Technische Universität Kaiserslautern

KOPS

Kaiserslautern Occasional Papers in Political Science

Silke Hans
Christoph Hönnige

Noughts and Crosses. Challenges in Generating Political Positions from CMP-Data

No. 2/2008

Politikwissenschaft der Technischen Universität Kaiserslautern

Noughts and Crosses

Challenges in Generating Political Positions from CMP-Data

Silke Hans
Christoph Hönnige

ISSN 1861-7018

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.ddb.de** abrufbar.

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at **http://dnb.ddb.de**.

Kaiserslautern Occasional Papers in Political Science (KOPS)

Wissenschaftliche Schriftenreihe der Politikwissenschaft
der Technischen Universität Kaiserslautern

Herausgeber: Politikwissenschaft
Prof. Dr. Jürgen Wilzewski
Postfach 3049
Technische Universität Kaiserslautern
D-67653 Kaiserslautern
http://www.uni-kl.de/politik/pubs/kops/

Satz: David Sirakov, M.A.

Verlag: Technische Universität Kaiserslautern

Druck: Technische Universität Kaiserslautern
ZBT – Abteilung Foto-Repro-Druck

D-386

Das vorliegende Papier ist bereits in der Reihe BSSE - Berliner Studien zur Soziologie Europas der FU Berlin im Juli 2008 erschienen.

Als Manuskript gedruckt. Printed in Germany.

ISSN 1861-7018

Authors

Silke Hans, Lecturer and Research Associate in Sociology, Freie Universität Berlin.

Christoph Hönnige, J.-Prof. Dr. rer. pol., Assistant Professor for Comparative Politics, Technische Universität Kaiserslautern.

Abstract

The *Comparative Manifesto Project* (CMP) dataset is the only dataset providing information about the positions of parties for comparative researchers across time and countries. This article evaluates its structure and finds a peculiarity: A high number of zeros and their unequal distribution across items, countries and time. They influence the results of any procedure to build a scale, but especially those using factor analyses.

The article shows that zeroes have different meanings: Firstly, there are substantial zeroes in line with saliency theory. Secondly, zeroes exist for non-substantial reasons: The length of a manifesto and the percentage of uncoded sentences, both strongly varying across time and country. We quantify the problem and propose a procedure to identify data points containing non-substantial zeroes. For the future comparative use of the dataset we plead for a theoretical selection of items combined with the information about the likelihood that zeroes are substantially meaningful.

Content

1. Introduction

The spatial analysis of politics is certainly one of the most powerful paradigms of modern political science. Spatial models were developed in great numbers to analyse government formation (Laver and Schofield 1990), government survival (Huber 1996), coalition-building (Mayhew 1974), bicameralism (Tsebelis and Money 1995), the relationship between bureaucracies and government (McCubbins, Noll, and Weingast 1987) or the political systems in general (Tsebelis 2002) – to give just a few examples.

If political scientists are not happy with country case studies and do not want so succumb to model Platonism either, they are faced with a basic problem: where to get data to test the increasingly elaborate models in a comparative manner? There are a few possibilities ranging from expert-surveys (Castles and Mair 1984; Laver and Hunt 1992; Laver and Benoit 2006) via opinion polls (World Values Survey, Eurobarometer) to the Comparative Manifesto Project data (CMP) finally published in 2001 for Western Countries and 2006 for Central and Eastern Europe. The CMP data are based on election manifestos of political parties and cover all free elections from 49 countries between 1945 and 2003. It is the only dataset available that provides data for a great number of countries over a long period of time. It enables us thus to use positional data for cross-sectional, time-series and pooled time-series analyses in comparative politics.

Over the last years there have been a few attempts to develop methods for extracting positions of political parties from the CMP data (e.g. Budge, Robertson, and Hearl 1987; Laver and Budge 1992; Laver and Garry 2000; Gabel and Huber 2000; Kim and Fording 1998, 2002; Franzmann and Kaiser 2006). They deliver entirely divergent results – not only with regard to distances between parties but also with respect to their ordinal ranking on a left-right scale. This is a somewhat alarming puzzle for comparative researchers using the dataset as a supply of information about the positions of parties.

We discovered a previously unmentioned peculiarity of the CMP-dataset: An extraordinary high number of zeros and their unequal distribution across items, countries and time. Additionally, there are

two types of zeroes in the dataset: substantial ones in line with saliency theory and non-substantial ones. The aim of this article is to show the effects of this peculiarity on existing scales explain the occurrence of non-substantial zeroes with the length of a manifesto and the percentage of uncoded sentences and propose a procedure to identify the size of the problem per data point.

The findings have major implications for the future use of the dataset and research on scale construction. Firstly, the uneven distribution of zeros across countries and time presents a special difficulty for comparative researchers. If she/he wants to set up a comparative study, a time-series analysis or a combination of both in a pooled times-series analysis, the data about positions of parties must be comparable across time and countries. Secondly, there are methodological concerns in regard of existing procedures using factor analysis to select items or generate positions, since the high number of zeros leads to a low correlation of items and thus a violation of a main assumption of factor analysis. Thirdly, researchers' construction scales have to take the different types of zeroes into account and find a solution to deal with them. Currently used procedures are unable to discriminate between those two types of zeros which may distort the results. We, fourthly, argue that items for scale building should be chosen on a theoretical base and propose a way to identify items and also data points that are affected. Researchers constructing scales are thus able to quantify the problem and tackle it.

We proceed as follows: Firstly, we give a short introduction on the basic ideas and assumptions of the saliency approach underlying the CMP-data. In a second step we classify and review previous attempts to extract a cross-national left-right-scale of political conflict and party positions from the CMP-data. We show that entirely different results are created by three well-known approaches: Laver and Budge 1992, Gabel and Huber 2000, and Kaiser and Franzmann 2006 constitute an alarming puzzle. Fourthly, we provide evidence that the CMP-data show a peculiarity due to the salience approach underlying it that has not been tackled so far. This peculiarity leads to theoretical and methodological pitfalls in analysing CMP data. Fifthly, we quantify the problem and propose a solution to identify data points suffering from non-substantial zeroes.

2. Basic Components of the Saliency Approach and the CMP Dataset

Party positions might be identified in various ways. The traditional way is to conduct expert surveys (Castles and Mair 1984; Laver and Hunt 1992; Benoit and Laver 2006). While they deliver valid data, they only give the position of parties at a certain point of time. It is not possible to track the movement of parties in political space. Alternatively, the position of a party can be identified via opinion polls (Eurobarometer, World Values Survey). Regularly conducted opinion polls can track those movements. However, the positions of parties are identified via the self-positioning of their electorate and give only implicit information about the actual parties' positions. Manifesto data are a third way to identify party locations in political space. One method to generate positions from a manifesto text is to use computer aided procedures such as wordscore or wordfish (Laver and Garry 2000; Laver, Benoit, and Garry 2003; Martin and Vanberg 2008; Slapin and Proksch 2008). Although highly reliable, they do not deliver comparative information and are dependent on the quality of the reference text and the reference position used to anchor parties. Another way is to use the Comparative Manifest Project (CMP) data set.

The basic idea behind the CMP data set is the saliency approach. It arises from the finding by Robertson (1976) that parties do not oppose each other directly issue by issue. Parties do not take pro and con positions and fight publicly on all issues as it would be assumed by a traditional Downsian model. On the contrary, "…they duck and weave, avoiding direct hits form their opponents..." (Budge 2001b: 60). This means there is only one tenable position per issue and each issue discussed in the manifesto should be clearly identifiable as left or right. If a party supports a topic it should mention it very often, if a party opposes the topic it will be ignored in the manifesto (Budge 2001a: 82). This way, a frequently mentioned issue gives information about the position of an actor. The more often it is mentioned the further it will

shift the actor's position to the left or right. The notorious Sir Humphrey claims this logic to be straightforward: never mention anything controversial, it might loose you votes or even the election (Lynn and Jay 1981: 142). Consequently, what we find in the CMP dataset are 56 items for all countries and a number of subcategories for CEE countries covering a whole range of policy areas. Each data point gives the number of quasi-sentences found in a manifesto per item, election and party in percent (Volkens 2001: 93ff). The dataset covers 49 countries with 3305 cases (parties / elections) in the period from 1945 until 2003.

