}$	0.7896*	1.9842*	-0.2311*	-0.3975+	-0.0187	0.0537	-0.0629	-0.2439
	(0.3976)	(0.823)	(0.0927)	(0.2233)	(0.3179)	(0.6285)	(0.2185)	(0.4307)
Δ Opposition Left-Right	5.2618**	15.4736*	-0.5651	-1.6794	1.9659	5.77	-1.324	-3.8675
	(1.689)	(7.4935)	(0.4534)	(1.7435)	(1.5502)	(6.32)	(0.9184)	(3.4186)
Opposition Left-Right $_{t-1}$	2.8481*	8.1898*	-0.4085	-1.1949	-1.2285	-3.5611	-0.0508	-0.2215
	(1.3275)	(3.6918)	(0.2917)	(0.9404)	(1.0866)	(3.0995)	(0.6678)	(1.6094)
Level of spending $_{t-1}$	-0.2054***	-0.2513***	-0.2134***	-0.2132***	-0.3333***	-0.3403***	-0.2733***	-0.2473**
	(0.0407)	(0.0592)	(0.0394)	(0.0515)	(0.0669)	(0.0611)	(0.0781)	(0.0926)
Δ GDP per capita	-20.0420***	-20.0804***	-3.6211***	-3.5780***	-11.5927***	-11.5661***	-3.1840**	-3.1442
	(2.3234)	(3.4115)	(0.6835)	(0.7701)	(1.9696)	(2.5205)	(1.1352)	(2.0174)
GDP per capita $_{t-1}$	2.4656**	2.1634*	0.8080***	0.9384**	-0.2842	-0.3723	-0.217	-0.0828
	(0.8264)	(0.9964)	(0.1938)	(0.3145)	(0.5968)	(0.8948)	(0.4119)	(0.5304)
Δ Output gap	-0.0292	-0.1904	-0.0108	0.0507	0.4856	0.3538	-0.455	-0.2854
	(0.5303)	(0.509)	(0.1319)	(0.1776)	(0.3359)	(0.401)	(0.2941)	(0.3011)
Output gap $_{t-1}$	-0.1752+	-0.2717+	0.0362	0.0379	0.0056	0.0393	0.0493	0.047
	(0.0901)	(0.1448)	(0.0229)	(0.0281)	(0.056)	(0.0726)	(0.0416)	(0.0529)
Δ Trade	-0.0352***	-0.0380*	-0.0076*	-0.0068	0.0034	0.0035	0.0087+	0.009
	(0.0099)	(0.0149)	(0.0033)	(0.0042)	(0.0094)	(0.0136)	(0.0045)	(0.0057)
Trade $_{t-1}$	-0.0319***	-0.0327*	-0.0096**	-0.0090*	0.0052	0.0111	0.0094*	0.0075
	(0.009)	(0.0133)	(0.003)	(0.0036)	(0.0099)	(0.0147)	(0.0044)	(0.006)
Δ Population over 65	-0.0423	-0.146	0.0271	0.034	0.243	0.2853	-0.1474	-0.171
	(0.4848)	(0.5392)	(0.1059)	(0.1191)	(0.3183)	(0.4102)	(0.3283)	(0.3309)
Population over 65 $_{t-1}$	0.1545	0.2243	-0.0254	-0.0349	-0.0069	-0.0115	-0.0574	-0.0544
	(0.1164)	(0.1396)	(0.0219)	(0.0248)	(0.0691)	(0.0907)	(0.0729)	(0.0777)
Δ Population 15 to 64	-0.0619	-0.1468	0.0915	0.0822	0.0601	0.0898	-0.0596	-0.0903
	(0.3457)	(0.4082)	(0.0858)	(0.0938)	(0.3019)	(0.3443)	(0.1543)	(0.2132)
Population 15 to 64 $_{t-1}$	-0.0824	-0.0611	-0.0141	-0.0241	-0.0398	-0.0683	0.0248	0.0166
	(0.0595)	(0.0649)	(0.0122)	(0.0172)	(0.042)	(0.0586)	(0.0324)	(0.0401)
Δ Population size	-2.6565	-5.3851	-5.608	-5.9919	7.6566	3.6861	-14.7827+	-12.213
	(16.9925)	(26.7243)	(5.5827)	(6.9101)	(17.8276)	(20.9934)	(7.9559)	(10.141)
Population size $_{t-1}$	-10.9961**	-13.1379*	-2.4331*	-2.328	-6.0448+	-3.6735	2.0193	1.0226
	(4.15)	(5.1717)	(1.0913)	(1.4677)	(3.0949)	(4.1053)	(1.5735)	(2.2669)
R-squared	0.4859		0.3593		0.3002		0.2235	
N	298	298	298	298	298	298	298	298

Table 3.2: *Estimation results for four spending functions.* Political variables are highlighted with grey. Country fixed effects are included in the estimation but not presented here. Significance codes: ‘+’ 0.10 ‘*’ 0.05 ‘**’ 0.01 ‘***’ 0.001

