Coder Reliability and Misclassification in the Human Coding of Party Manifestos[*]

Slava Mikhaylov
London School of Economics
v.mikhaylov@lse.ac.uk

Michael Laver
New York University
michael.laver@nyu.edu

Kenneth Benoit
Trinity College Dublin
kbenoit@tcd.ie

February 18, 2010

Abstract

The long time series of estimated party policy positions generated by the Comparative Manifesto Project (CMP) is the only such time series available to the profession and has been extensively used in a wide variety of applications. Recent work (e.g. Benoit, Laver, and Mikhaylov 2009; Klingemann et. al. 2006, chs. 4–5) focuses on non-systematic 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 dataset and coding process.

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

[*]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 of the volunteer test coders who completed the on-line 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.

1 Reliability versus Validity in Content Analysis

Sooner or later, anyone interested in measuring policy positions of political actors turns to the systematic analysis of political texts. Measuring policy positions by analyzing political texts satisfies a number of key scientific criteria. The act of measurement does not disturb what is being measured, a claim difficult to sustain for any type of survey, during which ideas can be put into people's heads by the very questions they are asked. Once deposited on record, as it has been from ancient times, text lasts forever, in contrast to surveys, which cannot be conducted in the past. Once captured, text does not change, and this makes it possible to devise computational techniques that always yield the same result when applied to the same text, a quality of reproducibility that will not apply to repeated questioning of survey respondents. Nor does text wear out, once captured, no matter how many times we torture it through repeated analysis. Once we have captured an expert survey panel, by contrast, we can only torture them so often without polluting our source of data.

Text is plentiful and cheap. Indeed, in contrast to the typical data problems facing researchers in the social sciences, the main problem with text data is that there is too much of it, not too little. This cornucopia of text confronts us with two particular problems. The first concerns *which* texts to analyze—essentially a problem of content validity. No matter how sophisticated our analysis of a given text, our results are of no value if the text is not a valid source of information about the matter under investigation. Putting aside this obvious though difficult matter, however, our focus in this paper concerns the second problem: how best to estimate policy positions from a valid text corpus. Since this corpus must be vastly reduced in complexity in order to be of any use to social scientists, this implies developing a systematic scheme for converting the text into usable quantities. The most common approach by far is to analyze the content of texts using a categorical scheme consisting of two steps (Krippendorff, 2004, 219). First, texts are parsed into smaller units relevant to the research question, such as words, sentences, or quasi-sentences, depending on the research design. Following this first step of *unitization*, a second step involves *coding* each unit by assigning a category from the coding scheme to each text unit. Both steps can be held to scrutiny not just in terms of the validity of the resulting information, but also in terms of the reliability of the procedure, two criteria that must often be traded off with one another in practice.

Whenever non-deterministic instruments—such as human beings—are used to unitize and code texts, then the content analysis procedure faces potential problems with *reliability*. Depending on how unreliable the procedure is, estimates constructed from the codings may also lack *validity* because of the level of noise or even bias introduced by the content analysis procedure. Reliability is no guarantee of validity, however, and in practice validity tends to suffer in the pursuit of maximizing reliability. Indeed, the debate over machine versus human coded content analysis largely revolves around the tradeoff between reliability and validity. Proponents of computerized schemes for estimating party positions from political manifestos (e.g. Laver, Benoit and Garry, 2003; Laver and Garry, 2000; Slapin and Proksch, 2008) cite perfect reliability in their favor, yet struggle to demonstrate validity. Hand-coded schemes such as the CMP claim validity as a central advantage but then devote huge resources to attempts to enhance reliability (see for instance Klingemann et al., 2006, chs.4–6).

As a thought experiment, suppose we want to estimate the position on a left-right scale of French president Nicolas Sarkozy, using as texts the complete set of speeches he made on the record during 2009. Leaving aside more nuanced ideological differences for the moment, assume we propose a new coding scheme consisting of two categories, “left” and “right”, and that the task is to tag each sentence from Sarkozy’s speeches according to this binary classification. To code the texts with this scheme, we could recruit a panel of scholars accepted within the profession as the world’s greatest experts on French politics, ask them to read the Sarkozy speeches and then classify each sentence as left or right. The experts must apply subjective judgments based on their interpretation of the each sentence’s meaning, and will surely disagree on how at least some sentences should be classified. Indeed, this ability to apply judgment is precisely why we choose trained coders over, say, chimpanzees, whose expected agreement would have been 25% through pure chance. Yet subjective judgments are at the very least subject to stochastic variation. Furthermore, our coders might judge that many sentences in Sarkozy's text have nothing to do with either left or right in French politics yet, asked to code these, may nonetheless assign some sentences to "left" or "right". Finally, our coders might tend to categorize ambiguous sentences as "right" given their contextual knowledge about Sarkozy—sentences that might have been classified as "left" if the identical words had been spoken by someone known to belong to a socialist or communist party. Our procedure will therefore be likely to yield different answers each time we repeat it. Part of the problem has arisen from the fundamentally indeterminate nature of human judgment, but this problem has been compounded an ambiguous coding scheme itself—two interrelated matters to which we return at length below.