Despite the plausibility of this approach criticism has been brought forward concerning the validity of saliency theory. (Budge 2001b: 52): The first questions the validity of the concept as a whole. Laver (2001) argues that salience of a topic means in fact the importance of an issue in political competition and importance is independent of the positions of actors. Moreover, the data only deliver stated policy positions and no information about ideal policy positions which we normally use in theoretical models. The second criticism questions the applicability of salience data to measure policy positions directly. Gabel and Huber (2000) state that there is no analysis of whether and under what conditions measurements lead to reasonable results. Laver and Garry (2000) and Janda, Harmel, and Goff (1995) doubt the validity of the coding scheme, Laver, Benoit, and Mikhaylov (2008) the reliability of the coder in the different countries.

However, if one accepts the basic assumptions of the saliency approach, the CMP dataset indeed seems excellent to test spatial models comparatively over time. The large number of items allows to find either one dimension which can be considered as left-right axis or a multiplicity of dimensions either part of the left-right-axis or independent from it.

3. An Alarming Puzzle: Different Ordinal Positions of Parties on Different Scales

3.1. Different Positions and Rankings from the Same Data

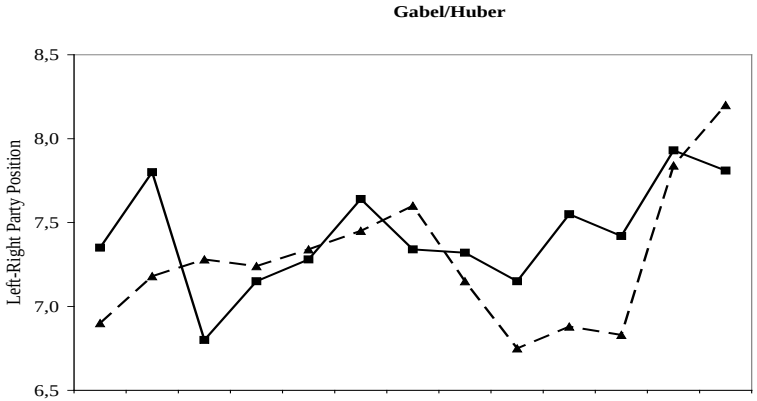
The CMP dataset might be a very powerful tool to assess the huge amount of simple and complex spatial models. In fact, most comparative researchers are probably not interested in the dataset itself but in actual party positions they need to test the hypotheses of their projects. The dataset contains a left-right scale developed by Laver and Budge (1992), but in the literature there are numerous other scales to be found. So, which one should a comparative researcher use? Faced with this question we decided to examine several scales, starting with a country where the results should be easy to compare and interpret because it has a two party system and is well known: the United States. Of the various approaches to construct a left-right-scale, the most prominent and widely discussed are Laver and Budge 1992, Gabel and Huber 2000 and Franzmann and Kaiser 2006; so we decided to test them comparatively across all items and parties for post 1945 elections.

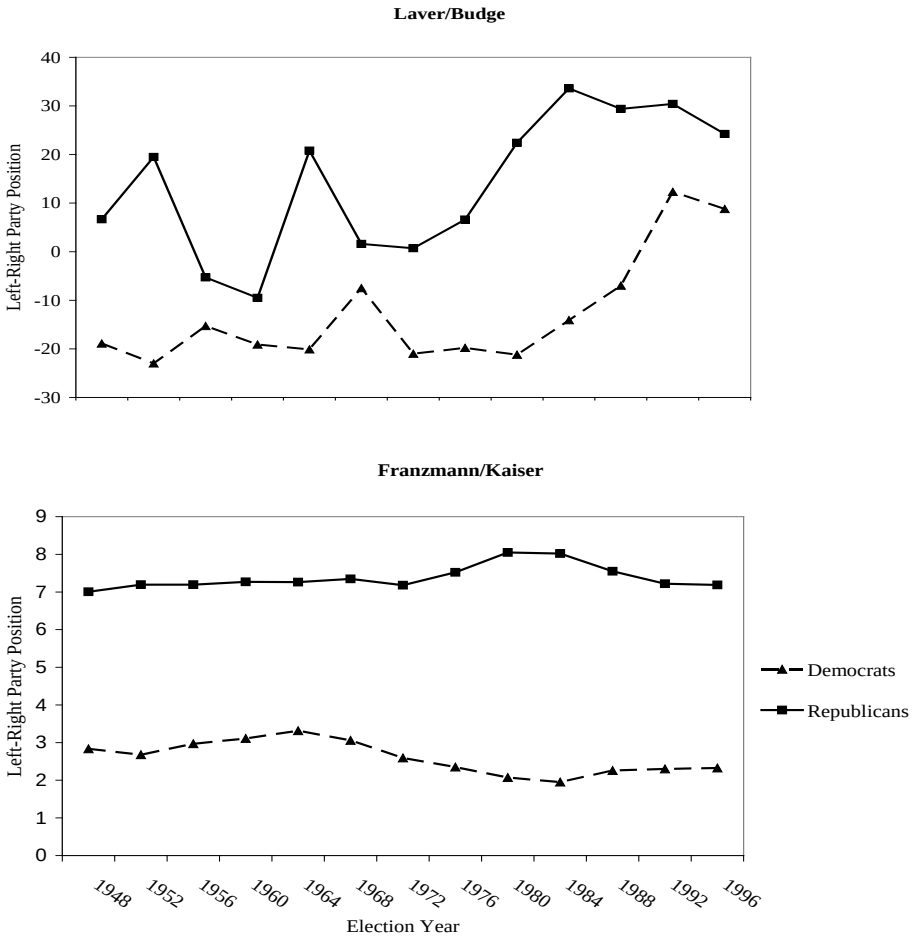
Laver and Budge start with an exploratory factor analysis to identify items loading together. They calculate 7 new variables out of 28 original items and add the number of saliency counts together to avoid loosing information – but only if it makes theoretical sense. They further include 13 of the original items which are not loading together and therefore receive 20 items. With these items they conduct a rotated exploratory factor analysis to identify 4 marker variables clearly indicating left and right positions. They compare all of the other items to those marker items to check if they were rather right or left-wing and excluded the not precisely identifiable ones. This leaves them with 13 left und 13 right positioned items on one dimension. After selecting items, original saliency counts are added up for all items indicating a left position and all items indicating a right position of a party. Right-wing items are then subtracted form the left-wing items.

Gabel and Huber run a rotated exploratory factor analysis pooled over the whole dataset and took all items loading on the first dimension as left-right-axis. Thus, they call their method *Vanilla Approach*: items are contained in their left-right-factor like pulp dots in a vanilla bean. To define positions of parties on the left-right-axis they used factor scores.

Franzmann and Kaiser understand saliency scores as signals of parties to communicate a movement of their position to the electorate rather than information about their true position. They suggest a complex 12-step procedure to extract party positions scores for multi-dimensional spaces. In short, Franzmann and Kaiser start by classifying CMP items as valence and positional issues on theoretical grounds, run a regression analysis to select the relevant items for their scale, and finally smooth the scale with a special procedure.

Figure 1: Left-Right Positions of Parties in the US According to Gabel / Huber (top), Laver/Budge (middle) and Franzmann/Kaiser (down)[a)]





a) All calculations are based on Budge et al. (2001) and Klingemann et al. (2006). The values for the Gabel/Huber scale were calculated by the authors according to Gabel/Huber (2000), for Laver/Budge they were taken from Budge et al. (2001) and Klingemann et al. (2006). The values for the Franzmann/Kaiser scale (2006) were provided by the authors. We would like to thank them.

The results of the three procedures are shown in figure 1. They differ heavily. Firstly, the distances between parties differ between the scales. They are much smaller in Gabel/Huber than in Laver/Budge and especially in Franzmann/Kaiser. Secondly, the Gabel/Huber scale

shows a strong movement of positions over time, Laver and Budge a medium one, and Franzmann and Kaiser a small one. Thirdly, the Gabel/Huber scale shows a crossing of the positions while Laver and Budge and Franzmann and Kaiser do not. So who is right? Validity is certainly difficult to measure. In regard of construct validity, all authors validate their concept by correlating them with other scales and survey data. But even though they correlate quite highly, the practical differences are significant.

