Crowd-sourced text analysis: reproducible and agile production of political data

CROWD-SOURCED TEXT ANALYSIS: REPRODUCIBLE AND AGILE PRODUCTION OF POLITICAL DATA*

Kenneth Benoit
London School of Economics
and Trinity College, Dublin

Drew Conway
New York University

Benjamin E. Lauderdale
London School of Economics

Michael Laver
New York University

Slava Mikhaylov
University College London

June, 2015

Abstract

Empirical social science often relies on data that are not observed in the field, but are transformed into quantitative variables by expert researchers who analyze and interpret qualitative raw sources. While generally considered the most valid way to produce data, this expert-driven process is inherently difficult to replicate or to assess on grounds of reliability. Using crowd-sourcing to distribute text for reading and interpretation by massive numbers of non-experts, we generate results comparable to those using experts to read and interpret the same texts, but do so far more quickly and flexibly. Crucially, the data we collect can be reproduced and extended transparently, making crowd-sourced datasets intrinsically reproducible. This focuses researchers' attention on the fundamental scientific objective of specifying reliable and replicable methods for collecting the data needed, rather than on the content of any particular dataset. We also show that our approach works straightforwardly with different types of political text, written in different languages. While findings reported here concern text analysis, they have far-reaching implications for expert-generated data in the social sciences.

* An earlier draft of this paper, with much less complete data, was presented at the third annual *Analyzing Text as Data* conference at Harvard University, 5-6 October 2012. A very preliminary version was presented at the 70th annual Conference of the Midwest Political Science Association, Chicago, 12-15 April 2012. We thank Joseph Childress and other members of the technical support team at CrowdFlower for assisting with the setup of the crowd-sourcing platform. We are grateful to Neal Beck, Joshua Tucker and five anonymous journal referees for comments on an earlier draft of this paper. This research was funded by the European Research Council grant ERC-2011-StG 283794-QUANTESS.

Political scientists have made great strides toward greater reproducibility of their findings since the publication of Gary King's influential paper *Replication, Replication* (King 1995). It is now standard practice for good professional journals to insist that authors lodge their data and code in a prominent open access repository. This allows other scholars to replicate and extend published results by reanalyzing the data, rerunning and modifying the code. Replication of an *analysis*, however, sets a far weaker standard than reproducibility of the *data*, which is typically seen as a fundamental principle of the scientific method. Here, we propose a step towards a more comprehensive scientific replication standard in which the mandate is to replicate data production, not just data analysis. This shifts attention from specific datasets as the essential scientific objects of interest, to the *published and reproducible method by which the data were generated.*

We implement this more comprehensive replication standard for the rapidly expanding project of analyzing the content of political texts. Traditionally, a lot of political data is generated by experts applying comprehensive classification schemes to raw sources in a process that, while in principle repeatable, is in practice too costly and time-consuming to reproduce. Widely used examples include:[1] the *Polity* dataset, rating countries on a scale "ranging from -10 (hereditary monarchy) to +10 (consolidated democracy)";[2] the *Comparative Parliamentary Democracy* data with indicators, of the "number of inconclusive bargaining rounds" in government formation and "conflictual" government terminations[3]; the *Comparative Manifesto Project* (CMP), with coded summaries of party manifestos, notably a widely-used left-right score[4]; and the *Policy Agendas*

[1] Other examples of coded data include: expert judgments on party policy positions of party positions (Benoit and Laver 2006; Hooghe et al. 2010; Laver and Hunt 1992); democracy scores from *Freedom House* and corruption rankings from *Transparency International*.
[2] http://www.systemicpeace.org/polity/polity4.htm
[3] http://www.erdda.se/cpd/data_archive.html
[4] https://manifesto-project.wzb.eu/

Project, which codes text from laws, court decisions, political speeches into topics and subtopics (Jones 2013). In addition to the issue of reproducibility, the fixed nature of these schemes and the considerable infrastructure required to implement them discourages change and makes it harder to adapt them to specific needs, as the data are designed to fit general requirements rather than a particular research question.

Here, we demonstrate a method of crowd-sourced text annotation for generating political data that is both *reproducible* in the sense of allowing the data generating process to be quickly, inexpensively, and reliably repeated, and *agile* in the sense of being capable of flexible design according to the needs of a specific research project. The notion of agile research is borrowed from recent approaches to software development, and incorporates not only the flexibility of design, but also the ability to iteratively test, deploy, verify, and if necessary, redesign data generation through feedback in the production process. In what follows, we apply this method to a common measurement problem in political science: locating political parties on policy dimensions using text as data. Despite the lower expertise of crowd workers compared to experts, we show that properly deployed crowd-sourcing generates results indistinguishable from expert approaches. Given the millions of available workers online, crowd sourced-data collection can also be *repeated* as often as desired, quickly and with low cost.. Furthermore, our approach is easily tailored to specific research needs, for specific contexts and time periods, in sharp contrast to large "canonical" data generation projects aimed at maximizing generality. For this reason, crowd-sourced data generation may represent a paradigm shift for data production and reproducibility in the social sciences. While, as a proof of concept, we apply our particular method for crowd-sourced data production to the analysis of political texts, the core problem of specifying *a reproducible data production process* extends to all subfields of political science.

In what follows, we first review the theory and practice of crowd sourcing. We then deploy an experiment in content analysis designed to evaluate crowd sourcing as a method for reliably and validly extracting meaning from political texts, in this case party manifestos. We compare expert and crowd-sourced analyses of the same texts, and assess external validity by comparing crowd-sourced estimates with those generated by completely independent expert surveys. In order to do this, we design a method for aggregating judgments about text units of varying complexity, by readers of varying quality,[5] into estimates of latent quantities of interest. To assess the external validity of our results, our core analysis uses crowd workers to estimate party positions on two widely used policy dimensions: "economic" policy (right-left) and "social" policy (liberal-conservative). We then use our method to generate "custom" data on a variable not available in canonical datasets, in this case party policies on immigration. Finally, to illustrate the general applicability of crowd-sourced text annotation in political science, we test the method in a multi-lingual and technical environment to show that crowd-sourced text analysis is effective for texts other than party manifestos and works well in different languages.

HARVESTING THE WISDOM OF CROWDS

The intuition behind crowd-sourcing can be traced to Aristotle (Lyon and Pacuit 2013) and later Galton (1907), who noticed that the average of a large number of individual judgments by fair-goers of the weight of an ox is close to the true answer and, importantly, closer to this than the typical individual judgment (for a general introduction see Surowiecki 2004). Crowd-sourcing is now understood to mean using the Internet to distribute a large package of small tasks to a large number of anonymous workers, located around the world and offered small financial rewards per

[5] In what follows we use the term "reader" to cover a person, whether expert, crowd worker or anyone else, who is evaluating a text unit for meaning.

task. The method is widely used for data-processing tasks such as image classification, video annotation, data entry, optical character recognition, translation, recommendation, and proofreading. Crowd-sourcing has emerged as a paradigm for applying human intelligence to problem-solving on a massive scale, especially for problems involving the nuances of language or other interpretative tasks where humans excel but machines perform poorly.

Increasingly, crowd-sourcing has also become a tool for social scientific research (Bohannon 2011). In sharp contrast to our own approach, most applications use crowds as a cheap alternative to traditional subjects for experimental studies (e.g. Lawson et al. 2010; Horton et al. 2011; Paolacci et al. 2010; Mason and Suri 2012). Using subjects in the crowd to populate experimental or survey panels raises obvious questions about external validity, addressed by studies in political science (Berinsky et al. 2012), economics (Horton et al. 2011) and general decision theory and behavior (Paolacci et al. 2010; Goodman et al. 2013; Chandler et al. 2014). Our method for using workers in the crowd to label *external* stimuli differs fundamentally from such applications. We do not care at all about whether our crowd workers represent any target population, as long as different workers, on average, make the same judgments when faced with the same information. In this sense our method, unlike online experiments and surveys, is a canonical use of crowd-sourcing as described by Galton.[6]

All data production by humans requires expertise, and several empirical studies have found that data created by domain experts can be matched, and sometimes improved at much lower cost, by aggregating judgments of non-experts (Alonso and Mizzaro 2009; Hsueh et al. 2009; Snow et al. 2008; Alonso and Baeza-Yates 2011; Carpenter 2008; Ipeirotis et al. 2013). Provided crowd workers are not systematically biased in relation to the "true" value of the latent quantity of interest, and it is important to check for such bias, the central tendency of even erratic

[6] We are interested in the weight of the ox, not in how different people judge the weight of the ox.

workers will converge on this true value as the number of workers increases. Because experts are axiomatically in short supply while members of the crowd are not, crowd-sourced solutions also offer a straightforward *and scalable* way to address reliability in a manner that expert solutions cannot. To improve confidence, simply employ more crowd workers. Because data production is broken down into many simple specific tasks, each performed by many different exchangeable workers, it tends to wash out biases that might affect a single worker, while also making it possible to estimate and correct for worker-specific effects using the type of scaling model we employ below.

Crowd-sourced data generation inherently requires a method for aggregating many small pieces of information into valid measures of our quantities of interest.[7] Complex calibration models have been used to correct for worker errors on particular difficult tasks, but the most important lesson from this work is that increasing the number of workers reduces error (Snow et al. 2008). Addressing statistical issues of "redundant" coding, Sheng et al. (2008) and Ipeirotis et al. (2010) show that repeated coding can improve the quality of data as a function of the individual qualities and number of workers, particularly when workers are imperfect and labeling categories are "noisy." Ideally, we would benchmark crowd workers against a "gold standard," but such benchmarks are not always available, so scholars have turned to Bayesian scaling models borrowed from item-response theory (IRT), to aggregate information while simultaneously assessing worker quality (e.g. Carpenter 2008; Raykar et al. 2010). Welinder and Perona (2010) develop a classifier that integrates data difficulty and worker characteristics, while Welinder et al. (2010) develop a unifying model of the characteristics of both data and workers, such as competence, expertise and bias. A similar approach is applied to rater evaluation in Cao

[7] Of course aggregation issues are no less important when combining any multiple judgments, including those of experts. Procedures for aggregating non-expert judgments may influence both the quality of data and convergence on some underlying "truth", or trusted expert judgment. For an overview, see Quoc Viet Hung et al. (2013).

et al. (2010) where, using a Bayesian hierarchical model, raters' judgments are modeled as a function of a latent item trait, and rater characteristics such as bias, discrimination, and measurement error. We build on this work below, applying both a simple averaging method and a Bayesian scaling model that estimates latent policy positions while generating diagnostics on worker quality and sentence difficulty. We find that estimates generated by our more complex model match simple averaging very closely.