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

> is *reliable* when it responds to the same phenomena in the same way regardless of the circumstances of its implementation...In content analysis, this means that the reading of textual data as well as of the research results is replicable elsewhere, that researchers demonstrably agree on what they are talking about. (Krippendorff, 2004, 211)[1]

In any content analysis scheme using the subjective judgments of human coders to apply a coding scheme with any degree of substantive meaning, however, perfect reliability is impossible. Our first task as data analysts, therefore, is to identify and characterize problems of validity and reliability, as well as potential consequences (such as bias) in our research procedure and resulting estimates. Absent this, our estimates are worthless. Indeed they are in a real sense worse than worthless since we have no idea at all how good or bad they are, completely undermining any procedural confidence in the results produced by the research. When it comes to interpreting data, an unreliable research procedure casts basic doubts on the substantive meaning of the data and on any analysis of these

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

(Krippendorff, 2004, 212). Our first priority should therefore be to characterize problems regarding validity, reliability, and bias in our research procedure; our second task to work as hard as we can to minimize their effects.

In what follows, our main substantive interest lies in measuring party policy positions on the basis of systematic analyses of party manifestos conducted by the longstanding 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. The task we set ourselves in this paper is to develop a more systematic characterization of some of these problems than has hitherto been attempted. In what follows, we set out a framework for reliability and misclassification in categorical content analysis, and apply framework this to the CMP coding scheme. To come to concrete terms with reliability and misclassification in the context of the CMP, we designed and carried out a series of coding experiments on texts for which the CMP has supplied a "correct" coding, and we report on these tests. Finally, we discuss these results and their implications for continued use of the CMP research. Our aim in doing this is to enhance our ability to draw reliable, valid and unbiased statistical inferences from the CMP data, which remains the profession's main source of text-based time series data on party policy positions.

2 The CMP Coding Scheme and Sources of Disagreement

Elsewhere (Benoit, Laver and Mikhaylov, 2009) we describe the full process generating the CMP dataset; here our focus is on the CMP coding scheme and the way that human coders assign coding categories to each text unit. CMP estimates of the policy position of a particular party on a particular matter at a particular election are generated by using a trained human coder to allocate every 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

[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 + uncoded standard CMP categories.

relative salience for the party of the policy area defined by each category as the percentage of all text units allocated to that category.

In what follows, we leave to others the important question of whether party manifestos are valid sources of information about the policy positions of political parties. We also leave for future work the potential for *coding bias*, which arises because human coders are inevitably aware of the authorship of the texts they are coding, a problem especially acute for highly self-referential documents such as party manifestos. We deal elsewhere (Benoit, Laver and Mikhaylov, 2009) with non-systematic measurement error in CMP data that arises from stochastic features of the *text generation* process. Here, we focus on error arising in CMP data from stochastic 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. We refer to this coding error in general terms as *misclassification*.

2.1 Coding differences from human "features"

CMP data are fundamentally susceptible to coding error because, of their essence, they derive from subjective judgments made by human coders. These days, indeed, human coding is preferred to machine coding in settings where it is explicitly felt that subjective coding by human experts is more valid than objective coding by machines. Coding error arises because different human coders at the same time, or the same human coder at different times, are likely to code the same text in somewhat different ways. This process may be unbiased, in the sense that we can think of an unobservable "true and certain" value of the quantity being measured, with each human text coding being a noisy 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.[3]

The CMP data, however, are generated by party manifestos coded once, and once only, by a single human coder. There is no variation in noisy 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

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

CMP dataset. We are very confident, however, on the basis of both anecdotal evidence and good old fashioned common sense that, if there were to be a series of independent codings of the same manifesto, then these would all differ at least somewhat from each other. Indeed, if someone reported that 1,000 highly trained coders had each coded 10,000 manifesto text units using the CMP's 57 category scheme, and that every single coder had coded every single text unit in precisely the same way, then our overwhelming suspicion would be that the data had been faked.

2.2 Coding differences from category ambiguities

CMP coders often report difficulties determining precisely which of the coding categories to assign to text units. Hence important sources of coder error are the ambiguities and overlap that exist in the way that some of the categories are defined. Consider the distinction between the following categories:

> "401: Free enterprise: 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 category definitions but it is easy to imagine text for which the coder's decision as to which category is most appropriate would be a knife-edge judgment, one that would be made in different ways by different coders. In contrast "501: Environmental protection" is essentially the only CMP coding category making explicit reference to the environment, so there is nowhere else in the scheme to allocate text units referring to the environment (a decision that, incidentally, renders anti-environmentalist statements uncodable by the CMP). Any text coding scheme must be viewed as a whole, taking into account overlaps and the sharpness of boundaries between categories as well as the definitions of each category on a stand-alone basis. However, we do expect some CMP coding categories to be more "reliable" (different coders tend to code the same text unit into the category in question) than others (different coders do not all use the category in question for the same text unit.) As we shall see, this is very much what we find in our coding experiments.

