Political Analysis Advance Access published December 11, 2011

doi:10.1093/pan/mpr047

Coder Reliability and Misclassification in the Human Coding of Party Manifestos

Slava Mikhaylov
Department of Political Science, University College London
e-mail: v.mikhaylov@ucl.ac.uk (corresponding author)

Michael Laver
Department of Politics, New York University
michael.laver@nyu.edu

Kenneth R. Benoit
Methodology Institute, London School of Economics and Political Science
kbenoit@lse.ac.uk

The Comparative Manifesto Project (CMP) provides the only time series of estimated party policy positions in political science and has been extensively used in a wide variety of applications. Recent work (e.g., Benoit, Laver, and Mikhaylov 2009; Klingemann et al. 2006) focuses on nonsystematic sources of error in these estimates that arise from the text generation process. Our concern here, by contrast, is with error that arises during the text coding process since nearly all manifestos are coded only once by a single coder. First, we discuss reliability and misclassification in the context of hand-coded content analysis methods. Second, we report results of a coding experiment that used trained human coders to code sample manifestos provided by the CMP, allowing us to estimate the reliability of both coders and coding categories. Third, we compare our test codings to the published CMP "gold standard" codings of the test documents to assess accuracy and produce empirical estimates of a misclassification matrix for each coding category. Finally, we demonstrate the effect of coding misclassification on the CMP's most widely used index, its left–right scale. Our findings indicate that misclassification is a serious and systemic problem with the current CMP data set and coding process, suggesting the CMP scheme should be significantly simplified to address reliability issues.

1 Reliability Versus Validity in Content Analysis

The systematic analysis of political text offers a source of plentiful and cheap data often unavailable to researchers by other means. To become useful data, however, text must first be converted into usable quantities according to a systematic scheme. The most common approach consists of two steps (Krippendorff 2004, 219). First, texts are parsed into units of analysis relevant to the research question, such as words, sentences, or quasi-sentences. Following this first step of *unitization*, each unit is coded by assigning it to a category from some coding scheme that is a core component of the text analysis project.

Authors' note: Previously presented at the 66th MPSA Annual National Conference, Palmer House Hilton Hotel and Towers, April 3–6, 2008. Our heartfelt thanks goes out to all the volunteer test coders who completed the online coder tests used in the research for this paper. We also thank Andrea Volkens for cooperation and assistance with details of the coding process, and Jouni Kuha, Michael McDonald, Michael Peress, Sven-Oliver Proksch, Jonathan Slapin for useful comments. For replication data and code, see Mikhaylov, Laver, and Benoit (2011). Supplementary materials for this article are available on the *Political Analysis* Web site.

Text coding can be conducted by machines—for example, when words or phrases are coded according to some predefined coding dictionary—or by humans, who are in essence asked to read text units for "meaning" and then categorize these according to some scheme. Whenever humans read texts for meaning, the content analysis procedure faces potential problems with *reliability* since different human readers may attach different meaning to the same text. Although automated text coding by machines will likely become the default method in this field, even automatic methods will never dispense entirely with the need for human coders since "supervised" and "semisupervised" machine coding methods typically depend on "reference" or "training" texts with policy positions for which there are independent expert estimates (e.g., Laver, Benoit, and Garry 2003; Hopkins and King 2010).[1] Ultimately, therefore, there will always be a fundamental need to validate and/or calibrate results generated by automated techniques against systematically collected and characterized input from human experts.

Our aim here is to evaluate the reliability of this human input process using coding experiments. To assess a typical and well-known project that relies on the human coding of political texts, we examine the dominant professional source of human expert text codings of party policy positions in political analysis: the long-standing Comparative Manifestos Project (CMP) (Budge et al. 2001; Klingemann et al. 2006). CMP data are widely used by third party researchers to measure policy positions of political parties on an election-by-election basis, indeed they are profession's primary source of such data. We know axiomatically that these data have problems of validity, reliability, and bias, just as all data do. Because of the critical qualitative element of human coding at the core of the CMP, however, the issue of whether this step can be performed reliably is one of key concern. Reliability can be tested using experiments, although astoundingly no systematic experiments on the widely used CMP coding process have yet been reported. In what follows, we set out a framework for reliability and misclassification in categorical content analysis and apply this framework to the CMP coding scheme. Using experiments with multiple coders, we test the reliability of the CMP scheme and estimate specific misclassification rates. Our findings have direct implications for future use of the CMP scheme but also illustrate how reliability can be tested for any research design that incorporates human coding.

2 The CMP Coding Scheme and Sources of Disagreement

To generate the data used to scale policy positions for particular parties for specific elections, the CMP employs trained human coders to allocate every sentence unit in the party's manifesto into one, and only one, of 57 policy coding categories (one of which is "uncoded").[2] The first CMP coding category, for example, is "101: Foreign special relationships: positive." Having counted text units allocated to each category, the CMP then uses its theoretical "saliency" model of party competition to inform a measurement model that defines the relative salience for the party of the policy area defined by each category as the percentage of all text units allocated to that category.

We have dealt elsewhere (Benoit, Laver, and Mikhaylov 2009) with nonsystematic measurement error in CMP data that arises from stochastic features of the *text generation* process. Here, we focus on error arising in CMP data from stochastic and systematic features of the *text coding* process—specifically, the potential failure for different coders to reliably apply the same codes to the same text, including the possibility that coders will make systematic errors in applying codes to text, a coding error to which we refer in general terms as *misclassification.*

2.1 *Sources of Intercoder Disagreement*

Data built on human-coded text units are fundamentally susceptible to coding error because, in their essence, they derive from subjective judgments made by human coders. Coding error arises because

[1] And even "unsupervised" methods depend on scaling intertext distance matrices derived from the "bag of words" in each text, but a posteriori interpretation of the resulting scales typically depends upon independent expert knowledge of the positions of particular known "marker" texts (Slapin and Proksch 2008).