Gabel and Huber on the one hand argue in a multi-trait multi-method validation of their scale that it goes overall much more in line with expert surveys (Castles and Mair 1984) and survey data of the electorate (World Values Survey/Eurobarometer) than other approaches. Franzmann and Kaiser also support their scale with survey data. On the other hand there is no reason to assume that the Democrats ever had a position to the right of the Republicans in post-war history – which supports Laver and Budge, and Franzmann and Kaiser. The Laver and Budge scale predicts historical events like the Republicans' shift to the right with the Reagan election far better than the other scales. Unfortunately, the US example is not unique. The diverse results created by different scales persist when they are applied to various countries and periods of time. This is a rather alarming puzzle, and still the question which procedure to use is not answered. Not only do the positions of parties differ, but also the distances between them and the ordinal ranking.

These discrepancies can lead to two conclusions: Firstly, one could question the validity of the saliency concept as a whole or in part (Laver 2001; Laver and Garry 2000; Janda, Harmel, and Goff 1995). Secondly, there might be a problem with the validity and reliability of the existing scales. Therefore we classify the scales before analysing the dataset and showing the effects the dataset has on them.

3.2. Previous Attempts to Solve the Puzzle

Since the CMP data have become available, an increasing number of researchers attempted to extract political positions of parties from the dataset. These include Budge, Robertson, and Hearl 1987, Laver and Budge 1992, Klingemann 1995, Laver and Garry 2000, Gabel and Huber 2000, McDonald and Mendes 2001, Kim and Fording 1998, 2002, and Franzmann and Kaiser 2006. Some of them try to define a

left-right-axis which can be generalized for all countries and across all years, others used only several countries or a limited period of time.

This variety shows the ordinary non CMP-researcher two things. First, there is no obvious "correct" or even best practice method for putting parties in their place. The CMP data don't seem to speak for themselves due to the enormous size and detail of the dataset. Second, CMP data and the method of calculation selected should be handled with care. But which of these approaches should a researcher who is in need of comparative data about the positions of parties choose? Every researcher using the CMP data and choosing a scale is essentially faced with two questions: (1) Which items should be included to construct a generalizeable left-right axis? (2) Which rules should be employed to put political parties in their place on this scale?

Previous approaches of selecting items range as a continuum from inductive to deductive techniques (Laver and Garry 2000: 627). The ideal type of an inductive technique is exploratory factor analysis. It can untangle complex structures and find in a great number of manifest variables a latent structure. It also can be used for scaling purposes: it is able to examine whether a number of variables can indeed be reduced to one dimension. To put it bluntly: the idea is to recognize the items necessary to find several clear cut cleavages explaining party competition or just one left-right dimension. The resulting factors are independent of each other so the result would fit the standard models of spatial theory. This technique is engaged in its pure form for example by Budge, Robertson, and Hearl 1987 and Gabel and Huber 2000.

At the other end of the spectrum is the ideal type of an entirely theoretical selection of items as a deductive technique. This concept assumes that a researcher has a clear idea about the policy content defining the left-right-dimension. In its pure form this approach was used by McDonald and Mendes 2001 who identified 32 items on two dimensions.

Most authors use a combination of both methods in several steps. Klingemann (1995) starts by choosing a number of items on theoretical grounds before conducting an unrotated factor analysis to identify different dimensions of conflict in a country-by-country analysis. This leaves him with 11 items. Laver and Budge (1992) use a combination of factor analysis and theoretical considerations as well, but start with the

factor analysis as first step to identify 26 items. Kim and Fording (1998, 2002) and Laver and Garry (2000) use the basics of the Laver and Budge approach as well. A deviation is Franzmann and Kaiser (2006) here using all 56 items to define the policy space.

In regard of the second question about which rules should be employed to define the position of political parties we found a number of approaches as well. Essentially, one has to decide how the chosen items contribute to the construction of the scale and thus to the positioning of the parties. We distinguish these rules in weighted and unweighted methods. McDonald and Mendes (2001: 94) mention two unweighted methods of choosing the rules of identifying the position of parties. The first one is a subtractive measure. Original saliency counts are added up for all items indicating a left position of a party and all items indicating a right position of a party. Then right-wing items are subtracted from the left items. This measure is employed by Laver and Budge. A variation of the subtractive measure is employed by Franzmann and Kaiser subtracting right issue scores from left issue scores but put them in relation to the overall item score. The second one is a ratio measure. Left and right items are also summarized but they are put in relation to the overall number of items. Ratio measures are used by Kim and Fording, and Laver and Garry. The basic idea behind all these methods is that all items should carry the same weight and there is no theoretical reason to assume that one item should be more influential than another one. Weighted measures are usually based on factor scores (Budge, Robertson, and Hearl 1987; Klingemann 1995; Gabel and Huber 2000). The factor scores indicate how well an item correlates with the factor computed. The basic idea behind this method is that not all items should carry the same weight to define the position of a party on the left-right-axis. They assume that some items are of greater importance than others. Table 1 shows a classification of attempts to answer both questions. The columns refer to the first question. They show the different possibilities of choosing relevant and reliable items to construct a left-right scale. The table discriminates dichotomously the first step of the investigation in exploratory factor analysis and theoretical consideration of some sort. Although some authors in the first quadrant use a combined method, they all start with factor analysis, which is the decisive point when it comes to the evaluation of the used techniques. In the rows the second question is

answered. They show the various kinds of assigning values to parties. We distinguish methods using factor scores versus unweighted additive or subtractive indices.

Table 1: Classification of Approaches to Measure Positions of Political Parties

		Choosing Items (only first Step)	
		Exploratory Factor Analysis	*Theoretical Considerations*
Measuring Position	*Additive Scores (unweighted)*	Laver/Budge 1992 Laver/Garry 2000 Kim/Fording 1998 Kim/Fording 2002	McDonald/Mendes 2001 Kaiser/Franzmann 2006
	Factor Scores (weighted)	Budge/ Robertson/Hearl 1987 Gabel/Huber 2000	Klingemann 1995

4. A Peculiarity of CMP-Data and its Consequences

4.1. Noughts and Crosses

In the process of analysing different procedures for extracting political positions from CMP data, we discovered a previously unmentioned peculiarity of the dataset: More than half of the entries in the data set are equal to zero. Saliency theory argues that parties compete not via conflicting issues but via occupying new issues and strengthening their importance to their electorate in the manifesto for every election. Empirically this leads to data where items are equal to zero if they are not mentioned in a given manifesto. Items that *are* mentioned range between any value greater than zero and 100 percent, according to the amount of space in a manifesto devoted to them. Accordingly, a certain amount of items in any data set based on the saliency approach should be zero; and the values of non-zero items should correspond to the importance given to the item in a manifesto.

Figure 2: Average Number of Items Not Mentioned in Party Manifestoes (Data Points Equal to Zero), in Percent

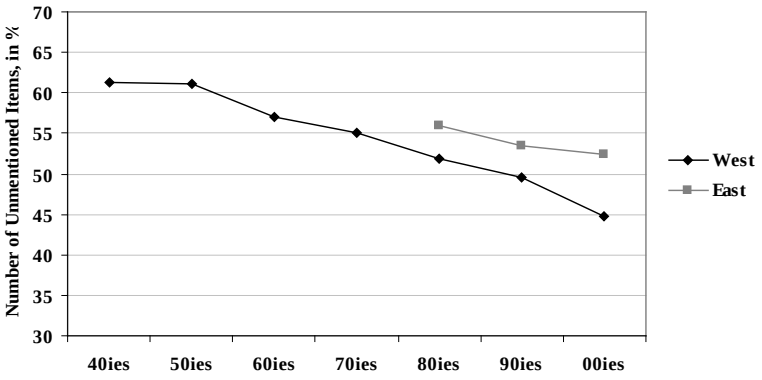


Figure 2 shows the average number of items that are equal to zero for the CMP data set for all countries over time. Although there is

obviously a decline in the average number of non-mentioned items over time, amounting to more than 60 percent in the 1940ies and 1950ies, even the most recent manifestoes do not mention almost 45 percent of all items in Western Europe and other OECD member countries. For Eastern European countries, more than half of the data set consists of zeroes. In fact, a few items are practically not mentioned by any party at all. For instance, item 415 (Marxist analysis) is mentioned only 159 times out of 3305 times possible – only 4.8 percent of the entries are unequal to zero. Item 504 (welfare state expansion), on the other hand, is mentioned in 2884 out of 3305 manifestoes, that is, 87 percent. Even more disturbing than the overall number of non-mentioned items in the data set is the fact that on average, more than 20 percent of the items are *not mentioned by any of the parties* competing in an election in a certain country and a certain year. This clearly means that a zero measured by the CMP data does not necessarily correspond to the meaning of a zero according to the saliency approach.