Variables	**Defence**		**General public services**		**Recreational, cultural, & religious affairs**		**Housing & community amenities**	
	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX
Δ Government Left-Right	-0.0963	-0.2599	-0.2035	-0.5377	0.0365	0.0726	0.2558+	0.5719
	(0.1075)	(0.249)	(0.1777)	(0.495)	(0.0295)	(0.0875)	(0.1486)	(0.3971)
Government Left-Right $_{t-1}$	-0.0591	-0.1709	0.067	0.0826	0.0452+	0.0973	0.0458	0.1692
	(0.0954)	(0.2656)	(0.1401)	(0.382)	(0.0266)	(0.06)	(0.1196)	(0.2623)
Δ Opposition Left-Right	-0.2434	-0.7016	-0.5486	-1.6435	-0.0158	-0.0492	0.6851	2.0116
	(0.578)	(1.0978)	(0.7528)	(2.6882)	(0.1257)	(0.43)	(0.5009)	(1.4066)
Opposition Left-Right $_{t-1}$	-0.325	-0.9218	0.3672	1.0142	0.0743	0.2161	0.2985	0.8303
	(0.3573)	(0.9498)	(0.5242)	(1.8943)	(0.0932)	(0.2351)	(0.375)	(0.9122)
Level of spending $_{t-1}$	-0.0971*	-0.0900*	-0.1809**	-0.1760*	-0.4184***	-0.4242***	-0.2609***	-0.2649***
	(0.0432)	(0.0429)	(0.0577)	(0.0694)	(0.0801)	(0.088)	(0.0604)	(0.0685)
Δ GDP per capita	-1.9323***	-1.8544**	-0.761	-0.7349	-0.2501	-0.2634	-1.7074**	-1.7571**
	(0.4833)	(0.6487)	(1.1155)	(1.2898)	(0.1738)	(0.1669)	(0.62)	(0.6093)
GDP per capita $_{t-1}$	0.0424	0.1162	0.3849	0.3907	0.1835**	0.1490*	-0.4957*	-0.5923*
	(0.233)	(0.242)	(0.3349)	(0.4745)	(0.0571)	(0.0694)	(0.221)	(0.2957)
Δ Output gap	0.2060*	0.2430+	-0.0465	-0.022	-0.0592+	-0.0731+	-0.1211	-0.1885
	(0.0846)	(0.1388)	(0.2046)	(0.2525)	(0.034)	(0.0422)	(0.1252)	(0.1639)
Output gap $_{t-1}$	0.0527**	0.0573*	-0.0159	-0.0232	-0.0245**	-0.0255*	-0.0449+	-0.0482
	(0.0185)	(0.0263)	(0.0364)	(0.0451)	(0.0078)	(0.01)	(0.0232)	(0.0312)
Δ Trade	-0.0032	-0.0032	-0.0088+	-0.0088	0.0026*	0.0027+	0.0071+	0.0069
	(0.0025)	(0.0036)	(0.0052)	(0.0064)	(0.001)	(0.0016)	(0.0038)	(0.0047)
Trade $_{t-1}$	-0.0022	-0.0019	-0.0053	-0.007	0.0024**	0.0023	0.0078*	0.008
	(0.0022)	(0.0027)	(0.0045)	(0.0061)	(0.0009)	(0.0016)	(0.0038)	(0.0057)
Δ Population over 65	0.0283	0.0258	-0.0796	-0.0972	0.0548	0.056	0.025	0.0314
	(0.0879)	(0.1013)	(0.181)	(0.1978)	(0.0362)	(0.037)	(0.1026)	(0.1415)
Population over 65 $_{t-1}$	-0.0003	-0.0066	-0.0495	-0.0446	-0.0113*	-0.0092	0.0673**	0.0748*
	(0.022)	(0.0212)	(0.0431)	(0.0462)	(0.0052)	(0.0057)	(0.023)	(0.0322)
Δ Population 15 to 64	0.042	0.0394	-0.0279	-0.0304	0.0249	0.0286	-0.1600+	-0.1518
	(0.0749)	(0.1005)	(0.1402)	(0.1933)	(0.0253)	(0.0271)	(0.0846)	(0.1095)
Population 15 to 64 $_{t-1}$	-0.0166	-0.0236	-0.0173	-0.0078	-0.0072*	-0.0052	-0.0058	0.0001
	(0.0127)	(0.0173)	(0.0215)	(0.0312)	(0.003)	(0.0041)	(0.0133)	(0.0171)
Δ Population size	-4.8223	-5.3598	0.7766	1.8102	-1.0685	-0.8156	14.9349*	15.1787*
	(4.8376)	(6.1292)	(8.5015)	(10.4088)	(1.3266)	(1.3531)	(6.4903)	(6.15)
Population size $_{t-1}$	-2.0182+	-1.7746	-0.2293	-0.9244	-0.3575	-0.3509	-0.8273	-0.7761
	(1.1563)	(1.5698)	(1.6715)	(2.5278)	(0.2978)	(0.3473)	(0.9035)	(1.3847)
R-squared	0.1614		0.1374		0.2735		0.2281	
N	298	298	298	298	298	298	298	298

Table 3.3: *Estimation results for four spending functions.* Political variables are highlighted with grey. Country fixed effects are included in the estimation but not presented here. Significance codes: '+' 0.10 '*' 0.05 '**' 0.01 '***' 0.001

