Political Studies

Political Studies Association

POLITICAL STUDIES: 2013 VOL 61(S1), 3–23

doi: 10.1111/1467-9248.12015

# What to Do (and Not to Do) with the Comparative Manifestos Project Data

**Kostas Gemenis**

*University of Twente*

The Comparative Manifestos Project (CMP) data are the most popular source for parties' positions on the left–right and other ideological and policy dimensions. Over the past few years, several researchers have identified various methodological problems in the CMP data, but third-party users rarely acknowledge them. This article classifies the problems associated with the CMP into four areas: (1) theoretical underpinnings of the coding scheme; (2) document selection; (3) coding reliability; and (4) scaling. The article reviews each area systematically and concludes with a set of recommendations regarding the use of the CMP data.

**Keywords:** Comparative Manifestos Project; policy positions; measurement quality; validity; reliability

The theoretical and empirical qualities of manifestos as representative documents of party ideologies have led researchers to estimate parties' policy positions through content analysis. To date, the most systematic attempt has been initiated by Ian Budge and his colleagues who established the Manifesto Research Group (MRG) in 1979. The MRG embarked on an ambitious task, namely to collect and code the manifestos of all major political parties in nineteen countries. Ten years later, the project was renamed the Comparative Manifestos Project (CMP) and under the direction of Hans-Dieter Klingemann its coverage was extended to include parties in Central and Eastern European countries. In 2009, the project was once more renamed as Manifesto Research on Political Representation (MARPOR) with plans to extend its coverage to political parties in Asia and Latin America under a twelve-year grant from the German Science Foundation (DFG).[1]

The project works under the assumption that parties compete against each other by emphasising different policy issues rather than taking opposing positions on the same issues. The CMP relies on trained coders to collect the manifestos, parse them into 'quasi-sentences' and assign each of these quasi-sentences into one of the CMP's 56 issue categories (Budge, 2001a). The results of this laborious process are presented in terms of percentage frequencies, which intend to measure each party's 'relative emphasis' on each of these 56 issues of the coding scheme. In addition, the CMP has been using a technique to scale these frequency data in order to estimate parties' positions on the left–right (L–R) dimension (Laver and Budge, 1992).

Despite the advent of various computerised approaches to content analysis of party manifestos and other political texts (Laver and Garry, 2000; Laver *et al.*, 2003; Proksch and Slapin, 2009), the CMP remains the most popular source for parties' policy positions. Undoubtedly, its popularity lies in the rich time-series data which run for more than two dozen countries since 1945 and include parties' positions on the L–R scale (Budge *et al.*,

2001; Klingemann *et al.*, 2006). As a consequence, the CMP data have been used in hundreds of PhD theses, monographs and journal articles to test important questions regarding political representation, government coalition formation and spatial models of voting behaviour, and have won the American Political Science Association's 2003 best data set award. Proponents of the project argue that the CMP data can be used as a 'gold standard' to validate the results of the computer coding of party manifestos (Pennings, 2011), while the legacy of the CMP in methodological terms can be found in other projects such as the Euromanifestos Project, the Regional Manifestos Project and the Comparative Agendas Project.

Undoubtedly, the popularity of the CMP data lies in its availability and extensive temporal and spatial coverage. Expert surveys, which emerged in the 1980s as the most visible methodological rival for estimating parties' policy positions, have not been conducted and disseminated in a systematic manner comparable to the CMP. Expert surveys are still conducted rather infrequently (Gabel and Huber, 2000, p. 94) and often use different question wordings and response scales. In terms of data collection and dissemination the CMP is simply unparalleled, but how valid and reliable are its data?

Over the years, numerous researchers have questioned the validity and reliability of the CMP data and especially of its L–R estimates. Although these criticisms have not been resolved, proponents of the project argue that its data are valid and reliable and that they should be accepted 'as is' simply because there is no alternative (Budge and Pennings, 2007b; Pennings, 2011;Volkens *et al.*, 2009). Consequently, many third-party users simply acknowledge but nonetheless dismiss some of the most important criticisms or treat the CMP data as entirely uncontroversial. This article offers a state-of-the-art review of the problems associated with the CMP data by classifying them into four areas: (1) theoretical underpinnings of the coding scheme; (2) document selection; (3) coding reliability; and (4) scaling. The aim of the article is not to challenge existing research by casting doubt on the validity of the CMP data, but rather to summarise the critical literature. If researchers wish to make valid and reliable inferences by using the CMP data, they ought to be aware not only of the project's strengths, but also of its weaknesses. The article therefore concludes with a set of suggestions about how researchers can make use of the CMP data in a much more efficient way.

## The Theory behind the Coding Scheme

The first strand in the critical literature has looked at the theoretical underpinnings and applicability of the 56-categories coding scheme which has been developed by the CMP. According to the CMP, party competition is characterised by the prevalence of valence issues, that is, issues in which there is a broad agreement about the desired outcome: 'low unemployment, low inflation, high educational standards, or good healthcare' (Clarke *et al.*, 2004, p. 8). According to Budge (2001a, p. 82), this means that 'all party programmes endorse the same position, with only minor exceptions' because endorsing the not-so-favoured position would result in electoral suicide (Budge, 2001b, pp. 212–3). This also implies that parties choose a set of issues which they consider they 'own' and emphasise them consistently 'in an attempt to increase the salience of these for voters' (Budge, 2001a, p. 82) and therefore win elections. The methodological implication of this 'salience theory of party

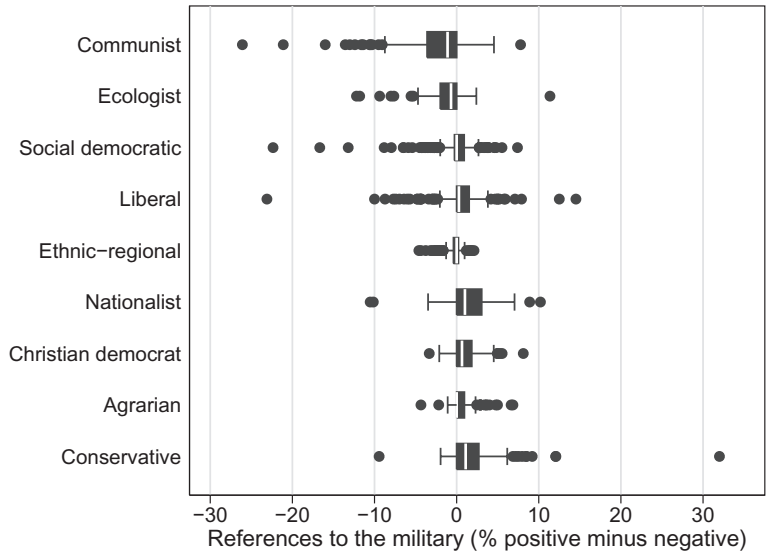
competition' is that 'policy differences between parties' are assumed to 'consist of contrasting emphases placed in different policy areas' (Budge, 2001a, p. 82). Thereby, the CMP data that measure the relative emphasis that parties put on the 56 issues in their manifestos can be used (through an appropriate scaling technique) to estimate parties' positions.

Unfortunately, the credibility of the salience theory of party competition as outlined by Budge and his colleagues in the CMP weakens in the face of a rigorous theoretical scrutiny. As Michael Laver (2001a, p. 7) pointed out, the CMP assumption that 'all party programmes endorse the same position, with only minor exceptions' is based on the 'false premise of majoritarian elections'. As is widely known, in multiparty systems not all parties endorse the same position in their programmes because proportional representation gives incentives for the creation of 'niche' parties that advocate not-so-popular positions. Laver (2001a, p. 7) also noted that the 'salience theory' was informed by studying the empirical data from Britain and the US, two countries with majoritarian party systems where valence issues are prevalent. It should be expected therefore that, in multiparty systems (or systems with a pronounced confrontational character in party competition), the CMP estimates might have much lower validity (Dinas and Gemenis, 2010; Franzmann and Kaiser, 2006; Pelizzo, 2003).

In addition, for some issues, like the legalisation of soft drugs or abortion, the important thing is not the emphasis, but the position of parties (Laver, 2001b, p. 66). 'If we win the next election all bankers will be taken out and shot': what matters in this statement is not how many times it will be repeated but its positional content (quoted in Laver, 2001a, p. 27). This is not taken into account by the CMP. Other issues such as support for the environment could be considered as valence issues in which all parties would endorse the same position. Although this is certainly true, it does not necessarily follow that parties cannot take positions that are practically anti-environmental (Gemenis *et al.*, 2012). Many parties, especially those on the radical right, have been routinely opposing green taxes and environmental standards, and have been challenging the view that humans have largely contributed to global warming. Contrary to what the salience theory of party competition would predict, advocating anti-environmental positions did not result in electoral suicide but rather contributed to the success of the radical right (Ivarsflaten, 2008, p. 8).