The peculiarity about the CMP data is thus the extraordinary large number of items equal to zero, that is, which were not mentioned in a manifesto. On average, over time and over all countries, 30 out of 56 data points (or 53.5 percent) are equal to zero. This is rather unusual for data not meant to be binary, i.e. distinguishing between zero and non-zero data points. Rather than containing truly metric data, then, the CMP data has the structure of noughts and crosses – items that are zero and that are something other than zero.

The question that arises from the large number of zeroes in the CMP data set is whether this has consequences for the use of the data in general, and the generation of party positions, i.e. on a left-right scale, in particular. As we will argue below, there are at least three pitfalls in the CMP data that researchers have to consider and overcome: First and foremost, the substantial meaning of the zeros in the data set is unclear. Secondly, the systematic variation of the number of zeroes over time and countries complicates the use of CMP data for comparative purposes, which is supposed to be one of its main assets. Thirdly, the large amount of zeroes impairs and in fact prohibits many of the methods of data analysis commonly used for extracting party positions from CMP data.

4.2. The Meaning of Zeroes in the Data Set

In part 4.1 above we have shown that the structure of the CMP data is, first and foremost, one of noughts and crosses, although the crosses may have different values according to the importance placed on an issue by a party. No matter what this value is – as long as an item is mentioned by a party, it is certain that the issue is more or less important to this party and that the issue therefore indicates something about the party's policy position.

Whereas mentioning an issue is a reflection of parties' policy preferences, this is not true for items that are not mentioned. In fact, the meaning of the zeroes in the data set is far from being clear. The zeroes can exist for two different reasons: Firstly, items might be zero because parties really do consider an issue unimportant. Secondly, zeroes may arise for non-substantial reasons and are in turn related to (a), the number of uncoded sentences in a manifesto and (b) the length of a manifesto. In the first case, there is a substantial meaning to items that are zero, indicating something about a party's position in political space. This is in line with the saliency approach described above. In the latter two cases, however, there is no genuine meaning to a zero. The distinction between real, meaningful, substantial zeroes and other ones can be compared to the distinction between negative responses and non-responses to items in survey data. Whereas the answer "yes" to a survey question (coded as "1" in the data) tells us something about the opinion of the person polled, the entry "0" could mean than the person responded with "no", refused to respond, or simply was not asked this particular question.[1]

Non-substantial zeroes arise in the case of items that do not structure political space in a particular country and/or at a particular time. Consider, for instance, a country without any substantial immigration within the last decade, such as Iceland, and also without any autochthonous minorities. For such a country, opinions on multiculturalism do not structure political competition. If a party in such a country does not include a positive reference on

[1] In most surveys, of course, a distinction between different forms of non-response and negative responses is made. Our intention here is to demonstrate the effects that arise when such a distinction is absent.

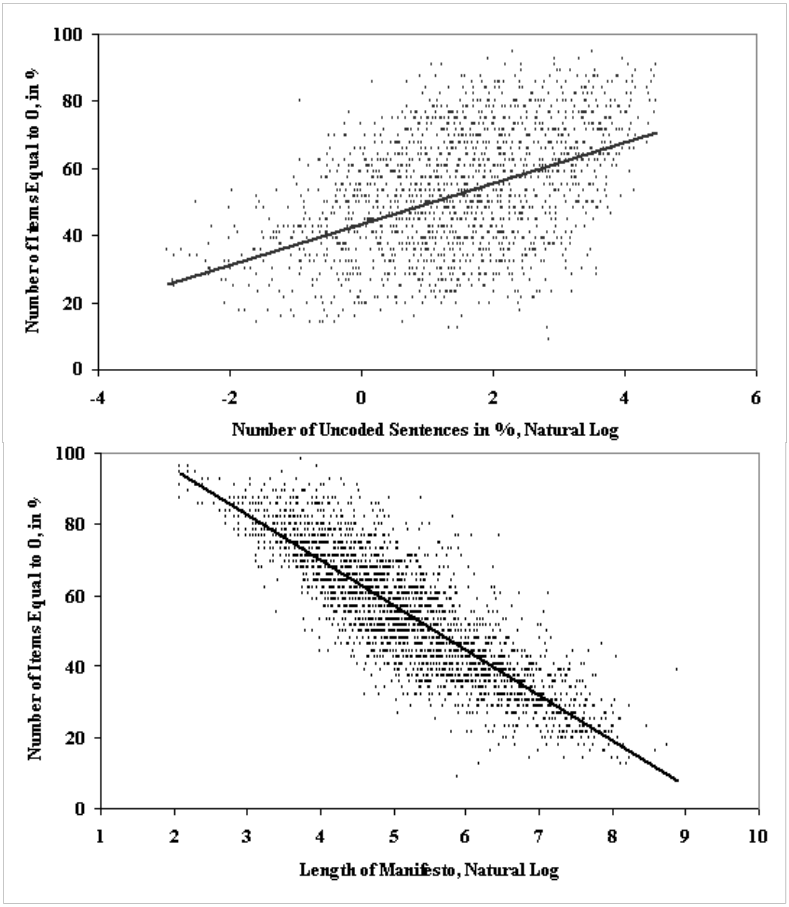
multiculturalism (items 607), which is generally thought to be an indicator of a "left" ("right") political position, in its manifesto, this is not an indication for a "right" position. The same holds true for items on environmental issues in earlier decades, resulting in a non-substantial zeroes in the data set.

There is empirical support for the assumption that items considered relevant for political competition by the Manifesto Research Group may not necessarily be considered important by parties. In most manifestos, a certain percentage of quasi-sentences did not fit into any of the 56 categories used for coding. The amount of sentences that could not be coded varies strongly by time and country – 14 percent remained uncoded in the 1940ies, but only just above 2 percent after 2000 – and it is strongly related to the number of items that are equal to zero. This relationship is shown in figure 3. The correlation between the (natural log of the) number of uncoded sentences and the number of items not mentioned in a manifesto is r=0.46. This is quite strong evidence for the fact that some of the "zeros" in the CMP data do not carry the meaning implied by the saliency approach and therefore indicate nothing about a party's policy preferences. Rather, the strong relationship between zero items and uncoded sentences implies that political competition in some countries is based on issues other than the ones included as coding categories by the Manifesto Research Group.

In addition to zeroes due to uncoded sentences, there is another group of zeros whose meaning is not the one implied by the saliency approach. They exist for the simple technical reason of manifesto length. If a manifesto includes less than 56 quasi-sentences, not all items can be referred to and it follows that some of them must be zero even though this indicates nothing whatsoever about a party's political position. Overall, a substantial amount of 15 percent of all manifestos is shorter than the minimum length required to include all 56 items. Short manifestos are common among parties in Israel, where the average length is only 46 quasi-sentences and Japan (65 quasi sentences), but also in Denmark, Finland, France, and Sweden (less than 200 sentences, on average). Spanish, Greek and Norwegian parties, on the other hand, prefer long manifestos with an average of more than 1000 quasi-sentences. Also, the length of manifestos has increased over time almost everywhere. Overall, the average length has more than

quadrupled from the 1940ies to the post-2000 elections. Why should this be important? Isn't it sufficient for a manifesto to exceed the length of 56 sentences in order to avoid non-substantial zeroes? By no means: Figure 3 demonstrates that there is an astonishingly strong relationship between the length of a manifesto and the number of items that are not mentioned. The correlation is r = -0,85. The longer a manifesto, the more items are referred to.