A METHOD FOR REPLICABLE CODING OF POLITICAL TEXT

We apply our crowd-sourcing method to one of the most wide-ranging research programs in political science, the analysis of political text, and in particular text processing by *human* analysts that is designed to extract meaning systematically from some text corpus, and from this to generate valid and reliable data. This is related to, but quite distinct from, spectacular recent advances in *automated* text analysis that in theory scale up to unlimited volumes of political text (Grimmer and Stewart 2013). Many automated methods involve *supervised machine learning* and depend on labeled training data. Our method is directly relevant to this enterprise, offering a quick, effective and above all *reproducible* way to generate labeled training data. Other, *unsupervised*, methods intrinsically require *a posteriori* human interpretation that may be haphazard and is potentially biased.[8]

Our argument here speaks directly to more traditional content analysis within the social sciences, which is concerned with problems that automated text analysis cannot yet address. This involves the "reading" of text by real humans who interpret it for meaning. These interpretations, if systematic, may be classified and summarized using numbers, but the underlying human interpretation is fundamentally qualitative. Crudely, human analysts are employed to engage in

[8] This human interpretation can be reproduced by workers in the crowd, though this is not our focus in this paper.

natural language processing (NLP) which seeks to extract "meaning" embedded in the syntax of language, treating a text as more than a bag of words. NLP is another remarkable growth area, though it addresses a fundamentally difficult problem and fully automated NLP still has a long way to go. Traditional human experts in the field of inquiry are of course highly sophisticated natural language processors, finely tuned to particular contexts. The core problem is that they are in very short supply. This means that text processing by human experts simply does not scale to the huge volumes of text that are now available. This in turn generates an inherent difficulty in meeting the more comprehensive scientific replication standard to which we aspire. Crowd-sourced text analysis offers a compelling solution to this problem. Human workers in the crowd can be seen, perhaps rudely, as generic and very widely available "biological" natural language processors. Our task in this paper is now clear. Design a system for employing generic workers in the crowd to analyze text for meaning in a way that is as reliable and valid as if we had used finely tuned experts to do the same job.

By far the best known research program in political science that relies on expert human readers is the long-running *Manifesto Project* (MP). This project has analyzed nearly 4,000 manifestos issued since 1945 by nearly 1,000 parties in more than 50 countries, using experts who are country specialists to label sentences in each text in their original languages. A single expert assigns every sentence in every manifesto to a single category in a 56-category scheme devised by the project in the mid-1980s (Budge et al. 1987; Laver and Budge 1992; Klingemann et al. 1994; Budge et al. 2001; Klingemann et al. 2006).[9] This has resulted in a widely-used "canonical" dataset that, given the monumental coordinated effort of very many experts over 30 years, is unlikely ever to be re-collected from scratch and in this sense is unlikely to be replicated. Despite low levels of inter-expert reliability found in experiments using the MP's

[9] https://manifesto-project.wzb.eu/

coding scheme (Mikhaylov et al. 2012), a proposal to re-process the entire manifesto corpus many times, using many independent experts, is in practice a non-starter. Large canonical datasets such as this, therefore, tend not to satisfy the deeper standard of reproducible research that requires the transparent repeatability of data generation. This deeper replication standard can however be satisfied with the crowd-sourced method we now describe.

A simple coding scheme for economic and social policy

We assess the potential for crowd-sourced text analysis using an experiment in which we serve up an identical set of documents, and an identical set of text processing tasks, to both a small set of experts (political science faculty and graduate students) and a large and heterogeneous set of crowd workers located around the world. To do this, we need a simple scheme for labeling political text that can be used reliably by workers in the crowd. Our scheme first asks readers to classify each sentence in a document as referring to economic policy (left or right), to social policy (liberal or conservative), or to neither. Substantively, these two policy dimensions have been shown to offer an efficient representation of party positions in many countries.[10] They also correspond to dimensions covered by a series of expert surveys (Benoit and Laver 2006; Hooghe et al. 2010; Laver and Hunt 1992), allowing validation of estimates we derive against widely used independent estimates of the same quantities. If a sentence was classified as economic policy, we then ask readers to rate it on a five-point scale from very left to very right; those classified as social policy were rated on a five-point scale from liberal to conservative. Figure 1 shows this scheme.[11]

[10] See Chapter 5 of Benoit and Laver (2006) for an extensive empirical review of this for a wide range of contemporary democracies.

[11] Our instructions—fully detailed in the supplementary materials (section 6)—were identical for both expert and non-experts, defining the economic left-right and social liberal-conservative policy dimensions we estimate and providing examples of labeled sentences.

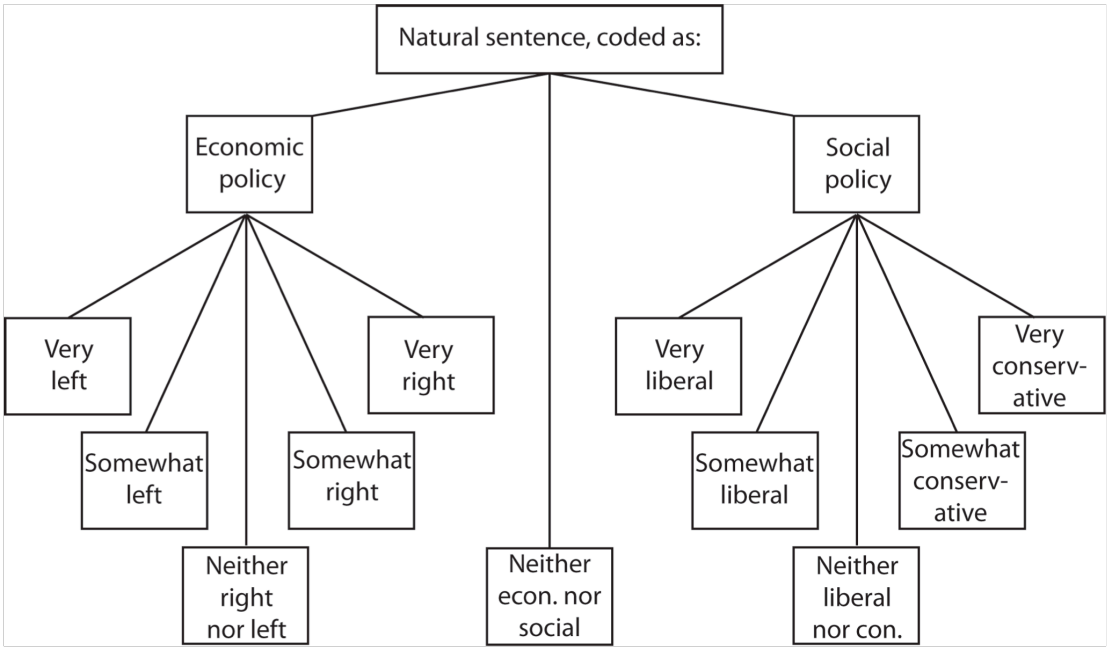

Figure 1: Hierarchical coding scheme for two policy domains with ordinal positioning.

We did not use the MP's 56-category classification scheme, for two main reasons. The first is methodological: complexity of the MP scheme and uncertain boundaries between many of its categories were major sources of unreliability when multiple experts applied this scheme to the same documents (Mikhaylov et al. 2012). The second is practical: it is impossible to write clear and precise instructions, to be understood reliably by a diverse, globally distributed, set of workers in the crowd, for using a detailed and complex 56-category scheme quintessentially designed for highly trained experts. This highlights an important trade-off. There may be data production tasks that cannot feasibly be explained in clear and simple terms, sophisticated instructions that can only be understood and implemented by highly trained experts. Sophisticated instructions are designed for a more limited pool of experts who can understand and implement them and, for this reason, imply less scalable and replicable data production. Such tasks may not be suitable for crowd-sourced data generation and may be more suited to traditional methods. The striking alternative now made available by crowd-sourcing is to break

down complicated data production tasks into simple small jobs, as happens when complex consumer products are manufactured on factory production lines. Over and above the practical need to have simple instructions for crowd workers, furthermore, the scheme in Figure 1 is motivated by the observation that most scholars using manifesto data actually seek simple solutions, typically estimates of positions on a few general policy dimensions; they do not need estimates of these positions in a 56-dimensional space.

Text corpus

While we extend this in work we discuss below, our baseline text corpus comprises 18,263 natural sentences from British Conservative, Labour and Liberal Democrat manifestos for the six general elections held between 1987 and 2010. These texts were chosen for two main reasons. First, for systematic external validation, there are diverse independent estimates of British party positions for this period, from contemporary expert surveys (Laver and Hunt 1992; Laver 1998; Benoit 2005, 2010) as well as MP expert codings of the same texts. Second, there are well-documented substantive shifts in party positions during this period, notably the sharp shift of Labour towards the center between 1987 and 1997. The ability of crowd workers to pick up this move is a good test of external validity.

In designing the breakdown and presentation of the text processing tasks given to both experts and the crowd, we made a series of detailed operational decisions based on substantial testing and adaptation (reviewed in the Appendix). In summary, we used natural sentences as our fundamental text unit. Recognizing that most crowd workers dip into and out of our jobs and would not stay online to code entire documents, we served target sentences from the corpus in a random sequence, set in a two-sentence context on either side of the target sentence, without identifying the text from which the sentence was drawn. Our coding experiments showed that

these decisions resulted in estimates that did not significantly differ from those generated by the classical approach of reading entire documents from beginning to end.