Furthermore, the credibility of the salience theory of party competition weakens in the face of empirical scrutiny. As Michael McDonald and Silvia Mendes (2001b, pp. 91–2) observed, the CMP coding scheme of 56 categories is inconsistent with the salience theory. A close look at the CMP coding scheme reveals that the 56 categories were not really designed to measure the salience given to issues, but rather the emphasis within a given pro or con position. Twenty-four of the CMP categories are explicitly positional 'pro versus con' categories, whereas another twenty-eight are implicitly positional and censored, in the sense that for a pro category there is no corresponding con category and *vice versa*. This implies that the CMP is asking its coders to assign quasi-sentences in what is evidently a confrontational and not salience coding scheme. Budge (2001a, p. 83) argued that the original coding employed by the MRG included the fifteen 'pro versus con' categories devised to capture some confrontational aspects of party manifestos, revealing thus their 'gut feeling that party competition consists in direct confrontation between pro and con positions on each specific issue' but that, 'in practice', the CMP coding scheme works like

**Figure 1: Differences in Positions towards the Military among Party Families (EU-27 Countries)**




a pure saliency one (Budge, 2001b, p. 211). The CMP later used this feature of the coding scheme to assess the construct validity of the data. By contrasting the 'pro' with the 'con' dyads within the CMP data they found 'the overwhelming number of references going to one of the possible positions' (Robertson, 1987, pp. 50–1). As a consequence, the CMP has taken this finding as evidence in favour of the salience theory.

This analysis, however, was performed on the aggregated data across 24 countries. The aggregated data tell us little if anything about the patterns of emphasis within individual party systems. The CMP data re-analysis performed by Laver (2001a, pp. 12–4) showed that the aggregate analysis masked important cross-country and cross-party variation and effectively showed that most issues have a clear positional character. An example of variation in positions across party families can be seen from the box plots in Figure 1, which present the support for the military by subtracting the negative from the positive references. Clear differences emerge as the majority of communist and ecologist parties make more negative than positive statements about the military in their manifestos, whereas the majority of Christian democrat and conservative parties make more positive than negative statements. Similar patterns are observed in cross-national comparisons among other positional issues using the CMP data.

The argument against the salience theory of party competition was later confirmed by the analysis of Simon Franzmann and André Kaiser (2006) who differentiated the valence from positional issues in each of the party systems in the CMP data set. The analysis points out that party competition in most countries is characterised by the prevalence of positional issues, whereas there are temporal and spatial differences regarding which issues can be considered as valence. Issues have a life cycle and may appear as valence or positional

depending on the particular context. For instance, Franzmann and Kaiser (2006, pp. 169–70) found that the environment changed from a valence to a positional issue in 1983 with the entry of the greens in the German legislature. Most recently, research has challenged the notion that valence issues necessarily imply consensus (Pardos-Prado, 2012) while voting advice applications have successfully estimated parties' positions cross-nationally by explicitly adopting a confrontational coding scheme (Gemenis, 2012b). Taken together, these findings reveal that the CMP data are based on a coding scheme that may not truly reflect party competition in the issue market (Franzmann, 2011) and this has implications for how the percentage emphasis on issue categories can be scaled in order to measure parties' positions on dimensions of interest.

## Document Selection

Most recently, researchers began to question the document collection and selection procedure of the CMP. In a rather illustrative article, Martin Hansen (2008) focused on the Danish documents that have been used by the CMP. He observed that many of the CMP data are based on the coding of documents other than national election manifestos: party leader speeches, drafts of manifestos, local election manifestos, advertisements in newspapers, speeches by non-party leaders and, in one case, a text from a think tank reporting on a rival party's policy goals (Hansen, 2008, pp. 208–10). Similarly, Sven-Oliver Proksch and Jonathan Slapin (2009, p. 330) found that the German party documents collected by the CMP included short election proclamations, party congress speeches and, in one case, an action programme published two years after the election for which it was used. Moreover, for Japan the CMP coding was based exclusively on 'rapid-fire pre-election interviews by a national daily newspaper' which are extremely short and constrained in their range of issues (Proksch *et al*., 2010, p. 115), whereas for Israel the CMP collection is almost entirely based on articles, advertisements and interviews in national newspapers. It would be comforting to find that these four countries are deviant cases inasmuch as the document collection is concerned. A recent review (Gemenis, 2012a), however, revealed that in many countries the CMP has coded a wide variety of documents including regional manifestos, election flyers, party leader speeches, programme summaries in newspapers and hand-written documents. Several researchers have questioned whether these documents can be considered to be equivalent to election manifestos as they differ in terms of size, style and authoritativeness.

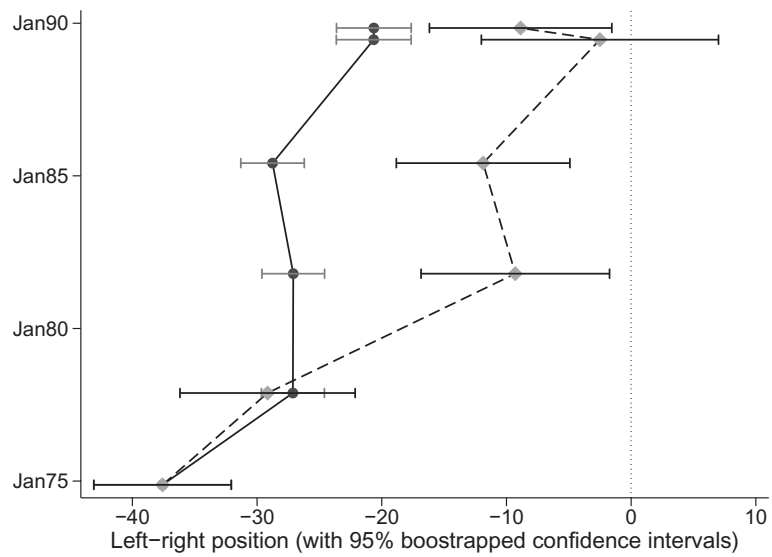
The case for coding such proxy documents was presented in the individual chapters of Budge *et al*. (1987) where the authors argued that these documents can be considered as being equivalent to manifestos. This is questionable, however, for several reasons. First, there is the scepticism of the CMP investigators expressed rather covertly in the chapters of Budge *et al*. (1987) and more openly in their earlier correspondence with Budge (Gemenis, 2012a, p. 596). Second, the inclusion of material such as reports in newspapers, speeches, pamphlets and leaflets was born out of necessity, rather than a well-thought strategy based on evidence showing that the different types of document are equally comparable to manifestos (Gemenis, 2012a, p. 596).

Although some of the problems associated with document collection were outlined in the first edited volume by the CMP (Budge *et al*., 1987), the subsequent CMP publications

did not pay much attention to issues surrounding the selection of documents (Budge *et al*., 2001; Klingemann *et al*., 2006). As a consequence, the problems related to document selection were never seriously addressed until Kenneth Benoit *et al*. (2009) considered the stochastic process of generating party manifestos and argued that manifesto-based policy estimates should come with associated measures of error reflecting the differences between documents. Nevertheless, the estimates they provided only consider document length as a source of error. Benoit *et al*. (2009, p. 497) rightly argued that longer documents are more authoritative and therefore less error prone, but did not consider the differences in the types of document. What happens when the CMP compares different types of document?

We could argue that proxy documents, such as party leader speeches, are not drafted through the same processes that election manifestos are subject to and therefore might not be representative of the party as a unitary actor. Moreover, some of the proxy documents that have been used by the CMP often have different roles in election campaigns and may not cover all policy issues. Controlled, qualitative and statistical comparisons of coded documents showed that considerable differences emerge when proxy documents are coded in place of election manifestos (Gemenis, 2012a, pp. 597–601). More specifically, the non-pledge content typically found in proxy documents such as party leader speeches introduces measurement error which often appears as a centrist bias in parties' L–R estimates. An illustrative example can be found in Figure 2, which shows the position of the social democratic party in Greece (PASOK). From 1981 onwards, the position of PASOK would have shifted by 20 points towards the centre if the CMP had coded party leader speeches instead of the PASOK manifestos. As is clearly shown from Figure 2, estimating

**Figure 2: PASOK's Position using Manifestos (Solid Line) and Party Leader Speeches (Dashed Line)**



measurement error through bootstrapping the coded documents (Benoit *et al.*, 2009) is an approach that cannot always account for such differences among documents. Conversely, scaling methods such as the one proposed by Will Lowe *et al.* (2011) are able to adjust for the bias introduced by the coding of proxy documents, but only at the expense of introducing considerable noise which compromises the value of the resulting party policy estimates when those are compared to expert surveys. These findings imply that third-party users of the CMP data should be aware that the estimates are often based on the coding of proxy documents and should discuss whether this has implications for their inferences.