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

2.3 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. Furthermore, as a matter of practical fact, most third-party users of CMP policy time series data are looking for something much simpler; nearly all of them, indeed, are looking for party positions on a single left-right scale.

In response to these interlocking demands, the CMP is best known for its left-right "Rile" scale, which the CMP itself calls its "crowning achievement" (Budge et al., 2001, 19). This is a simple additive index based 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 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-canceling.[4] In our tests below, we critically examine this claim.

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

2.4 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"—and also by constant communication and interaction with the supervisor in Berlin (Volkens, 2001*b*, 94), (see also Volkens, 2001*a*, 37-40). Specifically the CMP has done this by setting out to train all CMP coders to code the same two manifestos in the same way as a CMP "gold standard" coding that is taken to reflect a "certain truth" about the policy positions expressed in those manifestos used as training documents.

The CMP has invested great effort into improving the quality of its process for training coders. Based on the first evaluation of test results, a new version of coding instructions was produced (Volkens, 2007, 118).[5] The revised instructions draw particular attention to three specific ambiguities in the CMP coding scheme affecting coding reliability: when no category seems to apply to the quasi-sentence, when more than one category seems to apply, and when the statement in the quasi-sentence is unclear (Klingemann et al., 2006, 170). When the statement seems unclear the coder is advised to seek cues from the context and/or contact the supervisor in Berlin.

Other investigations of reliability have specifically targeted possible error in the the 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 some pretty strong theoretical assumptions and assumptions about the latent reliability structure, they conclude that "Rile" is effectively very close to being perfect (Klingemann et al., 2006, 103). Such tests focus on very different issues from those of stability and reproducablity 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 tendencies for misclassification.

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

3 A Framework for Stochastic Misclassification of Text Categories

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

Our discussion here follows the framework of Kuha, Skinner 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 non-negative 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 words, the CMP's "per" variables. When Θ contains non-zero 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. King and Lu, 2008; Rogan and Gladen, 1978). 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.[6] A coding scheme with a high sensitivity will mean that text units will tend, with a high degree of reliability, to be assigned to the category the to which the text units do in truth belong. Our testing framework allows us to estimate specificity directly, and this forms our focus in

[6] The converse of sensitivity is *specificity*, the rate at which $\Pr(A_i^* \neq k | A_i \neq k)$, or in our context, the probability that a sentence is not classified as a given category when it really does not belong to that category, or the rate of true negatives. Specificity's converse is often expressed as α, the false positive rate, known in hypothesis testing as the probability of a Type I error. If the null hypothesis were that $A \neq k$, then a Type I error would be the conclusion that $A = k$, or a false positive.

the tests that follow.

4 An Experiment to Assess Coder Agreement

4.1 Methods and Data

Our method for evaluating misclassification and reliability in the CMP coding procedure was to perform a simple experiment: to see how much agreement could be obtained by multiple coders applying the CMP scheme to the same texts. Our experiment employed two texts, both taken from the "Manifesto Coding Instructions" provided in Appendix II to Klingemann et al. (2006). Apart from detailed instructions for coders, Appendix II also contains two fully coded sample texts designed to serve as examples. Using these two texts held several key advantages. First, each text had already been "officially" parsed into quasi-sentences by the CMP, meaning that we could take the 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 tester agreement in our experiments.

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

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 from "Manifesto Coding Instructions" in Appendix II to Klingemann et al. (2006). Coders were asked to select for each text unit an appropriate category from a scroll-down menu. We also collected some minimal information on coder identifiers and previous experience in coding manifestos. Only completed manifestos could be submitted into the system. Going for a mix of experience and youth we sent out invitations to participate in our experiment to the majority of trained CMP coders[7] and a selection of usual suspects: staff and postgraduates at several European and North American universities. We ended up with a list of 172 names with active emails who were randomly assigned to one of the two test documents.

Our response set consisted of 39 coders, but some of these results were discarded. To be as fair as possible to the CMP, we discarded the bottom fourth of test coders in terms of their reliability, while dropping none from the top. Overall, the New Zealand manifesto was completed by 12 coders and the UK manifesto by 17. The coders whose results are reported here had a 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 reliabily those with less experience.

4.2 Methods of Assessing Agreement

Previous analysis of inter-coder variation, coder bias, and misclassification can only be characterized as limited. The CMP measured the extent to which coder training was successful by correlating 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. For 23 coders that were trained from the the second version of coding manual, their average correlation with the "gold standard" was reported to be 0.83. Of these coders fourteen were new hires taking the test for the first time. Their average correlation with the master copy is 0.82. Nine coders on the second contract took the test again with results for this group going up from 0.70 in the first round to 0.85 in the second round (Volkens, 2007, 118). Klingemann et al. (2006, 107) report that coders on another contract retaking the test showed an average correlation coefficient of 0.88. These reported results are collected in Table 1.