SCALING DOCUMENT POLICY POSITIONS FROM CODED SENTENCES

Our aim is to estimate the policy positions of entire documents: not the code value of any single sentence, but some aggregation of these values into an estimate of each document's position on some meaningful policy scale while allowing for reader, sentence, and domain effects. One option is simple averaging: identify all economic scores assigned to sentences in a document by all readers, average these, and use this as an estimate of the economic policy position of a document. Mathematical and behavioral studies on aggregations of individual judgments imply that simpler methods often perform as well as more complicated ones, and often more robustly (e.g. Ariely et al. 2000; Clemen and Winkler 1999). Simple averaging of individual judgments is the benchmark when there is no additional information on the quality of individual coders (Lyon and Pacuit 2013; Armstrong 2001; Turner et al. 2013). However, this does not permit direct estimation of misclassification tendencies by readers who for example fail to identify economic or social policy "correctly," or of reader-specific effects in the use of positional scales.

An alternative is to model each sentence as containing information about the document, and then scale these using a measurement model. We propose a model based on item response theory (IRT), which accounts for both individual reader effects and the strong possibility that some sentences are intrinsically harder to interpret. This approach has antecedents in psychometric methods (e.g. Baker and Kim 2004; Fox 2010; Hambleton et al. 1991; Lord 1980), and has been used to aggregate crowd ratings (e.g. Ipeirotis et al. 2010; Welinder et al. 2010; Welinder and Perona 2010; Whitehill et al. 2009).

We model each sentence, j, as a vector of parameters, θ_{jd}, which corresponds to sentence attributes on each of four latent dimensions, d. In our application, these dimensions are: latent *domain propensity* of a sentence to be labeled economic (1) and social (2) versus none; latent *position* of the sentence on economic (3) and social (4) dimensions. Individual readers i have potential *biases* in each of these dimensions, manifested when classifying sentences as "economic" or "social", and when assigning positions on economic and social policy scales. Finally, readers have four *sensitivities*, corresponding to their relative responsiveness to changes in the latent sentence attributes in each dimension. Thus, the latent coding of sentence j by reader i on dimension d is:

$$\mu^*_{ijd} = \chi_{id}\left(\theta_{jd} + \psi_{id}\right) \qquad (1)$$

where the χ_{id} indicate relative *responsiveness* of readers to changes in latent sentence attributes θ_{jd}, and the ψ_{id} indicate relative *biases* towards labeling sentences as economic or social ($d = 1,2$), and rating economic and social sentences as right rather than left ($d = 3,4$).

We cannot observe readers' behavior on these dimensions directly. We therefore model their responses to the choice of label between economic, social and "neither" domains using a multinomial logit given μ^*_{ij1} and μ^*_{ij2}. We model their choice of scale position as an ordinal logit depending on μ^*_{ij3} if they label the sentence as economic and on μ^*_{ij4} if they label the sentence as social.[12] This results in the following model for the eleven possible combinations of labels and scales that a reader can give a sentence:[13]

[12] By treating these as independent, and using the logit, we are assuming independence between the choices and between the social and economic dimensions (IIA). It is not possible to identify a more general model that relaxes these assumptions without asking additional questions of readers.

[13] Each policy domain has five *scale* points, and the model assumes proportional odds of being in each higher scale category in response to the sentence's latent policy positions θ_3 and θ_4 and the coder's sensitivities to this association. The cutpoints ξ for ordinal scale responses are constrained to be symmetric around zero and to have the

$$p(none) = \left(\frac{1}{1 + \exp(\mu^*_{ij1}) + \exp(\mu^*_{ij2})}\right)$$

$$p(econ; scale) = \left(\frac{\exp(\mu^*_{ij1})}{1 + \exp(\mu^*_{ij1}) + \exp(\mu^*_{ij2})}\right)\left(logit^{-1}(\xi_{scale} - \mu^*_{ij3}) - logit^{-1}(\xi_{scale-1} - \mu^*_{ij3})\right)$$

$$p(soc; scale) = \left(\frac{\exp(\mu^*_{ij2})}{1 + \exp(\mu^*_{ij1}) + \exp(\mu^*_{ij2})}\right)\left(logit^{-1}(\xi_{scale} - \mu^*_{ij4}) - logit^{-1}(\xi_{scale-1} - \mu^*_{ij4})\right)$$

The primary quantities of interest are not sentence level attributes, θ_{jd}, but rather aggregates of these for entire documents, represented by the $\bar{\theta}_{k,d}$ for each document *k* on each dimension *d*. Where ϵ_{jd} are distributed normally with mean zero and standard deviation σ_d, we model these latent sentence level attributes θ_{jd} hierarchically in terms of corresponding latent document level attributes:

$$\theta_{jd} = \bar{\theta}_{k(j),d} + \epsilon_{jd}$$

As at the sentence level, two of these (*d=1,2*) correspond to the overall frequency (importance) of economic and social dimensions relative to other topics, the remaining two (*d=3,4*) correspond to aggregate left-right positions of documents on economic and social dimensions.

This model enables us to generate estimates of not only our quantities of interest for the document-level policy positions, but also a variety of reader- and sentence- level diagnostics concerning reader agreement and the "difficulty" of domain and positional coding for individual sentences. Simulating from the posterior also makes it straightforward to estimate Bayesian credible intervals indicating our uncertainty over document-level policy estimates.[14]

---

same cutoffs in both social and economic dimensions, so that the latent scales are directly comparable to one another and to the raw scales. Thus, $\xi_2 = \infty$, $\xi_1 = -\xi_{-2}$, $\xi_0 = -\xi_{-1}$, and $\xi_{-3} = -\infty$.

[14] We estimate the model by MCMC using the JAGS software, and provide the code, convergence diagnostics, and other details of our estimations in section 2 of the supplementary materials.

Posterior means of the document level $\bar{\theta}_{kd}$ correlate very highly with those produced by the simple averaging methods discussed earlier: 0.95 and above, as we report below. It is therefore possible to use averaging methods to summarize results in a simple and intuitive way that is also invariant to shifts in mean document scores that might be generated by adding new documents to the coded corpus. The value of our scaling model is to estimate reader and sentence fixed effects, and correct for these if necessary. While this model is adapted to our particular classification scheme, it is general in the sense that nearly all attempts to measure policy in specific documents will combine domain classification with positional coding.

BENCHMARKING A CROWD OF EXPERTS

Our core objective is to compare estimates generated by workers in the crowd with analogous estimates generated by experts. Since readers of all types will likely disagree over the meaning of particular sentences, an important benchmark for our comparison of expert and crowd-sourced text coding concerns levels of disagreement between experts. The first stage of our empirical work therefore employed multiple (four to six)[15] experts to independently code each of the 18,263 sentences in our 18-document text corpus, using the scheme described above. The entire corpus was processed twice by our experts. First, sentences were served in their natural sequence in each manifesto, to mimic classical expert content analysis. Second, about a year later, sentences were processed in random order, to mimic the system we use for serving sentences to crowd workers. Sentences were uploaded to a custom-built, web-based platform that displayed sentences in context and made it easy for experts to process a sentence with a few mouse clicks. In all, we harvested over 123,000 expert evaluations of manifesto sentences, about seven per

[15] Three of the authors of this paper, plus three senior PhD students in Politics from New York University processed the six manifestos from 1987 and 1997. One author of this paper and four NYU PhD students processed the other 12 manifestos.

sentence. Table 1 provides details of the 18 texts, with statistics on the overall and mean numbers of evaluations, for both stages of expert processing as well as the crowd processing we report below.

Manifesto	Total sentences in manifesto	Mean expert evaluations: natural sequence	Mean expert evaluations: random sequence	Total expert evaluations	Mean crowd evaluations	Total crowd evaluations
Con 1987	1,015	6.0	2.4	7,920	44	36,594
LD 1987	878	6.0	2.3	6,795	22	24,842
Lab 1987	455	6.0	2.3	3,500	20	11,087
Con 1992	1,731	5.0	2.4	11,715	6	28,949
LD 1992	884	5.0	2.4	6,013	6	20,880
Lab 1992	661	5.0	2.3	4,449	6	23,328
Con 1997	1,171	6.0	2.3	9,107	20	11,136
LD 1997	873	6.0	2.4	6,847	20	5,627
Lab 1997	1,052	6.0	2.3	8,201	20	4,247
Con 2001	748	5.0	2.3	5,029	5	3,796
LD 2001	1,178	5.0	2.4	7,996	5	5,987
Lab 2001	1,752	5.0	2.4	11,861	5	8,856
Con 2005	414	5.0	2.3	2,793	5	2,128
LD 2005	821	4.1	2.3	4,841	5	4,173
Lab 2005	1,186	4.0	2.4	6,881	5	6,021
Con 2010	1,240	4.0	2.3	7,142	5	6,269
LD 2010	855	4.0	2.4	4,934	5	4,344
Lab 2010	1,349	4.0	2.3	7,768	5	6,843
Total	18,263	91,400	32,392	123,792		215,107

Table 1. Texts and sentences coded: 18 British party manifestos

External validity of expert evaluations

Figure 2 plots two sets of estimates of positions of the 18 manifestos on economic and social policy: one generated by experts processing sentences in natural sequence (*x*-axis); the other generated by completely independent expert surveys (*y*-axis).[16] Linear regression lines summarizing these plots show that expert text processing predicts independent survey measures

[16] These were: Laver and Hunt (1992); Laver (1998) for 1997; Benoit and Laver (2006) for 2001; Benoit (2005, 2010) for 2005 and 2010.

very well for economic policy (R= 0.91), somewhat less well for the noisier dimension of social policy (R=0.81). To test whether coding sentences in their natural sequence affected results, our experts also processed the entire text corpus taking sentences in random order. Comparing estimates from sequential and random-order sentence processing, we found almost identical results, with correlations of 0.98 between scales.[17] Moving from "classical" expert content analysis to having experts process sentences served at random from anonymized texts makes no substantive difference to point estimates of manifesto positions. This reinforces our decision to use the much more scalable random sentence sequencing in the crowd-sourcing method we specify.