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

different human coders at the same time, or even the same human coder at different times, are likely to code the same text in different ways. This process may be unbiased in the sense that we can think of an unobservable "true and certain" value of the quantity being measured, with each human text coding being a noisy realization of this. Assuming unbiased coding, we can take the mean of the noisy realizations as an estimate of the unobservable latent quantity and the variation in these observations as a measure of the uncertainty of this estimate.

The CMP data, however, are for the most part generated by party manifestos coded once, and once only, by a single human coder. There is no variation in noisy realizations of the unobservable underlying quantity and thus no estimate can be formed of the uncertainty of CMP estimates arising from coding errors. In a nutshell, we have no way of knowing whether subsequent codings of the same manifesto would be exactly the same as or completely different from the recorded coding that goes into the CMP data set. We are confident, however, that a series of independent codings of the same manifesto would all differ at least somewhat from each other. Indeed, if someone reported that 1000 highly trained coders had each coded 10,000 manifesto text units using the CMP's 57 category scheme and that every single coder had coded every single text unit in precisely the same way, then our overwhelming suspicion would be that the data had been faked.

CMP coders often report difficulties determining which of the coding categories to use for a particular text unit. Important sources of coder error are thus the ambiguities and overlap that exist in the way coding categories are defined. Consider the distinction between the following CMP categories:

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

There is of course a difference between these definitions but it is easy to imagine text for which the coder's decision would be a knife-edge judgment, one made in different ways by different coders or even by the same coder at different times. In contrast "501: Environmental protection" is essentially the only CMP coding category making explicit reference to the environment, so there is nowhere else in the scheme to allocate text units referring to the environment (a decision that, incidentally, renders antienvironmentalist statements uncodable by the CMP). Any text coding scheme must be viewed as a whole, taking into account overlaps and the sharpness of boundaries between categories as well as the definitions of each category on a stand-alone basis. However, we do expect some CMP coding categories to be more "reliable" (different coders tend to code the same text unit into the category in question) than others (different coders do not all use the category in question for the same text unit.)[3] As we shall see, this is very much what we find in our coding experiments.

2.2 *From Categories to Scales*

One response to overlapping or vague boundaries between text coding categories is to combine these to produce a more reliable aggregate category. In addition, what amounts to the 56-dimensional policy space measured by the CMP manifesto codings is cumbersome to use as an operationalization of specific models of party competition. In practice, therefore, most third-party users of CMP data are looking for something much simpler, and many are looking for party positions on a single left–right scale.

This is why the CMP is best known for its left–right "Rile" scale, which it calls its "crowning achievement" (Budge et al. 2001, 19). This is a simple additive index aggregating 13 coding categories seen as being on the "left", 13 seen as being on the "right", and subtracting the percentage of aggregated left categories from those of the right. The theoretical range of this scale is thus [−100, 100], although in practice nearly all Rile scores span the scale's middle range of [−50, 50]. The aggregate Rile scale is potentially

[3]In practice, the full 56-category coding scheme is never deployed on any one manifesto and the norm is that far fewer than the full set of categories are used in the coding of a typical manifesto. Analysis of the CMP-provided data set shows that the typical manifesto coding uses only 25 categories, less than half of those available. Coding category usage ranges from startlingly monothemed manifestos such as the 1951 Australian National Party manifesto which consisted of 42 text units all assigned to a single category ("703: Farmers Positive") to a maximum of 51 different categories used to code the 365 text units found in the 1950 British Conservative Party manifesto.

more reliable than any single coding category since it is likely that most of the stochastic variation in text coding will result from different coders allocating the same text unit to different categories on the left or the right. From the perspective of the left–right scale that most third-party users are interested in, such coding "errors" are in effect self-cancelling.[4] In our tests below, we critically examine this claim.

2.3 *Strategies to Maximize Reliability*

Previous work investigating the reliability of the CMP scales has focused on different and quite specific aspects of the issue. The CMP's approach to coding reliability is to focus on procedures of coding used in data production (Klingemann et al. 2006, 107). Possible problems of coding error that we discuss below are approached by emphasizing rigorous training—"setting and enforcing central standards on coders"—by setting out to train all CMP coders to code the same two manifestos in the same way as a CMP gold standard coding that is taken to reflect a "certain truth" about the policy positions expressed in those manifestos used as training documents. The CMP also sought to improve reliability by revising its coding instructions after the project had begun and also by constant communication and interaction with the supervisor in Berlin, including the possibility of a coder seeking cues from the project headquarters when a text unit's category seems unclear (Volkens 2001b, 94; see also Volkens 2001a, 37–40).

Other investigations of reliability have specifically targeted possible error in the aggregated indexes, namely Rile. McDonald and Mendes (2001) and Klingemann et al. (2006, chapter 5) focus on the issue of measurement error in the Rile scale as an approach to assessing reliability. Exploiting the panel structure of the data set and using the Heise measurement model (Heise 1969), the authors claim to be able to sift out measurement error from real change. From the results, and making strong theoretical assumptions about the latent reliability structure, they conclude that Rile is close to being perfect (Klingemann et al. 2006, 103). Such tests focus on very different issues from those of stability and reproducibility faced here, however, where our primary concern is whether coders can reliably implement the CMP coding instructions without serious misclassification errors. Only a direct comparison of different coders on the same text, as well as to a "gold standard", offers the possibility of a true test of coding reliability and the potential for systematic misclassification.

