Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions

Kenneth Benoit Trinity College
Michael Laver New York University
Slava Mikhaylov Trinity College

Political text offers extraordinary potential as a source of information about the policy positions of political actors. Despite recent advances in computational text analysis, human interpretative coding of text remains an important source of text-based data, ultimately required to validate more automatic techniques. The profession's main source of cross-national, time-series data on party policy positions comes from the human interpretative coding of party manifestos by the Comparative Manifesto Project (CMP). Despite widespread use of these data, the uncertainty associated with each point estimate has never been available, undermining the value of the dataset as a scientific resource. We propose a remedy. First, we characterize processes by which CMP data are generated. These include inherently stochastic processes of text authorship, as well as of the parsing and coding of observed text by humans. Second, we simulate these error-generating processes by bootstrapping analyses of coded quasi-sentences. This allows us to estimate precise levels of nonsystematic error for every category and scale reported by the CMP for its entire set of 3,000-plus manifestos. Using our estimates of these errors, we show how to correct biased inferences, in recent prominently published work, derived from statistical analyses of error-contaminated CMP data.

Text as a Source of Information about Policy Positions

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

Kenneth Benoit is Professor of Quantitative Social Sciences, Department of Political Science, Trinity College, Dublin 2, Ireland (kbenoit@tcd.ie). Michael Laver is Professor of Politics, Department of Politics, New York University, 19 W. 4th Street, New York, NY 10012 (michael.laver@nyu.edu). Slava Mikhaylov, Department of Political Science, Trinity College, Dublin 2, Ireland (mikhailv@tcd.ie).

This research was partly supported by the European Commission Fifth Framework (project number SERD-2002-00061) and by the Irish Research Council for Humanities and the Social Sciences. We thank Andrea Volkens for generously sharing her experience and data regarding the CMP; Thomas Daubler for research assistance; and Thomas Daubler, Gary King, Michael D. McDonald, Oli Proksch, and Jon Slapin for comments. We also thank James Adams, Garrett Glasgow, Simon Hix, Abdoul Noury, and Sona Golder for providing and assisting with their replication datasets and code.

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

American Journal of Political Science, Vol. 53, No. 2, April 2009, Pp. 495–513

 ISSN 0092-5853

a *sine qua non* for serious empirical research in the social sciences (King, Keohane, and Verba 1994, 9).

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

While our arguments below relate to any type of text, we focus in particular on a set of political texts that has been extensively studied: party manifestos. A huge number of manifestos have been analyzed, using human interpretative coders, by the Comparative Manifestos Project (CMP).[2] First reported in 1987 (Budge, Robertson, and Hearl 1987), a hugely expanded version of this dataset was reported in the project's core publication, *Mapping Policy Preferences* (Budge et al. 2001, hereafter *MPP*), to have covered thousands of policy programs, issued by 288 parties, in 25 countries over the course of 364 elections during the period 1945–98. The dataset has recently been extended, as reported in the project's most recent publication, *Mapping Policy Preferences II* (Klingemann et al. 2006, hereafter *MPP2*), to incorporate 1,314 cases generated by 651 parties in 51 countries in the OECD and central and eastern Europe (CEE). Commendably, these data are freely available and have been very widely used, as can be seen from over 800 Google Scholar citations by third-party researchers of core CMP publications.[3] The CMP data are particularly attractive to scholars seeking long time-series of party policy positions in many different countries, for whom this dataset is effectively the only show in town. Despite their pervasive use by the profession, however, these data come with no associated measures of uncertainty. The *reliability* of many CMP scales, especially the left-right scale, has been investigated (e.g., Hearl 2001; *MPP2*, chap. 5; McDonald and Mendes 2001b), as has the *validity* of CMP scales in comparison with external measures (e.g., Hearl 2001; *MPP2*, chap. 4; McDonald and Mendes 2001a). But there is no estimate of *uncertainty* that accompanies the very precise point estimates of policy emphasis that are the essential payload of the CMP and form the basis of any scales estimated from the CMP dataset.

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

We address this problem by decomposing stochastic elements in the data generation process underlying interpretative content analysis by humans. This has two essential components: text generation and text coding. In this article, we focus on measurement uncertainty

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

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

arising from the stochastic nature of political text itself. Any observed text is but one of a huge number of *possible* texts that could have been generated by an author intent on conveying the same message. Characterizing stochastic text generation allows us to systematize the blindingly obvious but hitherto neglected intuition that *longer texts tend to contain more information than shorter ones.* Thus there is huge variation in the length of texts analyzed by the CMP; some coded texts are more than 200 times longer than others. Astonishing as this seems the moment we think about it, all published work using CMP data assumes all texts are equally informative.

We proceed as follows. First, we describe the CMP dataset and the processes that led to its generation. Focusing on stochastic text generation and the impact of text length on measurement uncertainty, we show two different ways to calculate standard errors for each estimate in the CMP dataset; one relies on analysis, one on simulation. Analyzing these error estimates we find that many CMP quantities, *even assuming perfectly reliable human coders*, should be associated with substantial uncertainty. We show how these error estimates can be used to distinguish substantive change from measurement error in both time-series and cross-sectional comparisons of party positions. Finally, we suggest ways to use our error estimates to correct analyses that use CMP data as covariates, rerunning and correcting some prominent analyses reported in recent literature. In a companion article, we focus on measurement uncertainty arising from stochastic variation in the coding of a given observed text by human coders. While our approach allows us to calculate precise estimates of nonsystematic "text generation" error associated with every reported CMP measure of party policy, it can be adapted to other datasets in which quantitative codings are derived by humans on the basis of reading texts.

From Policy Positions to Coded Dataset

Before we characterize error in the CMP dataset, we must understand the processes by which this error arises. These are essentially the same processes that underlie any human interpretative coding based, wholly or partially, on text sources. They therefore apply more generally to the many social science datasets that include variables generated by humans who read some text and then record a quantitative coding conditioned on this. To aid exposition, however, we focus on the data generation processes underlying the CMP. These are summarized in Figure 1.