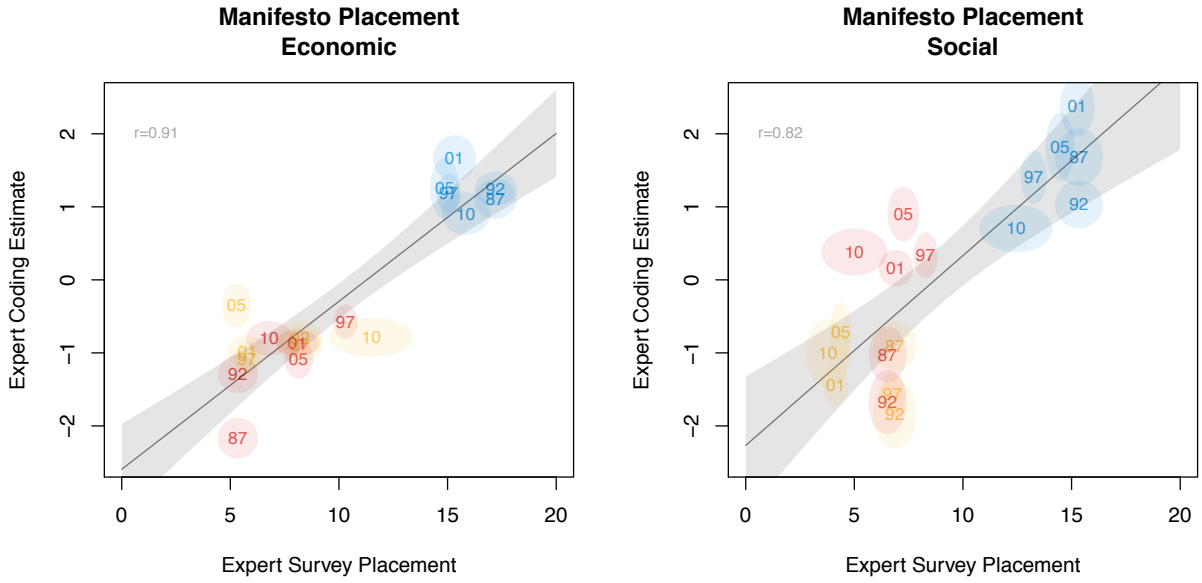

Figure 2. British party positions on economic and social policy 1987 – 2010; sequential expert text processing (vertical axis) and independent expert surveys (horizontal).
(Labour red, Conservatives blue, Liberal Democrats yellow, labeled by last two digits of year)

[17] Details provided in supplementary materials, section 5.

Internal reliability of expert coding

Agreement between experts

As might be expected, agreement between our experts was far from perfect. Table 2 classifies each of the 5,444 sentences in the 1987 and 1997 manifestos, all of which were processed by the same six experts. It shows how many experts agreed the sentence referred to economic, or social, policy. If experts are in perfect agreement on the policy content of each sentence, either all six label each sentence as dealing with economic (or social) policy, or none do. The first data column of the table shows a total of 4,125 sentences which all experts agree have no social policy content. Of these, there are 1,193 sentences all experts also agree have no economic policy content, and 527 that all experts agree do have economic policy content. The experts disagree about the remaining 2,405 sentences: some but not all experts label these as having economic policy content.

Experts Assigning Economic Domain	*Experts Assigning Social Policy Domain*							
	0	1	2	3	4	5	6	Total
0	**1,193**	196	67	59	114	190	**170**	1,989
1	326	93	19	11	9	19	-	477
2	371	92	15	15	5	-	-	498
3	421	117	12	7	-	-	-	557
4	723	68	10	-	-	-	-	801
5	564	31	-	-	-	-	-	595
6	**527**	-	-	-	-	-	-	527
Total	4,125	597	123	92	128	209	170	5,444

Table 2: Domain classification matrix for 1987 and 1997 manifestos: frequency with which sentences were assigned by six experts to economic and policy domains. (Shaded boxes: perfect agreement between experts.)

The shaded boxes show sentences for which the six experts were in unanimous agreement – on economic policy, social policy, or neither. There was unanimous expert agreement on about 35

percent of the labeled sentences. For about 65 percent of sentences, there was disagreement, even about the policy area, among trained experts of the type usually used to analyze political texts.

Scale reliability

Despite substantial disagreement among experts about individual sentences, we saw above that we can derive externally valid estimates of party policy positions if we aggregate the judgments of all experts on all sentences in a given document. This happens because, while each expert judgment on each sentence is a noisy realization of some underlying signal about policy content, the expert judgments taken as a whole scale nicely – in the sense that in aggregate they are all capturing information about the same underlying quantity. Table 3 shows this, reporting a scale and coding reliability analysis for economic policy positions of the 1987 and 1997 manifestos, derived by treating economic policy scores for each sentence allocated by each of the six expert coders as six sets of independent estimates of economic policy positions.

Item	*N*	*Sign*	*Item-scale correlation*	*Item-rest correlation*	*Cronbach's alpha*
Expert 1	2,256	+	0.89	0.76	0.95
Expert 2	2,137	+	0.89	0.76	0.94
Expert 3	1,030	+	0.87	0.74	0.94
Expert 4	1,627	+	0.89	0.75	0.95
Expert 5	1,979	+	0.89	0.77	0.95
Expert 6	667	+	0.89	0.81	0.93
Overall					0.95
k policy domain					0.93

Table 3. Inter-expert scale reliability analysis for the economic policy, generated by aggregating all expert scores for sentences judged to have economic policy content.

Despite the variance in expert coding of the policy domains as seen in Table 2, overall agreement as to the policy domain of sentences was 0.93 using Fleiss' kappa, a very high level of

inter-rater agreement (as κ ranges from 0-1.0).[18] A far more important benchmark of reliability, however, focuses on the construction of the scale resulting from combining the coders' judgments, which is of more direct interest than the codes assigned to any particular fragment of text. *Scale* reliability, as measured by a Cronbach's alpha of 0.95, is "excellent" by any conventional standard.[19] We can therefore apply our model to aggregate the noisy information contained in the combined set of expert judgment at the sentence level to produce coherent estimates of policy positions at the document level. This is the essence of crowd-sourcing. It shows that our experts are really a small crowd.

DEPLOYING CROWD-SOURCED TEXT CODING

CrowdFlower: a crowd-sourcing platform with multiple channels

Many online platforms now distribute crowd-sourced micro-tasks (Human Intelligence Tasks or "HITs") via the Internet. The best known is Amazon's Mechanical Turk (MT), an online marketplace for serving HITs to workers in the crowd. Workers must often pass a pre-task qualification test, and maintain a certain quality score from validated tasks that determines their status and qualification for future jobs. However, MT has for legal reasons become increasingly difficult to use for non-US researchers and workers, with the result that a wide range of alternative crowd-sourcing channels has opened up. Rather than relying on a single crowdsourcing channel, we used CrowdFlower, a service that consolidates access to dozens of

[18] Expert agreement for the random order coding as to the precise scoring of positions within the policy domains had κ =0.56 for a polarity scale (left, neutral, right) and κ =0.41 for the full five-point scale. For position scoring agreement rates can be estimated only roughly, however, as sentences might have been assigned different policy domains by different raters, and therefore be placed using a different positional scale.

[19] Conventionally, an alpha of 0.70 is considered "acceptable". Nearly identical results for social policy are available in supplementary materials (section 1d). Note that we use Cronbach's alpha as a measure of scale reliability across readers, as opposed to a measure of inter-reader agreement (in which case we would have used Krippendorf's alpha).

channels.[20] CrowdFlower not only offers an interface for designing templates and uploading tasks that look the same on any channel but, crucially, also maintains a common training and qualification system for potential workers from any channel before they can qualify for tasks, as well as cross-channel quality control while tasks are being completed.

Quality control

Excellent quality assurance is critical to all reliable and valid data production. Given the natural economic motivation of workers in the crowd to finish as many jobs in as short a time as possible, it is both tempting and easy for workers to submit bad or faked data. Workers who do this are called "spammers". Given the open nature of the platform, it is vital to prevent them from participating in a job, using careful screening and quality control (e.g. Kapelner and Chandler 2010; Nowak and Rger 2010; Eickhoff and de Vries 2012; Berinsky et al. forthcoming). Conway used coding experiments to assess three increasingly strict screening tests for workers in the crowd. (Conway 2013).[21] Two findings directly inform our design. First, using a screening or qualification test *substantially* improves the quality of results; a well-designed test can screen out spammers and bad workers who otherwise tend to exploit the job. Second, once a suitable test is in place, increasing its difficulty *does not* improve results. It is vital to have a filter on the front end to keep out spammers and bad workers, but a tougher filter does not necessarily lead to better workers.

[20] See http://www.crowdflower.com.

[21] There was a baseline test with no filter, a "low-threshold" filter where workers had to correctly code 4/6 sentences correctly, and a "high-threshold" filter that required 5/6 correct labels. A "correct" label means the sentence is labeled as having the same policy domain as that provided by a majority of expert coders. The intuition here is that tough tests also tend to scare away good workers.

The primary quality control system used by CrowdFlower relies on completion of "gold" HITs: tasks with unambiguous correct answers specified in advance.[22] Correctly performing "gold" tasks, which are both used in qualification tests and randomly sprinkled through the job, is used to monitor worker quality and block spammers and bad workers. We specified our own set of gold HITs as sentences for which there was unanimous expert agreement on both policy area (economic, social or neither), and policy direction (left or right, liberal or conservative), and seeded each job with the recommended proportion of about 10% "gold" sentences. We therefore used "natural" gold sentences occurring in our text corpus, but could also have used "artificial" gold, manufactured to represent archetypical economic or social policy statements. We also used a special type of gold sentences called "screeners", (Berinsky et al. forthcoming). These contained an exact instruction on how to label the sentence,[23] set in a natural two-sentence context, and are designed to ensure coders pay attention throughout the coding process.

Specifying gold sentences in this way, we implemented a two-stage process of quality control. First, workers were only allowed into the job if they correctly completed 8 out of 10 gold tasks in a qualification test.[24] Once workers are on the job and have seen at least four more gold sentences, they are given a "trust" score, which is simply the proportion of correctly labeled gold. If workers get too many gold HITs wrong, their trust level goes down. They are ejected from the job if their trust score falls below 0.8. The current trust score of a worker is recorded with each HIT, and can be used to weight the contribution of the relevant piece of information to some aggregate estimate. Our tests showed this weighting made no substantial difference, however, mainly because trust scores all tended to range in a tight interval around a mean of

[22] For CrowdFlower's formal definition of gold, see http://crowdflower.com/docs/gold.
[23] For example, "Please code this sentence as having economic policy content with a score of very right."
[24] Workers giving wrong labels to gold questions are given a short explanation of why they are wrong.