[TABLE 1 ABOUT HERE]

Several serious issues with these reported results become immediately apparent to anyone who has ever used the CMP data. The key issue in reliability tests taken by the CMP coders is whether they agree on unitization and categorization of text units with the "gold standard". There is a clear distinction, however, between measuring *agreement* and measuring *association*. Strong association

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

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 CMP's 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 tend to be used. The test manifesto used by the CMP to assess reliability is no exception, and since the correlation vectors from the CMP's reliability are indexed by category, this means a majority of the elements in the correlation vectors will have zeroes. The effect is to register high correlations based not on how well coders agree on applicable categories, but how well they agree on categories that clearly do not apply (such clear agreement on the absence of any EU-category quasi-sentences in the 1966 New Zealand training document).

Beyond the measures of association there are standard measures of agreement that are used extensively in the literature on content analysis. One standard measure is Krippendorff's α, which is "the most general agreement measure with appropriate reliability interpretations in content analysis" (Krippendorff, 2004, 221). Outside the content analysis literature by far the most widely used method of statistical analysis of agreement for categorical variables is the κ measure (Roberts, 2008, 811).[8] Hayes and Krippendorff (2007) compare Krippendorff's α and Fleiss' κ and suggest that they are very similar. We also find that in most practical contexts both measures produce essentially identical coefficients. Both α and κ coefficients have a range from zero (perfect disagreement) to one (perfect agreement). Both measures also take into account the fact that some agreement may occur purely by chance.

It should be noted that there are two major issues with applying any measure of agreement and

[8] 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). The kappa coefficient was proposed by Cohen (1960) and extended to multiple raters by Fleiss (1971).

association to the CMP reliability results. First, since unitization differs between coders, as it does +/-10% in the tests reported in Table 1, it is not clear on what, if anything, the coders are supposed to agree on. Second, coders report only aggregate percentages for each category leaving open the question whether coders actually agreed on codes applied to individual text units. Only by fixing the units and analyzing agreement at the category level as in our experiment can true reliability be assessed, something which our test controls for.

The CMP group prefers to focus on reliability of composite indicators on the basis of their performance within the data set (Klingemann et al., 2006, 107). Reliability results for individual estimates are viewed of limited importance with the emphasis placed on general tendencies and patterns (Klingemann et al., 2006, 108). 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 is 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 Inter-coder Agreement

According to Hayes and Krippendorff (2007, 78), 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." Applied to the CMP, reliability refers to the extent that different coders, coding the same manifesto independently, are able to agree on the categories to which each quasi-sentence belongs. In what follows we therefore report the simplest and easiest to test indicator of the CMP coding's reliability: how well different test coders agreed with one another when assigning categories to each quasi-sentence. Note that in assessing this form of reliability, we need make no reference to the master or "true" coding at all. If we find significant coder disagreement, then we can directly conclude that misclassification is occurring, since by necessity not every disagreeing coder can be correctly classifying each text unit.

Perfect reliability is never to be expected, but 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.[9]

[TABLE 2 ABOUT HERE]

The results of our reliability scores from test coder results are summarized in Table 2. The table reports results for the British manifesto, the New Zealand manifesto, and the two combined. The first column reports κ for all coders by category. In theory, each quasi-sentence could have been rated by each coder as belonging to any one of the 56 policy categories or classified as "uncoded", although in practice many categories were never used by any coder.

Because each category also plays a role in the definition of the CMP's centrally important "Rile" index—being one of the 13 left or 13 right categories, or one of the 31 categories that is neither—we also compared the "Rile category" assigned by each coder to the quasi-sentences, reported in the second column ("By RILE") of Table 2. This allowed us to test whether reliability could be improved—as expected by the CMP—when only this reduced set of three categories was used. By this view, two coders assigning "403" and "404" to the same quasi-sentence would be viewed in perfect agreement, since both of these categories are classified as "left" in the Rile scale.

Finally, for the categories that the CMP's master coding identified as being present in 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. The results are broadly consistent with the summary results, although several exceptionally unreliable categories stand out. From the left side of the "Rile" scale, "202: Democracy Positive" is extremely poor, with $\kappa = 0.18$, as are "701: Labour Groups: Positive" and "Economic Planning: Positive". On the Right, "605: Law and Order: Positive" and especially "305: Political Authority: Positive" are flagged by our experiment as being extremely unreliable. In general, categories identifying broad policy objectives such as "economic goals" seem to be very highly prone to inter-coder disagreement when it comes to assigning them to specific text units.

Overall, these results show that regardless of whether coders are compared in the full category tests or on the reduced three-fold "Rile" classification, rater agreement is exceptionally poor by

[9] 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.

conventional standards: 0.35-.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, and even when at least one-third of our test coders have previous experience with coding manifestos for the CMP, reliability for the CMP scheme is significantly below conventionally acceptable standards.

