A pairwise comparison framework for fast, flexible, and reliable human coding of political texts*

David Carlson
Washington University in St. Louis

Jacob M. Montgomery
Washington University in St. Louis

ABSTRACT

Scholars are increasingly utilizing online workforces to encode latent political concepts embedded in written or spoken records. In this letter, we build on past efforts by developing and validating a crowdsourced pairwise comparison framework for encoding political texts that combines the human ability to understand natural language with the ability of computers to aggregate data into reliable measures while ameliorating concerns about the biases and unreliability of non-expert human coders. We validate the method with advertisements for U.S. Senate candidates and with State Department reports on human rights. The framework we present is very general, and we provide free software to help applied researchers interact easily with online workforces to extract meaningful measures from texts.

*We thank Burt Monroe, John Freeman, and Brandon Stewart for providing comments on a previous version of this paper. We are indebted to Ryden Butler, Dominic Jarkey, Jon Rogowski, Erin Rossiter, and Michelle Torres for their assistance with this project. We particularly wish to thank Matt Dickenson for his programming assistance. We also appreciate the assistance in the R package development from David Flasterstein, Joseph Ludmir, and Taishi Muraoka. We are grateful for the financial support provided by the Weidenbaum Center on the Economy, Government, and Public Policy. Finally, we wish to thank the partner-workers at Amazon's Mechanical Turk who make this research possible.

Recent work in political science has demonstrated the vast potential of using online workforces to efficiently content code political texts (e.g., Henderson 2015; Benoit et al. 2016). This crowdsourcing approach builds on the intuition that we can combine the superior ability of humans to read and understand the meaning of natural language with the superior ability of computers to aggregate human judgements into reliable measures of latent traits. This combination provides a scalable method for transforming texts into measures of latent concepts that are valid, can be specified in advance by the researcher, exist on a continuous scale, and are highly reliable.

In this letter, we extend previous work on crowdsourced coding by developing and validating a more general-purpose framework for encoding natural language by dividing the larger undertaking into thousands of simple micro-tasks (viz., binary pairwise comparisons) that can be easily completed by a trained but non-expert online workforce. By sending thousands of these simple comparisons to online workers, we circumvent issues of unreliable human coding associated with absolute scales (e.g., Likert scales). This makes the use of crowdsourced judgements significantly more flexible, applicable to more substantive domains, and ameliorates potential issues related to the education, background, and ability of coders. The result is a highly flexible, reliable, and replicable method for creating valid estimates of latent traits within documents. Further, we provide easy to use open-source software that allows applied researchers to smoothly interact with online workers when encoding texts.

After providing the details of our method, we evaluate it using texts from congressional advertisements and State Department reports. In each case, we compare our estimates to existing measures of researcher-specified latent traits embedded within the texts to show that our measures are not only highly reliable, but also valid measures of the underlying latent traits of interest. We provide software and practical guidelines to help researchers interact with online workforces, and our Supplemental Information (SI) Appendix includes extensive diagnostics, further applications, and detailed examples that will aid researchers with minimal technical backgrounds wishing to use crowdsourced judgments in their own work.

THE SENTIMENTIT PLATFORM

The system we designed, which we label SentimentIt, is based on the following principals. First, it leverages human ability to understand language and socially constructed political concepts. Thus, our focus is measuring researcher-defined characteristics embedded within text (e.g., positivity). However, *we exclude* explicitly subjective characteristics (e.g., persuasiveness).