Spending functions	**Rate of return to equilibrium**		**Short-term effect of change in government left-right position**		**Long-term multiplier of change in government left-right position**		**Short-term effect of change in opposition left-right position**		**Long-term multiplier of change in opposition left-right position**	
	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX	PCSE	SIMEX
Social security and welfare	-0.2054*	-0.2513*	1.3566*	3.5135*	3.8442*	7.8957*	5.2618*	15.4736*	13.8661*	32.5897*
	(0.0407)	(0.0592)	(0.4642)	(1.5794)	(1.7251)	(2.5704)	(1.689)	(7.4935)	(6.1782)	(13.1304)
Education	-0.2134*	-0.2132*	-0.0431	-0.1812	-1.0829*	-1.8644+	-0.5651	-1.6794	-1.9142	-5.6046
	(0.0394)	(0.0515)	(0.1181)	(0.4116)	(0.475)	(1.0722)	(0.4534)	(1.7435)	(1.4029)	(4.5071)
Economic affairs and services	-0.3333*	-0.3403*	0.7894+	1.7749	-0.0561	0.1578	1.9659	5.77	-3.6859	-10.4646
	(0.0669)	(0.0611)	(0.4586)	(1.6195)	(0.9549)	(1.8416)	(1.5502)	(6.32)	(3.3215)	(9.3723)
Health care	-0.2733*	-0.2473*	-0.4128	-1.1005	-0.2302	-0.9863	-1.324	-3.8675	-0.1859	-0.8957
	(0.0781)	(0.0926)	(0.2987)	(1.0235)	(0.8108)	(1.8061)	(0.9184)	(3.4186)	(2.4447)	(6.5644)
Defence	-0.0971*	-0.0900*	-0.0963	-0.2599	-0.6087	-1.8989	-0.2434	-0.7016	-3.3471	-10.2422
	(0.0432)	(0.0429)	(0.1075)	(0.249)	(0.9793)	(3.0408)	(0.578)	(1.0978)	(3.922)	(11.4619)
General public services	-0.1809*	-0.1760*	-0.2035	-0.5377	0.3704	0.4693	-0.5486	-1.6435	2.0299	5.7625
	(0.0577)	(0.0694)	(0.1777)	(0.495)	(0.8003)	(2.1777)	(0.7528)	(2.6882)	(3.0635)	(11.1997)
Recreational, cultural and religious affairs	-0.4184*	-0.4242*	0.0365	0.0726	0.108+	0.2294+	-0.0158	-0.0492	0.1776	0.5094
	(0.0801)	(0.088)	(0.0295)	(0.0875)	(0.0623)	(0.1251)	(0.1257)	(0.43)	(0.2204)	(0.5333)
Housing and community amenities	-0.2609*	-0.2649*	0.2558+	0.5719	0.1755	0.6387	0.6851	2.0116	1.1441	3.1344
	(0.0604)	(0.0685)	(0.1486)	(0.3971)	(0.4541)	(0.9234)	(0.5009)	(1.4066)	(1.437)	(3.3819)

Table 3.4: *Estimates of political effects by spending function.* Comparison of SIMEX error correction and naive estimates. Significance codes: '+' 0.10, '*'p<0.05

References

Abrevaya, J. & J.A. Hausman. 2004. “Response error in a transformation model with an application to earnings-equation estimation.” *The Econometrics Journal* 7(2):366–388.

Adams, James. 2001. “A Theory of Spatial Competition with Biased Voters: Party Policies Viewed Temporally and Comparatively.” *British Journal of Political Science* 31(1):121–158.

Adams, James, M. Clark, L. Ezrow & G. Glasgow. 2006. “Are Niche Parties Fundamentally Different from Mainstream Parties? The Causes and the Electoral Consequences of Western European Parties’ Policy Shifts, 1976–1998.” *American J of Political Science* 50(3):513–529.

Agresti, A. 1996. *An introduction to categorical data analysis*. Wiley New York.

Alt, James & David Lassen. 2006. “Transparency, Political Polarization, and Political Budget Cycles in OECD Countries.” *American Journal of Political Science* 50(3):530–550.

Altman, D.G. & J.M. Bland. 1983. “Measurement in medicine: the analysis of method comparison studies.” *Statistician* 32(3):307–17.

Ansolabehere, Stephen, Jonathan Rodden & James M. Snyder. 2008. “The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting.” *American Political Science Review* 102(2):215–232.

Arellano, M. 2003. *Panel Data Econometrics*. Oxford University Press.

Baron, D.P. 1991. “A Spatial Bargaining Theory of Government Formation in Parliamentary Systems.” *The American Political Science Review* 85(1):137–164.

Bartolini, S. & P. Mair. 1990. *Identity, Competition, and Electoral Availability: The Stabilisation of European Electorates 1885-1985*. Cambridge University Press.

Baumgartner, Frank, Martial Foucault & Abel François. 2006. "Punctuated equilibrium in French budgeting processes." *Journal of European Public Policy* 13(7):1086–1103.

Bäuml, Karl-Heinz. 2002. "Semantic Generation Can Cause Episodic Forgetting." *Psychological Science* 13(4):356–360.