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. Indeed, this is the standard benchmark applied by the CMP in previous tests of reliability (e.g. Volkens, 2007; Klingemann et al., 2006, 107). If the training process has succeeded and coders are successfully able to apply the coding scheme to actual text units, then their agreement with the master coding should be high. Agreement with the master coding can also be taken as a measure of the errors introduced by the difficulty of the coding scheme.

[FIGURE 1 ABOUT HERE]

The results of our tests were not encouraging. For the British manifesto test, the New Zealand manifesto test, and combined tests respectively, the median κ test coders' agreement with the master were 0.43, 0.54, and 0.46 respectively. The best coder agreed 0.74 with the master, and the worst 0.22. The full results are portrayed in Figure 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 black line indicates the median results (0.42 and 0.54) from each test. For comparison with the conventional minimum level of acceptable reliability, we have also plotted a dashed 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 or framework for misclassification, we are able to use the empirical 57×57 matrix of true versus observed codings to estimate the misclassification matrix Θ_{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-canceling," and also focuses attention on important errors, such as whether a category that is really "left" will be classified as one which is considered "right" in the CMP's "Rile" scale, and vice-versa. Because the "Rile" index—as are all other quantities in the CMP dataset—are considered as proportions of all text units, we also consider misclassifications into the "Other" category that is neither left nor right.

Full misclassification probabilities are reported in the Appendix, for each CMP coding category. Categories are sorted so that the 13 "Rile-left" categories are listed first, the 13 "Rile-right" categories second, followed by the "Rile-other" categories. The probability that an individual policy category will be classified as belonging to its own "Rile" classification are highlighted in boldface. For quasi-sentences that really belong to "202: Democracy: Positive" for instance—a relative high-frequency category at 3.55% of all CMP quasi-sentences in the combined dataset—the probability is only 0.50 from our tests that it will be assigned a CMP code that is one of the 13 "Rile-left" categories. The probability is almost even (0.47) that it will be coded as a category that is not part of the "Rile" index, and just 0.03 that it will coded as a "Rile-right" category. Similar interpretations can be made for each of the other CMP coding categories listed in the Appendix. (The limited set of categories in our test documents meant that we could only report misclassification probabilities for the 22 categories identified by the CMP's Master coding.)

[TABLE 3 ABOUT HERE]

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 1,668 text unit classifications, which we could identify from the CMP's master coding as belonging to a left, right, or neither "Rile" category. Comparing these to the Rile categories of the coding category that our testers identified, we see significant frequencies in the off-diagonal cells. "Left" text units in particular were prone to misclassification, as

0.35 or 35% of the time these were assigned a category that was not in the "Rile" scale. Conversely, about 19% of the text units that were not in a category found in the "Rile" scheme were classifies instead as "left". Overall, the highest diagonal proportion—equivalent to the *sensitivity* or true positive rate defined previously—was just .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 $\beta = (1 - .70) = .30$. Full sensitivities on a category-by-category basis are listed in the Appendix (represented by the bold figures equivalent to the bold diagonals from Table 3). The results across the board are very discouraging. Some categories in the test had abysmally high false negative rates, such as .82 for 404: Economic Planning Positive, and .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 ($\beta = .55$). 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 $\beta \leq .20$. 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.

[FIGURE 2 ABOUT HERE]

A graphical summary of the misclassification probabilities (presented in the Appendix) is to use a ternary plot. This method also clearly singles out visually the worst categories from the standpoint of misclassification. 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 in 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.)[10] If no misclassification existed, then all categories of the same color 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

[10] 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 .2. Reading from the bottom side, the probability of coding category 402 as "Other" is just under .10. Finally, reading from the right side, the probability of 403 being coded as "Right" is just above .7.

305 (truly “right”) and 404 (truly “left”). Located towards the middle of the triangle, these categories are not only severely prone to misclassification, but also their misclassification occurs to either 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 non-systematic, that are likely to be present in the CMP’s reported Rile estimates. From the combined CMP dataset, we know that the population proportions of the “Rile” left, right, and neither text units are roughly 0.25, 0.25, and 0.50 respectively. Our range of “Rile” therefore fixes the other category at 0.50 and lets the other frequencies vary so that we can observe “Rile” values from -50 to +50, once again a range taken from the empirical range in the combined CMP dataset.[11]

[FIGURE 3 ABOUT HERE]

The results of simulated misclassification are shown in Figure 3. Here we have manually manipulated the misclassification matrix to be symmetric and to produce reliabilities of (reading from top left to right) $\kappa = 0.90, 0.80, 0.70, 0.60$, and 0.50. The last panel (lower left) shows the effect of simulating error using the misclassification probabilities from Table 3, and having a median reliability of 0.47. A faint cross-hair indicates the origin, and a dashed line shows the identity point at which $A_i^* = A_i$.