## Coding Reliability

The third strand in the critical literature has focused on the reliability of the coding procedure, where documents are parsed into quasi-sentences and quasi-sentences are assigned the issue categories of the CMP coding scheme. When coding reliability is concerned in the context of content analysis, we can distinguish between stability, reproducibility and accuracy (Krippendorff, 2004a, pp. 214–6). Stability is the weakest form of coding reliability and can be measured when the same text is coded by the same coder more than once. Reproducibility is measured by the degree of agreement among independent coders. Put otherwise, we can infer the coding reliability by measuring the coder agreement (Krippendorff, 2004b, p. 414). Finally, accuracy is the strongest form of coding reliability and is measured by the agreement between coders and a given standard (Krippendorff, 2004a, p. 216). Coding reliability has been a concern for the CMP since its early stages, as the investigators have been asked to mention the specific reliability tests they employed.

It has been argued that the CMP data 'have been subjected to standard stability and inter-coder reliability tests' (Budge and Bara, 2001, p. 14), that the CMP's data 'reliability ha[s] been extensively examined' (Budge and Pennings, 2007a, p. 125) and that inter-coder agreement tests were carried out 'with satisfactory results' (Klingemann *et al.*, 2006, p. 107). Third-party users of the CMP data tend to accept these claims, although more critical evaluations tend to agree that the reliability of the CMP data has been overstated. One of the publications cited by Budge and Bara (2001, p. 14) which purportedly tested the reliability of the CMP data is Budge and Farlie (1977, pp. 421–3). This work, however, was published before the MRG was instituted and the only mention of reliability refers to David Robertson's (1976) coding of British manifestos under two different coding schemes, concluding that 'the generally close correspondence of these results with his previous codings was taken as a sufficient guarantee of their reliability and validity' (Budge and Farlie, 1977, p. 422). This test, however, cannot be taken as a reliability measure as defined above because it refers to the coding by a coder using two different coding schemes. In all three forms of coding reliability defined above it is assumed that coders use a common coding scheme.

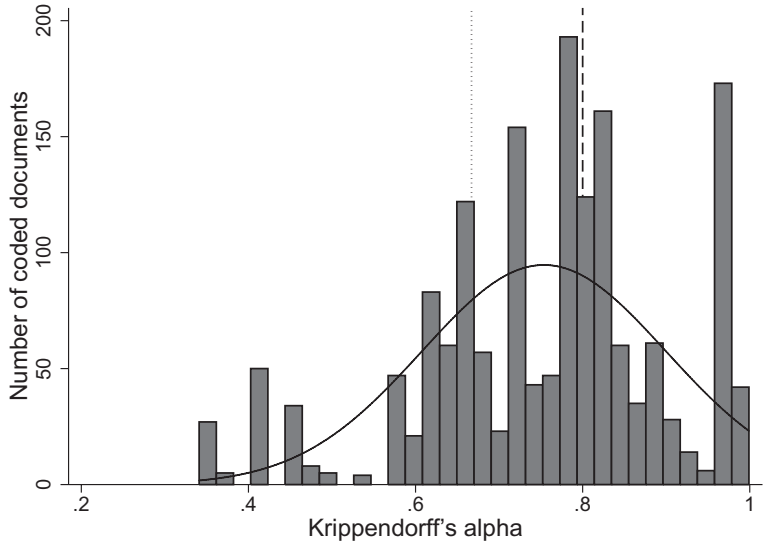
Another widely cited source (Budge and Bara, 2001, p. 14; Budge and Pennings, 2007a, p. 125; Klingemann *et al.*, 2006, p. 107) which purportedly tested the reliability of the CMP data is Budge *et al.* (1987, pp. 23–4). This source in turn refers to the individual chapters in the book. For instance, Judith Bara (1987) presented as 'reliability checks' the coding of some documents by the same coder. This is a measure of stability, the weakest form of coding reliability but, even so, the results of the checks were not reported. David Hearl

(1987, p. 237) mentioned that when some Belgian programmes were coded by two different coders, there were few discrepancies, although the discrepancies were 'not quantified'. Elsewhere, the CMP attempted to measure reliability by comparing the coding results before and after 1983. The tests performed by Hearl (2001) included a comparison of means and a comparison of factor analyses between the two parts of the data set. Such tests, however, are uninformative with regard to coder reliability. In addition, the oft-cited reliability testing conducted by McDonald and Mendes (2001a, pp. 138–40) and Klingemann *et al.* (2006, pp. 86–104) investigated the reliability of scaling techniques rather than coding reliability which involves different assumptions, measures and techniques (Krippendorff, 2004b; Lombard *et al.*, 2002).

Because it was considered that 'it is misleading to regard intercoder reliability as a direct measure of the reliability of the indicators themselves' (Klingemann *et al.*, 2006, p. 107; Volkens, 2001, p. 98), the project placed most emphasis on the 'quality control of the data' in the sense of 'setting and enforcing central standards on coders – first by getting them to conform to a prescribed English-language standard and secondly by very close interaction between them and the central supervisor' (Volkens, 2001, p. 94). To be sure, since 1989 the CMP has rigorously assessed the coding reliability of its coding scheme in terms of its accuracy. All the new coders who have been recruited to perform work for the CMP since 1989 have been asked to do a training test which involves coding a manifesto that has been chosen by the CMP in terms of its difficulty (Volkens, 2001, p. 99). The results are then compared to the 'standard' coding of the manifesto that has been agreed by the CMP investigators. Andrea Volkens (2001, pp. 99–100) reported that 'coders deviate an average of 10 percentage points from each other', that they 'also deviate by this amount on average from the "correct" solution' and that 'the average Pearson correlation is above 0.7 between all pairs of coders taking the test and between individual coding decisions and the "correct" solution'. There is a consensus among statisticians, however, that measures such as 'percent agreement' or correlation coefficients such as Pearson's product–moment correlation coefficient are misleading when used as a measure of coding reliability (Hayes and Krippendorff, 2007, pp. 79–80; Krippendorff, 2004a, pp. 428–9; Lombard *et al.*, 2002, pp. 590–1).

Such criticisms have prompted the project to report detailed data regarding the reliability scores of the CMP coders who have taken the training test using measures such as Krippendorff's alpha, the statistic that is generally agreed to be the 'measure with appropriate reliability interpretations in content analysis' (Krippendorff, 2004a, p. 221). The latest release of the data set contains the test results for 81 coders who have coded 1,687 of the documents in the data set, where the units of analysis are the issue categories and agreement between experts and the 'correct solutions' are expressed using the ratio version of Krippendorff's alpha. Figure 3 shows the distribution of these documents according to the coders' test results. The figure shows that approximately 17.4 per cent of the coded documents for which we have information have been coded by people who scored below 0.67, the minimum standard for making tentative conclusions (Krippendorff, 2004a, p. 241) indicated by the dotted line, and about 42.4 per cent were coded by people who scored above 0.8, the lower acceptable limit for good reliability (indicated by the dashed line). These findings are important, considering that the CMP test scores were often obtained after the second attempt at the test and only after coders received extensive feedback on

**Figure 3: The Distribution of Coded Documents According to the Coders' Reliability Score in the Training Test (Dashed Line at $\alpha = .8$, Dotted Line at $\alpha = .667$)**




their first-round mistakes. So does the problem lie with the coders or with the coding scheme?