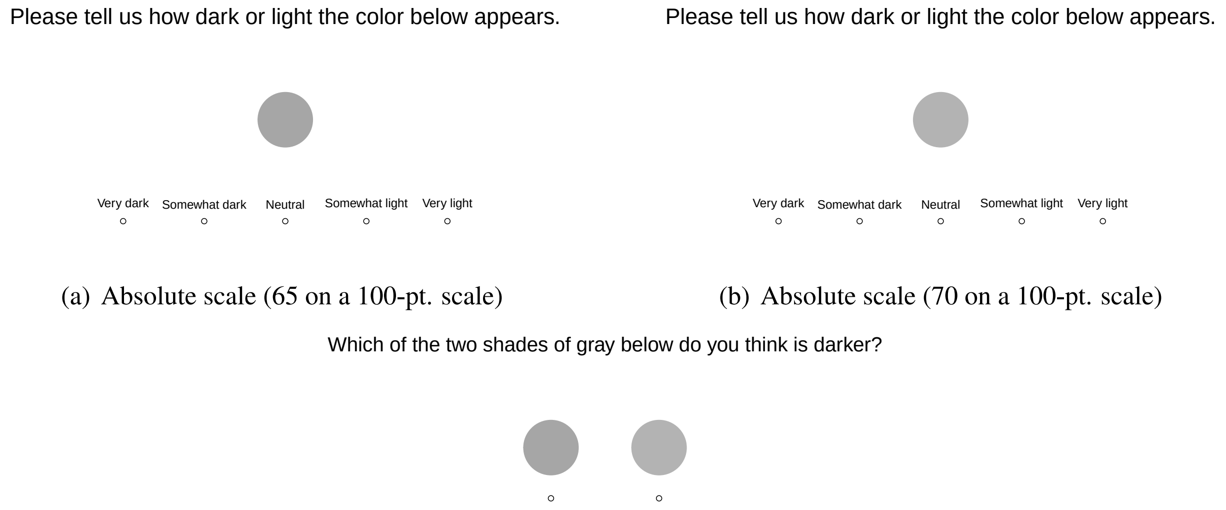
Second, we designed the task structure to be cognitively appropriate for non-experts. Specifically, we ask workers to conduct pairwise comparisons of texts, simply indicating which text is more extreme along a single dimension of interest (e.g., "Which text is more positive?"). A significant body of research indicates that pairwise comparisons can reduce the cognitive burden for respondents, improve the reliability of responses, and eliminate problems such as differential item functioning and reference group effects that plague alternative question formats such as Likert scales or sliders (e.g., King et al. 2004; Oishi et al. 2005).

The advantages of the pairwise comparison framework are illustrated in Figure 1, which shows how two visual stimuli can be presented to workers on either absolute or relative scales. The top images exemplify how tasks are typically presented to workers. Completing such tasks requires workers to continuously maintain in their memory how previous documents were coded and remember detailed rules dictating how stimuli are placed into categories. Worse, to obtain reliable scores, all workers must follow the same set of rules equally or else researchers must hope that coder-specific biases "wash out" when judgements are aggregated.

These problems are significantly ameliorated when using a comparison framework shown in Figure 1c. If some coder tends to rate stimuli higher on the latent scale, this is not relevant due to the pairwise comparison framework. So long as we are asking for only relative evaluations, a general bias towards perceiving darker colors is irrelevant. In addition, pairwise comparisons are generally easier to complete, allow for more subtle distinctions between stimuli, and are thus coded with high reliability (see the SI Appendix).

Third, we relied on paid online workers recruited through Amazon's Mechanical Turk (AMT). While it is now common for researchers to use AMT workers as research subjects (Berinsky,

Figure 1: Placing stimuli on an absolute scale versus paired comparisons


Please tell us how dark or light the color below appears.

Please tell us how dark or light the color below appears.

(a) Absolute scale (65 on a 100-pt. scale)

(b) Absolute scale (70 on a 100-pt. scale)

Which of the two shades of gray below do you think is darker?

(c) Comparison (65 and 70 on a 100-pt. scale)

Huber, and Lenz 2012), the majority of jobs posted at AMT are small HITs (human intelligence tasks). AMT workers are generally well-educated, part-time workers who are highly experienced at completing micro-tasks for small amounts of money (Hitlin 2016).

Finally, we provide *quality assurance* via training, redundancy, and statistical monitoring. In order to qualify for our micro-tasks, workers must complete a training module that explains the task, provides detailed decision rules, and includes example HITs with discussion of difficult cases. Workers must then correctly complete a preset number of tasks before they are “certified”. Further, previous research shows that *aggregated* judgments by non-experts are often comparable to those provided by subject experts (e.g., Benoit et al. 2016; Sheng, Provost, and Ipeirotis 2008). Therefore, each document is included in multiple (at least 20) pairwise comparisons. In addition, our pairwise-comparison framework allows us to easily evaluate the data quality from individual workers as part of our statistical processing (discussed below).