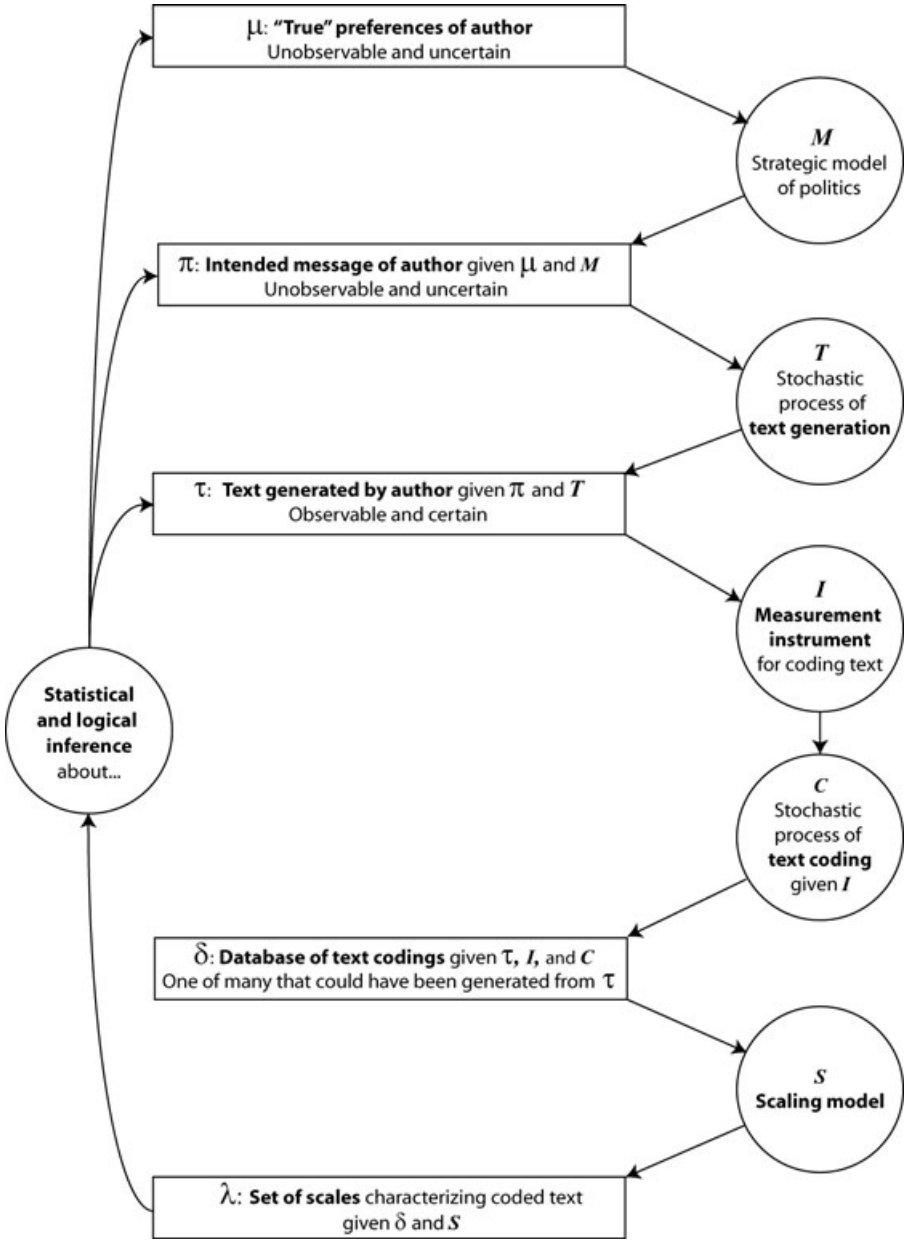
The premise of all content analysis is that there is something to be analyzed. Here, we think of this as the *true policy position*, μ, of the author of some text. This is fundamentally unobservable even, arguably, to the author. If the author is not a hermit, she may want to send signals about this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

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

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

Given an observed text τ and a realized coding scheme *I*, expert human readers interpret text units in τ and allocate these to coding categories in *I*. This coding process

FIGURE 1 Overview of the Positions to Text to Coded Data Process



has both subjective and stochastic elements. The same human reader at different times, or a different human reader at the same time, may well allocate the same text unit to different coding categories. There is thus a *stochastic text coding process C* that, given *I*, maps τ into δ, a database of text codings. Given the stochastic processes we have outlined above, the codings in δ are associated with considerable uncertainty.[4]

The analyst wants the database of text codings in the first place because she wants to estimate something about the text's author. This involves scaling the data, using some *scaling model S*. Clearly, there are many different scaling models that could be applied to the same database of text codings. The result of applying scaling model *S* to the database of text codings in δ will be a set of scales λ. In relation to the CMP, a very well-known scale is the

[4]There is also a serious potential problem with *systematic* coder error, a problem acknowledged by Klingemann et al. (2006, 112) and explored directly through experiments in Mikhaylov, Laver, and Benoit (2008).

left-right scale called "rile." This is the feature of the scaled CMP dataset that is overwhelmingly the most commonly used in published work. There are, of course, many *different possible sets of scales* λ that could be developed by applying scaling model *S* to database δ.

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

We have been very explicit about all of this because it is important to focus carefully on particular features of the long process of causal inference summarized in Figure 1. Lack of clarity about this can, for example, lead to misplaced criticisms of the CMP data. Many of the alleged shortcomings attributed to the estimation of party positions from manifestos, for instance, concern the validity of using manifestos as unbiased, observable implications of true party positions. It is frequently argued, for example, that party manifestos are strategic documents that do not convey the "true" party position, in effect that $\mu \neq \pi$. But this is not a measurement issue. Assuming we can measure the intended message π from the observed text τ in an unbiased way, this is a matter of specifying the correct strategic model *M* that maps μ into π. The claim that manifestos are strategic documents does not therefore have any bearing on CMP text codings, but rather on the logical inferences that are drawn from these about unobservable "true" policy positions μ. The solution to this problem is not better text codings in δ but a better strategic model of politics, *M*. Similarly, it is perfectly reasonable to argue that the CMP's additive left-right scale "rile" is flawed and that other left-right scales using the same data, for example those proposed by Gabel and Huber (2000), or by Kim and Fording (1998), are more valid bases for drawing inferences about the policy positions, μ or π, of text authors. Again, this does not concern the database of CMP text codings, δ, but rather the validity of the scaling model *S* that maps these into a set of derived scales λ. The solution to this problem is a better scaling, not better text codings.