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 model—in fact this is the single most common usage of the CMP dataset by far—this means that such models

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

are likely to have biased estimates (for a fuller discussion see Benoit, Laver and Mikhaylov, 2009). Second, all of 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 asymmetic 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 over-classify texts as "right", leading to a systematic bias towards the right as well as to the general attenuation bias caused simply by unreliable human coding.

7 Conclusions

We know with absolute certainty, from information published by the CMP itself and summarized in Table 1, that CMP coders disagree with CMP master codings when assigning text units to CMP coding categories. Since different coders all have different correlations with the CMP master codings, we also know with absolute certainty that different CMP coders disagree with each other when coding the master documents. In this paper, we characterize this disagreement as stochastic coding error and set out to derive estimates of the scale of this. This is crucially important since each point in the widely used CMP time series is based on a single coding by a human coder and comes with no estimate of associated error. Before we can draw statistically valid inferences from these data, however, we need estimates of the error associated with their generation.

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

We also found that some categories in the CMP scheme are much more susceptible to coding error than others. Findings on this are summarized in Figure 2 and given in more detail in the Appendix. We see for example that CMP coding categories "305: Political authority" and "404: economic planning: positive" generate coding errors on a very frequent basis. More worryingly for users of the CMP left right scale, they often generate coding errors that assign text units "master coded" as right (305) or left (404) to a coding category on the "wrong" side of the left right scale. This in turn means that problems arising from coding error are not solved by using the 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 fact, the results of our coding experiment, using the best group of coders from our sample and working with well-known test documents, show that not only did the coding process fail to meet conventionally acceptable standards of reliability, but also fails for almost every category to meet acceptable risk levels for misclassification error.

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 +/-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 towards 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.

The coding experiments we report above strongly reinforce the conclusion that the CMP data, based on human interpretative coding of party manifestos, are very unreliable because they are highly prone to misclassification by human coders, even trained and experienced coders. 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. Here, our study has been limited in scope since it is based on limited multiple codings of only two English-language manifestos. We used the master documents coded by the CMP in this limited exercise because we wanted to have some sense of how the multiple codings we generated compare with the CMP's own view of the "true and certain" position of each document. What is clearly now indicated, however, is a project that would procure multiple independent codings of a much larger sample of CMP documents, for which no master coding exists, to allow more confident conclusions to be drawn about the extent of unsystematic inter-coder (un)reliability and the biasing effects of systematic coder misclassification. The work we report above establishes a strong *prima facie* case that this is a problem to be taken very seriously indeed by third-party users of the CMP's estimates of time-series of party policy positions.

References

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. New York NY: 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(2, April):495–513.

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

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(5):378–382.

Fleiss, Joseph L., B. Levin and M.C. Paik. 2003. *Statistical Methods for Rates and Proportions*. 3 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(1):77.

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

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

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*. New York: John Wiley & Sons.

Kuha, Juni, C. Skinner and J. Palmgren. 2000. Misclassification error. In *Encyclopedia of Epidemiologic Methods*, ed. M. Gail and J. Benichou. Wiley pp. 578–585.

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

Laver, M. and J. Garry. 2000. "Estimating policy positions from political texts." *American Journal of Political Science* 44(3):619–634.

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

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*, ed. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara and Eric Tanenbaum. Oxford University Press.

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

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

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(3):705–722.

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

Volkens, Andrea. 2001*b*. Quantifying the Election Programmes: Coding Procedures and Controls. In *Mapping Policy Preferences: Parties, Electors and Governments: 1945-1998: Estimates for Parties, Electors and Governments 1945-1998*, ed. Ian Budge, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald and Silvia Mendes. Oxford: Oxford University Press.

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

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

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

Table 1: *Coder reliability test results reported by CMP.* Sources are (Klingemann et al. 2006; Volkens 2001a, 2007); figures reported are Pearson's R for the aggregate percentage measured across 56 coding categories for the test document found in *MPP2*, pp181–186.

	Fleiss's κ	
Reliability Test	**By Category**	**By RILE**
British Manifesto Test (107 text units, 17 coders)	0.35	0.36
New Zealand Manifesto Test (72 text units, 12 coders)	0.40	0.47
Combined Manifestos Test Results (144 text units, 24 coders)	0.31	0.32
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	

Table 2: *Reliability Results from Coder Tests.* The (L) or (R) desginates whether a CMP category was part of the "Rile" left or right definition, respectively.

		True Rile Category			Total
		Left	Right	None	
	Left	430	100	188	718
		0.59	0.11	0.19	
Coded	Right	41	650	115	806
Rile		0.06	**0.69**	0.11	
	None	254	193	712	1,159
		0.35	0.20	**0.70**	
	Total	725	943	1,015	1,668
	"False negative" rate	.41	.31	.30	
	"False positive" rate	.15	.09	.27	

Table 3: *Misclassification matrix for true versus observed* `Rile`. 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, while the false positive rate is $1-$specificity.