3 A Framework for Stochastic Misclassification of Text Units

Our discussion here follows the framework of Kuha, Skinner, and Palmgren (2000) and Bross (1954). In formal terms, let the true categories of each text unit i be represented by A_i, whose values are well defined and fixed but classified with error as A_i^*. Misclassification occurs through a stochastic process

$$\Pr(A_i^* = j | A_i = k) = \theta_{jk}, \tag{1}$$

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

If a coding scheme could be applied to text units perfectly, then Θ would consist of an $m \times m$ identity matrix. To the extent that there are off-diagonals in Θ, however, then misclassification will produce biased estimates of the true proportions of A_i, depending on the degree of systematic errors present in misclassification as well as purely stochastic errors applied to unequal A_i proportions. Through experiments and with comparison to a gold standard, we can estimate this degree of bias. Following Kuha and Skinner (1997), for a text, let N_j^A be the number of text units for which $A_i = j$, and let $P_j^A = N_j^A / N$, where $N = \sum N^A$ is the total number of text units. Our objective is to estimate the vector $\mathbf{P}^A = (P_1^A, \ldots, P_m^A)'$ of proportions of each category of manifesto code from the coding scheme for our given text—in other

[4]This problem, which the CMP has termed "coding seepage" (Klingemann et al. 2006, 112), is thought to mainly take place between categories within the same "Rile" divisions. Because the components of the Rile index "combine closely related categories, the coding errors created by ambiguity between these are eliminated. The overall measures are thus more stable and reliable than any one of their components" (Klingemann et al. 2006, 115).

words, the CMP's "per" variables. When Θ contains nonzero off-diagonal elements, we will observe only the misclassified proportions P^{A*} for which

$$\mathrm{E}(\mathbf{P}^{A*}) = \Theta\mathbf{P}^{A}. \tag{2}$$

The bias from misclassification will then be expressible as

$$\mathrm{Bias}(\mathbf{P}^{A*}) = (\Theta - \mathbf{I})\mathbf{P}^{A}, \tag{3}$$

where $\mathbf{I}$ is the $m \times m$ identity matrix. Our task in assessing misclassification and the unreliability of the coding procedure that follows, therefore, is to obtain estimates of the misclassification matrix Θ. To the extent that this misclassification matrix differs from identity, then the observed (and misclassified) proportions of coding categories will be unreliable and generally biased estimates of the true proportions of the textual content.

Misclassification is also frequently expressed in terms of the *sensitivity* of a test (as well as the related concept of *specificity*) (see, e.g., Rogan and Gladen 1978; King and Lu 2008). Sensitivity refers to $\Pr(A_i^* = k|A_i = k)$ or in our context, for example, the ability of the coding process to classify a given text unit to its correct coding category. In the three-part Rile classification, for example, sensitivity is the probability that a sentence is coded as left when it really is left or is coded as right when it really is right or coded as "neither" when it really is neither left nor right. Sensitivity can also be expressed as the true positive rate, or conversely, in terms of the *false negative rate*. In the language of hypothesis testing, the false negative rate β represents the probability of a Type II error—here, the probability of coding a sentence into a a wrong category $\sim k$ when it really belongs to a category k.

Our testing framework allows us to estimate specificity directly, and this forms our focus in the tests that follow.

4 An Experiment to Assess Coder Agreement

4.1 *Methods and Data*

To evaluate misclassification and reliability in the CMP coding procedure, we performed a simple experiment: We asked multiple coders to code two political manifesto excerpts in order to observe intercoder reliability. Both texts were taken from the "Manifesto Coding Instructions" provided in Appendix II to Klingemann et al. (2006), where these two texts are provided fully unitized and coded to serve as examples alongside the coder instructions. Using these two texts held several key advantages. First, each text had already been "officially" parsed into quasi-sentences by the CMP, meaning that we could take the unitization step as given and focus in the experiment only on the assignment of codes to each quasi-sentence. Second, because each text was also officially coded by the CMP, the CMP codings serve as a gold standard for comparing to tester codings. Finally, since these two texts had been chosen for their clarity and codeability to be instructional examples, they also made good texts for comparing coder agreement in our experiments.

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

Our test was set up on a dedicated web page containing digitized versions of sample texts, already divided into quasi-sentences. Each page also contained detailed instructions adapted directly from Manifesto Coding Instructions in Appendix II to Klingemann et al. (2006). Coders were asked to select for each text unit an appropriate category from a scroll-down menu. We also collected some minimal information on coder identifiers and previous experience in coding manifestos. Only completed manifestos could be submitted into the system. To ensure we recruited experienced CMP coders, we invited the majority of

Table 1 Coder reliability test results reported by CMP

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

Note. Sources are Klingemann et al. (2006), Volkens (2001a, 2007); figures reported are Pearson's *R* for the aggregate percentage measured across 56 coding categories for the test document found in *MPP2*, 181–186.

trained CMP coders to participate[5] as well as selection of usual suspects: faculty and postgraduates at several European and North American universities. This yielded a list of 172 names with active email addresses who were randomly assigned to one of the two test documents.

Our response set consisted of 39 coders, but some of these results were discarded. To be as fair as possible to the CMP, we discarded the bottom fourth of test coders in terms of their reliability while dropping none from the top. Overall, the New Zealand manifesto was completed by 12 coders and the U.K. manifesto by 17. The coders whose results are reported here had a range of prior experience with coding manifestos using the CMP scheme. Although we do not focus on the relationship between coder characteristics and reliability here, it is worth noting that we found no evidence in our experiments that experienced coders performed more reliably than those with less experience.

4.2 *Methods of Assessing Agreement*

The only previous reported analysis of CMP coding reliability relied on correlating the percentages coded into each category by a given trainee with percentages coded into the same categories in the CMP gold standard coding of a test manifesto. Depending on which test we are talking about, reported correlations range from 0.70 to 0.80.[6] These reported results are collected in Table 1.