Figure 1 also helps us focus on features of the CMP dataset that are indeed intrinsic to the data collection project itself, further distinguishing between problems that can be fixed without recourse to additional data collection and those that cannot be addressed without new data on the coding of party manifestos. Thus little attempt has been made to take account of the fact that the CMP's core measurement instrument *I*, its 57-category coding scheme, is but one realization of the many possible coding schemes that could have been devised.[5] Clearly the CMP coding scheme is an utterly integral feature of the CMP dataset. Equally clearly, assessing the implications of this involves recoding the same documents using different schemes, and thus a major new data collection enterprise.

Very little attempt has been made, furthermore, to characterize the stochastic *coding* process, *C*, by estimating the extent of variation between coders in applying the same coding scheme *I* to the same text τ. This cannot be investigated without conducting multiple human codings of the same document using the same coding scheme and thus also involves a major new data collection enterprise. Considerable attention has, however, been paid to the reliability and validity of scales derived from the CMP database of text codings, reflected in extensive discussion of the validity of the CMPs "rile" scale.[6] Such discussions about scaling do not hinge on the collection of a new database of new text codings, δ, but rather on how a given dataset should be scaled.[7]

We are not concerned here with building scales from the CMP data, but with another aspect of the CMP manifesto dataset that can be addressed without a major new data collection exercise. This concerns the fact that there is a stochastic text generation process, *T*, that maps the intended message μ into an observed

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

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

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

text τ. We model this process below, using both analytical techniques and simulations, allowing us to formalize the intuition that longer political texts, other things being equal, convey more information about their authors.

Characterizing the Stochastic Process of Text Generation

In what follows, we want to estimate the level of uncertainty in CMP estimates of party policy positions that arises from the stochastic process of text generation. Before going forward, therefore, it is important to be clear about which of the processes mapped in Figure 1 we are going to hold constant. Taking things from the top, we are not concerned with modeling the text authors' strategic incentives to dissemble. We thus in effect assume that $\mu = \pi$. Readers who do not believe this must specify a strategic model M of politics, mapping μ into π, that we do not consider here. Nor are we concerned here with the stochastic process, C, of human text coding, although this is something we directly estimate in a companion article. What we do assume here is that this stochastic process is unbiased. We take the CMP's 57-category coding scheme as given and do not concern ourselves with the datasets that alternative coding schemes might have produced. While the scaling model S that has been applied to the database of CMP codings clearly raises crucial issues, we take two core features of this as given in what follows. The first is the scaling assumption that measures a text's relative emphasis on a CMP coding category as the percentage of coded text units assigned to that category. The second is the precise definition of the CMP's "rile" scale. What we *do* focus on in what follows is the stochastic process T that maps text authors' unobservable policy positions μ ($=\pi$) into observable text deposits τ.

For a given policy category j, define π_{ij} as the *true* but *unobservable* intended policy message from the text's author, represented as country-party-date unit i. The j categories in this case are the 56 policy categories in the CMP coding scheme, plus an additional category for "uncoded," giving a total of $k = 57$ categories. Since, according to the CMP's measurement model, true policy positions are represented by relative or "contrasting" emphases on different policy categories within the manifesto, these policy positions are relative proportions, with $\sum_{j=1}^{k} \pi_j = 1$.[8] For example, party i's emphasis, for a given election, on the 20th issue category in the CMP coding scheme (401: Free Enterprise), is represented as π_{i20}.

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

$$\mathrm{E}(p_{ij}) = \pi_{ij} \tag{1}$$

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

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

$$\Pr(X_j = x_j, \ldots, X_k = x_k) = \begin{cases} \dfrac{n!}{x_j! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k} & \text{when } \displaystyle\sum_{j=1}^{k} x_j = n \\ \text{when } \sum x_j = n \text{ and 0 otherwise.} \end{cases} \tag{2}$$

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

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

manifesto, what is actually reported is $x_{ij}/n_{ij}100$, or the estimate of manifesto *i*'s "true" percentage ($\pi_{ij}100$) of the quasi-sentences from category *j*. We have no additional information that might lead us to conclude there is a systematic function mapping (in a biased way) the true position to a different expected observed position—already expressed by equation (1). Our concern here is with nonsystematic (unbiased) error, which is the extent to which Var $(p_{ij}) > 0$, even though π_{ij} is fixed at a single, unvarying point.[9]

So far we have considered only the case of a "given" manifesto, but of course the combined CMP dataset deals with many such units—a total of 3,018 separate units representing different combinations of country, election date, and political parties for the combined (*MPP* + *MPP2*) datasets.[10] If we are to fully characterize the error from the stochastic process whereby texts are generated, then this will mean estimating Var(p_{ij}) for every manifesto *i* for all $k = 57$ categories.[11]

The lengths (n_i) of the coded manifestos underlying the CMP dataset vary significantly, although this valuable information is almost never referred to by subsequent users of CMP data. About 30% of all coded manifestos had fewer than 100 quasi-sentences, coded into one of 56 categories. Some had fewer than 20 quasi-sentences; some had more than 2,000. Despite very wide variation in the amount of policy information in different manifestos, policy positions estimated from CMP data are almost always treated in the same way, regardless of whether they are derived from coding 20 text units or 2,000.[12] The total number of text units found in a manifesto appears to be, absent systematic information or prior expectation on this matter, unrelated to any political variable of interest. Yet, while assuming that the proportions π_{ij} remain the same regardless of document length, increasing the length of a manifesto does increase confidence in our estimates of these proportions. This reflects one of the most fundamental concepts in statistical measurement: uncertainty about an estimate should decrease as we add information to that estimate.[13] Given that our characterization of the stochastic process that produces observed text categories depends directly on the length of the text, we show next how to use this information to produce error estimates directly reflecting this basic uncertainty principle.