Figure 3: Relationship between the Number of Zeroes, Uncoded Sentences (Percent, natural log), and the Length of the Manifesto



The majority of the zeroes in the data set are obviously not the result of substantial considerations of parties about their preferred policies.[2] On the contrary, they can to a large extent be attributed to the mere length of a manifesto and the – partly arbitrary – choice of issues to include in the coding scheme. The number of zeroes in a particular manifesto can be predicted by a linear regression model including manifesto length and the percentage of uncoded sentences (both transformed to their natural logarithms):

$$\gamma_1 = \beta_0 + \beta_1 * ln(uncoded\ sentences) + \beta_2 * ln(manifesto\ length) + \varepsilon$$

The results shown in table 2 (model 1) demonstrate that almost three quarters of the variation in the number of zeroes in different manifestos can be attributed to these two explanatory variables alone - without any substantial considerations on policy preferences. If our assumption about the existence of non-substantial zeroes in the CMP data set was wrong, there should be no statistical relationship between manifesto length, uncoded sentences and the number of zeroes. Yet these two variables predict the occurrence of zeroes in a manifesto almost perfectly. The difficulty for any application of the saliency approach to CMP data is now to determine which of the zeroes are meaningful and which are not. Before we provide some hints for this task in part 5 of this article, another pitfall of CMP data remains to be explained: its comparative applications.

Table 2: Linear Regression Model – Number of Zeroes vs. Length of Manifesto, Uncoded Sentences, Time and Country[a)]

	Model 1		*Model 2*	
	Unstandardized Coefficient	*Standardized Coefficient*	*Unstandardized Coefficient*	*Standardized Coefficient*
Ln (Manifesto Length)	-12.22*	-0.80*	-11.36*	-0.75*
Ln (Percent Uncoded)	1.55*	0.12*	1.89*	0.14*

[2] Of course one might argue that the length of a manifesto is a result of a conscious decision by parties which items to include. In this case there should be no such correlation for very long manifestos. However, even for manifestoes longer than 560 quasi-sentences – leaving room to refer to all items at least ten times – there is still a strong negative relationship between the length and then number of zeroes of r = -0,51.

Country Year		Coefficients are not reported, but variables are controlled for
Constant	116.18*	116.18*
N	2167	2167
R^2	0.74	0.80

a) Because the range of the dependent variable is restricted to values from 0 to 100, we fitted alternative models using the natural log of the percentage of zeroes as a dependent variable. The results are virtually the same – the explained variation these model is $R^2 = 0.73$ and 0.79 (instead of 70.74 and 0.79 respectively for the original dependent variable).

* $p < 0.001$

4.3. Pitfalls of Comparative Applications

One of the presumed assets of the CMP data is its usability for comparative analyses. As we have shown in section 3.2, several authors have attempted to derive a general, time- and country-invariant left-right scale from the data that can be used to place parties in political space and compare their positions both internationally and over time. However, the methodological complications arising from the number of zeroes in the data set and their unknown substantial content are further increased if the data is used comparatively. The reason for this is quite simple: The zeros in the data set are distributed unequally not only over items, but also over countries and time. This also applies to the factors explaining the occurrence of zeroes, such as manifesto length. As a consequence, it is questionable if the same items should be chosen for a left-right-scale in all countries.

Table 3: Average Number of Items Mentioned by Parties – By Country and Time

Country	*40s*	*50s*	*60s*	*70s*	*80s*	*90s*	*00s*	*Total*
Cyprus						33,0	22,8	28,9
Spain				39,6	31,5	25,8		30,4
Malta						33,5		33,5
United States	45,5	38,4	35,7	33,3	36,2	29,5	41,1	34,9
Norway	53,4	46,7	39,8	36,8	31,9	23,7	28,3	35,4
Netherlands	51,6	46,4	42,4	36,2	28,3	24,8	35,3	35,8
Great Britain	36,6	48,4	40,3	38,9	26,3	35,0	39,4	37,7
Turkey	59,8	59,8	33,1	33,9	40,1	39,2	45,8	39,9

Greece				49,2	41,8	35,2		40,0
Belgium	55,2	48,0	50,9	49,6	38,2	32,7	30,4	41,5
Portugal				51,1	63,2	29,8	32,1	46,8
Germany	69,6	63,8	50,0	45,0	43,8	40,7	25,0	48,3
New Zealand	47,8	60,7	51,8	42,7	47,4	44,1	50,0	48,5
France	54,8	48,0	50,6	43,8	47,0	54,9	47,0	49,9
Austria	59,5	61,5	56,7	52,0	40,5	51,3	45,0	51,0
Luxembourg	65,2	65,6	54,0	48,4	41,5	40,5		51,3
Switzerland	48,6	53,6	51,3	45,9	47,5	55,5	47,9	51,5
Canada	46,9	57,3	51,1	51,1	48,2	54,4		52,4
Ireland	72,1	77,4	65,6	57,7	55,6	39,0	42,9	56,2
Italy	74,7	59,7	55,4	54,8	51,1	59,4	44,5	57,3
Australia	62,8	65,2	66,9	61,7	49,5	53,1	50,0	58,6
Iceland	60,7	66,3	56,9	66,1	56,0	52,8	65,4	59,8
Sweden	67,1	66,9	67,7	58,2	66,9	65,3	54,3	64,7
Denmark	71,4	73,2	64,1	72,5	65,1	55,6	66,7	66,7
Finland	65,9	68,8	69,7	70,6	74,8	67,5	59,6	68,8
Japan		60,3	66,7	71,1	66,8	75,5	68,9	70,4
Israel	76,5	66,8	74,0	76,4	79,7	79,5		76,6
WESTERN - ALL	*61,3*	*61,1*	*57,0*	*55,0*	*52,6*	*51,0*	*47,3*	*53,5*
Czech Republic					52,5	40,6	32,1	41,4
Bulgaria					48,2	41,4	36,4	41,5
Lithuania						42,9		42,9
Slovakia					56,1	44,5	41,7	46,0
GDR					48,2			48,2
Croatia					58,5	45,5	55,4	49,2
Moldova						50,0		50,0
Russia						51,5	44,1	50,1
Macedonia					50,0	52,3	44,6	51,0
Albania						52,5	53,9	52,7
Hungary					56,7	51,1	50,4	52,7
Romania					59,8	47,8		53,1
Ukraine						52,4	57,7	53,3
Armenia						53,4	55,7	54,2
Georgia						56,1		56,1
Slovenia					59,2	57,5		57,9
Belarus						58,3		58,3

Montenegro					56,7	57,9	94,6	59,0
Latvia						60,8	61,7	61,0
Estonia						63,0	53,9	61,1
Serbia					62,2	61,4		61,6
Azerbaijan						62,3		62,3
Bosnia-Herz.					67,0	62,3	58,6	62,5
Poland						62,5	66,6	63,2
CEE - ALL					*55,9*	*53,4*	*52,4*	*53,7*
TOTAL	*61,3*	*61,1*	*57,0*	*55,0*	*52,6*	*51,0*	*47,3*	*53,5*

Table 3 shows the average percentage of empty items in a given country and decade. Apparently, the amount of zeros – no matter whether they have a substantial meaning in line with the saliency approach or they do not – varies greatly both by time and country. The fact that the number of items unmentioned decreases in with time (see figure 2) also holds for individual countries. In Germany for instance, almost 70 percent of all data points are zero in the 1940ies. From the 1950ies onwards, the number of zeros decreases with each decade, falling below 50 percent in the 1970ies and amounting to only 25 per cent in the post-2000 elections. Overall, the smallest number of zeros – less than 40 percent - can be found in the United States, in the Netherlands, in the UK, and in Turkey.[3] The countries with the largest number of zeroes are Israel (76.6 percent) and Japan (70.4 percent), followed closely by Finland, Denmark and Sweden.

To explain the occurrence of zeroes and their variation, time period and dummy variables for individual countries are added to our original regression model:

$$y = \beta_0 + \beta_1 * ln(uncoded\ sentence)s + \beta_2 * ln(manifesto\ length) + \beta_3 * time + \beta_4 * country + \varepsilon$$

Time and country alone explain 47 percent of the variation of the number of items equal to zero in a given manifesto. This is quite a large amount considering the fact that we are using data that still varies

[3] Even less zeroes can be found in Cyprus (28.9 %), Spain (30.4 %), and Malta (33.5 %). However, data for these countries are only available for recent elections, in which more items are mentioned in all countries.

within each country at a specific time period. If the length of the manifesto and the percentage of uncoded sentences are included simultaneously, 80 percent of the variation can be explained (see table 2, model 2). The largest coefficient results from the length of a manifesto, followed by the number of uncoded sentences.