The results of some recent rigorous empirical tests of coding reliability in the context of the CMP can offer some insights. Slava Mikhaylov *et al.* (2012) asked a large number of participants to code the two manifestos that have been used by the CMP in the training test. The participants were recruited through a list of coders who have worked for the CMP and the Euromanifestos Project, as well as a selection of staff and postgraduates in European and North American universities. In all, 172 participants were invited and 39 completed the test online. Mikhaylov *et al.* (2012) used each quasi-sentence in the text as the unit of analysis and assessed agreement between the coders with Fleiss' kappa, a statistic that approximates Krippendorff's alpha in large (coders by units) samples. In order to be 'fair', Mikhaylov *et al.* (2012, p. 83) 'discarded the bottom fourth of the test coders in terms of their reliability while dropping none from the top'. The results showed that the obtained reliability coefficients were 'exceptionally poor by conventional standards', ranging from 0.35 to 0.47 (Mikhaylov *et al.*, 2012, p. 84). The results showed that the coders systematically miscoded quasi-sentences, often by coding 'left' issues as 'right' and *vice versa*, and the mistakes did not tend to 'cancel out' as had previously been argued by the CMP (Klingemann *et al.*, 2006, pp. 112–5). Similarly, when the Euromanifestos Project coding scheme was subjected to similar inter-coder reliability tests, the results confirmed this pattern (Braun *et al.*, 2010, pp. 27–32). Mikhaylov *et al.* (2012, p. 83) found no significant differences among coders with different levels of experience although they reported substantial differences in reliability among different coding categories. Coders generally agree on assigning (quasi-) sentences to some categories (such as 703: Farmers), while others (such as 303: Government and

Administrative Efficiency) are plagued by misclassification (Mikhaylov *et al.*, 2012, p. 85). These results imply that the problem of unreliability does not lie with the coders but with the complex nature of the CMP coding scheme, which could benefit from adjustments or revisions. For instance, Thomas Däubler *et al.* (2012) presented a good argument for switching the coding from quasi-sentences to natural sentences which coders can define exogenously. Even though this move will not improve the inter-coder reliability substantially (Braun *et al.*, 2010, p. 31), it can save time and associated costs without affecting the validity of the results (Däubler *et al.*, 2012).

## The 'Standard' CMP Scaling Technique

Without a doubt, the most criticised aspect of the project is its 'standard' scale measuring parties' and governments' L–R positions (aka 'rile' in the data sets). In the early days of the MRG, the favoured technique to estimate the L–R positions of parties involved a two-stage factor analysis (Budge, 1987, pp. 28–30). By factor analysing the data for each country separately, the policy content of the L–R scale was assumed to be different in each country. As Wouter Van der Brug (2001, pp. 120–1) and others (Elff, 2012; Franzmann, 2012; Hans and Hönnige, 2008) have argued, however, factor analysis is not the appropriate multivariate technique to be used with the CMP data. The CMP data are best conceived as proximity relations between parties and issues and, as such, the frequencies of the coding categories will not be linearly related as is assumed by factor analysis. Parties that tend to give a positive emphasis to an issue are less likely to give a negative emphasis to the same issue and *vice versa*. The relationship between contrasting issue categories that make up most of the CMP coding scheme is not linear and therefore factor analysis will yield misleading results.

Subsequently, Laver and Budge (1992) proposed a new L–R scale which was used in their book on government coalitions and subsequently adopted as the 'standard' L–R scale by the CMP (aka 'rile'). Their approach differed radically from the previous attempt at scaling because it followed the logic of the summated rating scale. In this scale, the sum of the emphasis on a fixed set of 'right' issues is subtracted from another fixed set of 'left' issues. To determine which of the 56 issues would be included in each of these two sets, Laver and Budge (1992) reorganised the 56 issues into 20 more general policy dimensions and established 'marker' items for the left and right dimensions. With the use of factor analysis they were able to see which of the issues loaded consistently highly with either of the left or right dimensions. In addition, some issues were added intuitively in the left and right components.

Even though the ubiquitous CMP L–R scale has been hailed as the project's 'crowning achievement' (Budge and Klingemann, 2001, p. 19), it is also one of the most controversial and criticised aspects of the project. To begin with, the move from country-specific factor analyses to a common L–R scale for all countries in Europe and beyond runs counter to well-established research which argues that the meaning of L–R differs across countries (Benoit and Laver, 2006, pp. 132–6; Fuchs and Klingemann, 1990). In this respect, Ryan Bakker (2007, p. 17) and Elias Dinas and Kostas Gemenis (2010) challenged the notion that L–R is made of two reliable 'left' and 'right' scales by showing that some scale items do not 'fit' in the underlying 'left' and 'right' dimensions.

In a critical review essay, Laver (2001a) cautioned third-party users of the CMP data to be aware of some undesirable properties of the 'standard' L–R scale. It was shown that party movement on the L–R scale is not only caused by changes in emphasis to the categories included in the left and right components but also by changes in emphasis to all the other categories excluded from the scale (Laver, 2001a, p. 22), including the 57th category of 'uncoded' quasi-sentences. This could possibly explain the CMP's tendency to portray extreme right or left parties, spuriously, as centrist (Dinas and Gemenis, 2010; Pelizzo, 2003). To deal with this problem, HeeMin Kim and Richard Fording (1998, pp. 78–9) have modified the 'standard' CMP scale by dividing the difference between the left and right components, not by the total number of quasi-sentences in the manifesto, but by the total number of quasi-sentences included in the L–R scale. This adjusted scale, however, has the tendency of moving parties towards the extremes (Lowe *et al.*, 2011, pp. 129–30).

The approaches by Klingemann (1995) and Matthew Gabel and John Huber (2000) marked a return to factor analysis. Klingemann (1995) deductively excluded some issues and performed country-by-country factor analyses before arriving at a common scale, while Gabel and Huber (2000) followed an inductive approach by using principal components analysis to extract one single dimension which would represent a 'super issue' equivalent to the encompassing L–R dimension. The assumption that a common L–R scale can be used to estimate parties' positions across diverse countries has been shown to be problematic (Dinas and Gemenis, 2010; Franzmann and Kaiser, 2006). Moreover, both approaches are controversial inasmuch as they use factor analysis on proximity data (Van der Brug, 2001) that contain a high number of issue categories with zero frequencies (Hans and Hönnige, 2008).

Constructing scales after discarding the zero frequency categories or after filling them with 'reasonable values' from the posterior distribution of a latent variable under a Bayesian approach did not give very useful results (Bakker, 2007, pp. 17–8, pp. 26–7). Franzmann and Kaiser (2006) therefore proposed a more elaborate approach to scaling the CMP data which is based on a series of country-specific regression analyses with partial fixed effects and a smoothing technique. The aim was to differentiate between 'valence' and 'positional' left or right issues in each party system and to eliminate the extreme zigzagging of parties in the political space that is often observed in CMP L–R estimates. Even though the results seem to attain a good deal of validity in situations where the 'standard' CMP scale fails (Dinas and Gemenis, 2010), they have been criticised for being 'too good' because they are based on tautological assumptions regarding which issues can be considered left or right (Meyer, 2010, pp. 222–4).

More recently, Eric Linhart and Susumu Shikano (2009) proposed another scaling technique. By using the German CMP data, they categorised issues as positional L–R or valence and, after assigning scores (–1 and 1 to position issues and 0 to valence issues), aggregated them within each policy dimension and applied a logarithmic transformation. The idea behind this logarithmic transformation of the CMP data is to decrease the marginal effect of every additional quasi-sentence coded into each category. This is in line with the framework previously laid out by Laver (2001a) and McDonald and Mendes, (2001b), namely to build scales based on the opposing (pro/con) categories in the CMP coding scheme and take into account that, once a position is taken, its repetition is of less

importance. Lowe *et al.* (2011) took this idea further and argued that the construction of log ratio scales using the CMP data follows the idea behind the Weber-Fechner law of psychophysics regarding people's perception of differences between quantities. In practice this works simply by applying a logarithmic transformation to the left and right components of the 'standard' L–R scale or any other pro/con pair of issue categories in the CMP coding scheme. Moreover, Lowe *et al.* (2011, pp. 144–8) provided bootstrapped confidence intervals which intend to measure the degree of uncertainty of individual estimates, as well as evidence that the new scales perform better compared to the ones proposed by the CMP and others (Kim and Fording, 1998).

As is made evident by recent attempts to apply a similar scaling procedure to the Euromanifestos Project data (Veen, 2011), this contribution is often regarded as the most promising attempt to scale the CMP data. Thomas Meyer (2010, pp. 225–7), however, showed that the log ratio scales may produce odd results when one is interested in measuring policy shifts, while Franzmann (2012) argued that such transformations may be trying to correct what parties have been doing on purpose: if we are interested in estimating parties' positions as parties intend to present them, it should not matter how citizens perceive them. In such cases, psychophysics laws are irrelevant. Additionally, applying a logarithmic transformation to the CMP data comes at a cost. The resulting scales are expressed in log ratios, which means that measurement is shifted from ratio to interval level (Lowe *et al.*, 2011, p. 131). As such, the zero point in the resulting scales does not have a substantive interpretation in itself but represents an identification constraint imposed by the logit model used for scaling. Moreover, the scales do not have predefined end points. These issues therefore constrain the usability of the scales when researchers are interested in making substantive interpretations based on the values of their variables.