Estimating Error in Manifesto Generation

Analytical Error Estimation

One way to assess the error variance of estimated percentages of text units in any of the CMP's 56 coding categories is through the analytic calculation of variance for the multinomial distribution we have used to model category counts. The goal is to determine the variance of each of the policy ("PER") categories reported by the CMP, which in the language described above represent $\hat{\pi}_{ij}100$ for each category *j* and each manifesto *i*. Here we assume no coding bias (by equation 1), where each π_{ij} represents the *true* but *unobservable* position of country-party-date unit *i* on issue *j*.

Returning to the definition of the multinomial distribution in equation (2), for any multinomial count X_{ij}, the variance is defined as

$$\text{Var}(X_{ij}) = n_i\, p_{ij}(1 - p_{ij}) \tag{3}$$

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

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

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

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

[13]Experience from the CMP has also found that human coders tend to divide the texts into quasi-sentences in a less than perfectly reliable fashion, although this is an aspect of coder variance that we do not deal with here. An analysis of results from repeated codings of the training document used by the CMP to initiate new coders by Volkens (2001) gives us insight into deviation by different coders from the "correct" quasi-sentence structure, as seen by the CMP. Volkens reports that average deviation from the "master" quasi-sentence length by 39 coders employed in the CMP was around 10%. In the CMP coding tests we have analyzed ourselves, which involve 59 different CMP coders in the course of training, coders identified between 127 and 211 text units in the same training document, with a SD of 19.17 and an IQR of (148, 173).

With a bit of algebraic manipulation[14] we can express the variance of the proportion p_{ij}, and the rescaled percentage (used by the CMP as):

$$\text{Var}(p_{ij}) = \frac{1}{n_i}\, p_{ij}(1 - p_{ij}) \tag{4}$$

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

$$\text{SD}(p_{ij}) \propto \frac{1}{\sqrt{n_i}} \tag{5}$$

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

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

[14] Dropping the manifesto index i for simplicity:

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

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

$$\begin{aligned}
10,000\text{Var}(p_j) &= \frac{10,000}{n}\, p_j(1 - p_j) \\
\text{SD}(p_j100) &= \frac{100}{\sqrt{n}}\sqrt{p_j(1 - p_j)}
\end{aligned}$$

Estimating Error Through Simulation

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

The bootstrapping procedure is straightforward. Since the CMP dataset contains percentages of total manifesto sentences coded into each category, as well as the

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

One problem that is not addressed by bootstrapping the CMP manifesto codings is that, as anyone who has a close acquaintance with this dataset knows, many CMP coding categories are typically empty for any given manifesto—resulting in zero scores for the variable concerned. No matter how large the number we multiply by zero, we get zero. Thus a user of CMP data dealing with a 20-sentence manifesto that populates only 10 coding categories out of 56 must in effect assume that, had the manifesto been 20,000 sentences long, it would still have populated only 10 categories. *In extremis*, if some manifesto populated only a single CMP coding category, then every sampled manifesto would be identical. We cannot get around this problem with the CMP data by bootstrapping, unless we make some very interventionist assumptions about probability distributions for nonobserved categories. We prefer to assume that zero categories—for example, zero mentions of the European Union by Australian party manifestos in 1966—reflect a real intention of the text author not to refer to the matter at issue. We thus, for want of better information, take zero categories at face value. In addition, tests using simple methods to deal with observed zeros—e.g., "add-one" smoothing (Jurafsky and Martin 2000, chap. 6.3)—showed no noticeable differences to our results.[15]

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

The results of this bootstrapping provide error variances that decline as exponential functions of text length, something that holds true both for single categories and for additive scales such as the CMP "right-left." In addition, comparing bootstrapped error variance with variance computed analytically (per equation 5), we get nearly identical results.[16] The near equivalence of these two very different methods for estimating standard errors adds to our confidence in both the analytical derivation of CMP error variance and the method of bootstrapping text units in manifestos. In particular, it suggests that the violation of the assumption of independence between coding category probabilities across text units does not seem to be a serious problem, although this assumption deserves attention in future work. It also adds confidence to our belief that the number of text units identified is not systematically related to the coding of these units into policy categories. When we apply our new error estimates to specific empirical research problems in the next section, we use the bootstrap-estimated error as our best approximation of overall nonsystematic error in the CMP's reported estimates.

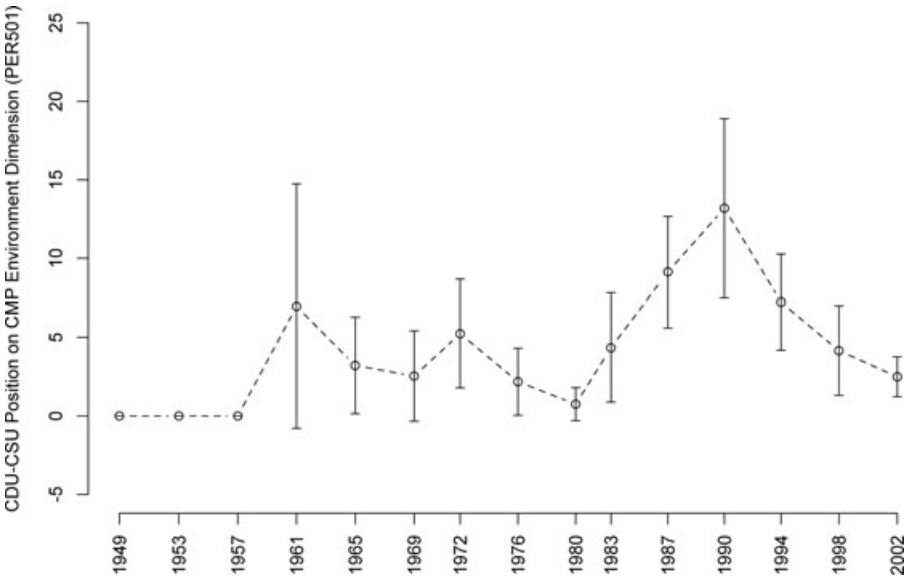
Using CMP Error Estimates in Applied Research

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

[15]Add-one smoothing is one of several methods for dealing with empty observed categories in text analysis and natural language processing, but since these modifications systematically affect the likelihoods, they relate more to systematic than the purely nonsystematic error which forms our focus here.