There is a clear distinction, however, between measuring *agreement* and measuring *association*. Strong association is required for strong agreement, but the reverse is not true (Agresti 1996, 243). The association measure reported by the CMP is the Pearson product–moment correlation that measures the degree of *linear trend* between two (at least) ordinal variables: the degree to which values of one variable predict values of the other variable. Measures of agreement, on the other hand, gauge the extent to which one variable equals the other. If a coder consistently miscategorizes quasi-sentence of a particular type, then association with the gold standard will be strong even though the strength of agreement is poor. Moreover, the Pearson product–moment correlations are not applicable for nominal-level data, which is the case in the analysis of (mis)coding of individual text units. For these reasons, correlations should be avoided since "in content analysis their use is seriously misleading" (Krippendorff 2004, 245).

Another problem with the CMPs coder reliability data concerns the issue of zero-category inflation. As discussed earlier, for any given manifesto, only a small subset of the available categories tends to be used. The test manifesto used by the CMP to assess reliability is no exception, and since the correlation vectors from the CMPs reliability are indexed by category, this means a majority of the elements in the

[5]Andrea Volkens kindly provided us with a list of names of 84 CMP coders of which 60% were matched with e-mail addresses. We also used publicly available e-mail addresses of coders trained by the CMP for a separate *Euromanifestos Project* (see Wüst and Volkens 2003).

[6]For 23 coders that were trained from the the second version of coding manual, their average correlation with the gold standard was reported to be 0.83. Of these coders, 14 were new hires taking the test for the first time. Their average correlation with the master copy is 0.82. Nine coders on the second contract took the test again with results for this group going up from 0.70 in the first round to 0.85 in the second round (Volkens 2007, 118). Klingemann et al. (2006, 107) report that coders on another contract retaking the test showed an average correlation coefficient of 0.88.

correlation vectors will have zeroes. The effect is to register high correlations based not on how well coders agree on applicable categories, but how well they agree on categories that clearly do not apply (such clear agreement on the absence of any EU-category quasi-sentences in the 1966 New Zealand training document).

To gauge intercoder agreement, we reported Fleiss's κ as our measure of intercoder agreement, one of the most widely used methods of statistical analysis of agreement for categorical variables (Roberts 2008, 811).[7] The κ coefficient ranges from 0 (perfect disagreement) to 1.0 (perfect agreement) and takes into account the fact that some agreement may occur purely by chance. The CMP group prefers to focus on reliability of composite indicators on the basis of their performance within the data set (Klingemann et al. 2006, 107) rather that on individual agreement. Although it has been declared that "the data-set as a whole is reliable" (Klingemann et al. 2006, 108), we believe that reliability can only be assessed by data that are additional to the data whose reliability is in question (Krippendorff 2004, 212). In the case of the CMP, this means analyzing reliability data obtained through duplication of coding exercise by several independent coders.

5 Results of the Coding Experiment

5.1 *Intercoder Agreement*

Our tests of intercoder agreement target the basic definition of reliability offered by Hayes and Krippendorff (2007, 78), wherein reliability "amounts to evaluating whether a coding instrument, serving as common instructions to different observers of the same set of phenomena, yields the same data within a tolerable margin of error. The key to reliability is the agreement among independent observers." Perfect reliability is never to be expected, but widely agreed guidelines for interpreting our primary reliability measure κ hold 0.80 to be the threshold above which a research procedure is considered to have an acceptable reliability. In the context of content analysis, Krippendorff (2004, 241) suggests not to rely on variables with reliabilities below $\kappa = 0.80$ and to consider variables with reliabilities between $\kappa = 0.667$ and $\kappa = 0.80$ only for drawing tentative conclusions.[8]

Table 2 summarizes the intercoder agreement from our results.[9] For each manifesto plus the combined results, the first column reports κ for all coders by category for agreement over all 56 policy categories (plus uncoded). To test whether individual disagreements cancel each other out in aggregation—as asserted by the CMP—we also compared the "Rile category" assigned by each coder to the quasi-sentences, reported in the second column ("By RILE") of Table 2. By this view, two coders assigning "403" and "404" to the same quasi-sentence would be viewed in perfect agreement since both these categories are classified as left in the Rile scale. Finally, for the categories that the CMPs master coding identified as being present in the test manifestos, we are also able to report individual κ statistics for the reliability of each category. These figures are shown in the bottom part of Table 2, indicating how well different coders could agree on quasi-sentences being designated as specific categories by category.

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

[7] The kappa coefficient is measured as $\kappa = \frac{p_o - p_e}{1 - p_e}$, where p_o is the overall proportion of observed agreement and p_e is the overall proportion of agreement expected by chance (Fleiss, Levin, and Paik 2003, 605). It ranges from 0 to 1.0. The kappa coefficient was proposed by Cohen (1960) and extended to multiple raters by Fleiss (1971).

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

[9] For replication data and code, see Mikhaylov, Laver, and Benoit (2011).

Table 2 Reliability results from coder tests

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

Note. The (L) or (R) designates whether a CMP category was part of the "Rile" left or right definition, respectively.

From the tests reported by individual category, several exceptionally unreliable categories stand out. From the left side of the Rile scale, "202: Democracy Positive" is extremely poor, with $\kappa = 0.18$, as are "701: Labour Groups: Positive" and "Economic Planning: Positive." On the right, "605: Law and Order: Positive" and especially "305: Political Authority: Positive" are flagged by our experiment as being extremely unreliable. In general, categories identifying broad policy objectives such as "economic goals" seem to be very highly prone to intercoder disagreement when it comes to assigning them to specific text units.