0.84.[25] *Many* more potential HITs than we use here were rejected as "untrusted", because the workers did not pass the qualification test, or because their trust score subsequently fell below the critical threshold. Workers are not paid for rejected HITs, giving them a strong incentive to perform tasks carefully, as they do not know which of these have been designated as gold for quality assurance. We have no hesitation in concluding that a system of thorough and continuous monitoring of worker quality is necessary for reliable and valid crowd sourced text analysis.

Deployment

We set up an interface on CrowdFlower that was nearly identical to our custom-designed expert web system and deployed this in two stages. First, we over-sampled all sentences in the 1987 and 1997 manifestos, because we wanted to determine the number of judgments per sentence needed to derive stable estimates of our quantities of interest. We served up sentences from the 1987 and 1997 manifestos until we obtained a minimum of 20 judgments per sentence. After analyzing the results to determine that our estimates of document scale positions converged on stable values once we had five judgments per sentence—in results we report below—we served the remaining manifestos until we reached five judgments per sentence. In all, we gathered 215,107 judgments by crowd workers of the 18,263 sentences in our 18 manifestos, employing a total of 1,488 different workers from 49 different countries. About 28 percent of these came from the US, 15 percent from the UK, 11 percent from India, and 5 percent each from Spain, Estonia, and Germany. The average worker processed about 145 sentences; most processed between 10 and 70 sentences, 44 workers processed over 1,000 sentences, and four processed over 5,000. [26]

[25] Our supplementary materials (section 4) report the distribution of trust scores from the complete set of crowd-codings by country of the worker and channel, in addition to results that scale the manifesto aggregate policy scores by the trust scores of the workers.

[26] Our final crowd-coded dataset was generated by deploying through a total of 26 CrowdFlower channels. The most common was Neodev (Neobux) (40%), followed by Mechanical Turk (18%), Bitcoinget (15%), Clixsense (13%),

CROWD-SOURCED ESTIMATES OF PARTY POLICY POSITIONS

Figure 3 plots crowd-sourced estimates of the economic and social policy positions of British party manifestos against estimates generated from analogous expert text processing.[27] The very high correlations of aggregate policy measures generated by crowd workers and experts suggest both are measuring the same latent quantities. Substantively, Figure 3 also shows that crowd workers identified the sharp rightwards shift of Labour between 1987 and 1997 on both economic and social policy, a shift identified by expert text processing and independent expert surveys. The standard errors of crowd-sourced estimates are higher for social than for economic policy, reflecting both the smaller number of manifesto sentences devoted to social policy and higher coder disagreement over the application of this policy domain.[28] Nonetheless Figure 3 summarizes our evidence that the crowd-sourced estimates of party policy positions can be used as substitutes for the expert estimates, which is our main concern in this paper.

---

and Prodege (Swagbucks) (6%). Opening up multiple worker channels also avoided the restriction imposed by Mechanical Turk in 2013 to limit the labor pool to workers based in the US and India. Full details along with the range of trust scores for coders from these platforms are presented in the supplementary materials (section 4).

[27] Full point estimates are provided in the supplementary materials, section 1.

[28] An alternative measure of correlation, Lin's concordance correlation coefficient (Lin 1989, 2000), measures correspondence as well covariation, if our objective is to match the values on the identity line, although for many reasons here it is not. The economic and social measures for Lin's coefficient are 0.95 and 0.84, respectively.

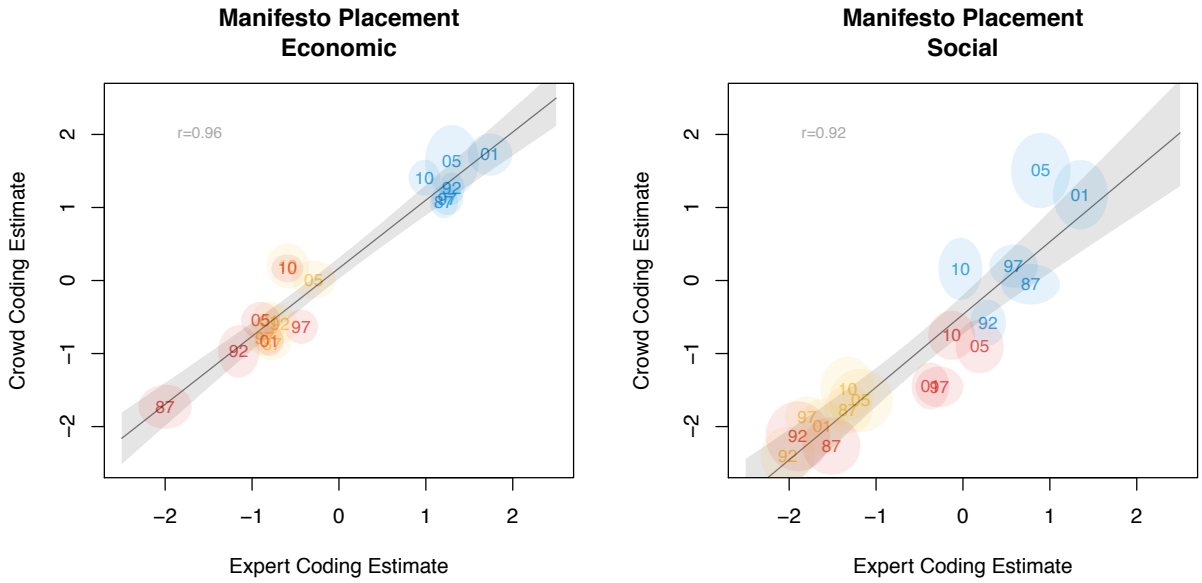

Figure 3. Expert and crowd-sourced estimates of economic and social policy positions.

Our scaling model provides a theoretically well-grounded way to aggregate all the information in our expert or crowd data, relating the underlying position of the political text both to the "difficulty" of a particular sentence and to a reader's propensity to identify the correct policy domain, and position within domain.[29] Because positions derived from the scaling model depend on parameters estimated using the full set of coders and codings, changes to the text corpus can affect the relative scaling. The simple mean of means method, however, is invariant to rescaling and always produces the same results, even for a single document. Comparing crowd-sourced estimates from the scaling model to those produced by a simple averaging of the mean of mean sentence scores, we find correlations of 0.96 for the economic and 0.97 for the social policy positions of the 18 manifestos. We present both methods as confirmation that our scaling method has not "manufactured" policy estimates. While this model does allow us to take proper account of reader and sentence fixed effects, it is also reassuring that a simple mean of means produced substantively similar estimates.

[29] We report more fully on diagnostic results for our coders on the basis of the auxiliary model quantity estimates in the supplementary materials (section 1e).

We have already seen that noisy expert judgments about sentences aggregate up to reliable and valid estimates for documents. Similarly, crowd-sourced document estimates reported in Figure 3 are derived from crowd-sourced sentence data that are full of noise. As we already argued, this is the essence of crowd-sourcing. Figure 4 plots mean expert and against mean crowd-sourced scores *for each sentence*. The scores are highly correlated, though crowd workers are substantially less likely to use extremes of the scales than experts. The first principal component and associated confidence intervals show a strong and significant statistical relationship between crowd sourced and expert assessments of individual manifesto sentences, with no evidence of systematic bias in the crowd-coded sentence scores.[30] Overall, despite the expected noise, our results show that crowd workers systematically tend to make the same judgments about individual sentences as experts.

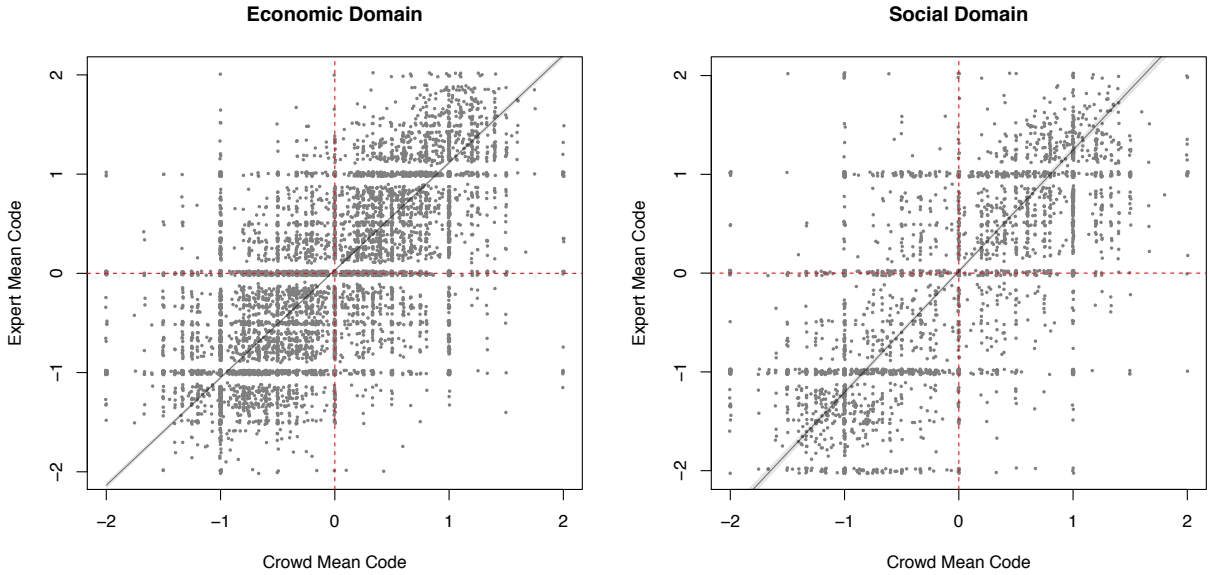

Figure 4. Expert and crowd-sourced estimates of economic and social policy codes of individual sentences, all manifestos. Fitted line is the principal components or Deming regression line.

[30] Lack of bias is indicated by the fact that the fitted line crosses the origin.