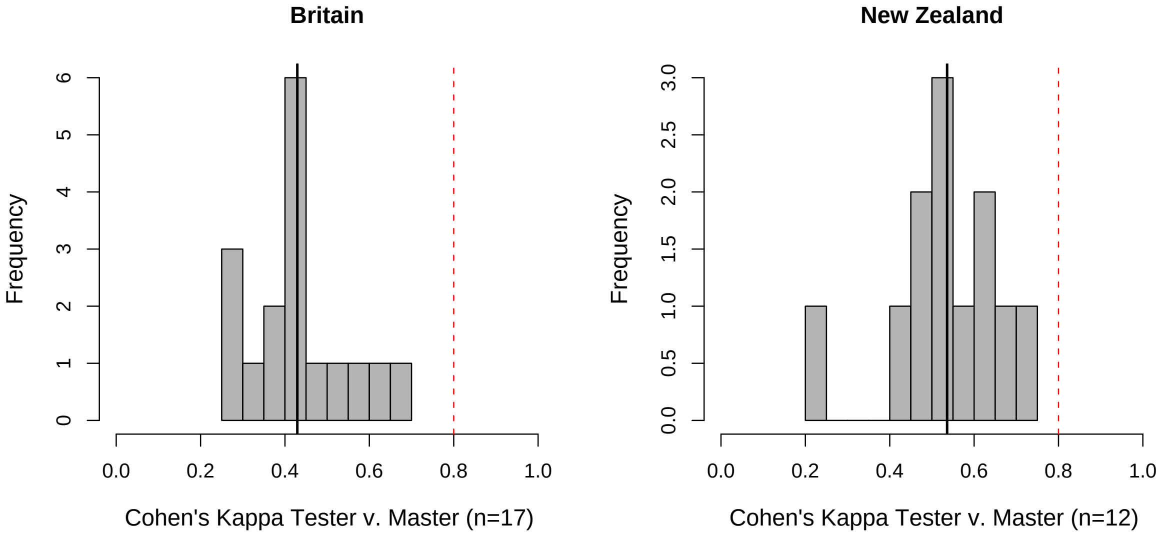

Figure 1: *Summary of Coder Reliabilities Compared to Master, Cohen's* κ*.* The dashed red line indicates the conventional lower bound as to what is considered "reliable" in interpretations of Cohen's κ. The solid black lines are the median value of κ from all the coders completing the tests.

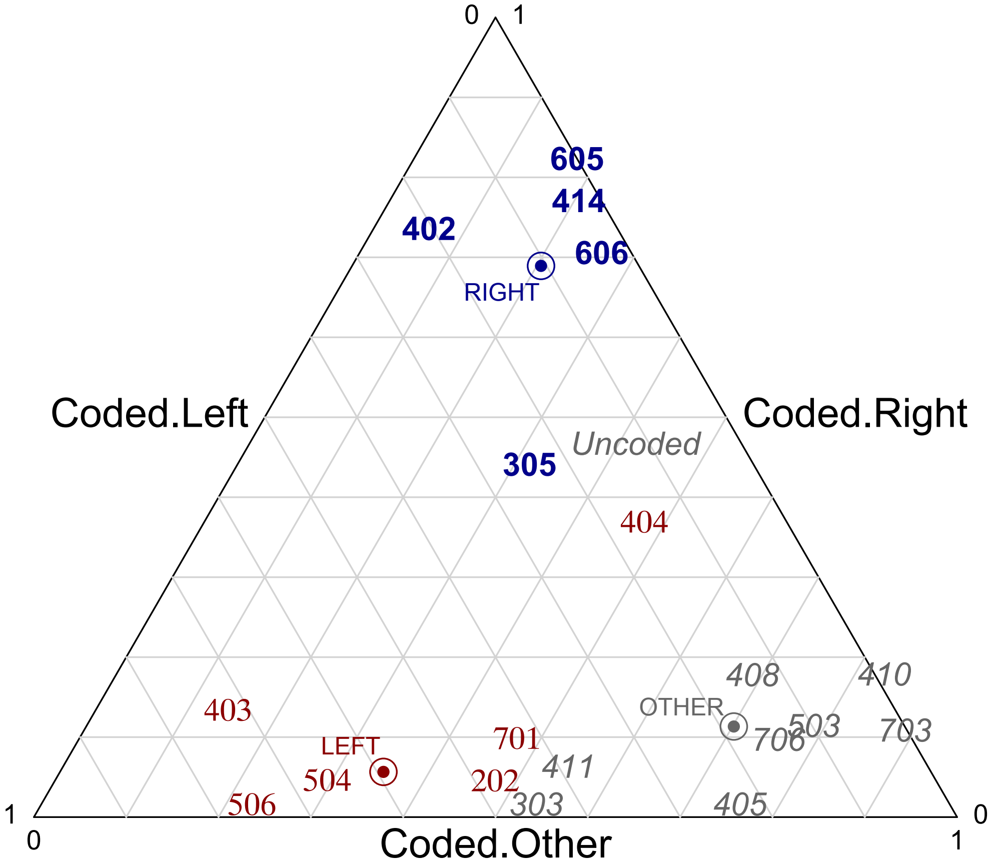

Figure 2: *Misclassification into Left, Right, or Other by coding category, from experiments.* The dark circles in hollow points represent the misclassification for the 3 × 3 left-right-other misclassification matrix. Each number plotted identifies the probability of this category being coded as a left, right, or other category, where the red numbers are really left, the blue numbers really right, and the gray numbers really other categories. If no misclassification existed, all numbers would cluster together into their respective corners, which clearly does not happen.