5.2 *Coder Agreement with the Master*

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

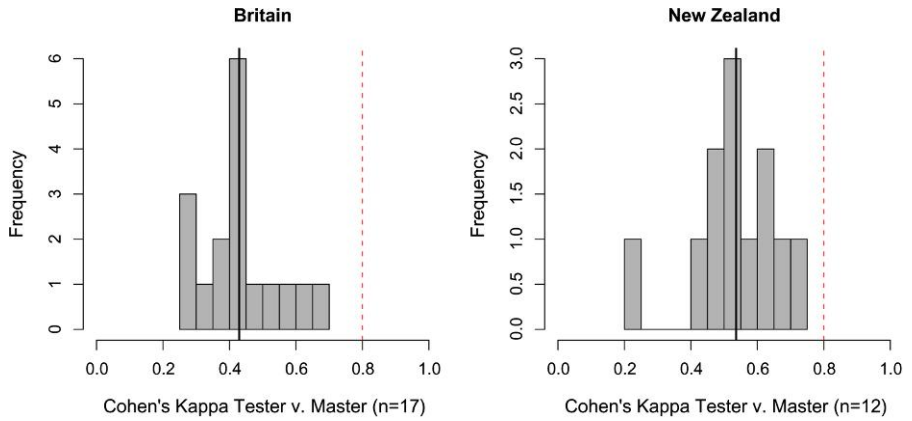

Fig. 1 Summary of coder reliabilities compared to Master, Cohen's κ. The dashed line indicates the conventional lower bound as to what is considered reliable in interpretations of Cohen's κ. The solid lines are the median value of κ from all the coders completing the tests.

The results of our tests are not encouraging. For the British manifesto test, the New Zealand manifesto test, and combined tests, respectively, the median κ test coders' agreement with the master were 0.43, 0.54, and 0.46, respectively. The best coder agreed 0.74 with the master and the worst 0.22. The full results are portrayed in Fig. 1. This histogram shows the frequency of different levels of κ for coder-master agreement from the 17 and 12 coders for the British and New Zealand texts, respectively. The solid line indicates the median results (0.42 and 0.54) from each test. For comparison with the conventional minimum level of acceptable reliability, we have also plotted a dashed line indicating the conventional 0.80 cutoff for acceptable reliability. As can be clearly seen, the main density of the distribution of results for individual coders was well below standard levels of reliability on both test documents.

5.3 *Misclassification*

Comparing the different coders' categorizations of the same text units not only allows us to estimate reliability but also allows us to characterize precisely the nature of this misclassification. Using the master codings as an external validation sample, we are able to determine for each "true" category, what the probabilities were that test coders would assign a text unit to the correct categories versus incorrect categories. In the earlier language of our framework for misclassification, we are able to use the empirical 57×57 matrix of true versus observed codings to estimate the misclassification matrix Θ_{jk}. By equation (3), we know that the size of the off diagonals (or $\hat{\Theta} - \mathbf{I}$) will estimate the difference between the true categories A_i and the observed categories A_i^*.

In order to make the misclassification matrix manageable, we have reduced the focus to the probability that individual categories will be misclassified in terms of the three-fold Rile classification. Looking at misclassification in this way tests the CMP assertion that errors in classification will be "self-cancelling" and also focuses attention on important errors, such as whether a category that is really left will be classified as one which is considered right in the CMPs Rile scale and vice versa. Because the Rile index—as are all other quantities in the CMP data set—are considered as proportions of all text units, we also consider misclassifications into the other category that is neither left nor right.[10]

Table 3 provides the most reduced summary of this misclassification according to a 3×3 table. The coders from our two tests provided a total of 1668 text unit classifications, which we could identify from the CMP's master coding as belonging to a left, right, or neither Rile category. Comparing these to the Rile categories of the coding category that our testers identified, we see significant frequencies in the

[10]Full misclassification probabilities are reported in supplementary materials on the *Political Analysis* Web site for each CMP coding category.

Table 3 Misclassification matrix for true versus observed Rile

		True Rile category			
		Left	*None*	*Right*	*Total*
	Left	430	188	100	718
		0.59	0.19	0.11	
Coded	None	254	712	193	1159
Rile		0.35	**0.70**	0.20	
	Right	41	115	650	806
		0.06	0.11	**0.69**	
	Total	725	1015	943	1668
	False negative rate	0.41	0.30	0.31	
	False positive rate	0.15	0.27	0.09	

Note. The top figure in each cell is the raw count; the bottom figure is the column proportion. The figures are empirically computed from combined British and New Zealand manifesto tests. The false negative rate is 1—sensitivity, whereas the false positive rate is 1—specificity.

off-diagonal cells. Left text units in particular were prone to misclassification, as 0.35 or 35% of the time, these were assigned a category that was not in the Rile scale. Conversely, about 19% of the text units that were not in a category found in the Rile scheme were classifies instead as left. Overall, the highest diagonal proportion—equivalent to the sensitivity or true positive rate defined previously—was just 0.70, indicating that 30% or more of the text units were classified into a wrong Rile category. Put another way, the probability of a "false negative" assignment of the other category is $\beta = (1 - 0.70) = 0.30$.

The results across the board are very discouraging.[11]. Some categories in the test had abysmally high false negative rates, such as 0.82 for 404: Economic Planning Positive, and 0.56 for 305: Political Authority Positive. Coders were also extremely unlikely to declare a text unit uncoded when according to the gold standard, it was in fact uncoded ($\beta = 0.55$). But even better performing categories typically failed to reach levels at which by most accepted standards we would be willing to accept the risk of false negatives: Only three categories from those tested reached levels of $\beta \leqslant 0.20$. The conclusion from these tests is quite clear: Even the better group of coders from our tests, including those trained and retrained by the CMP itself, are unable to apply the coding instructions to the training texts without a degree of misclassification that would be considered unacceptable by any conventional standard.