Since much of the argument put forward by Lowe *et al.* (2011) rests on the successful validation of their scales against expert survey estimates, it would be useful to re-examine their evidence. The four scatter plots at the top of Figure 4 compare the 'standard' ('rile') L–R scale to its log ratio version (Lowe *et al.*, 2011) against two common benchmarks: the L–R scale from the 2006 Chapel Hill expert survey (Hooghe *et al.*, 2010) and a modified version of the EU Profiler L–R scale (Gemenis, 2012b), two measures that have been shown to fare well in terms of validity and reliability. The four scatter plots at the bottom do so for the 'standard' EU integration scale and its log ratio version proposed by Lowe *et al.* (2011). In all scatter plots, the 2009 EU Profiler and 2006 expert survey data are matched to the closest observation of the 2005–10 period in the CMP data set. To be able to make comparisons across different measurement approaches, all scales have been rescaled to range from zero to one, while each scatter plot features two lines. A solid line indicates the line of perfect concordance ($y = x$), and a dashed line indicates the fit of a Deming regression, a technique that allows for errors in both $x$ and $y$ variables. The deviation of the reduced major axis from the line of perfect concordance indicates the degree of systematic error (or bias), which can also be expressed by Lawrence Lin's (1989) bias correction factor $C_b$ coefficient, while non-systematic error (noise) can be expressed by the well-known Pearson product–moment correlation coefficient $r$.

As is evident from the four scatter plots at the top, the log ratio version of the 'standard' CMP L–R scale only slightly improves the concordance between the scales and the

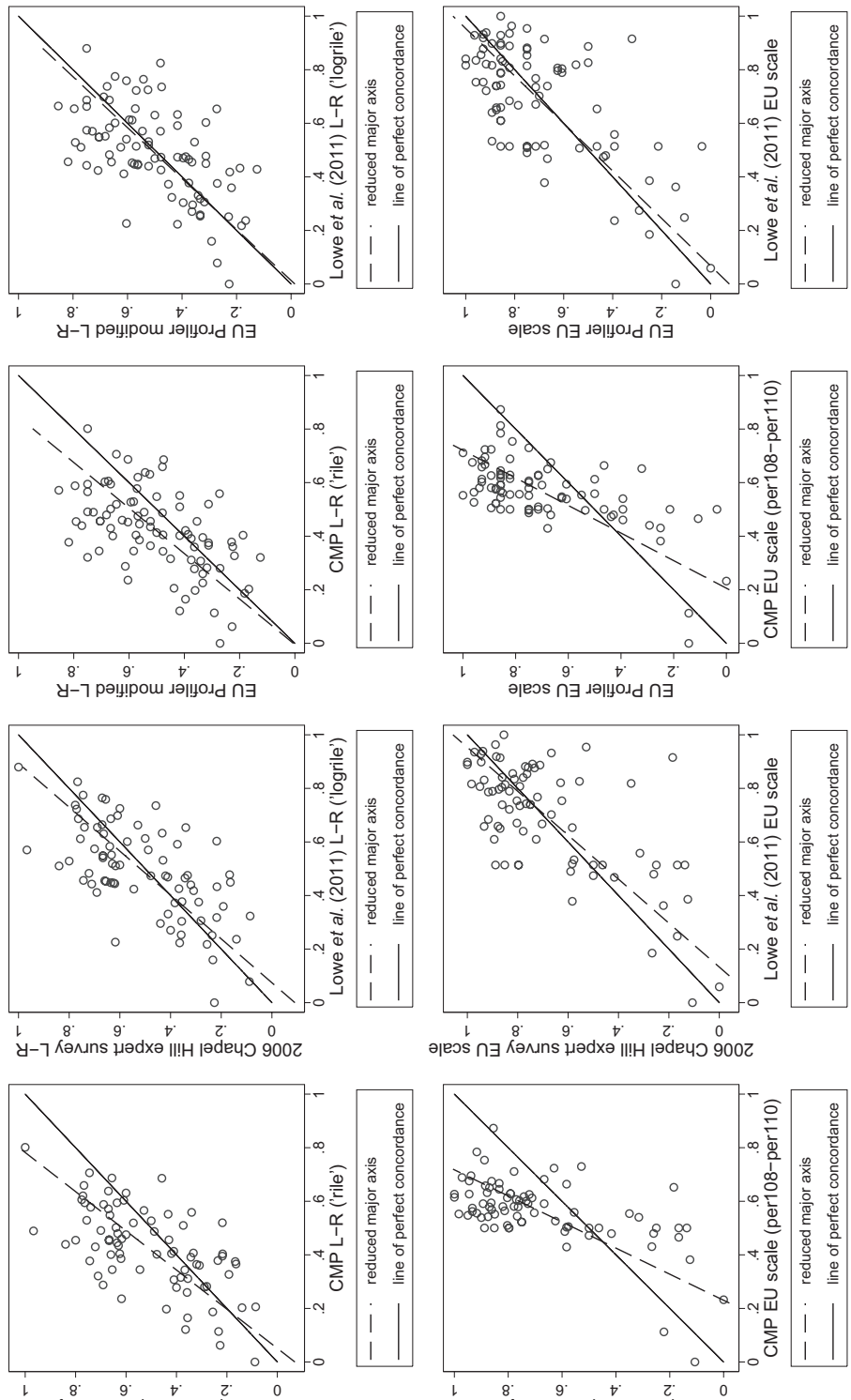



**Figure 4: Comparing the CMP L–R and EU Scales to the Corresponding Lowe *et al.* (2011) Scales**

**Table 1: Evaluating the Improvements Introduced by the Lowe *et al.* (2011) Log Ratio Scales**

| | *Concordance correlation coefficient (95% confidence intervals)* | *Bias correction $C_b$* | *Noise correction r* | *n* |
|---|---|---|---|---|
| Left–right (aka 'rile') | | | | |
| CHES/CMP | 0.539 (0.391, 0.660) | 0.860 | 0.627 | 79 |
| CHES/Lowe *et al.* | 0.633 (0.486, 0.745) | 0.976 | 0.648 | 79 |
| EU Profiler/CMP | 0.477 (0.316, 0.611) | 0.884 | 0.540 | 85 |
| EU Profile/Lowe *et al.* | 0.569 (0.407, 0.697) | 0.997 | 0.571 | 85 |
| EU integration | | | | |
| CHES/CMP | 0.403 (0.277, 0.516) | 0.666 | 0.606 | 79 |
| CHES/Lowe *et al.* | 0.669 (0.533, 0.772) | 0.980 | 0.683 | 79 |
| EU Profiler/CMP | 0.417 (0.296, 0.524) | 0.661 | 0.630 | 85 |
| EU Profiler/Lowe *et al.* | 0.669 (0.535, 0.770) | 0.992 | 0.675 | 85 |
| Planned vs. market economy | | | | |
| CHES/CMP | 0.373 (0.207, 0.518) | 0.816 | 0.456 | 79 |
| CHES/Lowe *et al.* | 0.461 (0.292, 0.603) | 0.885 | 0.521 | 79 |
| State involvement in the economy | | | | |
| CHES/CMP | 0.442 (0.253, 0.599) | 0.969 | 0.456 | 79 |
| CHES/Lowe *et al.* | 0.459 (0.300, 0.593) | 0.871 | 0.527 | 79 |
| Environmental protection | | | | |
| Benoit-Laver/CMP | 0.215 (0.133, 0.295) | 0.392 | 0.549 | 75 |
| Benoit-Laver/Lowe *et al.* | 0.230 (0.046, 0.398) | 0.813 | 0.283 | 75 |
| EU Profiler/CMP | 0.063 (0.021, 0.106) | 0.193 | 0.329 | 84 |
| EU Profiler/Lowe *et al.* | 0.181 (–0.022, 0.369) | 0.947 | 0.191 | 84 |

alternative sources. As noted before (Elff, 2012; Franzmann, 2012; Meyer, 2010), the 'logrile' scale scores are not much different to the 'rile' scores to begin with. The improvement introduced by the logarithmic transformation consists primarily of a small reduction in bias, which is also indicated by the move of the reduced major axis towards the line of perfect concordance. The improvement over the CMP EU integration scale, however, is more considerable. Table 1 quantifies these improvements brought to the log ratio scales proposed by Lowe *et al.* (2011).