Extending previous work

Our system offers three distinct advantages relative to previous work in political science using crowdsourced coding. We specifically contrast our approach with Benoit et al. (2016) —the most comprehensive work in political science on the subject. First, we do not rely on "gold standard" tasks, which require workers to provide judgements about pre-coded texts to continuously evaluate the quality of their work. These judgements, which researchers must pay for and workers must complete, are of no direct use other than to monitor worker quality and effort. Yet, Benoit et al. (2016, p. 286) recommend that 10% of all evaluations be focused exclusively on quality control.

In contrast, our method relies on a pairwise-comparisons framework that allows for seamless supervision of workers' outputs in the very process of collecting the data. Indeed, as we discuss at length in our SI Appendix, our comparison framework generally works to make coding easier for workers, improve data quality, and reduce the influence of coder-specific biases. One consequence is that while Benoit et al. (2016, p. 286) indicate that they had to exclude "many" judgements from untrusted workers from their analyses, in our applications we were forced to remove only one worker (out of 256) and still generated highly reliable and valid scores. In all, our approach simultaneously improves the reliability of the estimates, reduces overhead costs, and makes the method more applicable to subject domains where researchers might worry about coder biases.

Second, interacting with online workforces, setting up tasks, ensuring worker quality, and providing compensation can be a time-consuming and difficult process and in some cases requires a high level of technical acumen. Our system offers a simplified approach to recruiting, training, monitoring, and aggregating online workers. The software we provide offers a suite of tools for researchers to easily set up tasks, manage and train workers, and analyze data, all within the widely used `R` computing environment.

Finally, the procedures outlined in Benoit et al. (2016) require either the use of custom-built interfaces or working through the `CrowdFlower` platform. The former approach is beyond the abilities of most applied researchers. For the latter option, privacy requirements or legal obligations may prevent scholars from making their text and data publicly available, transferable, and non-exclusive (a requirement for academic accounts on CrowdFlower)[1] while the costs of commercial accounts may be prohibitive (at least $1500/month). Our system instead provides direct access to a high quality online workforce (Amazon's Mechanical Turk), includes free tools for worker management, but allows researchers to keep private their own data and texts.[2]

Workflow

The core functionality of the SentimentIt platform is a freely available cloud-based web application that interacts smoothly with AMT to post jobs, certify workers, store responses, and generally reduce the burden of interacting with AMT. Researchers access the functionality of SentimentIt via application program interfaces (APIs) that can be called from any computing environment or platform (Python, Java, etc.). However, all of the functionality described below is fully integrated into our R package, making the process for researchers accustomed to the R language especially straightforward. The complete workflow is depicted in Figure 2.