Misclassification probabilities for each CMP category used in the test documents, as well as the tripartite Rile left, right, and neither categories can be fully portrayed using a ternary plot. Figure 2 plots each category according to its probability of (mis)classification into the three-fold Rile set of left, right, or other. The categories that are truly left are plotted by their numeric category identifiers in normal typeface. Those that are truly right are in bold type and those that are neither are in italics. In addition, the mean misclassification probabilities for each of the three categories are shown as labelled points with a circle (these correspond to the proportions in Table 3.)[12] This method clearly singles out, visually, the worst categories from the standpoint of misclassification. If no misclassification existed, then all categories of the same "right–left" designation would be clustered in the corners of the triangle, which as can be clearly seen does not happen. Some categories almost equally split between two of the Rile categories, such as "truly left" categories 701 and 202, which are are almost equally likely to be coded as other, although these were almost never miscoded as right. Yet, other categories suffer from even more severe misclassification in particular categories 305 (truly right) and 404 (truly left). Located toward the middle of the triangle, these categories are not only severely prone to misclassification but also their misclassification occurs to either

[11]Full sensitivities on a category-by-category basis are available in supplementary materials for this article from the *Political Analysis* Web site

[12]Locating plot coordinates on a ternary plot begins with moving from the corner marked "0" toward the corner marked "1," and using the ruled lines at 60 degrees left to read the value for that side. For category 402, for instance, the probability of it being coded left is 0.2. Reading from the bottom side, the probability of coding category 402 as other is just under 0.10. Finally, reading from the right side, the probability of 403 being coded as right is just above 0.7.

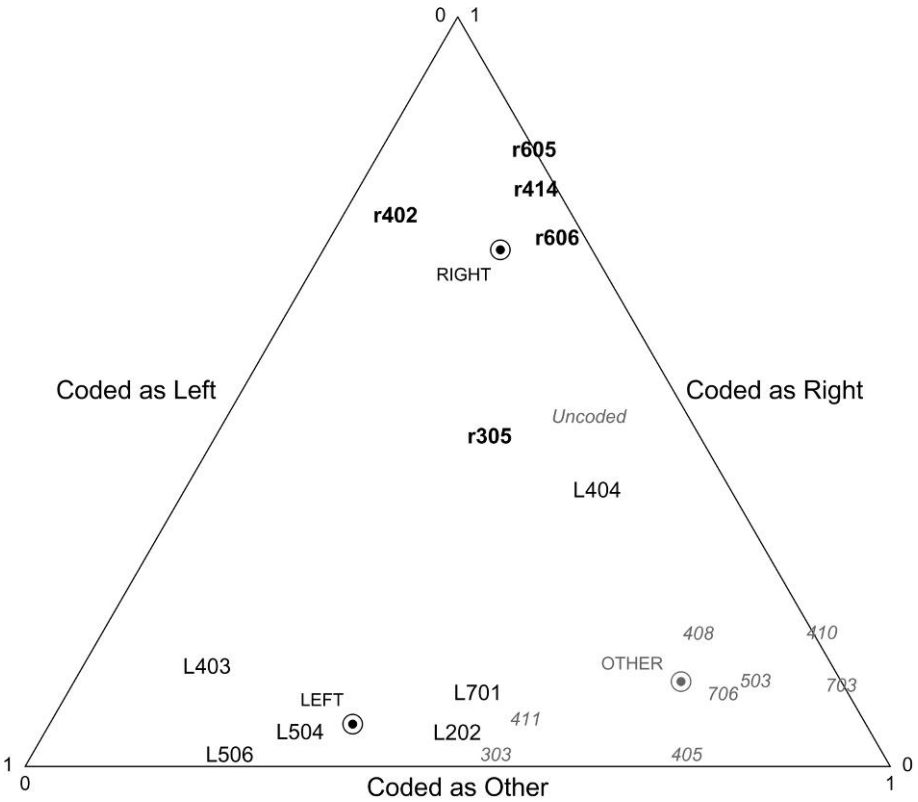

Fig. 2 Misclassification into left, right, or other by coding category, from experiments. The solid circles in hollow points represent the misclassification for the 3 × 3 left–right–other misclassification matrix. Each plotted number identifies a CMP category from Table 2, with its position indicating the probability of this category being coded as a left (prefixed with "L" on the plot), right (prefixed with "r" on the plot), or other category. If no misclassification existed, all plotted symbols and numeric labels would cluster together into their respective corners, which clearly does not happen.

of the two "wrong" right–left categories. Taken as a whole, these results are compelling evidence against the notion that coding mistakes tend to wash out when aggregated into left–right (or Rile) categories.

6 Demonstrating the Effects of Misclassification

We know from just the reduced 3 × 3 Rile misclassification matrix (estimated in Table 3) that the probabilities of misclassification into the wrong overall left–right categories are quite high. The question for practical purposes is: just how badly will this affect our resulting estimates?

To answer this question, we use simulation of the type of misclassification identified in our results above. By simulating the effect of stochastic misclassification on a range of Rile values at different levels of reliability, we can assess the degree of error, both systematic and nonsystematic, that are likely to be present in the CMP's reported Rile estimates. From the combined CMP data set, we know that the population proportions of the Rile left, right, and neither text units are roughly 0.25, 0.25, and 0.50, respectively. Our range of Rile therefore fixes the other category at 0.50 and lets the other frequencies vary so that we can observe Rile values from −50 to +50, once again a range taken from the empirical range in the combined CMP data set.[13]