[16]Full supplemental results are available from http://www.politics.tcd.ie/cmp/.

FIGURE 2 Movement on Environmental Policy of German CDU-CSU over Time



Movement of dashed line is % environment with 95% CI; dotted line is the number of quasi-sentences per manifesto coded PER501.

positions may either be different noisy estimates of the same underlying signal, or accurate estimates of different signals.

Estimating Valid Differences

A substantial part of the discussion found in *MPP* and *MPP2* of the face validity of the CMP data comes in early chapters of each book, during which policy positions of specific parties are plotted over time. Sequences of estimated party policy movements are discussed in detail and held to be substantively plausible, with this substantive plausibility taken as evidence for the face validity of the data. But are these vaunted changes in party policy "real" or just measurement noise? We illustrate how to answer this question with a specific example related to environmental policy in Germany, a country where environmental policy is particularly salient, and also where the CMP has been based for many years. Figure 2 plots the time series of the estimated positions of the CDU-CSU, for a long time Germany's largest party, on PER501 (*Environment: Positive* in the CMP coding scheme). The dashed line shows CMP estimates; error bars show our bootstrapped 95% confidence intervals around these estimates.

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

TABLE 1 Comparative Over-Time Mapping of Policy Movement on Left-Right Measure, Taking into Account Statistical Significance of Shifts

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

Table 1 reports the result of extending this anecdotal discussion in a much more comprehensive way. It deals with observed "changes" of party positions on the CMP's widely used left-right scale (RILE) and thus systematically summarizes all of the information about policy movements that is used anecdotally, in the early chapters of *MPP* and *MPP2*, to justify the face validity of the CMP data. The table reports, considering all situations in the CMP data in which the same party has an estimated position for two adjacent elections, the proportion of cases in which the estimated policy "change" between one election

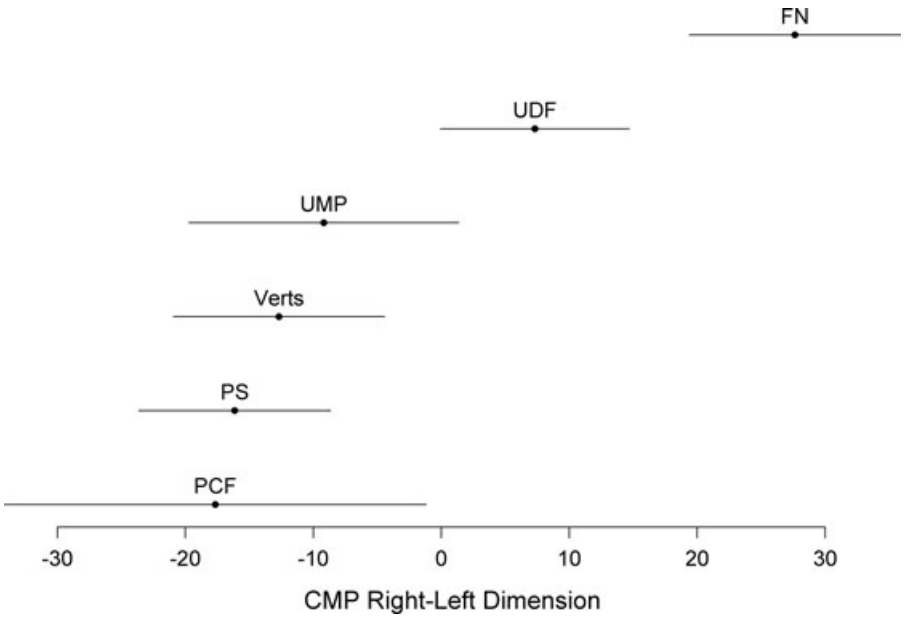
to the next is statistically significant. These results should be of considerable interest to all third-party researchers who use the CMP data to generate a time series of party positions. They show that observed policy "changes" are statistically significant in only 38% of relevant cases. We do not of course conclude from this that CMP estimates are invalid. We do conclude that many policy "changes" hitherto used to justify the content validity of CMP estimates are not statistically significant and may be noise. More generally, we argue that, if valid statistical (and hence logical) inferences are to be drawn from "changes" over time in party policy positions estimated from CMP data, it is essential that these inferences are based on valid measures of uncertainty in CMP estimates, which have not until now been available.

While one of the CMP's biggest attractions is undoubtedly the time-series data it appears to offer, another common CMP application involves comparing different parties at the same point in time. Considering a static spatial model of party competition, realized by estimating positions of actual political parties at some time point, many model implications depend on differences in policy positions of different parties. It is crucial, therefore, when estimating a cross-section of party policy positions, to know whether estimated positions of different parties do indeed differ from each other in a statistical sense. Figure 3 illustrates this problem, showing estimates of French party positions in 2002, on the CMP left-right scale. Taking into account the uncertainty of these estimates, four quite different parties—the Communists, Socialists, Greens, and Union for a Popular Movement (UMP)—have statistically indistinguishable estimated positions, even though the CMP point estimates seem to indicate differences. Only the far-right National Front had an estimated left-right position that clearly distinguishes it from other parties. On the basis of these estimates we simply cannot say, notwithstanding CMP point estimates, whether the Greens (*Verts*) were to the left or the right of the Socialists (*PS*) in 2002. The role of uncertainty in cross-sectional comparisons will differ according to context, but the French case demonstrates—for a major European multiparty democracy—that inferences of difference from CMP point estimates can be ill informed without considering measurement error.