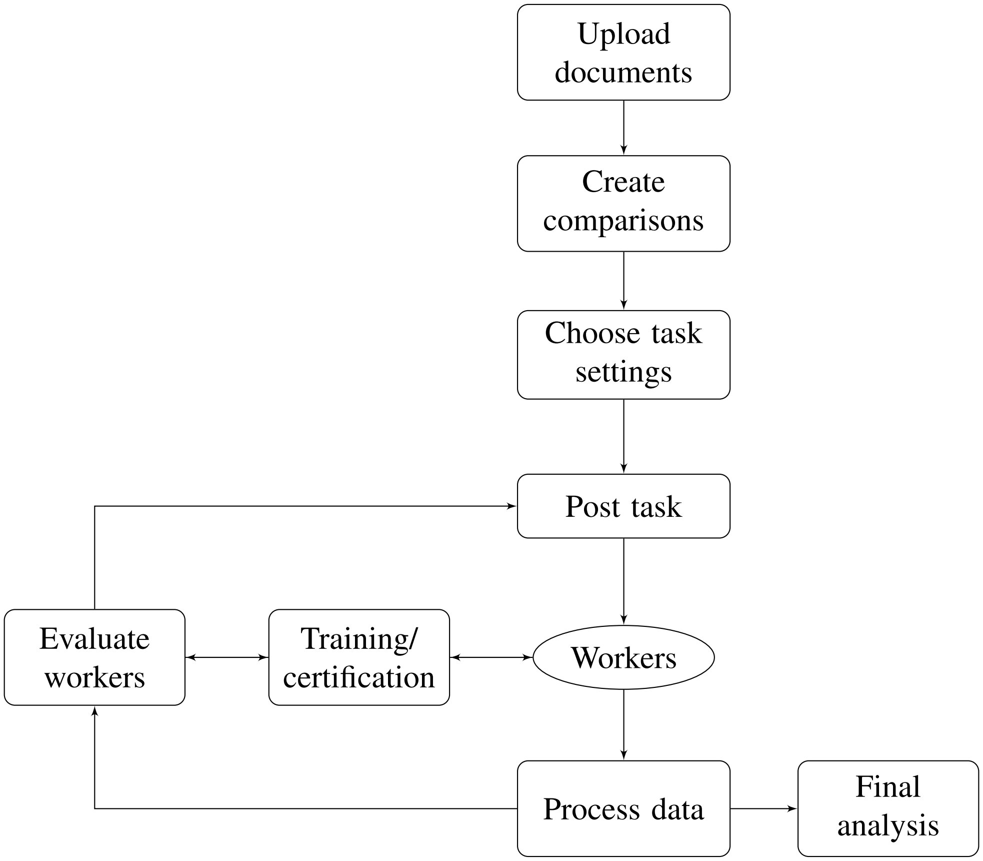
First, we pre-process the texts ensuring they are machine-readable and, where necessary, breaking the document into shorter, meaningful parts, such as paragraphs. The texts are then passed to SentimentIt and randomly paired into a series of comparisons. For example, if we want 20 comparisons per document for 500 documents, we randomly create 5,000 unique comparisons ($500 \times 20/2 = 5,000$). We then send paired document identification numbers and an associated question (e.g., "Which statement is more positive?") via API.

At this stage, we determine the task settings, dictating how much we pay workers and whether we require certification. We then send the comparisons out to the workers. In most cases, the complete universe of micro-tasks should not be posted simultaneously. We find that posting jobs in batches of 1,000 tasks allows us to keep track of how quickly the tasks are accomplished, and, importantly, gives us the opportunity to analyze the quality of the responses. If we determine that specific workers are providing poor data, we can revoke their certification and prevent them from

[1] A particular concern is that scholars may not have the legal authority to distribute specific texts taken from websites and social media platforms. In many cases, these texts are the intellectual property of the authors, the outlet, or both.

[2] In future work, we hope to extend the platform to allow for worker recruitment through alternative channels.

Figure 2: `SentimentIt` workflow



Note: See Appendix SI-10 for a detailed example.

further contaminating the data.

Once a sufficient number of the comparisons are complete, we can download the data via API. The data simply indicate which of the two documents was selected, the unique worker ID, and the time the task was completed. We then process the data using a random utility model (closely related to standard item response models) that creates document-level estimates along the dimension of interest. Specifically, we model the probability that one document would be chosen over another, while estimating worker reliability given the choices made by that worker. Let i and j index documents in a comparison and k index the worker. The model is then:

$$\Pr(y_{ijk} = j) = \frac{\exp(b_k(a_j - a_i)}{1 + \exp(b_k(a_j - a_i))} \tag{1}$$

The model is completed by specifying the following priors:

$$a_j \sim \mathcal{N}(0,1) \qquad b_k \sim tr\mathcal{N}(0,\sigma^2) \qquad \sigma \sim tr\mathcal{N}(0,3),$$

where $\mathcal{N}$ refers to the normal distribution, and $tr\mathcal{N}$ refers to the normal distribution truncated at zero to only support positive values.[3] We estimate the model using Hamiltonian Markov Chain Monte Carlo sampling using Stan (Carpenter et al. 2016). The model produces posterior estimates for the documents' positions on the latent scale of interest (a_j) as well as the workers' reliability (b_k).