Beck, Nathaniel & Jonathan N. Katz. 1995. "What to do (and not to do) with time-series cross-section data." *American Political Science Review* 89(3):634–647.

Benoit, Kenneth & Michael Laver. 2006. *Party Policy in Modern Democracies*. London: Routledge.

Benoit, Kenneth & Michael Laver. 2007*a*. "Benchmarks for Text Analysis: A Reply to Budge and Pennings." *Electoral Studies* 26:130–135.

Benoit, Kenneth & Michael Laver. 2007*b*. "Estimating Party Policy Positions: Comparing Expert Surveys and Hand Coded Content Analysis." *Electoral Studies* 26(1):90–107.

Benoit, Kenneth, Michael Laver & Slava Mikhaylov. 2009. "Treating Words as Data with Error: Estimating Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53(2).

Black, Sheila R. 2001. "Semantic Satiation and Lexical Ambiguity Resolution." *The American Journal of Psychology* 114(4):493–510.

Black, Sheila R. 2003. Review of Semantic Satiation. In *Advances in Psychology Research*, ed. Serge P. Shohov. Vol. 26 Nova Publishers chapter 4, pp. 63–74.

Bland, J.M. & D.G. Altman. 1986. "Statistical methods for assessing agreement between two methods of clinical measurement." *Lancet* 1(8476):307–310.

Bland, J.M. & D.G. Altman. 1995. "Comparing methods of measurement: why plotting difference against standard method is misleading." *Lancet* 346:1085–1085.

Bräuninger, Thomas. 2005. "A partisan model of government expenditure." *Public Choice* 125(3):409–429.

Brender, Adi & Allan Drazen. 2005. "Political budget cycles in new versus established democracies." *Journal of Monetary Economics* 52(7):1271–1295.

Breunig, Christian. 2006. "The more things change, the more things stay the same: a comparative analysis of budget punctuations." *Journal of European Public Policy* 13(7):1069–1085.

Breunig, Christian & Chris Koski. 2006. "Punctuated Equilibria and Budgets in the American States." *Policy Studies Journal* 34(3):363–379.

Bross, I. 1954. "Misclassification in 2 × 2 Tables." *Biometrics* 10:488–495.

Browne, E.C. & Mark N. Franklin. 1973. "Aspects of Coalition Payoffs in European Parliamentary Democracies." *American Political Science Review* 67(2):453–69.

Budge, Ian. 1994. "A new spatial theory of party competition: Uncertainty, ideology and policy equilibria viewed comparatively and temporally." *British Journal of Political Science* 24(4):443–467.

Budge, Ian. 2001. "Validating Party Policy Placements." *British Journal of Political Science* 31(1):179–223.

Budge, Ian, David Robertson & Derek Hearl. 1987. *Ideology, Strategy and Party Change: Spatial Analyses of Post-War Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara & Eric Tanenbaum. 2001. *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford: Oxford University Press.

Budge, Ian & Richard I. Hofferbert. 1990. "Mandates and Policy Outputs: US Party Platforms and Federal Expenditures." *American Political Science Review* 84(1):111–131.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski & Ciprian M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Number 105 *in* "Monographs on Statistics and Applied Probability" 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Carruba, C. J. & C. Volden. 2000. "Coalitional Politics and Logrolling in Legislative Institutions." *American Journal of Political Science* 44(2):261–277.

Chomsky, Noam. 1959. "Review of BF Skinner (1957) Verbal Behavior." *Language* 35(1):26–58.

Clinton, Joshua, Simon Jackman & Douglas Rivers. 2004. "The Statistical Analysis Of Roll Call Voting." *American Political Science Review* 98(2, May):355–370.

Converse, Philip E. & Roy Pierce. 1986. *Political representation in France*. Belknap Press of Harvard University Press.

Courant, Richard, Herbert Robbins & Ian Stewart. 1996. *What is mathematics?: an elementary approach to ideas and methods*. Oxford University Press.

Cox, G.W. 1997. *Making Votes Count: Strategic Coordination in the World's Electoral Systems*. Cambridge University Press.

Cox, N.J. 2006. "Assessing agreement of measurements and predictions in geomorphology." *Geomorphology* 76(3-4):332–346.

Cusack, Thomas R. 1997. "Partisan politics and public finance: Changes in public spending in the industrialized democracies, 1955–1989." *Public Choice* 91(3):375–395.

Cusack, Thomas R. & Lutz Engelhardt. 2003. The PGL File Collection: File Structures and Procedures. Data set WZB.

Dahl, R.A. 1971. *Polyarchy: Participation and Opposition*. Yale University Press.

Dahl, Robert A. 1956. *A Preface to Democratic Theory*. University Of Chicago Press.

De Boef, Suzanna & Luke Keele. 2008. "Taking Time Seriously." *American Journal of Political Science* 52(1):184–200.

Duso, T. & L.H. Röller. 2003. "Endogenous Deregulation: Evidence from OECD Countries." *Economics Letters* 81(1):67–71.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7(1):1–26.

Efron, Bradley & Robert Tibshirani. 1994. *An Introduction to the Bootstrap.* New York: Chapman and Hall/CRC Hall.