Although the country- and time-specific variation is a nuisance for researchers interested in comparative analyses of political space, it gives some additional insights into the reasons for the large number of zeroes in some countries and at certain times. In part 4.2. above we argued that the amount of items equal to zero is to a large part determined by the length of a manifesto and the number of sentences that could not be coded, indicating references to items not included by the Manifesto Research Group. The large number of zeroes in Israel and Japan suggests that the party system determines the structure of the data. In countries with a highly fractionalised or centralized party system without much competition or many single-issue parties, many of the items chosen as categories by the Manifesto Research Group will not be referred to. In countries with a small effective number of parties, such as the USA and the UK, the items seem to work quite well.

This dependence on the party system of a country does not challenge the validity of the saliency approach or the CMP data. However, there seem to be other factors at work that are not in line with the saliency approach and that seriously threaten the comparative use of the CMP data when items are reduced to one or a few central dimensions of political competition. First and foremost, the number of items not referred to in a manifesto seems to increase with temporal and cultural distance from the researchers who originally chose the 56 items according to which all manifestos are coded. The issues that structured political space and political competition in Great Britain in the early 1980ies might not be relevant in New Zealand or Sweden in the 1990ies. A look at the data for Great Britain in table 3 shows this quite nicely. Whereas on average only 26 percent of all items were not mentioned in the 1980ies, the number of zeroes increases both before and after this period.[4] Also, the fact that many items are not referred to

[4] The coding scheme was first developed for the United Kingdom by Robertson (1976) and then adapted for other countries (Klingemann, Volkens, Bara, Budge, and McDonald 2006: xviii)

in Denmark, Finland and Sweden suggests that the specific relationship between states and markets and the kind of welfare state implemented in a country affects the adequacy of the CMP items. Cultural factors might be influential as well: Many of the 56 items do not seem to be relevant at all in countries that are culturally very distant from Western Europe and not quite well known to many researchers from Western Europe, such as Japan and Israel. That is, they do not seem to mirror the political competition in these countries. Surprisingly, not even the inclusion of new items specific for the situation of Eastern European countries after 1990 does seems to alter much: On average 90 percent of them are equal to zero, ranging between 75 percent in Moldova and 99 percent in Azerbaijan.

The same reasoning applies to the number of items not mentioned by any party running in a particular election - we find a similar variation by time and country. For instance, less than half of the items were mentioned by any party in the 1998 Australian elections, the 1973 Irish elections and the 1954 elections in Luxembourg. On the other hand, only 2 items remained unmentioned in Spain in 1996 and 2000 and in Switzerland in 1999.

Notwithstanding the reasons for the unequal distribution of zeroes over time periods an countries: Considering the methodological pitfalls and the unknown substantial meaning of the zero entries, it is clear that this unequal distribution renders the one- and multidimensional scales derived from CMP data difficult for comparative purposes.

4.4. Methodological Pitfalls of Scale-Building

Apart from the question of the meaning of the zeroes, what effects could the large number of empty data points have methodologically, specifically, for the methods of data analysis used to place parties in political space? In the following paragraphs we show that scales on political positions extracted from CMP data are severely affected by the large amount of zeroes, irrespective of their meaning. The following arguments apply even if all empty data points were meaningful and in line with the saliency approach.

Our classification (see table 1) distinguished between procedures of choosing items and of actually ascribing positions to political parties. There were two common methods for choosing items: exploratory factor analysis and theoretical considerations. For a purely theoretical

choice of items the effects of the unmentioned items are negligible. The only effect is that one might end up with a number of items sparsely populated and unable to distinguish between left and right parties. Theoretical considerations usually result in scales with plenty of items because the researcher needs densely populated items as well as sparsely populated ones to distinguish between parties. This effect shows in the McDonald and Mendes (2001) scale which uses 6 items more than comparable scales.

Entirely different are the effects when using factor analysis: The results will be severely influenced by the large amount of zeroes. One of the basic requirements for a factor analyses – or in fact any other procedure to reduce the dimensionality of the data - is that variables have to be correlated. In the case of CMP data, a large amount of constant cases (cases equal to zero) causes a suppression of item variance and inter-item covariances. The large number of zeros in data thus leads to a low correlation among variables, making the data unsuitable for factor analysis. On the other hand, a correlation where there is in fact none may appear due to a large amount of zeroes in addition to otherwise metric data. Either way the results of a factor analysis will be distorted.

Imagine a dataset with two items and two parties over a period of 10 elections. Imagine both items would measure something extremely similar, such as *free enterprise* and *economic orthodoxy*. No item is referred to by a party during the first seven elections, but both are referred to during the last three campaigns. However, one party mentions only *free enterprise*, while the other only mentions *economic orthodoxy*. The salience for *free enterprise* is 2, 4, and 9 percent, respectively. The saliency for *economic orthodoxy* is 2.5, 4.5, and 9.5 percent. Because of this, there should be a perfect correspondence between these two items for substantial arguments based on the saliency approach. Yet, although the items are practically identical, and mentioned in the same years with about the same salience, they correlate only weakly (r=0,12) and not significantly. This is entirely due to the high number of zeros in the dataset. If the items are added up in advance this leads to a highly significant correlation of r=0,99. This small example demonstrates the suppressing effect on correlations due to the high number of zeros. Empirically, the likelihood of data factoring nicely can be measured with the Kaiser-Meyer-Olkin-Criterion (KMO). This measure tells

researchers if their sample is adequate for factor analysis and ranges from 0 to 1, with a lowest recommended KMO-level of 0.6. If the KMO is below this value, items with the lowest individual KMO should be dropped.

If one aims at choosing appropriate items for a scale of political positions via factor analysis, it is essential *not* to drop items a priori for reasons other than theoretical considerations. An item with a low individual KMO based on a low correlation with other items because it is mentioned in almost none of the manifestoes (or in almost all of them) may be one that distinguishes quite well between left and right parties *precisely because* it is mentioned so rarely (or so frequently). If one includes all CMP items in an exploratory factor analysis, the overall KMO is only 0,498. It is not much higher when Eastern European countries, for which a number of other issues may be relevant, are excluded from the analysis (KMO = 0,512). The KMO for individual items ranges from 0,23 for item 703 (positive mention of farmers) to 0,72 for item 411 (positive mention of infrastructure). According to common measures of sampling adequacy, the CMP data, with its structure of noughts and crosses instead of being truly metric or truly binary, should not be used for factor analysis. Otherwise, results will be seriously distorted.

The second part of our classification (see table 1) concerned the rules governing the combination of items. We stated a distinction between unweighted indices of left and right items, which are subtracted from each other, and a weighted measure based on factor scores. It is clear that in case of an approach using factor scoring, the previously mentioned problems of unstable solutions of factor analyses applies to the factor scores as well. Both factor loadings and the resulting positions of parties (factor scores) will be severely affected by the number of zeroes in the data set.

Unfortunately, this also applies to purely additive scales. The reason is simple: Even when items are chosen theoretically, the percentage of a manifesto devoted to any single item depends on how many items are referred to at all. Imagine that, for instance, item 403 (positive reference to market regulation) is included in a left-right scale as an indicator for a "left" position, as in the case of the RILE-scale suggested by Laver and Budge 1992. Also, imagine that Party A devotes 1 percent of its manifesto to item 403, and Party B, 5 percent. According to the

additive approach, Party B should be positioned more to the left than Party B. If however, Party B referred to only 10 out of 56 items in its manifesto, and Party A to all 56 items, it is quite clear that a 4-percent-reference to item 403 is much harder to reach for Party B. Once an item is referred to, its actual value strongly depends on the overall number of items referred to in a manifesto.[5] The less items are referred to (the larger the number of zeroes), the larger the value of the items that *are* referred to. This shows that the CMP data are not truly metric in nature. Consequences apply to all scales that are based on less than all 56 items.[6]

To sum up, the number of non-mentioned items in the CMP data has virtually no effects when items are chosen theoretically. However, it prohibits the use of factor analysis because correlations among items are too low and may be distorted. As a consequence, many scales extracting political positions from CMP data are not reliable, in particular those using factor analysis. These methodological pitfalls are even more disturbing considering the uneven distribution of zeroes across countries and time, no matter whether they are substantial or non-substantial. As a consequence, scales based on CMP data will be less reliable for some countries and time periods than for others, again prohibiting comparative applications.