Calibrating the number of crowd judgments per sentence

A key question for our method concerns *how many* noisier crowd-based judgments we need to generate reliable and valid estimates of fairly long documents such as party manifestos. To answer this, we turn to evidence from our over-sampling of 1987 and 1997 manifestos. Recall we obtained a minimum of 20 crowd judgments for each sentence in each of these manifestos, allowing us to explore what our estimates of the position of each manifesto would have been, had we collected fewer judgments. Drawing random subsamples from our over-sampled data, we can simulate the convergence of estimated document positions as a function of the number of crowd judgments per sentence. We did this by bootstrapping 100 sets of subsamples for each of the subsets of n=1 to n=20 workers, computing manifesto positions in each policy domain from aggregated sentence position means, and computing standard deviations of these manifesto positions across the 100 estimates. Figure 5 plots these for each manifesto as a function of the increasing number of crowd workers per sentence, where each point represents the empirical standard error of the estimates for a specific manifesto. For comparison, we plot the same quantities for the expert data in red.

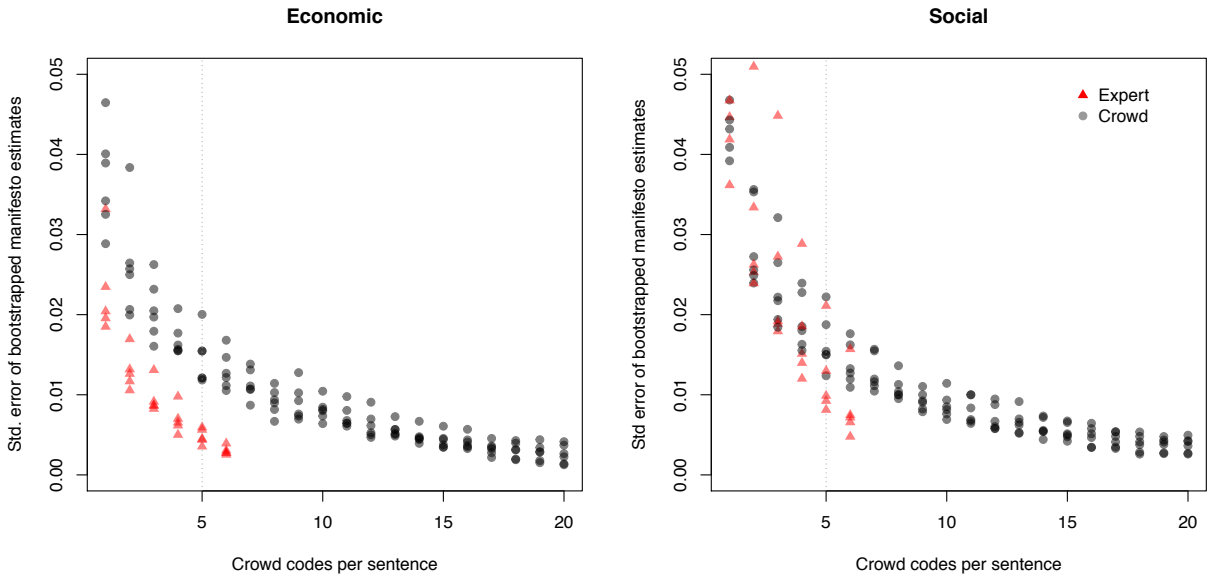

Figure 5. Standard errors of manifesto-level policy estimates as a function of the number of workers, for the oversampled 1987 and 1997 manifestos. Each point is the bootstrapped standard deviation of the mean of means aggregate manifesto scores, computed from sentence-level random *n* sub-samples from the codes.

The findings show a clear trend: uncertainty over the crowd-based estimates collapses as we increase the number of workers per sentence. Indeed, the only difference between experts and the crowd is that expert variance is smaller, as we would expect. Our findings vary somewhat with policy area, given the noisier character of social policy estimates, but adding additional crowd-sourced sentence judgments led to convergence with our expert panel of 5-6 coders at around 15 crowd coders. However, the steep decline in the uncertainty of our document estimates leveled out at around five crowd judgments per sentence, at which point the absolute level of error is already low for both policy domains. While increasing the number of unbiased crowd judgments will always give better estimates, we decided on cost-benefit grounds for the second stage of our deployment to continue coding in the crowd until we had obtained five crowd judgments per sentence. This may seem a surprisingly small number, but there are a number of important factors to bear in mind in this context. First, the manifestos comprise about

1000 sentences on average; our estimates of document positions aggregate codes for these. Second, sentences were randomly assigned to workers, so each sentence score can be seen as an independent estimate of the position of the manifesto on each dimension.[31] With five scores per sentence and about 1000 sentences per manifesto, we have about 5000 "little" estimates of the manifesto position, each a representative sample from the larger set of scores that would result from additional worker judgments about each sentence in each document. This sample is big enough to achieve a reasonable level of precision, given the large number of sentences per manifesto. While the *method* we use here could be used for much shorter documents, the *results* we infer here for the appropriate number of judgments per sentence might well not apply, and would likely be higher. But, for large documents with many sentences, we find that the number of crowd judgments *per sentence* that we need is not high.

CROWD-SOURCING DATA FOR SPECIFIC PROJECTS: IMMIGRATION POLICY

A key problem for scholars using "canonical" datasets, over and above the replication issues we discuss above, is that the data often do not measure what a modern researcher wants to measure. For example the widely-used MP data, using a classification scheme designed in the 1980s, do not measure immigration policy, a core concern in the party politics of the 21st century (Ruedin and Morales 2012; Ruedin 2013). Crowd-sourcing data frees researchers from such "legacy" problems and allows them more flexibly to collect information on their precise quantities of interest. To demonstrate this, we designed a project tailored to measure British parties' immigration policies during the 2010 election. We analyzed the manifestos of eight parties, including smaller parties with more extreme positions on immigration, such as the British National Party (BNP) and the UK Independence Party (UKIP). Workers were asked to label each

[31] Coding a sentence as referring to another dimension is a null estimate.

sentence as referring to immigration policy or not. If a sentence did cover immigration, they were asked to rate it as pro- or anti-immigration, or neutral. We deployed a job with 7,070 manifesto sentences plus 136 "gold" questions and screeners devised specifically for this purpose. For this job, we used an adaptive sentence sampling strategy which set a minimum of five crowd sourced labels per sentence, unless the first three of these were unanimous in judging a sentence *not* to concern immigration policy. This is efficient when coding texts with only "sparse" references to the matter of interest; in this case most manifesto sentences (approximately 96%) were clearly not about immigration policy. Within just five hours, the job was completed, with 22,228 codings, for a total cost of $360.[32]

We assess the external validity of our results using independent expert surveys by Benoit (2010) and the Chapel Hill Expert Survey (Marks 2010). Figure 6 compares the crowd-sourced estimates to those from expert surveys. The correlation with the Benoit (2010) estimates (shown) was 0.96, and 0.94 with independent expert survey estimates from the Chapel Hill survey.[33] To assess whether this data production exercise was as reproducible as we claim, we repeated the entire exercise with a second deployment two months after the first, with identical settings. This new job generated another 24,551 pieces of crowd-sourced data and completed in just over three hours. The replication generated nearly identical estimates, detailed in Table 4, correlating at the same high levels with external expert surveys, and correlating at 0.93 with party position estimates from the original crowd-coding.[34] With just hours from deployment to dataset, and for very little cost, crowd-sourcing enabled us to generate externally valid *and reproducible* data related to our precise research question.

[32] The job set 10 sentences per "task" and paid $0.15 per task.
[33] CHES included two highly correlated measures, one aimed at "closed or open" immigration policy another aimed at policy toward asylum seekers and whether immigrants should be integrated into British society. Our measure averages the two. Full numerical results are given in supplementary materials, section 3.
[34] Full details are in the supplementary materials, section 7.

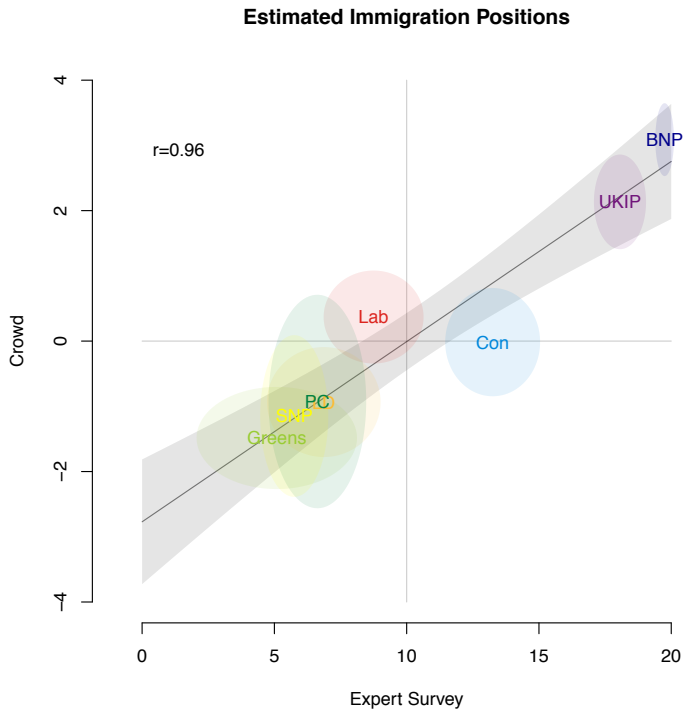

Figure 6. Correlation of combined immigration crowd codings with Benoit (2010) expert survey position on immigration.

	Wave		
	Initial	Replication	Combined
Total Crowd Codings	24,674	24,551	49,225
Number of Coders	51	48	85
Total sentences coded as Immigration	280	264	283
Correlation with Benoit expert survey (2010)	0.96	0.94	0.96
Correlation with CHES 2010	0.94	0.91	0.94
Correlation of results between waves			0.93

Table 4. Comparison results for Replication of Immigration Policy Crowd-Coding.