Erikson, R.S., M.B. Mackuen & J.A. Stimson. 2002. *The Macro Polity.* Cambridge University Press.

Esposito, Nicholas J. & Leroy H. Pelton. 1971. "Review of the measurement of semantic satiation." *Psychological Bulletin* 75(5):330–346.

Evans, Geoffrey. 2002. "European Integration, Party Politics and Voting in the 2001 Election." *Journal of Elections, Public Opinion & Parties* 12(1):95–110.

Evans, Geoffrey & Pippa Norris. 1999. *Critical Elections: British Parties & Voters in Long-term Perspective.* Sage Publications.

Fleiss, Joseph L., B. Levin & M.C. Paik. 2003. *Statistical Methods for Rates and Proportions.* 3 ed. New York: John Wiley.

Fowler, James H. & Oleg Smirnov. 2007. *Mandates, Parties, and Voters: How Elections Shape the Future.* Temple University Press.

Franzese, Robert J. 2002. *Macroeconomic Policies of Developed Democracies.* Cambridge University Press.

Franzese, Robert J. & Jude C. Hays. 2007. "Interdependence in Comparative Politics: Substance, Theory, Empirics, Substance." *Comparative Political Studies* 41(4-5):742–780.

Fuchs, Dieter & Hans-Dieter Klingemann. 1990. The Left-Right Schema. In *Continuities in Political Action: A Longitudinal Study of Political Orientations in Three Western Democracies*, ed. M. Kent Jennings & Jan W. van Deth. Berlin: Walter de Gruyter pp. 203–234.

Fuller, Wayne A. 1987. *Measurement error models*. New York: John Wiley and Sons.

Gabel, Matthew & John Huber. 2000. "Putting parties in their place: inferring party left-right ideological positions from party manifesto data." *American J of Political Science* 44:94–103.

Gallagher, Michael, Michael Laver & Peter Mair. 2006. *Representative Government in Modern Europe*. McGraw-Hill.

Gibbons, Matthew. 2004. "Review Article of Hans-Dieter Klingemann, Richard I. Hofferbert and Ian Budge's "Parties, Policies, and Democracy (Theoretical Lenses on Public Policy)"." *European Political Science* 3(3).

Golder, Sona N. 2006. "Pre-Electoral Coalition Formation in Parliamentary Democracies." *British Journal of Political Science* 36(2):193–212.

Grofman, Bernard. 2004. "Downs and Two-Party Convergence." *Annual Review of Political Science* 7:25–46.

Hallerberg, M., R. Strauch & J. von Hagen. 2001. "The Use and Effectiveness of Budgetary Rules and Norms in EU Member States." *Report Prepared for the Dutch Ministry of Finance by the Instititute of European Integration Studies in Bonn.* .

Hardin, J.W., H. Schmiediche & R.J. Carroll. 2003. "The simulation extrapolation method for fitting generalized linear models with additive measurement error." *The STATA Journal* 3(4):373–85.

Hausman, J. 2001. "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left." *The Journal of Economic Perspectives* 15(4):57–67.

Hayes, A.F. & K. Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1(1):77.

Hearl, Derek. 2001. Checking the Party Policy Estimates: Reliability. In *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*, ed. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara & Eric Tanenbaum. Oxford U. Press.

Heise, D.R. 1969. "Separating Reliability and Stability in Test-Retest Correlation." *American Sociological Review* 34(1):93–101.

Hix, Simon, Abdul Noury & Gérard Roland. 2006. "Dimensions of Politics in the European Parliament." *American Journal of Political Science* 50(2, April):494–511.

Hofferbert, Richard I., Ian Budge & Michael D. McDonald. 1993. "On Party Platforms, Mandates, and Government Spending: Response to King & Laver 1993." *American Political Science Review* 87(3):747–750.

Hopkins, Daniel & Gary King. 2007. "Extracting Systematic Social Science Meaning from Text." Harvard University manuscript. http://gking.harvard.edu/files/words.pdf.

Hyland, James L. 1995. *Democratic Theory: The Philosophical Foundations*. Manchester University Press.

Jakobovits, Leon A. 1967. "Semantic satiation and cognitive dynamics." *Journal of Special Education* 2:35–44.

Jakobovits, Leon A. & R. Hogenraad. 1967. "Some suggestive evidence on the operation of semantic generation and satiation in group discussions." *Psychological Reports* 20:1247–1250.

Jakobovits, Leon A. & W.E. Lambert. 1963. The effects of repetition in communication on meanings and attitudes. In *Television and human behavior*, ed. L. Arons & M.A. May. New York: Appleton-Century-Crofts pp. 167–176.

Janda, Kenneth, Robert Harmel, C. Edens & P. Goff. 1995. "Changes in Party Identity: Evidence from Party Manifestos." *Party Politics* 1(2):171–196.

Jones, Bryan D. & Christian Breunig. 2007. “Noah and Joseph Effects in Government Budgets: Analyzing Long-Term Memory.” *Policy Studies Journal* 35(3):329–348.

Jones, Bryan D. & Frank Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. University Of Chicago Press.