5 In addition to the length of the manifesto, as we have shown above.

6 This shows in low reliability scores as well. For instance, for the 26 items included in the additive RILE-scale (Laver and Budge 1992), Cronbachs alpha is only 0.394.

5. Avoiding the Pitfalls

The challenge for any researcher interested in extracting policy scales from CMP data is to determine which of the zeroes in the data set are meaningful and line with the saliency approach. If this can be achieved, it is possible to choose items with a low number of non-substantial zeroes and combine items in a way that avoids the methodological pitfalls mentioned. The analyses presented above provide some clues for these tasks.

Concerning the distinction between substantial and non-substantial zeroes, zeroes arising in very long manifestos are likely to be meaningful in terms of the saliency approach, whereas many of the zeroes in short manifestos are most likely due to technical reasons. Secondly, zero-items in manifestos that were completely coded are more likely to have a substantial meaning than items in manifestos with a large number of uncoded sentences. Thirdly, when there are many items in a single subject area obviously measuring the same concept, many of these items will be zero simply due to the fact that parties may prefer a certain vocabulary (such as *free enterprise* instead of *economic orthodoxy*). In this case, items in this subject area can simply be added up, as in case of the RILE scale. Fourthly, there are a few positional items included in the CMP data, e.g. positive and negative mention of the military (items 104 and 105). If one of these items is mentioned in a manifesto, the non-mentioning of its counterpart is almost certainly meaningful. If, on the other hand, none of the two items is mentioned, this might be coincidence.

In particular, the likelihood that a particular zero is due to some non-substantial reason (i.e. manifesto length) can be estimated using the following logistic regression model:[7]

[7] This approach is quite similar to the one applied by Slapin and Proksch (2008) to predict word score frequencies in manifestos. However, they use a model based on a Poisson distribution and do not include the percentage of uncoded sentences.

$$\pi = (e^{\beta 0 + \beta 1 * length + \beta 2 * uncoded + \beta 3 * country + \beta 4 * decade}) / (1 + e^{\beta 0 + \beta 1 * length + \beta 2 * uncoded + \beta 3 * country + \beta 4 * decade})$$

In this model, π indicates the probability that a particular item is mentioned in a manifesto. If this predicted probability is high, but the item is *not* mentioned, it is most likely that this particular zero is the result of the rejection of the item by the party and thus in line with the saliency approach. If, on the other hand, a zero goes along with a low predicted probability and thus a predicted classification as "zero", the non-mention of this item can be attributed to other factors and is not in line with the saliency approach.

Table 4: Occurrence of Item 501 (Environmental Protection) in Manifestos of German Parties Since 1949

	PDS		*(BD90)/Greens*		*SPD*		*FDP*		*CDU / CSU*	
	Ment.	*Prob.*	*Ment.*	*Prob.*	*Ment.*	*Prob.*	*Ment.*	*Prob.*	*Ment.*	*Prob.*
1949	-	-	-	-	no	0.13	no	0.21	no	0.24
1953	-	-	-	-	no	0.19	no	0.20	no	0.23
1957	-	-	-	-	no	0.18	no	0.18	no	0.14
1961	-	-	-	-	yes	0.67	yes	0.58	yes	0.49
1965	-	-	-	-	yes	0.91	yes	0.72	yes	0.61
1969	-	-	-	-	no	0.59	no	0.64	yes	0.59
1972	-	-	-	-	yes	0.89	yes	0.76	yes	0.88
1976	-	-	-	-	yes	0.90	yes	0.92	yes	0.89
1980	-	-	-	-	yes	0.94	yes	0.99	yes	0.93
1983	-	-	yes	0.94	yes	0.96	yes	0.97	yes	0.93
1987	-	-	yes	0.99	yes	0.97	yes	0.95	yes	0.97
1990	yes	0.98	yes	0.93	yes	0.97	yes	1.00	yes	0.94
1994	yes	0.97	yes	1.00	yes	0.97	yes	1.00	yes	0.96
1998	yes	0.99	yes	0.94	yes	0.99	yes	0.99	yes	0.95
2002	yes	0.98	yes	1.00	yes	1.00	yes	1.00	yes	1.00

Ment.: Item 501 mentioned in manifesto (yes or no); Prob.: Predicted probability that item was mentioned in manifesto according to logistic regression with predictors length of manifesto, percent uncoded sentences, country and decade.

As an example, table 4 summarizes predicted probabilities obtained by the above regression model for item 501 (environmental protection) for parties in the German parliament. It is obvious that the observed values for the mention of item 501 go in line with the predicted probabilities. Assuming that items are mentioned if the predicted probability is 50 percent or higher, there are only two cases of incorrect classifications: the Social Democrats and the Free Democrats in 1969. Both are expected to mention environmental protection in their manifesto, but in fact do not. All other cases can be explained by the regression model,

indicating a high incidence of non-substantial zeroes for this particular item, which becomes important in the 1970ies but does not structure political competition because it is equally important to all parties. This does not only apply to Germany, but other countries as well. Overall, the Pseudo-R^2 for the regression model (based on all parties in all countries) is 46 percent. More than 80 percent of all cases can be classified correctly as zero or non-zero simply based on the knowledge of manifesto length, uncoded sentences, decade and country – irrespective of the particular party. This item should therefore not be chosen for a scale on political positions.

Table 5: Logistic Regression: Item Mentioned vs. not Mentioned for 56 Items, Explanatory Variables: Country, Decade, Length of Manifesto, Uncoded Sentences

Item	*Zeroes*	*Pseudo R²*	*Correct Classi-fication*	*Item*	*Zeroes*	*Pseudo R²*	*Correct Classi-fication*
101	56.1 %	26.7 %	75.1 %	410	33.0 %	24.4 %	77.0 %
102	83.9 %	26,1 %	84.5 %	411	23.2 %	38.3 %	84.0 %
103	79.7 %	25.2 %	82.1 %	412	64.0 %	17.5 %	71.6 %
104	45.3 %	19.3 %	71.4 %	413	77.3 %	17.4 %	77.9 %
105	61.4 %	22.0 %	73.2 %	414	33.7 %	21.4 %	75.9 %
106	45.4 %	20.6 %	71.8 %	415	95.2 %	18.1 %	90.0 %
107	26.9 %	25.3 %	79.4 %	416	84.7 %	42.1 %	86.9 %
108	50.0 %	35.6 %	77.8 %	501	32.7 %	45.8 %	83.6 %
109	77.2 %	18.8 %	79.7 %	502	38.6 %	36.5 %	80.2 %
110	81.7 %	27.7 %	81.0 %	503	16.2 %	23.6 %	84.5 %
201	25.0 %	22.7 %	78.6 %	504	12.7 %	27.5 %	86.9 %
202	19.0 %	23.6 %	82.2 %	505	77.7 %	24.2 %	80.3 %
203	55.9 %	24.5 %	75.4 %	506	21.3 %	28.7 %	81.4 %
204	84.3 %	18.5 %	83.4 %	507	93.7 %	20.3 %	90.2 %
301	33.4 %	32.9 %	79.6 %	601	45.3 %	18.2 %	70.2 %
302	83.9 %	22.8 %	81.8 %	602	88.6 %	19.6 %	86.6 %
303	30.3 %	29.8 %	79.3 %	603	42.7 %	18.9 %	68.5 %
304	59.2 %	31.4 %	76.7 %	604	81.8 %	30.9 %	84.1 %
305	41.0 %	29.1 %	77.9 %	605	43.0 %	32.8 %	78.2 %

401	35.9 %	17.8 %	72.9 %	606	38.4 %	22.4 %	73.4 %
402	29.8 %	25.4 %	76.9 %	607	57.9 %	34.7 %	78.7 %
403	33.8 %	32.5 %	79.2 %	608	86.0 %	26.4 %	83.5 %
404	60.9 %	23.8 %	73.7 %	701	32.7 %	22.9 %	76.0 %
405	75.4 %	23.8 %	78.5 %	702	90.0 %	19.0 %	86.8 %
406	70.8 %	22.7 %	74.6 %	703	27.0 %	31.1 %	81.8 %
407	74.7 %	22.3 %	78.1 %	704	58.5 %	34.8 %	78.8 %
408	24.5 %	23.2 %	81.3 %	705	51.6 %	37.1 %	78.8 %
409	81.1 %	24.4 %	81.7 %	706	23.5 %	27.1 %	79.5 %

To provide some additional help for researchers facing the challenge of choosing items for their scale, table 5 summarizes the results of such regression models for all items. It is obvious that for most items, a large majority of cases can be classified correctly as mentioning or not mentioning the item based on the variables in the regression model. It follows that a large number of zeroes in the data set is due to non-substantial reasons. The larger the Pseudo-R^2 and the higher the amount of cases classified correctly, the more likely zeroes for a particular item are to be non-substantial and the less likely the item to structure political space. Therefore, researchers should choose items whose classifications can *not* be predicted simply by manifesto length and other technical factors.