Figure 5 presents some more evidence. Here, the four scatter plots at the top compare two different economic scales. The first economic scale is the 'Planned versus market economy' scale found in the CMP data sets. The second is the 'State involvement in the economy' scale which contains more issue categories and which was originally proposed by Benoit and Laver (2007). Both scales are compared to their log ratio versions using the economic L–R scale estimates of the 2006 Chapel Hill expert survey as a benchmark. The log ratio planned economy scale appears to have slightly less bias compared to the non-log version, whereas the log ratio state involvement scale actually increases the bias that it is supposed to correct (compare Benoit *et al.*, 2012). The four scatter plots at the bottom

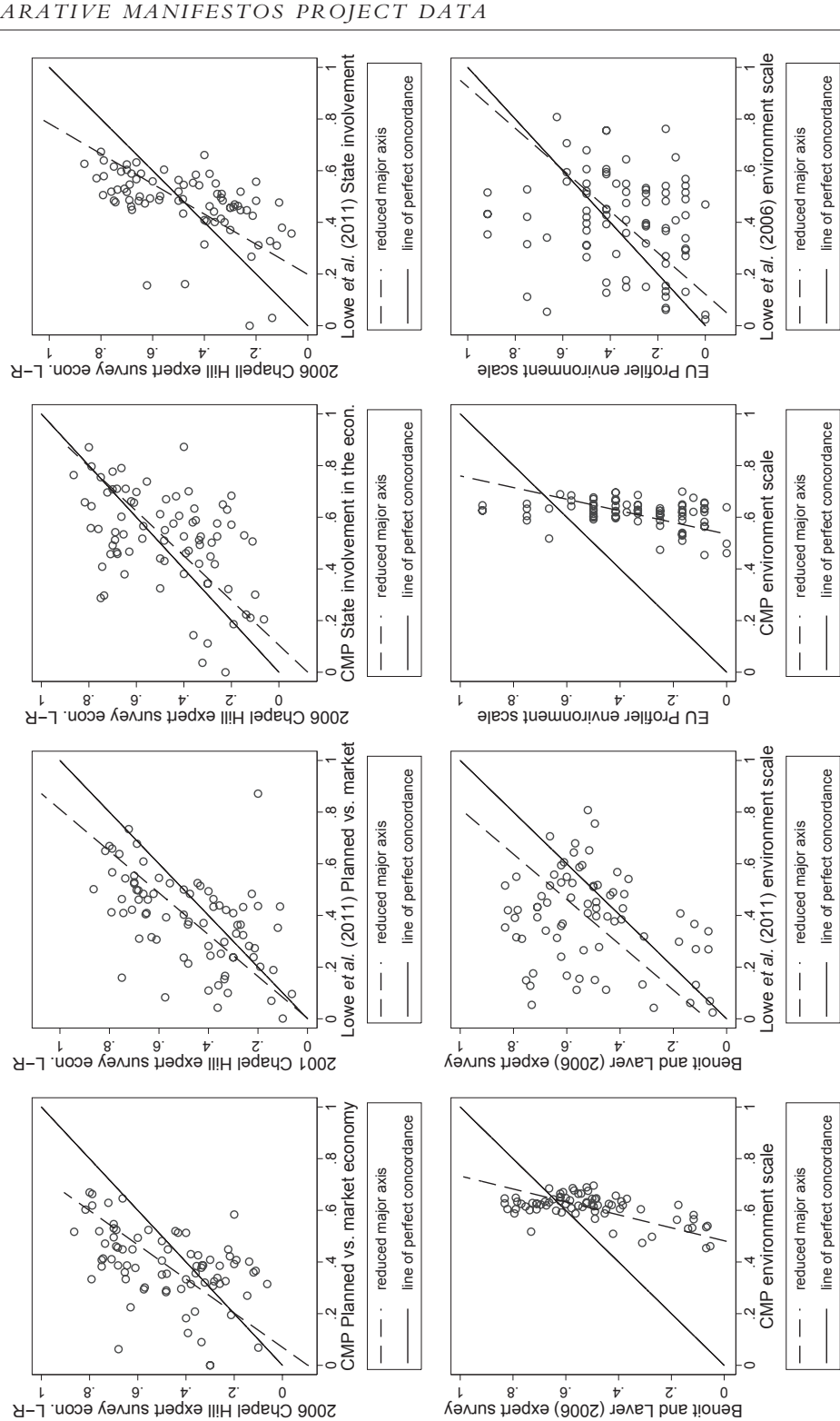



**Figure 5: Comparing the CMP Economic L–R and Environmental Scales to the Corresponding Lowe *et al.* (2011) Scales**

compare the CMP environment scale, originally proposed by Albert Weale *et al.* (2000) and reinvented by Lowe *et al.* (2011), to its log ratio version. As in Figure 4, the comparison is made against two benchmarks, in this case estimates from the Benoit and Laver (2006) expert survey and the EU Profiler environment scale (see Gemenis *et al.*, 2012). As already shown by Lowe *et al.* (2011), the logarithmic transformation successfully corrects for the centrist bias evident in the original scale. What the authors fail to acknowledge (at least explicitly), however, is that this is done at the expense of adding considerable noise in the scale as made evident by the decrease in Pearson's *r* and the dispersion of points in the scatter plots.

If we were to combine the measures of bias and noise in the multiplicative term known as the concordance correlation coefficient (Lin, 1989), we would conclude that the log ratio scales proposed by Lowe *et al.* (2011) do not substantially improve the concordance between scales based on the CMP data and alternative party position estimates. In each of the scales examined in Figures 4 and 5, the improvements in systematic error brought about by the transformation are largely counterbalanced by increases in random error or *vice versa* (compare Benoit *et al.*, 2012). With the exception of the EU integration scale, in all other scales examined here the overall improvement in concordance is not statistically significant when the *z*-transform confidence intervals of the concordance correlation coefficients are taken into account.

In addition to the approaches involving factor analysis and the construction of summated rating scales, others have proposed Bayesian approaches to CMP data scaling (Albright, 2008; Bakker, 2009, Elff, 2012). What is evident, however, is that alternative approaches to scaling yield considerably different results regarding parties' L–R positions (Dinas and Gemenis, 2010; Elff, 2012) which makes it somewhat disheartening to know that one's substantive inferences may be largely dependent on the choice of scaling method. It is therefore useful to argue about the appropriateness of a scaling approach in relation to its competitors but this 'search for the winner' may be elusive since scaling does not make the considerations about the validity and reliability of the CMP data as such redundant (Elff, 2012, pp. 36–7).

## Conclusions

The critical review in the preceding sections has identified some of the most important problems in the CMP approach to estimating parties' policy positions. On the one hand, if we were to summarise this review in a sceptical manner, we could say that the CMP uses a coding scheme that has not been empirically validated, applies it to many documents of questionable quality by using an unreliable hand-coding process, and scales the data into L–R estimates by using a technique that many researchers consider to be problematic. In light of this, we should conclude that third-party users should not treat the CMP L–R data as a 'gold standard' in validating parties' L–R positions obtained by other methods (compare Pennings, 2011). There is no evidence that the CMP data are more valid and reliable than those obtained by other methods, and validations that employ the CMP data should have a tentative character.

On the other hand, and given the lack of alternatives to the CMP data, we could summarise this review in an optimistic manner. The CMP is a unique and potentially valuable source of data on political parties. In particular, researchers could recognise that the

CMP estimates contain an unspecified amount of measurement error. Consequently, they can follow a strategy of separating what is valid and reliable in the data sets and using it in such a way that they can be confident about the robustness of their results. Following the review of the literature in the previous sections, I could summarise some recommendations in the following way.

Regarding the theoretical assumptions behind the CMP, third-party users need to place less emphasis on the 'salience theory' of the CMP. The theory cannot be validated empirically even by the CMP data. As Laver (2001a) suggested, the CMP data can be better conceptualised as 'relative emphasis' measures within a given (pro/con) position, or in case researchers are interested in saliency as such, they can combine the opposing categories of the coding scheme and create scales of policy importance (Lowe *et al.*, 2011, pp. 132–4). Nevertheless, it is also important to reiterate that it is sometimes difficult to measure parties' positions in specific policy areas by using the CMP data. For instance, Oleh Protsyk and Stela Garaz (2011, pp. 4–8) show how the multiculturalism positive and negative (607 and 608) categories of the CMP are poor proxies as they do not fully capture the intended concept.

Regarding the problems surrounding the collection of documents in the CMP, third-party users could follow a threefold approach. First, they should remove the so-called 'estimates' from the CMP data set which are essentially copy/pasted frequencies from the adjacent cells and treat these observations as missing data.[2] Second, if the CMP data are to be used as independent variables, third-party users can use the confidence intervals proposed by Benoit *et al.* (2009), which introduce an estimate of uncertainty based on the length of the coded documents. Third, since the proposed confidence intervals cannot control for different types of coded document in the CMP data sets, researchers need to check whether their findings are driven by influential observations based on suspect documents and check whether outliers can be explained in terms of data quality (Gemenis, 2012a). Moreover, researchers can enhance the robustness of their results by choosing the observations in the CMP data set that best fit the assumptions of their research project. For instance, if researchers are interested in making inferences on the assumption that the CMP data are based on election manifestos, they should proceed to exclude those coded proxy documents which were not produced with election campaigns in mind. Conversely, when the assumption that parties are unitary actors is essential for the argument, researchers can exclude the coded party leader speeches for parties that are known to be internally divided.