Correcting Estimates in Linear Models

When covariates measured with error are used in linear regression models, the result is bias and inefficiency when estimating coefficients on error-laden variables (Hausman 2001, 58). These coefficients are typically expected to suffer from "attenuation bias," meaning they are likely to be biased towards zero, underestimating the effect of relevant variables. This conclusion must, however, be qualified, since it depends on the relationship between the

FIGURE 3 Left-Right Placement of the Major French Parties in 2002. Bars Indicate 95% Confidence Intervals



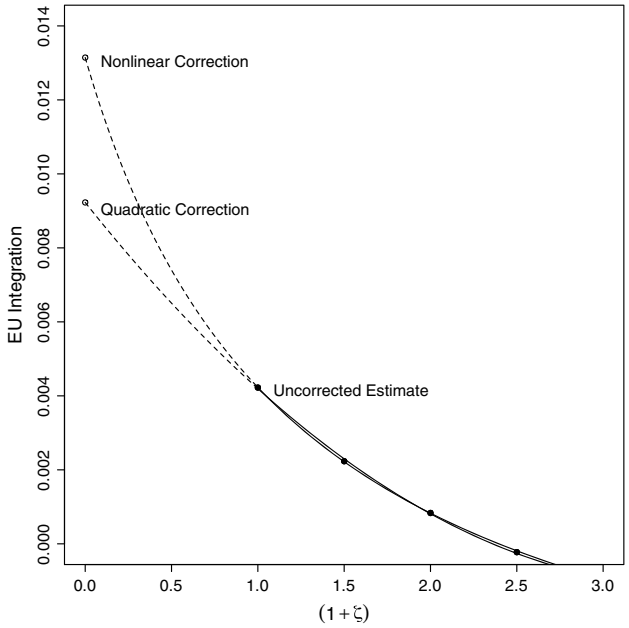
"true" predictor and the noisy proxy available to the researcher, and possibly other variables in the model. More precisely, the effect of measurement error depends on the estimation model and the joint distribution of measurement error and the other variables (Carroll et al. 2006, 41). In the case of linear regression the effects of measurement error can range from simple attenuation bias, to masking of real effects, appearance of effects in observed data that are not present in the error-free data, and even reversal of signs of estimated coefficients compared to the case in the absence of measurement error.

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

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

The basic idea behind SIMEX is fairly straightforward. If a coefficient is biased by measurement error, then adding more measurement error should increase the degree of this bias. By adding successive levels of measurement error in a resampling stage, it is possible to estimate the trend of bias due to measurement error versus the variance of the added measurement error. Once the trend has been established, it then becomes possible to extrapolate back to the case where measurement error is absent. Following Carroll et al. (2006, 98–100) the SIMEX algorithm can be succinctly described as a sequence of steps that we illustrate in Figure 4. The example taken is the *EU Integration* variable from Hix, Noury, and Roland (2006, Model 6) replicated fully below. First, in

FIGURE 4 SIMEX Error Correction in EU Integration with Quadratic and Nonlinear Extrapolant Functions, from Hix, Noury, and Roland (2006)



the simulation step additional random pseudo errors are generated from a normal distribution with mean 0 and variance $\zeta_m \sigma_u^2$ and added to the original data. Since m is known and chosen to satisfy $0 = \zeta_1 < \zeta_2 < \ldots < \zeta_M$ (we use typical values $\{0.0, 0.5, 1.0, 1.5, 2.0\}$), the simulation step creates m datasets with increasingly larger measurement error variances. The total measurement error variance in the m^{th} dataset is $\sigma_u^2 + \zeta_m \sigma_u^2 = (1 + \zeta_m)\sigma_u^2$. In the estimation step the model is fit on each of the generated error-contaminated datasets. The simulation and estimation steps are repeated a large number of times (500 times in our replication example), and the average is taken for each level of contamination. These averages are plotted against the values of ζ (the filled circles in Figure 4), and an extrapolant function is fit to the averaged, error-contaminated estimates. In terms of ζ_m an ideal, error-free dataset corresponds to $(1 + \zeta_m)\sigma_u^2 = 0$, i.e., $\zeta_m = -1$.[17] Extrapolation to the ideal case ($\zeta = -1$) yields the SIMEX estimate (the hollow circles in Figure 4).

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

The quadratic extrapolant function is usually preferred, since it has been shown to result in more conservative corrections for attenuation and is often more numerically stable than the alternative nonlinear function (also shown in Figure 4; Carroll et al. 2006; Hardin, Schmiediche, and Carroll 2003; Lederer and Küchenhoff 2006). (In our replications below, we report corrections based on the more conservative quadratic extrapolation.)

More complicated error corrections are of course possible, but here we deliberately chose a method that is simple, applicable to a wide class of generalized linear models, and for which freely available software is available that can be used with popular statistical packages.[18]

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

The first model analyzed in the original article and replicated here deals with whether mainstream and niche parties differently adjust their policies in response to public opinion shifts. Party policy shifts are operationalized as changes in a party's CMP left-right scale position in successive elections. This measure is regressed on public opinion shifts, a dummy variable for niche party status, the interaction of these two variables, lagged dependent variable, lagged vote share change, the interaction of these two terms, and a set of country dummies. The authors' expectation is that if the coefficient on *Public opinion shift* is positive and statistically significant, then mainstream parties are responsive to shifts in public opinion along the lines of the dynamic representation model. They also expect to find a negative and statistically significant coefficient on the *Niche Party* × *Public opinion shift* variable, providing evidence that niche parties are less responsive to public opinion shifts than mainstream parties, thereby supporting the main "policy stability" hypothesis of the article. In our replication of Adams et al. (2006, Table 1), we focus on the effect of measurement error in both the dependent variable on the left-hand side, its lagged value on the right-hand side, and an interaction of the lagged dependent variable and lagged change in vote share. In the classical measurement error (CME) domain, it is known that measurement error in the dependent variable, if uncorrelated with other covariates, will only inflate standard error of the regression (Abrevaya and Hausman 2004), while measurement error in independent variables will bias the results.[19] We assume here and in subsequent replications that all other covariates are measured without error. The error estimate in contaminated covariates is derived from our bootstrapped standard error.[20]

The second model in Adams et al. (2006) tests whether policy adjustments (shifts in policy towards the center of the voter distribution or away from it) affect parties' electoral support and whether this relationship differs between mainstream and niche parties. Key explanatory variables are constructed from the CMP and thus are expected to be error-prone: *Centrist policy shift, Noncentrist policy shift, Niche Party* × *Centrist policy shift, Niche Party* × *Noncentrist policy shift.* The first variable is measured as the absolute value of the change in a party's position on the CMP left-right scale when a leftist party shifts right or rightist party shifts left, and zero otherwise. The variable measuring the shift away from the center is similarly constructed. The next two variables pick up the differences in electoral effects for niche and mainstream parties in relation to centrist and noncentrist policy shifts.[21] Adams et al. (2006) expect mainstream parties to gain votes in the centrist policy shift and lose votes in noncentrist shift, thus leading to the expectation of a positive and statistically significant coefficient on *Centrist policy shift* and a negative and statistically significant coefficient on *Noncentrist policy shift.* The authors