Concerning the choice of items, are there any strategies to save exploratory factor analysis and other methods to reduce dimensionality or to create a unidimensional scale? After all, it is a very appealing method for generating multidimensional models of political space. Firstly, one could simply eliminate the items with a high number of zeros. However, apart from the fact that items essential in some countries but totally irrelevant in most others might be lost, it would have been difficult to define a percentage of occupied fields per item as an objective criterion. Secondly, especially sparsely populated items could simply be added up in a theoretically reasonable matter. While this would reduce the number of zeros in the dataset, it would pose another problem: A comparatively high number of items would be lost, especially in domain four including economic items. Lastly, any other reduction of parties and time periods – such as running separate analyses for single countries - would ruin the idea of a generalized left-

right scale at all. Factor analysis and related methods therefore cannot be recommended for CMP data. Rather, items should be chosen based on theoretical considerations.

The most sensible approach to avoid the pitfalls of CMP data therefore seems to be a theoretical choice of items, combined with the information about the likelihood that zeroes are substantially meaningful, and followed by a weighted or non-weighted additive approach to scale-building.

6. Conclusion

Using the CMP dataset is the only possibility for comparative political scientists to get positional data for political parties over time and for many countries. Researchers using this dataset are faced with certain challenges due to the structure of the dataset. This article analyses the pitfalls of the CMP dataset and proposes a solution to avoid them.

We disclose a distinctive feature of the CMP dataset affecting all procedures for extracting data. We found that with 53.5 percent of all observations, the dataset contains an unusually high number of zeroes. These zeroes vary strongly across items, election years and countries. While this in principle poses no problem with respect to saliency theory, it leads to a number of theoretical and methodological difficulties when dimensions of political conflicts are extracted and parties are positioned on these dimensions. Firstly, it is difficult to differentiate between zeroes that carry a substantial meaning in line with the theory on the one hand and empty items that result from technical reasons such as manifesto length and the percentage of uncoded sentences on the other hand Secondly, the fact that the empty items vary systematically across time and countries makes it difficult to extract data that can be used in comparative studies. Thirdly, all procedures using factor analysis either to select items or to measure positions are affected by the amount of zeroes. The high number of zeros results in a low correlation between items and thus violates a basic assumption of factor analysis.

Taking these problems into account, we give an indication which items are affected most and we deliver a possible way to identify these items. This allows researchers constructing a scale to exclude these items or single data points.

Notwithstanding these problems, the CMP data provide the only comparative dataset on the positions of political actors across a high number of countries and over a long period of time. This closes a gap between the still increasing use of spatial models in political science and our ability to empirically test them. Contradicting outcomes using various scales do not necessarily result in a lack of usability of the data

itself as some authors claim, they are rather a call for more sophisticated methods using the dataset.

7. References

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

Budge, Ian. 2001a. "Theory and Measurement of Party Policy Positions." In *Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998*, Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. Oxford: Oxford University Press.

Budge, Ian. 2001b. "Validating the Manifesto Research Group approach: theoretical assumptions and empirical confirmations." In *Estimating the Policy Positions of Political Actors*, Michael Laver. London: Routledge.

Budge, Ian. 1994. "A New Spatial Theory of Party Competition: Uncertainty, Ideology and Policy Equilibria View Comparatively and Temporally." *British Journal of Political Science* 24: 443-467.

Budge, Ian and Judith Bara. 2001. "Manifesto-Based Research: A Critical Review." In *Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998*, Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. Oxford: Oxford University Press.

Budge, Ian and Hans-Dieter Klingemann. 2001. "Finally! Comparative Over-Time Mapping of Party Policy Movement." In *Mapping Policy Preferences. Estimates for Parties, Electors, and Governments 1945-1998*, Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. Oxford: Oxford University Press.

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

Budge, Ian, David Robertson, and Derek Hearl. 1987. *Ideology, strategy and party change: spatial analyses of post-war election programmes in 19 democracies*. Cambridge: Cambridge University Press.

Castles, Francis G., and Peter Mair. 1984. "Left-Right Political Scales: Some Expert Judgements." *European Journal of Political Research* 12: 73-88.

Franzmann, Simon, and André Kaiser. 2006. "Locating Political Parties in Policy Space. Reanalysis of Party Manifesto Data." *Party Politics* 12 (2), 163-188.

Gabel, Matthew J. and John D. Huber. 2000. "Putting Parties in Their Place: Inferring Party Left-Right Ideological Positions from Party Manifestos Data." *American Journal of Political Science* 44(1): 94-103.

Janda, Kenneth, Robert Harmel, and Patricia Goff. 1995. "Changes in Party Identity: Evidence from Party Manifestos." *Party Politics* 1: 171-196.

Huber, John D. 1996. "The Vote of Confidence in Parliamentary Democracies." *American Political Science Review* 90(2): 269-282.

Kim, Heemin et al. 1998. "Voter Ideology in Western Democracies, 1946-1989." *European Journal of Political Research* 33: 73-79.

Kim, Heemin et al. 2002. "Government Partisanship in Western Democracies, 1945-1998." *European Journal of Political Research* 41: 187-206.

Klingemann, Hans-Dieter. 1995. "Party Positions and Voter Orientations." In Klingemann, Hans-Dieter and Fuchs, Dieter, *Citizens and the State*. Oxford: Oxford University Press.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael D. 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.

Laver, Michael. 2001. "Position and Salience in the Policies of Political Actors." In *Estimating the Policy Positions of Political Actors*, ed. Michael Laver. London: Routledge.

Laver, Michael, and Ian Budge. 1992. "Measuring Policy Distances and Modelling Coalition Formation." In *Party Policy and Government Coalitions*, Michael Laver, and Ian Budge. New York: St. Martin's Press.

Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3): 619-634.

Laver, Michael, and Ben W. Hunt,. 1992. *Policy and Party Competition*. New York, London: Routledge.

Laver, Michael, and Norman Schofield. 1990. *Multiparty Government: The Politics of Coalition in Europe*. Oxford: Oxford University Press.

Laver, Michael, John Garry. 2003. "Extracting policy positions from political texts using words as data" *American Political Science Review* 97, 311–31.

Laver Michael, Kenneth Benoit, and Slava Mikhaylov. 2008. "Estimating party policy positions with uncertainty, based on manifesto codings." *APSA Conference Paper*.

Lynn, Jonathan, and Antony Jay. 1981. *Yes, Minister*. London: BBC Books.

Martin, Lanny W., and Georg Vanberg. 2008. "A Robust Transformation Procedure for Interpreting Political Text" *Political Analysis* 16, 93-100.

McCubbins, Matthew D. et al. 1987. "Administrative Procedures as Instruments of Political Control." *Journal of Law, Economics, and Organization* 3(2): 243-277.

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

McDonald, Michael D., Silvia M. Mendes, and Ian Budge. 2004. "What are Elections For? Conferring the Median Mandate." *British Journal of Political Science* 34: 1-26.

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

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time Series Policy Positions from Texts." *American Journal of Political Science* 52(3): 705-722.

Tsebelis, George. 2002. *Veto Players. How Political Institutions Work*. Princeton: Princeton University Press.

Tsebelis, George, and Jeannette Money. 1995. "Bicameral Negotiations: The Navette System in France." *British Journal of Political Science* 25 (1): 101-129.

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

TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

Kaiserslautern Occasional Papers in Political Science (KOPS)
ISSN 1861-7018

Herausgeber
Prof. Dr. Jürgen Wilzewski

Druck
Technische Universität Kaiserslautern
Postfach 3049
D-67653 Kaiserslautern
http://www.uni-kl.de/politik