We can extend this model to allow a hierarchical structure. In our second application, we deal with large documents that require simplification to create suitable micro-tasks. We construct the pairwise comparisons using paragraphs rather than entire documents. To allow for a hierarchical structure of the data, let i and j still index paragraphs in a comparison and k index the worker. We now let m index the higher-level documents. The hierarchical model is still specified as in Equation (1). However, the a estimates are now centered at a higher-level-document mean for

[3] Note that we set the variance term for the prior of a_i and σ at 1 and 3 respectively to identify the scale of the latent distribution.

document m, denoted θ_m. Letting M be the set of paragraphs contained in document m, the priors are,

$$a_j \sim \mathcal{N}(\theta_m, \sigma_m^2) \forall j \in M \qquad b_k \sim tr\mathcal{N}(0,1) \qquad \sigma_m \sim tr\mathcal{N}(0,.5) \qquad \theta_m \sim \mathcal{N}(0,1).$$

While there are many alternative ways of modeling this data, in our experience the resulting document-level estimates are largely invariant to these choices. Indeed, in our appendix we show the correlations between the a_i estimates used in our applications below are correlated (Pearson's r) at $0.95 - 0.98$ with the arithmetic mean coder choice (where a document is coded as zero when it is not chosen and as one otherwise). Thus, while we feel that the model above is useful – particularly for identifying low-quality coders – the conclusions we draw below are robust to specific modeling choices and priors (see also Benoit et al. 2016).

In both models, the parameter estimates for the workers (b_k) give us an assessment of how well each worker performed. Intuitively, these estimates become lower for workers whose choices do not reflect how the documents are understood by other workers. If we estimate a worker as being a (low) outlier, we can ban the worker from future tasks. We find that revoking qualifications from these workers modestly improves the validity of our final estimates relative to benchmarks. After disqualifying problematic workers (if any), we can post further tasks and repeat as necessary.

Our `R` package can do as little or as much of this in an automated fashion as the researcher wants. If desired, researchers can execute each step specified above manually through `R`. Alternatively, it can be fully automated such that the software will create comparisons, post tasks, check if the tasks are completed, download the data, test for worker outliers, ban unwanted workers, and repeat the process until all of the desired data has been collected and analyzed.

APPLICATIONS

In this section we apply our procedure to texts from political ads and formal reports from a bureaucratic agency. (Several additional applications are shown in the SI Appendix.) In each, we demonstrate that our estimates are valid measures of underlying latent traits of interest by comparing them to relevant benchmarks. We show that our estimates correlate highly with these benchmarks and argue that where the measures disagree, `SentimentIt` is usually better at capturing underlying latent traits. Further, we demonstrate that the measures are reliable and replicable. Following the same procedure using the same settings in `SentimentIt` results in estimates that are highly correlated (Pearson's $r \geq 0.88$ in all cases).

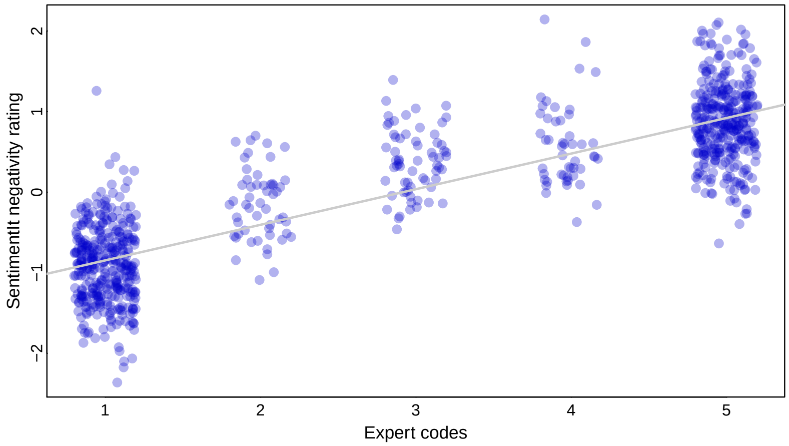
Advertisement tone

First we measure the "tone" of political ads, or their level of negativity. In the Wisconsin Advertising Project (WiscAds) dataset, ad tone is determined by expert coders who categorized ads as either promoting a single candidate, contrasting two candidates, or attacking a candidate. If the ad is contrasting, it is further categorized as either being more aimed at promoting than attacking, more attacking than promoting, or equally attacking and promoting. The result is a five-point scale of negativity ranging from one (positive) to five (attack). We apply `SentimentIt` to analyze all televised ads for the U.S. Senate in 2008 ($n = 942$) and compare our estimates to this five-point negativity scale.