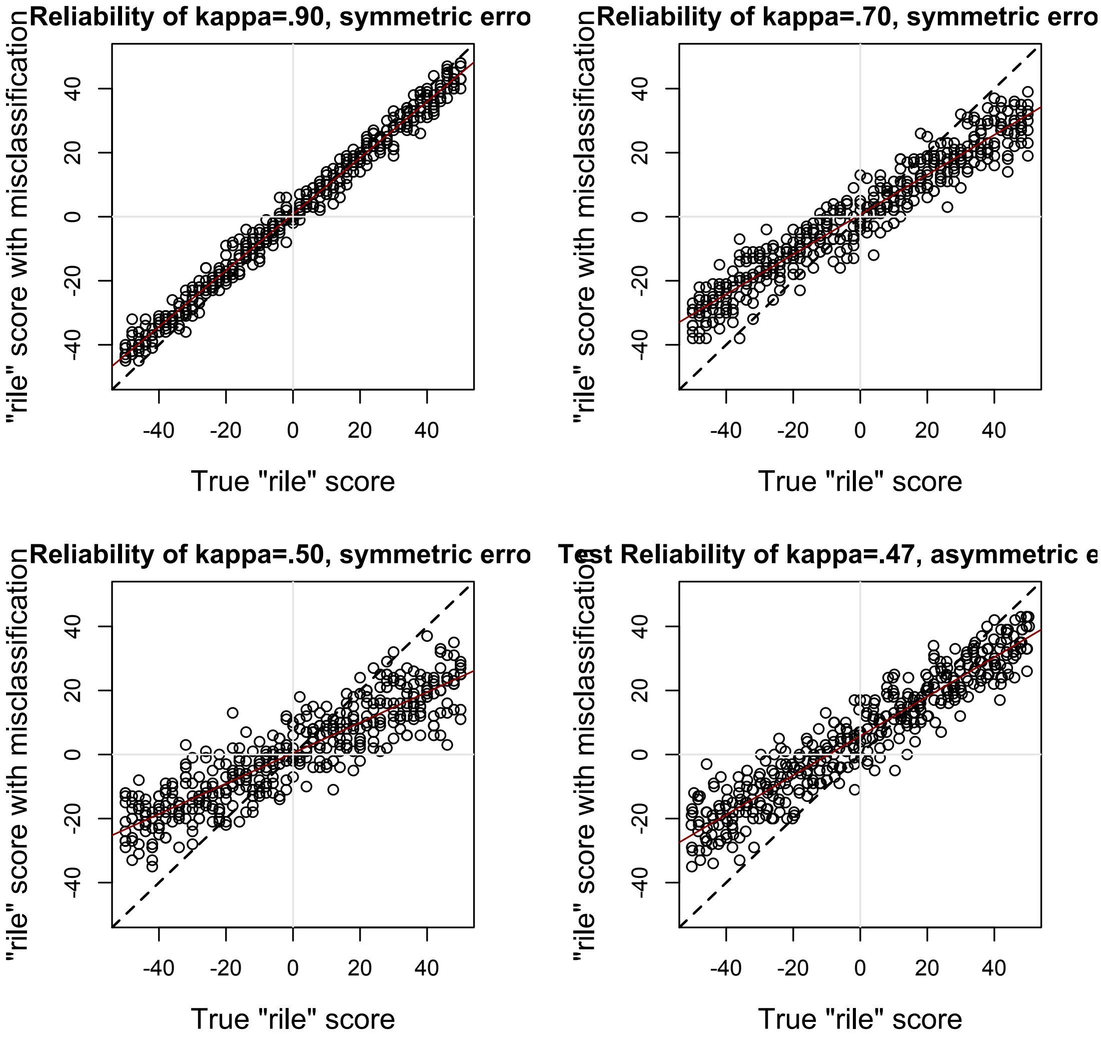

Figure 3: *Simulated Misclassification at Different Levels of* κ. The misclassification matrix Θ_{ji} is simulated from a manifesto with 50% uncoded content, for different levels of κ, except for the last panel, which uses $\hat{\Theta}_{ji}$ estimated from the coding experiments. Misclassification is simulated 8 times for each even-numbered true "Rile" score from -50 to 50.

APPENDIX: Complete Category Listing.

Detailed information on categories, reliability, and classification probabilities from tests.

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

Note: The “Overall % column refers to the proportion of coded quasi-sentences assigned to each category from the complete CMP dataset (from *MPP2*).