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

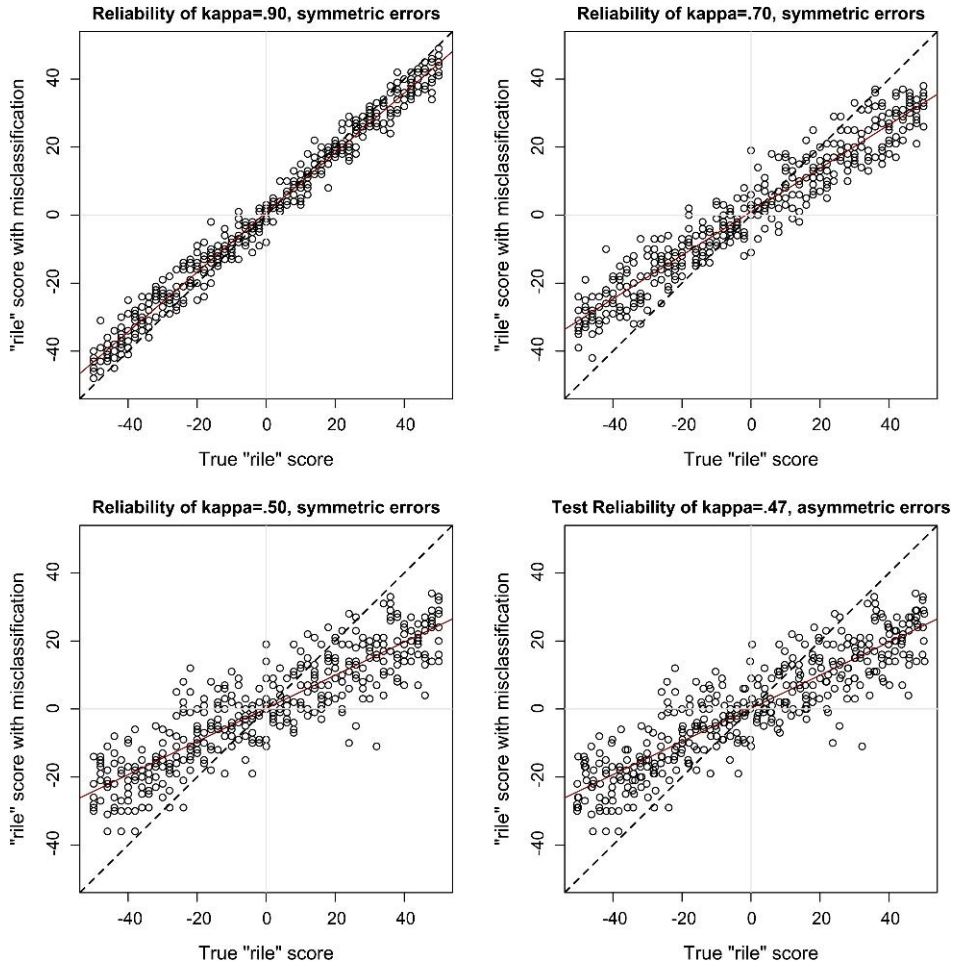

Fig. 3 Simulated misclassification at different levels of κ. The misclassification matrix Θ_{ji} is simulated from a manifesto with 50% uncoded content for different levels of κ, except for the last panel, which uses $\hat{\Theta}_{ji}$ estimated from the coding experiments. Misclassification is simulated eight times for each even-numbered true "Rile" score from -50 to 50.

The results of simulated misclassification are shown in Fig. 3. Here, we have manually manipulated the misclassification matrix to be symmetric and to produce reliabilities of (reading from top left to right) $\kappa = 0.90$, 0.70, and 0.50. The last panel (lower right) shows the effect of simulating error using the asymmetrical misclassification probabilities from Table 3 and having a median reliability of 0.47. For each panel, a faint cross-hair indicates the origin, and a dashed line shows the identity point at which $A_i^* = A_i$.

Two patterns clearly emerge from our simulation of misclassification. First, even at relatively high levels of reliability, misclassification adds significant noise to the resulting Rile estimates, meaning that any individual realization of the Rile index is likely to contain a significant degree of random error. Because Rile is most commonly used as an explanatory variable in political science models—in fact, this is the single most common usage of the CMP data set by far—this means that such models are likely to have biased estimates (for a fuller discussion see Benoit, Laver, and Mikhaylov 2009). Second, all the results tilt the observed values away from the identity line, making it flatter, and causing a centrist bias in the estimated Rile values even when the misclassification matrix is strictly symmetric. The reason is quite general: The more the true value consists of any single category, the greater the tendency of misclassification to dilute this category. (At the extreme of being, for instance, pure left, any misclassification can only move the estimate away from this extreme.) At the levels of reliability indicated by our tests—call it 0.50—this bias is quite severe, cutting the estimate of a "true" Rile value of -50 or 50 almost in half. The

effect on estimates when Rile is used as an explanatory variables is to compact the range of the variable further afflicting regression coefficients with attenuation bias. In the last plot, with asymmetrical error, we have used the actual misclassification matrix to simulate the error, leading to a shift to the right in the coded texts of between 20 and 10 points. This occurs because the misclassification tends to overclassify texts as right, leading to a systematic bias toward the right as well as to the general attenuation bias caused simply by unreliable human coding.

7 Conclusions and Recommendations

Our examination of coder disagreement using experimental recoding of core CMP documents clearly indicates that the CMP coding process is highly prone to misclassification and stochastic coding errors. Bearing in mind that the minimum standard conventionally deemed acceptable for the reliability coefficients reported in Table 2 is 0.8, the coefficients we find are worryingly low, almost all in the range [0.3, 0.5]. From this we infer that, had multiple independent human coders indeed been used to code every document in the CMP data set, then these codings would have been deemed unacceptably unreliable. Although this has previously been suspected on common sense grounds, it has not previously been demonstrated in a systematic way by analyzing multiple codings of the same document using the CMP coding scheme.