Jones, Bryan D., Tracy Sulkin & Heather A. Larsen. 2003. “Policy Punctuations in American Political Institutions.” *American Political Science Review* 97(1):151–169.

Jurafsky, Daniel & James H. Martin. 2000. *Speech and Natural Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.

Kastellec, Jonathan & Eduardo Leoni. 2007. “Using Graphs Instead of Tables in Political Science.” *Perspectives on Politics* 5(4):755–771.

Kim, Heemin & Richard C. Fording. 1998. “Voter Ideology in Western Democracies, 1946-1989.” *European Journal of Political Research* 33(1):73–97.

Kim, Heemin & Richard C. Fording. 2002. “Government partisanship in Western democracies, 1945-1998.” *European Journal of Political Research* 41(2):187–206.

King, Gary & Michael Laver. 1993. “On Party Platforms, Mandates, and Government Spending.” *American Political Science Review* 87(3):744–747.

King, Gary & Michael Laver. 1999. “Many publications, but still no evidence.” *Electoral Studies* 18(4):597–598.

King, Gary, Robert Keohane & Sidney Verba. 1994. *Designing Social Inquiry*. Princeton: Princeton University Press.

King, Gary & Ying Lu. 2008. “Verbal Autopsy Methods with Multiple Causes.” *Statistical Science* 23(1, February).

Kirkpatrick, Evron M. 1971. “Toward a More Responsible Two-Party System: political science, policy science, or pseudo-science.” *American Political Science Review* 65(4):965–990.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge & Michael McDonald. 2006. *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union and OECD 1990-2003*. Oxford: Oxford University Press.

Klingemann, Hans-Dieter, Richard I. Hofferbert & Ian Budge. 1994. *Parties, Policies, and Democracy*. Westview Press.

Kneller, R., M. F. Bleaney & N. Gemmell. 1999. "Fiscal policy and growth: evidence from OECD countries." *Journal of Public Economics* 74(2):171–190.

Kounios, John. 2007. Functional modularity of semantic memory revealed by event-related brain potentials. In *Neural basis of semantic memory*, ed. J. Hart & M.A. Kraut. Cambridge University Press pp. 65–104.

Kounios, John, Sonja A. Kotz & Phillip J. Holcomb. 2000. "On the locus of the semantic satiation effect: Evidence from event-related brain potentials." *Memory & Cognition* 28(8):1366–1377.

Krippendorff, Klaus. 1967. An examination of content analysis: A proposal for a general framework and an information calculus for message analytic situations. PhD thesis University of Illinois, Urbana.

Krippendorff, Klaus. 1970. Bivariate agreement coefficients for reliability of data. In *Sociological Methodology*, ed. E. F. Borgatta & G. W. Bohrnstedt. Vol. 2 San Francisco: Jossey-Bass pp. 139–150.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: Sage.

Kuha, Jouni & Chris Skinner. 1997. Categorical Data Analysis and Misclassification. In *Survey Measurement and Process Quality*. New York: John Wiley & Sons.

Landis, J.R. & G.G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33(1):159–174.

Laver, Michael. 2001. “On Mapping Policy Preferences Using Manifesto Data.” mimeo, Trinity College Dublin.

Laver, Michael & John Garry. 2000. “Estimating Policy Positions from Political Texts.” *American Journal of Political Science* 44(3):619–634.

Laver, Michael & Kenneth A. Shepsle. 1996. *Making and Breaking Governments: Cabinets and Legislatures in Parliamentary Democracies*. Cambridge University Press.

Laver, Michael, Kenneth Benoit & John Garry. 2003. “Estimating the policy positions of political actors using words as data.” *American Political Science Review* 97(2):311–331.

Laver, Michael & Norman Schofield. 1998. *Multiparty Governments: The Politics of Coalition in Europe*. Ann Arbor: The University of Michigan Press.

Lederer, Wolfgang & Helmut Küchenhoff. 2006. “A short Introduction to the SIMEX and MCSIMEX.” *R News* 6(4):26–31.

Lee, D. S., E. Moretti & M. J. Butler. 2004. “Do Voters Affect Or Elect Policies? Evidence From The US House.” *Quarterly Journal of Economics* 119(3):807–859.

Lin, L.I.-K. 1989. “A concordance correlation coefficient to evaluate reproducibility.” *Biometrics* 45(1):255–268.

Lin, L.I.-K. 2000. “A note on the concordance correlation coefficient.” *Biometrics* 56(1):324–325.

Mair, Peter. 1987. *The Changing Irish Party System: Organisation, Ideology and Electoral Competition*. Pinter.

Mansergh, Lucy E. & Rorbert Thomson. 2007. “Election Pledges, Party Competition, and Policymaking.” *Comparative Politics* 39(3):311–29.

Martin, Lanny W. W. & R. T. Stevenson. 2001. “Government Formation in Parliamentary Democracies.” *American Journal of Political Science* 45(1):33–50.

Mayhew, David R. 1991. *Divided We Govern: Party Control, Lawmaking, and Investigations, 1946–1990*. Yale: Yale University Press.

McDonald, Michael D. & Ian Budge. 2005. *Elections, Parties, Democracy: Conferring the Median Mandate*. Oxford University Press.