Regarding the problem of coding unreliability, there is little that third-party users can do to improve the use of the CMP data. As already mentioned, Mikhaylov *et al.* (2012, pp. 83–5) found no significant differences among coders with different levels of experience but reported substantial differences in reliability among different coding categories. Consequently, third-party users can consider using only the frequency categories for issues that are expected to have a small probability of being misclassified and combine the frequencies of issue categories in which classification 'seepage' is expected to occur during the coding process.

Regarding issues in the scaling of data, third-party users should consider using the L–R estimates from alternative scaling techniques to the 'standard' 'rile' provided in the CMP CD-ROMs. We now have a lot of published evidence where, too often to be ignored, the

'rile' scores in the CMP data sets do not provide valid and reliable estimates regarding parties' L–R positions. Moreover, the data for at least three alternative scaling techniques (Franzmann and Kaiser, 2006; Gabel and Huber, 2000; Lowe *et al.*, 2011) are easily accessible online. Given that the uncritical use of the CMP data can lead to faulty inferences, as Benoit *et al.* (2009, pp. 507–10) have shown, it is paramount for researchers to enhance the robustness of their results by performing sensitivity analyses using data from alternative scaling techniques and measures of L–R (e.g. mass or expert survey data).

Alternatively, researchers could use the CMP data for applications other than ideal-point estimation. Third-party users of the CMP data could avoid some of the aforementioned problems simply by estimating quantities other than parties' policy positions. In fact, the focus on parties' positions has obscured interesting applications in other areas of interest. To mention just a couple of instances, Kenneth Janda *et al.* (1995) have used the CMP data to study changes in party ideological profiles by comparing the frequency categories in adjacent elections, whereas Franzmann (2012) has proposed to apply the well-known Hirschmann/Herfindahl index to the CMP data in order to measure party policy concentration. This approach essentially implies that we need to find a replacement for the CMP's role in estimating the positions of political parties. To paraphrase Philip Schrodt (2010), who made a very similar point for international relations research, data collection should not be a monoculture. The idea is that, if all the proposed solutions for the identified problems are only partial and controversial on their own, resources need to be redirected from 'fixing problems' to 'building anew'.

As already noted, the literature that employs the CMP data to perform empirical tests in comparative politics theories is extensive and constantly growing. At the same time, however, the literature that points to pitfalls in the CMP approach and data is growing as well. Researchers and journal editors should not ignore the most prominent critical contributions summarised in this article. In this respect, the suggestions outlined in this concluding section are meant as a guide for good practice which could increase confidence in the published results employing CMP data.

(*Accepted*: 27 May 2012)

## About the Author

**Kostas Gemenis** is Assistant Professor of Research Methods at the Department of Public Administration of the University of Twente. His work has been published in journals including *Electoral Studies*, *Environmental Politics*, *Party Politics* and the *European Political Science Review*. He is currently involved in 'Preference Matcher' (http://www.preferencematcher.org), a consortium involving researchers who collaborate in developing voting advice applications and e-literacy tools designed to enhance voter education. Kostas Gemenis, Department of Public Administration, University of Twente, PO Box 217, 7500 AE, Enschede, the Netherlands; email: *k.gemenis@utwente.nl*

## Notes

A previous version of this article was presented at the 2011 EPOP Conference, University of Exeter, 9–11 September. I would like to thank the participants of the conference, the editor of *Political Studies* Cees van der Eijk, and two of the anonymous reviewers for their helpful comments. The usual disclaimer applies. Replication materials for the analyses presented in this article are available on the publisher's and author's websites.

1 The project's website can be accessed from: http://manifestoproject.wzb.eu/. Since the project is mostly known as the CMP, I will use this acronym throughout this article.

2 These observations can be identified in the CMP data sets using the 'progtype' variable (progtype = 3).

## References


Albright, J. J. (2008) 'Bayesian Estimates of Party Left–Right Scores', Society for Political Methodology Working Paper.

Bakker, R. (2007) *Re-measuring Left–Right: A Better Model for Extracting Left–Right Political Party Policy Preference Scores*. Unpublished PhD thesis, University of North Carolina-Chapel Hill.

Bakker, R. (2009) 'Re-measuring Left–Right: A Comparison of SEM and Bayesian Approaches for Extracting Latent Dimensions from Political Texts', *Electoral Studies*, 28 (3), 413–21.

Bara, J. (1987) 'Israel 1949–1981', in I. Budge, D. Robertson and D. J. Hearl (eds), *Ideology, Strategy and Party Change: Spatial Analyses of Post-war Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press, pp. 111–33.

Benoit, K. and Laver, M. (2006) *Party Policy in Modern Democracies*. London: Routledge.

Benoit, K. and Laver, M. (2007) 'Estimating Party Policy Positions: Comparing Expert Surveys and Hand-Coded Content Analysis', *Electoral Studies*, 26 (1), 90–107.

Benoit, K., Mikhaylov, S. and Laver, M. (2009) 'Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions', *American Journal of Political Science*, 53 (2), 495–513.

Benoit, K., Laver, M., Lowe, W. and Mikhaylov, S. (2012) 'How to Scale Coded Text Units without Bias: A Response to Gemenis', *Electoral Studies*, 31 (3), 605–8.

Braun, D., Mikhaylov, S. and Schmitt, H. (2010) 'Manifesto Study Documentation Advance Release', pre-release B, 22 July. Available from: http://www.piredeu.eu/ [Accessed 7 November 2012].

Budge, I. (1987) 'The Internal Analysis of Election Programmes', in I. Budge, D. Robertson and D. J. Hearl (eds), *Ideology, Strategy and Party Change: Spatial Analyses of Post-war Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Budge, I. (2001a) 'Theory and Measurement of Party Policy Positions', in I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds), *Mapping Policy Preferences: Estimates for pasties, Electors and Governments 1945–1998*. Oxford: Oxford University Press.

Budge, I. (2001b) 'Validating Party Policy Placements', *British Journal of Political Science*, 31 (1), 210–23.

Budge, I. and Bara, J. (2001) 'Introduction: Content Analysis and Political Texts', in I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford: Oxford University Press, pp. 1–16.

Budge, I. and Farlie, D. (1977) *Voting and Party Competition*. London: Wiley & Sons.

Budge, I. and Klingemann, H.-D. (2001) 'Finally! Comparative Overtime Mapping of Policy Movement', in I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford: Oxford University Press, pp. 19–50.

Budge, I. and Pennings, P. (2007a) 'Do They Work? Validating Computerised Word Frequency Estimates against Policy Series', *Electoral Studies*, 26 (1), 121–9.

Budge, I. and Pennings, P. (2007b) 'Missing the Message and Shooting the Messenger: Benoit and Laver's "Response" ', *Electoral Studies*, 26 (1), 136–41.

Budge, I., Robertson, D. and Hearl, D. J. (eds) (1987) *Ideology, Strategy and Party Change: Spatial Analyses of Post-war Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press.

Budge, I., Klingemann, H.-D., Volkens, A., Bara, J. and Tanenbaum, E. (eds) (2001) *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford: Oxford University Press.

Clarke, H. D., Sanders, D., Stewart, M. C. and Whiteley, P. (2004) *Political Choice in Britain*. Oxford: Oxford University Press.

Däubler, T., Benoit, K., Mikhaylov, S. and Laver, M. (2012) 'Natural Sentences as Valid Units for Coded Political Texts', *British Journal of Political Science*, 42 (4), 937–51.

Dinas, E. and Gemenis, K. (2010) 'Measuring Parties' Ideological Positions with Manifesto Data: A Critical Evaluation of the Competing Methods', *Party Politics*, 16 (4), 427–50.

Elff, M. (2012) 'Electoral Platforms, Political Positions and Their Evolution: Spatial Models of Coded Political Texts'. Paper presented at the MPSA Annual Conference, Chicago 12–15 April.

Franzmann, S. (2011) 'Competition, Contest, and Cooperation: The Analytic Framework of the Issue Market', *Journal of Theoretical Politics*, 23 (3), 317–43.

Franzmann, S. (2012) 'The Manifesto Project from the Perspective of Content Analysis: Logical and Substantial Inferences using MARPOR', prepared for A. Volkens *et al*. (eds), *Mapping Policy Preferences III*. Oxford: Oxford University Press.

Franzmann, S. and Kaiser, A. (2006) 'Locating Political Parties in Policy Space: A Reanalysis of Party Manifesto Data', *Party Politics*, 12 (2), 163–88.