We also found that some categories in the CMP scheme are much more susceptible to coding error than others. In particular, the CMP coding categories "305: Political authority" and "404: Economic planning: positive" are extremely prone to misclassification. More worryingly for users of the CMP left–right scale, they often generate coding errors that assign text units "master coded" as right (305) or left (404) to a coding category on the "wrong" side of the left–right scale. This in turn means that problems arising from coding error are not solved by using the CMP's aggregate left and right categories or the additive scale constructed from these. Text that should not be assigned to any category, in other words text that the gold standard declared "uncodeable," was also more likely than not to be wrongly assigned a policy category.

In addition to biasing the estimates of text proportions, misclassification will also add considerable noise to the CMP estimates, substantially more than estimated to arise from either the text generation process (described in Benoit, Laver, and Mikhaylov 2009) or of coder differences in unitization, estimated at $\pm 10\%$. In addition, the coder misclassification, by coding as left what should be right and vice versa, causes a centrist bias as a result of which extreme positions tend to be coded as more centrist than they "really" are. The additional noise plus the bias caused by misclassifications toward the middle are likely to cause additional attenuation bias of estimated causal effects when CMP quantities, especially Rile, are used as covariates in regression models.

Given the central importance of the CMP estimates to cross-national comparative research, our findings strongly indicate the need for further systematic work on this important matter. Several recommendations emerge from our analysis. First, any researchers proposing to use the CMP data set or apply its scheme to new manifesto tests should be fully aware of our findings, which indicate that the CMP coding process is prone to unacceptably high levels of unreliability. Aggregation of misclassified categories to coarser scales—such as the Rile scale of left–right policy—does not eliminate this problem.

Second, the propensity our findings demonstrate for misclassification by human coders, even trained and experienced coders, suggests a need for a much simplified coding scheme that would facilitate more reliable classification. Coding schemes must balance the researcher's desire to reflect accurately the *complexity* of the reality represented by a text, with the practical requirements of keeping coding schemes simple enough that they can be implemented by human coders *reliably*. This means extensive reliability testing to fine-tune coding categories at the design stage to balance these often competing objectives.

Finally, ongoing reliability benchmarking should form an integral part of any large-scale text coding project, using multiple independent codings on at least a subset of documents on an ongoing basis. Here, our study has been limited to using the master documents coded by the CMP in this limited exercise because we wanted to have some sense of how the multiple codings we generated compare with the CMP's own view of the "true and certain" position of each document. What is clearly now indicated, however, is a project that would procure multiple independent codings of a much larger sample of CMP documents for which no master coding exists to allow more confident conclusions to be drawn about the extent of unsystematic intercoder (un)reliability and the biasing effects of systematic coder misclassification.

References

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

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov. 2009. Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science* 53:495–513.

Bross, I. 1954. Misclassification in 2 × 2 tables. *Biometrics* 10:488–95.

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.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1):37.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76:378–82.

Fleiss, Joseph L., B. Levin, and M. C. Paik. 2003. *Statistical methods for rates and proportions*. 3rd ed. New York: John Wiley.

Hayes, A. F., and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding Data. *Communication Methods and Measures* 1:77.

Heise, D. R. 1969. Separating reliability and stability in test-retest correlation. *American Sociological Review* 34:93–101.

Hopkins, Daniel, and Gary King. 2010. A method of automated nonparametric content analysis for social science. *American Journal of Political Science* 54:229–47.

King, G., and Y. Lu. 2008. Verbal autopsy methods with multiple causes of death. *Statistical Science* 23(1):78–91.

Klingemann, Hans-Dieter, Andrea Volkens, Judith Bara, Ian Budge, and Michael McDonald. 2006. *Mapping policy preferences II: Estimates for parties, electors, and governments in eastern Europe, European Union and OECD 1990–2003*. Oxford: Oxford University Press.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. 2nd ed. Thousand Oaks, CA: Sage.

Kuha, Jouni, and Chris Skinner. 1997. Categorical data analysis and misclassification. In *Survey measurement and process quality*, eds. Lars E. Lyberg, Paul Biemer, Martin Collins, Edith D. De Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: John Wiley & Sons.

Kuha, Juni, C. Skinner, and J. Palmgren. 2000. Misclassification error. In *Encyclopedia of epidemiologic methods*, eds. M. Gail and J. Benichou, 578–85. New York: Wiley.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–74.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. Estimating the policy positions of political actors using words as data. *American Political Science Review* 97:311–31.

McDonald, Michael, and Silvia Mendes. 2001. Checking the party policy estimates: Convergent validity. In *Mapping policy preferences: Estimates for parties, electors, and governments 1945–1998*, eds. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, and Eric Tanenbaum. Oxford: Oxford University Press.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2011. Replication data for: Coder reliability and misclassification in the human coding of party manifestos. http://hdl.handle.net/1902.1/16863 UNF:5:/DiFWifTzUKbX0eH64QF9g==IQSS Dataverse Network [Distributor] V1 [Version].

Roberts, Chris. 2008. Modelling patterns of agreement for nominal scales. *Statistics in Medicine* 27:810–30.

Rogan, W. J., and B. Gladen. 1978. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* 107:71–6.

Slapin, J. B., and S.-O. Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science* 52:705–22.

Volkens, Andrea. 2001a. Manifesto research since 1979: From reliability to validity. In *Estimating the policy positions of political actors*, ed. Michael Laver, 33–49. London: Routledge.

———. 2001b. Quantifying the election programmes: Coding procedures and controls. In *Mapping policy preferences: parties, electors and gGovernments: 1945–1998: Estimates for parties, electors and governments 1945–1998*, eds. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. Oxford: Oxford University Press.

Volkens, Andrea. 2007. Strengths and weaknesses of approaches to measuring policy positions of parties. *Electoral Studies* 26: 108–120.

Wüst, Andreas M., and Andrea Volkens. 2003. *Euromanifesto Coding Instructions*. Mannheim, Germany: Mannheimer Zentrum für Europäische Sozialforschung.