McDonald, Michael D., Ian Budge & Richard I. Hofferbert. 1999. "Party mandate theory and time series analysis: A theoretical and methodological response." *Electoral Studies* 18:587–596.

McDonald, Michael & Silvia Mendes. 2001*a*. Checking the Party Policy Estimates: Convergent Validity. In *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*, ed. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara & Eric Tanenbaum. Oxford University Press.

McDonald, Michael & Silvia Mendes. 2001*b*. The policy space of party manifestos. In *Estimating the Policy Position of Political Actors*, ed. Michael Laver. London: Routledge.

Meguid, B. 2005. "Competition Between Unequals: The Role of Mainstream Party Strategy in Niche Party Success." *American Political Science Review* 99(3):347–359.

Mikhaylov, Slava, Michael Laver & Kenneth Benoit. 2008. "Coder Reliability and Misclassification in Comparative Manifesto Project Codings." Paper presented at the 66th MPSA Annual National Conference, Palmer House Hilton Hotel and Towers, April 3–6, 2008.

Miller, Warren E. & Donald E. Stokes. 1963. "Constituency Influence in Congress." *American Political Science Review* 57(1):45–56.

Milner, H.V. & B. Judkins. 2004. "Partisanship, Trade Policy, and Globalization: Is There a Left-Right Divide on Trade Policy?" *International Studies Quarterly* 48(1):95–120.

Monroe, Burt & Ko Maeda. 2004. "Talk's cheap: Text-based estimation of rhetorical idealpoints." Working paper: Michigan State University.

Peacock, Alan T. & Jack Wiseman. 1961. *The Growth of Public Expenditure in the United Kingdom*. Princeton University Press, Princeton, NJ.

Persson, Torsten & Guido Tabellini. 2003. *The Economic Effects of Constitutions*. MIT Press.

Petry, Françoise. 1988. "The Policy Impact of Canadian Party Programs: Public Expenditure Growth and Contagion from the Left." *Canadian Public Policy/Analyse de Politiques* 14(4):376–389.

Petry, Françoise. 1991. "Fragile Mandate: Party Programmes and Public Expenditures in the French Fifth Republic." *European Journal of Political Research* 20(2):149–171.

Petry, Françoise. 1995. "The Party Agenda Model: Election Programmes and Government Spending in Canada." *Canadian Journal of Political Science/Revue canadienne de science politique* 28(1):51–84.

Powell, G. Bingham. 2004. "Political Representation in Comparative Politics." *Annual Review of Political Science* 7(1):273–296.

Pynte, Joel. 1991. "The locus of semantic satiation in category membership decision and acceptability judgment." *Journal of Psycholinguistic Research* 20(4):315–335.

Rae, D.W. & H. Daudt. 1976. "The Ostrogorski paradox: a peculiarity of compound majority decision." *European Journal of Political Research* 4(4):391–398.

Ray, Leonard. 1999. "Measuring party orientations towards European integration: Results from an expert survey." *European Journal of Political Research* 36(2):283–306.

Ray, Leonard. 2007. "Validity of measured party positions on European integration: Assumptions, approaches, and a comparison of alternative measures." *Electoral Studies* 26(1):11–22.

Riker, William H. 1982. *Liberalism Against Populism: A Confrontation Between the Theory of Democracy and the Theory of Social Choice*. San Francisco: Freeman.

Riker, William H. 1996. *The Strategy of Rhetoric: Campaigning for the American Constitution*. Yale University Press. (ed. Calvert, Randall L. and Mueller, John and Wilson, Rick K.).

Roberts, Chris. 2008. "Modelling patterns of agreement for nominal scales." *Statistics in Medicine* 27(6):810–830.

Robertson, David. 1976. *A Theory of Party Competition*. London and New York: Wiley.

Robinson, S. E., F. Caver, K. J. Meier & L. J. O'Toole. 2007. "Explaining Policy Punctuations: Bureaucratization and Budget Change." *American Journal of Political Science* 51(1):140–150.

Rogan, W. J. & B. Gladen. 1978. "Estimating Prevalence from the Results of a Screening Test." *American Journal of Epidemiology* 107:71–76.

Sandler, T. & K. Hartley. 2001. "Economics of Alliances: The Lessons for Collective Action." *Journal of Economic Literature* 39(3):869–896.

Sanz, I. & Francisco J. Velázquez. 2003. "Fiscal illusions, fiscal consolidation and government expenditure composition in the OECD: a dynamic panel data approach.".

Schattschneider, E. E. 1950. "Toward a More Responsible Two-Party System: A Report of the Committee on Political Parties (1950)." *American Political Science Review* 44(3).

Schofield, Norman. 1993. "Political competitition and multiparty coalition governments." *European Journal of Political Research* 23(1):1–33.

Schofield, Norman & Itai Sened. 2006. *Multiparty Democracy: Elections and Legislative Politics*. Cambridge University Press.

Schumpeter, Joseph A. 2000[1943]. *Capitalism, Socialism and Democracy*. Routledge.

Sigelman, L. & H.B. Emmett. 2004. "Avoidance or Engagement? Issue Convergence in U.S. Presidential Campaigns, 1960-2000." *American Journal of Political Science* 48(4):650–661.