CROWD SOURCED TEXT ANALYSIS IN OTHER CONTEXTS AND LANGUAGES

As carefully designed official statements of a party's policy stances, election manifestos tend to respond well to systematic text analysis. In addition, manifestos are written for popular consumption and tend to be easily understood by non-technical readers. Much political information, however, can be found in texts generated from hearings, committee debates, or legislative speeches on issues that often refer to technical provisions, amendments, or other rules of procedure that might prove harder to analyze. Furthermore, a majority of the world's political texts are not in English. Other widely studied political contexts, such as the European Union, are multi-lingual environments where researchers using automated methods designed for a single language must make hard choices. Schwarz et al. (Forthcoming) applied unsupervised scaling methods to a multilingual debate in the Swiss parliament, for instance, but had to ignore a substantial number of French and Italian speeches in order to focus on the majority German texts. In this section, we demonstrate that crowd-sourced text analysis, with appropriately translated instructions, offers the means to overcome these limitations by working in any language.

Our corpus comes from a debate in the European Parliament, a multi-language setting where the EU officially translates every document into 22 languages. To test our method in a context very different from party manifestos, we chose a fairly technical debate concerning a Commission report proposing an extension to a regulation permitting state aid to uncompetitive coal mines. This debate concerned not only the specific proposal, involving a choice of letting the subsidies expire in 2011, permitting a limited continuation until 2014, or extending them

until 2018 or even indefinitely.[35] It also served as debating platform for arguments supporting state aid to uncompetitive industries, versus the traditionally liberal preference for the free market over subsidies. Because a vote was taken at the end of the debate, we also have an objective measure of whether the speakers supported or objected to the continuation of state aid.

We downloaded all 36 speeches from this debate, originally delivered by speakers from 11 different countries in 10 different languages. Only one of these speakers, an MEP from the Netherlands, spoke in English, but all speeches were officially translated into each target language. After segmenting this debate into sentences, devising instructions and representative test sentences and translating these into each language, we deployed the same text analysis job in English, German, Spanish, Italian, Polish, and Greek, using crowd workers to read and label the same set of texts, but using the translation into their own language. Figure 7 plots the score for each text against the eventual vote of the speaker. It shows that our crowd-sourced scores for each speech perfectly predict the voting behavior of each speaker, regardless of the language. In Table 5, we show correlations between our crowd-sourced estimates of the positions of the six different language versions of the same set of texts. The results are striking, with inter-language correlations ranging between 0.92 and 0.96.[36] Our text measures from this technical debate produced reliable measures of the very specific dimension we sought to estimate, and the validity of these measures was demonstrated by their ability to predict the voting behavior of the speakers. Not only are these results straightforwardly reproducible, but this reproducibility is invariant to the language in which the speech was written. Crowd-sourced text analysis does not only work in English.

[35] This was the debate from 23 November 2010, "State aid to facilitate the closure of uncompetitive coal mines." http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+CRE+20101123+ITEM-005+DOC+XML+V0//EN&language=EN

[36] Lin's concordance coefficient has a similar range of values, from 0.90 to 0.95.

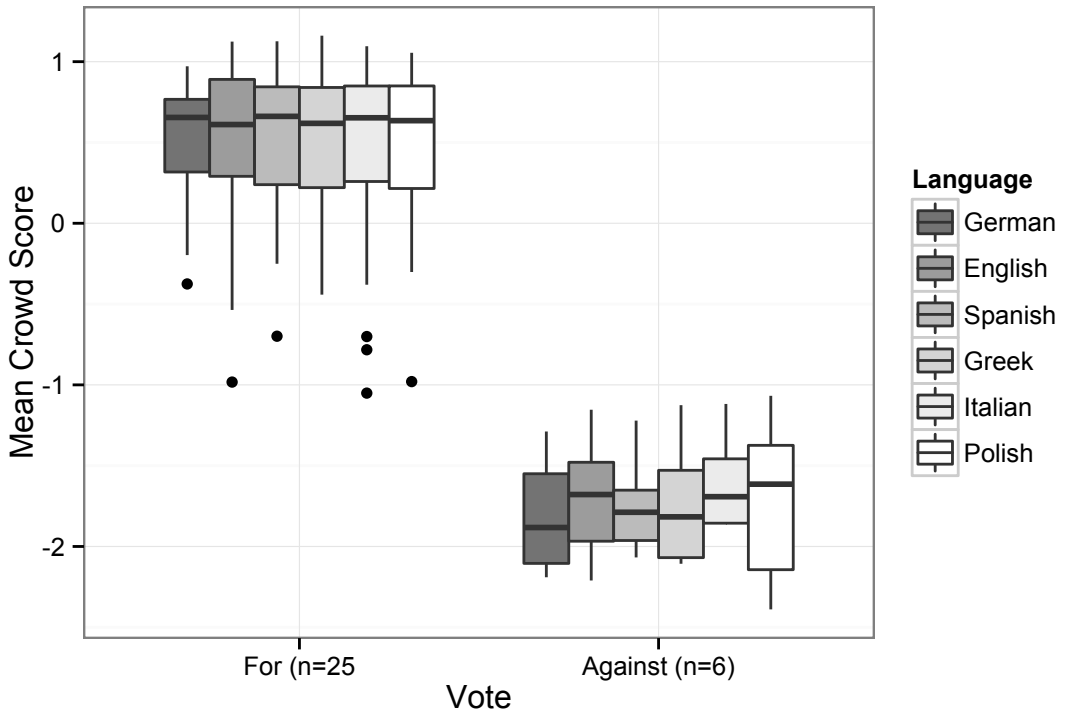

Figure 7. Scored speeches from a debate over state subsidies by vote, from separate crowd-sourced text analysis in six languages. Aggregate scores are standardized for direct comparison.

	Correlations of 35 speaker scores					
Language	English	German	Spanish	Italian	Greek	Polish
German	0.96	--	--	--	--	--
Spanish	0.94	0.95	--	--	--	--
Italian	0.92	0.94	0.92	--	--	--
Greek	0.95	0.97	0.95	0.92	--	--
Polish	0.96	0.95	0.94	0.94	0.93	--
Sentence *N*	414	455	418	349	454	437
Total Judgments	3,545	1,855	2,240	1,748	2,396	2,256
Cost	$109.33	$55.26	$54.26	$43.69	$68.03	$59.25
Elapsed Time (hrs)	1	3	3	7	2	1

Table 5. Summary of Results from EP Debate Coding in 6 languages

CONCLUSIONS

We have illustrated across a range of applications that crowd-sourced text analysis can produce valid political data of a quality indistinguishable from traditional expert methods. Unlike traditional methods, however, crowd-sourced data generation offers several advantages. Foremost among these is the possibility of meeting a replication standard far stronger than the current practice of facilitating reproducible *analysis*. By offering a published specification for feasibly replicating *the process of data generation*, the methods demonstrated here go much farther towards meeting a more stringent standard of *reproducibility* that is the hallmark of scientific inquiry. All of the data used in this paper are of course available in a public archive for any reader to reanalyze at will. Crowd-sourcing our data allows us to do much more than this, however. Any reader can take our publically available crowdsourcing code and deploy this code to *reproduce our data collection process and collect a completely new dataset*. This can be done many times over, by any researcher, anywhere in the world. This, to our minds, takes us significantly closer to a true scientific replication standard.

Another key advantage of crowd-sourced text analysis is that it can form part of an *agile* research process, precisely tailored to a specific research question rather than reflecting the grand compromise at the heart of the large canonical datasets so commonly deployed by political scientists. Because the crowd's resources can be tapped in a flexible fashion, text-based data on completely new questions of interest can be processed only for the contexts, questions, and time periods required. Coupled with the rapid completion time of crowd-sourced tasks and their very low marginal cost, this opens the possibility of valid text processing to researchers with limited resources, especially graduate students. For those with more ambition or resources, its inherent

scalability means that crowd-sourcing can tackle large projects as well. In our demonstrations, our method worked as well for hundreds of judgments as it did for hundreds of thousands.

Of course, retooling for any new technology involves climbing a learning curve. We spent considerable time pretesting instruction wordings, qualification tests, compensation schemes, gold questions, and a range of other detailed matters. Starting a new crowd-sourcing project is by no means cost-free, though these costs are mainly denominated in learning time and effort spent by the researcher, rather than research dollars. Having paid the inevitable fixed start-up costs that apply any rigorous new data collection project, whether or not this involves crowd-sourcing, the beauty of crowd-sourcing arises from two key features of the crowd. The pool of crowd workers is to all intents and purposes inexhaustible, giving crowd-sourcing projects a scalability and replicability unique among projects employing human workers. And the low *marginal* cost of adding more crowd workers to any given project puts ambitious high quality data generation projects in the realistic grasp of a wider range of researchers than ever before. We are still in the early days of crowd-sourced data generation in the social sciences. Other scholars will doubtless find many ways fortify the robustness and broaden the scope of the method. But, whatever these developments, we now have a new method for collecting political data that allows us to do things we could not do before.

APPENDIX: METHODOLOGICAL DECISIONS ON SERVING POLITICAL TEXT TO WORKERS IN THE CROWD

900 words

Text units: natural sentences

The MP specifies a "quasi-sentence" as the fundamental text unit, defined as "an argument which is the verbal expression of one political idea or issue" (Volkens). Recoding experiments by Däubler et al. (2012), however, show that using natural sentences makes no statistically significant difference to point estimates, but does eliminate significant sources of both unreliability and unnecessary work. Our dataset therefore consists of all natural sentences in the 18 UK party manifestos under investigation.[37]

Text unit sequence: random

In "classical" expert text coding, experts process sentences in their natural sequence, starting at the beginning and ending at the end of a document. Most workers in the crowd, however, will never reach the end of a long policy document. Processing sentences in natural sequence, moreover, creates a situation in which one sentence coding may well affect priors for subsequent sentence codings, so that summary scores for particular documents are not aggregations of independent coder assessments.[38] An alternative is to randomly sample sentences from the text corpus for coding—with a fixed number of replacements per sentence across all coders—so that each coding is an independent estimate of the latent variable of interest. This has the big advantage in a crowdsourcing context of *scalability*. Jobs for individual coders can range from very small to very large; coders can pick up and put down coding tasks at will; every little piece

[37] Segmenting "natural" sentences, even in English, is never an exact science, but our rules matched those from Däubler et al. (2012), treating (for example) separate clauses of bullet pointed lists as separate sentences.
[38] Coded sentences do indeed tend to occur in "runs" of similar topics, and hence codes; however to ensure appropriate statistical aggregation it is preferable if the codings of those sentences are independent.

of coding in the crowd contributes to the overall database of text codings. Accordingly our method for crowd-sourced text coding serves coders sentences randomly selected from the text corpus rather than in naturally occurring sequence. Our decision to do this was informed by coding experiments reported in the supplementary materials (section 5), and confirmed by results reported below. Despite higher variance in individual sentence codings under random sequence coding, there is no systematic difference between point estimates of party policy positions depending on whether sentences were coded in natural or random sequence.