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

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

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

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

suggest that niche parties are electorally penalized for policy adjustments regardless of the direction of this adjustment (centrist or noncentrist) in what they call the "costly policy shift" hypothesis. This leads to the expectation of statistically significant and negative coefficients on both *Niche Party* × *Centrist policy shift* and *Niche Party* × *Noncentrist policy shift*. At the same time another hypothesis put forward by Adams et al. (2006) states that niche parties lose votes in comparison to mainstream parties for moderating their policy stance ("costly policy moderation" hypothesis). In turn this results in the expectation of a negative and statistically significant coefficient only on the *Niche Party* × *Centrist policy shift* variable.

Table 2 presents results of our error correction for both models, taken from the two regression tables of Adams et al. (2006). For each model, we compare our replication of the published results with SIMEX estimates.[22] The most profound effect of the SIMEX correction of Model 1 is the expected inflation of the standard error of the regression and drop in explained variance as the consequence of measurement error in the dependent variable. The effect of error correction in the covariates decreases the key explanatory variables in size but they remain statistically significant. The full extent of SIMEX error correction effects can be gleaned from the changes in coefficients and standard errors presented in Table 2. The results show support for the hypothesis that niche parties' policy programs are less responsive to shifts in public opinion compared to mainstream parties (the grayed row in Model 1). Evidence for this claim, however, is drawn from a model with much weaker explanatory power.

In the original article, the negative and statistically significant coefficient on *Niche Party* × *Centrist policy shift* (Model 2) is meant to support the "costly policy moderation" hypothesis that, in comparison to mainstream parties, niche parties are penalized by voters for moderating their policy positions. Results in the original article substantively mean that a one-unit shift closer to the center of the voter distribution along the 1–10 left-right scale, results, ceteris paribus, in niche parties' electoral loss of nearly 4% (i.e., approximately $-5.67 + 1.45$, see 523). Evidence for this conclusion weakens as the result of the SIMEX correction. The coefficient on *Niche Party* × *Centrist policy shift* becomes smaller in size and remains statistically significant only at the 0.10 level. In turn, depending on the take on statistical significance cutoff points, this may force the rethinking of some of the theoretical implications of the article. The conclusion that for niche parties "*both* vote-seeking and policy-seeking objectives motivate a stand-pat strategy" because moderation in policy positions is penalized by voters (525, emphasis in original) is not supported by empirical evidence based on the error-corrected estimates at the conventional 0.05 level of significance.

Moreover, Adams et al. (2006, 525) claim that their empirical results support the "costless spatial mobility" assumption typically used in spatial modeling—i.e., that political parties are not electorally penalized for shifting positions in policy space—with respect to mainstream parties. In fact, as Table 2 shows, the corrected coefficient for *Noncentrist policy shift* almost doubles as the result of the SIMEX correction. Indeed, if a one-tailed hypothesis test were applied to the coefficients for both *Noncentrist policy shift* and *Niche Party* × *Centrist policy shift*, both would be considered statistically significant. In terms of the conclusions of the original article, the error-corrected results challenge its categorical conclusion that mainstream parties are not penalized for shifting policies away from the center—suggesting that this effect occurs with at least as much confidence as the conclusion that niche parties are punished for shifting their policies to the center.

Hix, Noury, and Roland (2006). Hix, Noury, and Roland (2006) are concerned with the content and character of political dimensions in the European Parliament (EP). Following an inductive scaling of roll-call votes in the EP from 1979 and 2001, Hix, Noury, and Roland (2006) set out to validate their interpretation of the derived policy dimensions by regressing the mean position of each national party's delegation of MEPs on two sets of independent variables. The first set includes exogenous measures of national party positions on the left-right, social and economic left-right, and pro-/anti-EU dimensions. The second set relates to government-opposition dynamics and consists of categorical variables describing whether a national party was in government and whether the party had a European Commissioner, as well as dummy variables for each European party group, each EU member state, and each (session of) European Parliament. Measures of national party positions are taken directly from the CMP dataset or constructed from it. National party positions on the EU are taken as the difference between positive (category PER108) and negative (category PER110) mentions of the EU. Party positions on economic and social policy are also constructed from the CMP categories (see Laver and Garry 2000, 628–29). The authors expect that national party ideal point estimates on the first

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

TABLE 2 Results of SIMEX Error Correction in Adams, Clark, Ezrow, and Glasgow (2006)

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

(*continued*)

Table 2 Continued

	Model 1		Model 2	
	OLS		OLS	
Variable	**Replication**	**SIMEX**	**Replication**	**SIMEX**
RMSE	0.59067	0.66800	4.39424	4.45137
R^2	0.35780	0.17871	0.26220	0.24292
N	154	154	122	122

Country dummies are included in the estimations but not reported here. Italicized variables are error-corrected as follows: (model 1) policy shift (dependent variable) = .36606, previous policy shift = .36988, interaction of previous policy shift and previous change in vote share = .36988; (model 2) centrist policy shift = 0.19421, noncentrist policy shift = 0.15219, party policy convergence = 0.73376, interaction of niche party and centrist policy shift = 0.0603, interaction of niche party and noncentrist policy shift = 0.03977, interaction of party policy convergence and peripheral party = 0.33419. Coefficients in bold are statistically significant at the $p \leq .05$ level; SIMEX standard errors are based on jackknife estimation; parenthetical values are standard errors.

dimension will be explained by the exogenous left-right policy positions, while exogenous policy positions on the EU Integration dimension explain national party ideal point estimates on the second dimension (501). The expectation then is roughly that the first dimension is predominantly about left-right and the second dimension is about Europe.