Skinner, B.F. 1957. *Verbal behavior*. Appleton-Century-Crofts, New York.

Slapin, Jonathan & Sven-Oliver Proksch. 2007. "A Scaling Model for Estimating Time-Series Policy Positions from Texts." Paper presented at the annual meeting of the Midwest Political Science Association, April 12.

Soroka, Stuart N. & Christopher Wlezien. 2005. "Opinion–Policy Dynamics: Public Preferences and Public Expenditure in the United Kingdom." *British Journal of Political Science* 35(4):665–689.

Soroka, Stuart N., Christopher Wlezien & Iain McLean. 2006. "Public expenditure in the UK: how measures matter." *Journal of the Royal Statistical Society Series A* 169(2):255–271.

Stefanski, L. A. & J. R. Cook. 1995. "Simulation-Extrapolation: The Measurement Error Jackknife." *Journal of the American Statistical Association* 90(432, December):1247–1256.

Stokes, Donald E. & Warren E. Miller. 1962. "Party Government and the Saliency of Congress." *Public Opinion Quarterly* 26(4):531–546.

Strom, K. & J.Y. Lejpart. 1989. "Ideology, Strategy, and Party Competition in Postwar Norway." *European Journal of Political Research* 17(3):263–288.

The World Bank. 2008. *World Development Indicators*. Washington D.C.: The World Bank.

Thomassen, Jacques. 1994. Empirical research into political representation: Failing democracy or failing models. In *Elections at home and abroad: Essays in honor of Warren E. Miller*, ed. M. Kent Jennings & Thomas E. Mann. Ann Arbor: University of Michigan Press pp. 237–264.

Thomassen, Jacques. forthcoming. The legitimacy of the European Union after enlargement. In *The legitimacy of the European Union after enlargement*, ed. Jacques Thomassen. Oxford University Press.

Thome, Helmut. 1999. "Party mandate theory and time-series analysis: a methodological comment." *Electoral Studies* 18(4):569–585.

Thomson, Rorbert. 1999. The Party Mandate: Election Pledges and Government Actions in the Netherlands, 1986-1998 PhD thesis Rijksuniversiteit Groningen.

Thomson, Rorbert, J. Boerefijn & F. Stokman. 2004. "Actor alignments in European Union decision making." *European Journal of Political Research* 43(2):237–261.

Treier, Shawn & Simon Jackman. 2008. "Democracy as a latent variable." *American Journal of Political Science* 52(1):201–217.

Tridimas, George. 2001. "The Economics and Politics of the Structure of Public Expenditure." *Public Choice* 106(3):299–316.

Tridimas, George & S.L. Winer. 2005. "The political economy of government size." *European Journal of Political Economy* 21(3):643–666.

True, J. L., B. D. Jones & F. R. Baumgartner. 1999. "Punctuated-Equilibrium Theory: Explaining Stability and Change in American Policymaking." *Theories of the Policy Process* pp. 97–115.

van der Brug, Wouter & Cees van der Eijk. 2007. *European elections & domestic politics: lessons from the past and scenarios for the future*. University of Notre Dame Press.

van der Brug, Wouter, M. Fennema & J. Tillie. 2005. "Why Some Anti-Immigrant Parties Fail and Others Succeed: A Two-Step Model of Aggregate Electoral Support." *Comparative Political Studies* 38(5):537–573.

van der Eijk, Cees & Mark N. Franklin. 1996. *Choosing Europe?: The European Electorate and National Politics in the Face of Union*. University of Michigan Press.

van der Eijk, Cees, Mark N. Franklin & Wouter van der Brug. 1999. Policy Preferences and Party Choice. In *Political Representation and Legitimacy in the European Union*, ed. Hermann Schmitt & Jacques Thomassen. Oxford University Press pp. 161–85.

Volden, C. & C. J. Carrubba. 2004. "The Formation of Oversized Coalitions in Parliamentary Democracies." *American Journal of Political Science* 48(3):521–537.

Volkens, Andrea. 2001*a*. Manifesto Research Since 1979. From Reliability to Validity. In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge pp. 33–49.

Volkens, Andrea. 2001*b*. Quantifying the Election Programmes: Coding Procedures and Controls. In *Mapping Policy Preferences: Parties, Electors and Governments: 1945-1998:*

Estimates for Parties, Electors and Governments 1945-1998, ed. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald & Silvia Mendes. Oxford: Oxford University Press.

Volkens, Andrea. 2007. "Strengths and Weaknesses of Approaches to Measuring Policy Positions of Parties." *Electoral Studies* 26(1):108–120.

Webb, P.D. 2000. *The Modern British Party System.* Sage Publications.

Wildavsky, Aaron. 1964. *The Politics of the Budgetary Process.* Boston: Little, Brown.

Wlezien, Christopher & Stuart N. Soroka. 2003. "Measures and Models of Budgetary Policy." *Policy Studies Journal* 31(2):273–286.

Woldendorp, Jaap, Hans Keman & Ian Budge. 2000. *Party Government in 48 Democracies (1945-1998): Composition-Duration-Personnel.* Dordrecht, Kluwer Academic Publishers.