Text authorship: anonymous

In classical expert coding, coders typically know the authorship of the document they are coding. Especially in the production of political data, coders likely bring non-zero priors to coding text units. Precisely the same sentence ("we must do all we can to make the public sector more efficient") may be coded in different ways if the coder knows this comes from a right- rather than a left-wing party. Codings are typically aggregated into document scores as if coders had zero priors, even though we do not know how much of the score given to some sentence is the coder's judgment about the content of the sentence, and how much a judgment about its author. In coding experiments reported in supplementary materials (section 5), semi-expert coders coded the same manifesto sentences both knowing and not knowing the name of the author. We found slight systematic coding biases arising from knowing the identity of the document's author. For example, we found coders tended to code precisely the same sentences from Conservative manifestos as more right wing, if they knew these sentences came from a Conservative manifesto. This informed our decision to withhold the name of the author of sentences deployed in crowd-sourcing text coding.

Context units: +/- two sentences

Classical content analysis has always involved coding an individual text unit in light of the text surrounding it. Often, it is this context that gives a sentence substantive meaning, for example because many sentences contain pronoun references to surrounding text. For these reasons, careful instructions for drawing on context have long formed part of coder instructions for content analysis (see Krippendorff 2013). For our coding scheme, on the basis of pre-release coding experiments, we situated each "target" sentence within a context of the two sentences either side in the text. Coders were instructed to code target sentence not context, but to use context to resolve any ambiguity they might feel about the target sentence.

REFERENCES

Alonso, O., and R. Baeza-Yates. 2011. "Design and Implementation of Relevance Assessments Using Crowdsourcing." In *Advances in Information Retrieval*, ed. P. Clough, C. Foley, C. Gurrin, G. Jones, W. Kraaij, H. Lee and V. Mudoch: Springer Berlin / Heidelberg.

Alonso, O., and S. Mizzaro. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. Paper read at Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation.

Ariely, D., W. T. Au, R. H. Bender, D. V. Budescu, C. B. Dietz, H. Gu, and G. Zauberman. 2000. "The effects of averaging subjective probability estimates between and within judges." *Journal of Experimental Psychology: Applied* 6 (2):130-47.

Armstrong, J.S., ed. 2001. *Principles of Forecasting: A Handbook for Researchers and Practitioners*: Springer.

Baker, Frank B, and Seock-Ho Kim. 2004. *Item response theory: Parameter estimation techniques*: CRC Press.

Benoit, Kenneth. 2005. "Policy positions in Britain 2005: results from an expert survey." London School of Economics.

———. 2010. "Expert Survey of British Political Parties." Trinity College Dublin.

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

Berinsky, A., G. Huber, and G. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis*.

Berinsky, A., M. Margolis, and M. Sances. forthcoming. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science*.

Bohannon, J. 2011. "Social Science for Pennies." *Science* 334:307.

Budge, Ian, Hans-Dieter Klingemann, Andrea Volkens, Judith Bara, Eric Tannenbaum, Richard Fording, Derek Hearl, Hee Min Kim, Michael McDonald, and Silvia Mendes. 2001. *Mapping Policy Preferences: Parties, Electors and Governments: 1945-1998: 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.

Cao, J, S. Stokes, and S. Zhang. 2010. "A Bayesian Approach to Ranking and Rater Evaluation: An Application to Grant Reviews." *Journal of Educational and Behavioral Statistics* 35 (2):194-214.

Carpenter, B. 2008. "Multilevel Bayesian models of categorical data annotation."

Chandler, Jesse, Pam Mueller, and Gabriel Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers." *Behavior Research Methods* 46 (1):112-30.

Clemen, R., and R. Winkler. 1999. "Combining Probability Distributions From Experts in Risk Analysis." *Risk Analysis* 19 (2):187-203.

Conway, Drew. 2013. Applications of Computational Methods in Political Science, Department of Politics, New York University.

Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2012. "Natural sentences as valid units for coded political text." *British Journal of Political Science* 42 (4):937-51.

Eickhoff, C., and A. de Vries. 2012. "Increasing cheat robustness of crowdsourcing tasks." *Information Retrieval* 15:1-17.

Fox, Jean-Paul. 2010. *Bayesian item response modeling: Theory and applications*: Springer.

Galton, F. 1907. "Vox Populi." *Nature* 75:450-1.

Goodman, Joseph, Cynthia Cryder, and Amar Cheema. 2013. "Data Collection in a Flat World: Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26 (3):213-24.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis*.

Hambleton, Ronald K, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*: Sage.

Hooghe, Liesbet, Ryan Bakker, Anna Brigevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny, Marco Steenbergen, and Milada Vachudova. 2010. " Reliability and Validity of Measuring Party Positions: The Chapel Hill Expert Surveys of 2002 and 2006." *European Journal of Political Research.* 49 (5):687-703.

Horton, J., D. Rand, and R. Zeckhauser. 2011. "The online laboratory: conducting experiments in a real labor market." *Experimental Economics* 14:399-425.

Hsueh, P., P. Melville, and V. Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. Paper read at Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing.

Ipeirotis, Panagiotis G., Foster Provost, Victor S. Sheng, and Jing Wang. 2013. "Repeated labeling using multiple noisy labelers." *Data Mining and Knowledge Discovery*:1-40.

Ipeirotis, Panagiotis, F. Provost, V. Sheng, and J. Wang. 2010. "Repeated Labeling Using Multiple Noisy Labelers." NYU Working Paper.

Jones, Frank R. Baumgartner and Bryan D. 2013. "Policy Agendas Project."

Kapelner, A., and D. Chandler. 2010. Preventing satisficing in online surveys: A `kapcha' to ensure higher quality data. Paper read at The World's First Conference on the Future of Distributed Work (CrowdConf 2010).

King, Gary. 1995. "Replication, replication." *PS: Political Science & Politics* 28 (03):444-52.

Klingemann, Hans-Dieter, Richard I. Hofferbert, and Ian Budge. 1994. *Parties, policies, and democracy*. Boulder: Westview 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.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. 3rd ed: Sage.

Laver, M. 1998. "Party policy in Britain 1997: Results from an expert survey." *Political Studies* 46 (2):336-47.

Laver, Michael, and Ian Budge. 1992. *Party policy and government coalitions*. New York, N.Y.: St. Martin's Press.

Laver, Michael, and W. Ben Hunt. 1992. *Policy and party competition*. New York: Routledge.

Lawson, C., G. Lenz, A. Baker, and M. Myers. 2010. "Looking Like a Winner: Candidate appearance and electoral success in new democracies." *World Politics* 62 (4):561-93.

Lin, L. 1989. "A concordance correlation coefficient to evaluate reproducibility." *Biometrics* 45 (255-268).

———. 2000. "A note on the concordance correlation coefficient." *Biometrics* 56:324 - 5.

Lord, Frederic. 1980. *Applications of item response theory to practical testing problems*: Routledge.

Lyon, Aidan, and Eric Pacuit. 2013. "The Wisdom of Crowds: Methods of Human Judgement Aggregation." In *Handbook of Human Computation*, ed. P. Michelucci: Springer.

Mason, W, and S Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44 (1):1-23.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2012. "Coder reliability and misclassification in comparative manifesto project codings." *Political Analysis* 20 (1):78-91.

Nowak, S., and S. Rger. 2010. How reliable are annotations via crowdsourcing? a study about inter-annotator agreement for multi-label image annotation. Paper read at The 11th ACM International Conference on Multimedia Information Retrieval, 29-31 Mar 2010, at Philadelphia, USA.

Paolacci, Gabriel, Jesse Chandler, and Panagiotis Ipeirotis. 2010. "Running experiments on Amazon Mechanical Turk." *Judgement and Decision Making* 5:411-9.

Quoc Viet Hung, Nguyen, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. "An Evaluation of Aggregation Techniques in Crowdsourcing." In *Web Information Systems Engineering – WISE 2013*, ed. X. Lin, Y. Manolopoulos, D. Srivastava and G. Huang: Springer Berlin Heidelberg.

Raykar, V. C., S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bo- goni, and L. Moy. 2010. "Learning from crowds." *Journal of Machine Learning Research* 11:1297-322.

Ruedin, Didier. 2013. "Obtaining Party Positions on Immigration in Switzerland: Comparing Different Methods." *Swiss Political Science Review* 19 (1):84-105.

Ruedin, Didier, and Laura Morales. 2012. "Obtaining Party Positions on Immigration from Party Manifestos."

Schwarz, Daniel, Denise Traber, and Kenneth Benoit. Forthcoming. "Estimating Intra-Party Preferences: Comparing Speeches to Votes." *Political Science Research and Methods*.

Sheng, V., F. Provost, and Panagiotis Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. Paper read at Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Snow, R., B. O'Connor, D. Jurafsky, and A. Ng. 2008. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. Paper read at Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Surowiecki, J. 2004. *The Wisdom of Crowds*. New York: W.W. Norton & Company, Inc.

Turner, Brandon M., Mark Steyvers, Edgar C. Merkle, David V. Budescu, and Thomas S. Wallsten. 2013. "Forecast aggregation via recalibration." *Machine Learning*:1-29.

Volkens, Andrea. 2001. "Manifesto Coding Instructions, 2nd revised ed." In *Discussion Paper (2001), p. 96.*, ed. W. Berlin.

Welinder, P., S. Branson, S. Belongie, and P. Perona. 2010. The multidimensional wisdom of crowds. Paper read at Advances in Neural Information Processing Systems 23 (NIPS 2010).

Welinder, P., and P. Perona. 2010. Online crowdsourcing: rating annotators and obtaining cost-effective labels. Paper read at IEEE Conference on Computer Vision and Pattern Recognition Workshops (ACVHL).

Whitehill, J., P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. Paper read at Advances in Neural Information Processing Systems 22 (NIPS 2009).