Table 3 contrasts coefficients from our replications of the models using CMP variables in Hix, Noury, and Roland (2006) with error-corrected measurements based on our bootstrapped variances. (Due to space constraints we present replications of only the two models that related to the structure of the first dimension in the European Parliament.) Model 3 aims to explain the mean positioning of political parties on the first derived EP dimension in terms of their positions on the economic left-right, social left-right, and European Integration dimensions; categorical variables relating to whether a party was in government and had a European Commissioner; and dummy variables for each session of the EP. Model 6 extends Model 3 to also include dummy variables for each European party group.

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

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

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

Concluding Remarks and Recommendations

Bodies of text are data. We can analyze these data using well-known statistical tools. The implications of this are deep and general. Our discussions in this article apply to the analysis of most bodies of text, and in particular to analyses of text based on interpretative coding by human experts. While we focus here on text observed in party manifestos and analyzed by the CMP, the problems we identify and set out to correct apply to any dataset based on human interpretative coding. Our focus on the CMP reflects the very widespread use of this dataset within the profession, generating a large number of publications in

TABLE 3 Results of SIMEX Error Correction in Hix, Noury, and Roland (2006, 503–504, Table 4)

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

All models include dummies for parliament, but these are not shown. Italicized variables are error-corrected as follows: social L-R = 1.9907, economic L-R = 1.88742, EU integration = 1.69393. Coefficients in bold are statistically significant at the $p \leq .05$ level; SIMEX standard errors are based on jackknife estimation; parenthetical values are standard errors.

the best professional journals. These publications never take account of the fact that the data analyzed clearly contain measurement error and that this measurement error can clearly bias research findings.

We approach this problem by considering ways in which manifestos provide systematic information about the policy positions of their authors, in the form of text units deposited as random variables in a process of authorship that is inherently stochastic, even when the author's underlying position is fixed. We simulate this process, thereby computing error estimates for the entire CMP dataset, and show how such errors affect

descriptive and causal inferences based on CMP measures. Building on this method, we offer a "corrected" version of the CMP dataset with bootstrapped standard errors for all key estimates, available our website from http://www.politics.tcd.ie/cmp/.

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

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

The importance of estimating and making use of uncertainty in political science data, of course, is not limited to manifesto coding and the CMP dataset. Many commonly used measurements, such as survey data, roll-call votes, expert surveys of party policy (Benoit and Laver 2006), categories of legislation (e.g., Mayhew 1991), the democraticness of regime type (see Bollen and Jackman 1989), and a myriad of other commonly used variables are measured with levels of error. Even when estimates of measurement error are provided—as is the case with surveys, expert surveys, and more recently, roll-call votes (e.g., Clinton, Jackman, and Rivers 2004)—political scientists rarely, if ever, make use of these estimates in the ways we encourage here.

While we have taken an important first step towards providing a practical and theoretically supported means to estimate nonsystematic measurement error in CMP estimates, the solution we provide here is hardly the last word on the topic. Analyses of coder differences and/or coder error, for example, could uncover systematic error leading to bias, something not addressed in this article but acknowledged to be a problem warranting serious attention. In ongoing work using coder experiments with multiple independent codings of the same texts, we have found strong evidence that coding is not only stochastic but also appears to suffer from systematic forms of error. While we have chosen to focus purely on nonsystematic error in this article, a full accounting for error in the CMP ought to consider both stochastic features of textual data as well as systematic and nonsystematic errors from the coding of that text. Finally, other means of implementing error correction models are certainly possible, including Bayesian-MCMC methods that can take into account the unit-specific nature of error in our error estimates. Indeed, we hope our focus on error in the widely used CMP estimates will stimulate a broader dialogue on measurement error in many of the most commonly used measures in political science, such as opinion survey results, expert survey measures, or other computed quantities. Given our knowledge of measurement error and the wide availability of techniques for dealing with this, there is no longer any excuse for scholars to use error-prone measures as if these were error free.

References

Abrevaya, J., and J. A. Hausman. 2004. "Response Error in a Transformation Model with an Application to Earnings-Equation Estimation." *The Econometrics Journal* 7(2): 366–88.

Adams, James, M. Clark, L. Ezrow, and G. Glasgow. 2006. "Are Niche Parties Fundamentally Different from Mainstream Parties? The Causes and the Electoral Consequences of Western European Parties' Policy Shifts, 1976–1998." *American Journal of Political Science* 50(3): 513–29.

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

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

Benoit, Kenneth, and Michael Laver. 2007. "Benchmarks for Text Analysis: A Reply to Budge and Pennings." *Electoral Studies* 26: 130–35.

Bollen, Kenneth A., and Robert W. Jackman. 1989. "Democracy, Stability, and Dichotomies." *American Sociological Review* 54(August): 612–21.

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

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

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

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Voting." *American Political Science Review* 98(2): 355–70.

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

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

Gabel, Matthew, and John Huber. 2000. "Putting Parties in Their Place: Inferring Party Left-right Ideological Positions from Party Manifesto Data." *American Journal of Political Science* 44: 94–103.

Hardin, J. W., H. Schmiediche, and R. J. Carroll. 2003. "The Simulation Extrapolation Method for Fitting Generalized Linear Models with Additive Measurement Error." *The STATA Journal* 3(4): 373–85.

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

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: Oxford University Press, 111–25.

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

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

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

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

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

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.

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): 311–31.

Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3): 619–34.

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

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

McDonald, Michael, and Silvia Mendes. 2001a. "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: Oxford University Press, 127–41.

McDonald, Michael, and Silvia Mendes. 2001b. "The Policy Space of Party Manifestos." In *Estimating the Policy Position of Political Actors*, ed. Michael Laver. London: Routledge, 90–114.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2008. "Coder Reliability and Misclassification in Comparative Manifesto Project Codings." Presented at the annual meeting of the Midwest Political Science Association.

Monroe, Burt, and Ko Maeda. 2004. "Talk's Cheap: Text-based Estimation of Rhetorical Ideal-Points." Working paper. Michigan State University.

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

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

Stefanski, L. A., and J. R. Cook. 1995. "Simulation-Extrapolation: The Measurement Error Jackknife." *Journal of the American Statistical Association* 90(432): 1247–56.

Volkens, Andrea. 2001. "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 Tanenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. Oxford: Oxford University Press, 93–109.