Fuchs, D. and Klingemann, H.-D. (1990) 'The Left–Right Schema', in K. M. Jennings (ed.), *Continuities in Political Action*. Berlin: De Gruyter, pp. 203–34 (DOI: 10.1057/ap.2011.10).

Gabel, M. J. and Huber, J. D. (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.

Gemenis, K. (2012a) 'Proxy Documents as a Source of Measurement Error in the Comparative Manifestos Project', *Electoral Studies*, 31 (3), 594–604.

Gemenis, K. (2012b) 'Estimating Parties' Positions through Voting Advice Applications: Some Methodological Considerations', *Acta Politica*.

Gemenis, K., Katsanidou, A. and Vasilopoulou, S. (2012) 'The Politics of Anti-environmentalism: Positional Issue Framing by the European Radical Right'. Paper presented at the MPSA Annual Conference, Chicago 12–15 April.
Hans, S. and Hönnige, C. (2008) 'Noughts and Crosses: Challenges in Generating Political Positions from CMP-Data', Kaiserslautern Occasional Papers in Political Science No. 2.
Hansen, M. E. (2008) 'Back to the Archives? A Critique of the Danish Part of the Manifesto Dataset', *Scandinavian Political Studies*, 31 (2), 201–16.
Hayes, A. F. and Krippendorff, K. (2007) 'Answering the Call for a Standard Reliability Measure for Coding Data', *Communication Methods and Measures*, 1 (1), 77–89.
Hearl, D. J. (1987) 'Belgium 1946–1981', in I. Budge, D. Robertson and D. J. Hearl (eds), *Ideology, Strategy and Party Change: Spatial Analyses of Post-war Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press, pp. 230–53.
Hearl, D. J. (2001) 'Checking the Party Policy Estimates: Reliability', in I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford: Oxford University Press, pp. 111–25.
Hooghe, L., Bakker, R., Brigevich, A., de Vries, C., Edwards, E., Marks, G., Rovny, J., Steenbergen, M. and Vachudova, M. (2010) 'Reliability and Validity of the 2002 and 2006 Chapel Hill Expert Surveys on Party Positioning', *European Journal of Political Research*, 49 (5), 687–703.
Ivarsflaten, E. (2008) 'What Unites Right-Wing Populists in Western Europe? Re-examining Grievance Mobilization Models in Seven Successful Cases', *Comparative Political Studies*, 41 (1), 3–23.
Janda, K., Harmel, R., Edens, C. and Goff, P. (1995) 'Changes in Party Identity: Evidence from Party Manifestos', *Party Politics*, 1 (2), 171–96.
Kim, H.-M. and Fording, R. C. (1998) 'Voter Ideology in Western Democracies, 1946–1989', *European Journal of Political Research*, 33 (1), 73–97.
Klingemann, H.-D. (1995) 'Party Positions and Voter Orientations', in H.-D. Klingemann and D. Fuchs (eds), *Citizens and the State*. Oxford: Oxford University Press, pp. 183–206.
Klingemann, H.-D., Volkens, A., Bara, J., Budge, I. and McDonald, M. D. (2006) *Mapping Policy Preferences II: Estimates for Parties, Electors and Governments in Central and Eastern Europe, European Union and OECD, 1990–2003*. New York: Oxford University Press.
Krippendorff, K. (2004a) *Content Analysis: An Introduction to its Methodology*, second edition. Thousand Oaks CA: Sage.
Krippendorff, K. (2004b) 'Reliability in Content Analysis: Some Common Misconceptions', *Human Communication Research*, 30 (3), 411–33.
Laver, M. (2001a) 'On Mapping Policy Preferences using Manifesto Data', unpublished paper, Trinity College Dublin.
Laver, M. (2001b) 'Position and Salience in the Policies of Political Actors', in M. Laver (ed.), *Estimating the Policy Position of Political Actors*. London: Routledge, pp. 66–75.
Laver, M. and Budge, I. (1992) 'Measuring Policy Distances and Modelling Coalition Formation', in M. Laver and I. Budge (eds), *Party Policy and Government Coalitions*. Basingstoke: Macmillan, pp. 15–40.
Laver, M. and Garry, J. (2000) 'Estimating Policy Positions from Political Texts', *American Journal of Political Science*, 44 (3), 619–34.
Laver, M., Benoit, K. and Garry, J. (2003) 'Estimating the Policy Positions of Political Actors using Words as Data', *American Political Science Review*, 97 (2), 311–31.
Lin, L. I.-K. (1989) 'A Concordance Correlation Coefficient to Evaluate Reproducibility', *Biometrics*, 45 (1), 255–68.
Linhart, E. and Shikano, S. (2009) 'Ideological Signals of German Parties in a Multi-dimensional Space: An Estimation of Party Preferences using the CMP Data', *German Politics*, 18 (3), 301–22.
Lombard, M., Snyder-Duch, J. and Bracken, C. C. (2002) 'Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability', *Human Communication Research*, 28 (4), 587–604.
Lowe, W., Benoit, K., Mikhaylov, S. and Laver, M. (2011) 'Scaling Policy Positions from Hand-Coded Political Texts', *Legislative Studies Quarterly*, 36 (1), 123–55.
McDonald, M. D. and Mendes, S. M. (2001a) 'Checking the Party Policy Estimates: Convergent Validity', in I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford: Oxford University Press, pp. 127–41.
McDonald, M. D. and Mendes, S. M. (2001b) 'The Policy Space of Party Manifestos', in M. Laver (ed.), *Estimating the Policy Position of Political Actors*. London: Routledge, pp. 90–114.
Meyer, T. (2010) *Party Competition over Time: How Voters and Intraparty Structure Constrain Party Policy Shifts*. Unpublished PhD thesis, University of Mannheim.
Mikhaylov, S., Laver, M. and Benoit, K. (2012) 'Coder Reliability and Misclassification in the Human Coding of Party Manifestos', *Political Analysis*, 20 (1), 78–91.
Pardos-Prado, S. (2012) 'Valence beyond Consensus: Party Competence and Policy Dispersion from a Comparative Perspective', *Electoral Studies*, 31 (2), 342–52.
Pelizzo, R. (2003) 'Party Positions or Party Direction? An Analysis of Party Manifesto Data', *West European Politics*, 26 (2), 67–89.
Pennings, P. (2011) 'Assessing the "Gold Standard" of Party Policy Placements: Is Computerized Replication Possible?', *Electoral Studies*, 30 (3), 561–70.
Proksch, S.-O. and Slapin, J. B. (2009) 'How to Avoid Pitfalls in Statistical Analysis of Political Texts: The Case of Germany', *German Politics*, 18 (3), 323–44.

Proksch, S.-O., Slapin, J. B. and Thies, M. F. (2010) 'Party System Dynamics in Post-war Japan: A Quantitative Content Analysis of Electoral Pledges', *Electoral Studies*, 30 (1), 114–24.
Protsyk, O. and Garaz, S. (2011) 'Politicization of Ethnicity in Party Manifestos', *Party Politics* [online]. doi: 10.1177/1354068811398058.
Robertson, D. (1976) *A Theory of Party Competition*. London: John Wiley.
Robertson, D. (1987) 'Britain, Australia, New Zealand and the United States, 1946–1981: An Initial Comparative Analysis', in I. Budge, D. Robertson and D. J. Hearl (eds), *Ideology, Strategy and Party Change: Spatial Analyses of Post-war Election Programmes in 19 Democracies*. Cambridge: Cambridge University Press, pp. 39–73.
Schrodt, P. A. (2010) 'Seven Deadly Sins of Contemporary Quantitative Political Analysis', Society for Political Methodology Working Paper.
Van der Brug, W. (2001) 'Analysing Party Dynamics by Taking Partially Overlapping Snapshots', in M. Laver (ed.), *Estimating the Policy Position of Political Actors*. London: Routledge, pp. 115–32.
Veen, T. (2011) 'Positions and Salience in European Union Politics: Estimation and Validation of a New Dataset', *European Union Politics*, 12 (2), 267–88.
Volkens, A. (2001) 'Quantifying the Election Programmes: Coding Procedures and Controls', in I. Budge, H.-D. Klingemann, A. Volkens, J. Bara and E. Tanenbaum (eds), *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945–1998*. Oxford: Oxford University Press, pp. 93–109.
Volkens, A., Bara, J. and Budge, I. (2009) 'Data Quality in Content Analysis: The Case of the Comparative Manifestos Project', *Historical Social Research*, 34 (1), 234–51.
Weale, A., Pridham, G., Cini, M., Konstadakopulos, D., Porter, M. and Flynn, B. (2000) *Environmental Governance in Europe: An Ever Closer Ecological Union?* Oxford: Oxford University Press.

## Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

**Table S1:** Replication dataset
**Table S2:** Replication .do file