For each ad, we created 20 pairwise comparisons for a total of 9,420 tasks. Workers were required to complete an extensive training module and were paid $0.06 per comparison. Workers were instructed to select the ad that was "most negative towards the candidate(s) mentioned, or least positive about the candidate(s) mentioned." In all, 123 workers participated in the task and none were banned.

Figure 3 shows our estimates plotted against the five-point negativity scale. The `SentimentIt` scores are correlated with the expert codings at $r = 0.85$. This is a very high correlation, but there are some disagreements. Table 1 shows the largest disagreements between the `SenitmentIt` measures and those generated by the expert coders. The first example is coded as positive by the expert coders, but the ad is negative in tone and appears to be a strongly negative contrasting ad that was mis-coded. The second ad is coded as contrast (2 = more positive than attack), but `SentimentIt` estimates it as having almost no negative content. In fact, the ad does not mention an opposing candidate or party and is another clear mis-coding. The third ad is coded as contrasting

Figure 3: `SentimentIt` negativity rating relative to five-point expert codes



Note: Using the five-point scale of human codes, we can see that as human codes increase, so do our estimates. In general, the differentiation between codes is evident. The largest disagreements are happening in the middle categories, as discussed in the text.

by expert coders because it mentions both candidates by name. However, the ad mostly consists of strong and negative language, causing our estimate of negativity to be quite high (1.39). Finally, the last example is coded as an attack by human coders because it does not provide positive information about a named candidate. However, the ad only mentions the target of the attack (Collins) in one sentence and otherwise conveys positive information about the Employee Free Choice Act. These last examples illustrate that the strict coding rules designed to improve the reliability of content analysis can obscure the underlying latent trait of interest.

To determine the reliability of our measure, we repeated the exercise 35 days later with 20 more comparisons on a random sample of half the ads ($n = 467$) using the same procedure as before. In all, 52 workers participated in this exercise. We revoked the certification from only one worker. The correlation between the two runs is 0.90.

Human rights reports

Next, we demonstrate how `SentimentIt` can be generalized to larger documents. Specifically, we turn to the sobering task of coding the section entitled "Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment" from all of the U.S. State Department Human Rights Reports issued annually for nearly every country in the world. In this application, we use reports

Table 1: Example ads where SentimentIt disagreed with expert coders

Advertisement	SentimentIt	Experts
[Priscilla Lord Faris]: "Early in this campaign I believed that Al Franken could defeat Norm Coleman. But, no matter how many millions he spends it is clear that his history of pornography, degrading women and minorities, and his questionable financial transactions will continue to be the focus of blistering Republican attack ads. I represent real Minnesota values as a mother, a teacher, a volunteer, and an advocate. I'm Priscilla Lord Faris, I approve this message, and ask for your vote September 9th."	1.26	Positive (1)
[Steve Novick]: "I'm Steve Novick and I approve this message." [John Kitzhaber]: "I'm John Kitzhaber and I approve of Steve Novick. Negative politics as usual or something different? Steve Novick is not a typical politician and he's not running a typical campaign. Steve is standing up for principle and that's why Oregon Democrats are standing up for Steve. Oregon teachers are supporting Steve, so are papers across the state. And I think Steve Novick is the only candidate we can count on for real healthcare reform. Steve Novick, the cure for politics as usual."	-1.09	Contrast (2)
[Announcers]: This isn't complicated. Roger Wicker serves with honor and integrity. Ronnie Musgrove. His ethics? Shameful. Roger Wicker. Supported by Thad Cochran, the VFW, the NRA. Ronnie Musgrove. Supported by pro-abortion, pro-gay marriage groups. Roger Wicker. Never voted for a pay raise, always supports Social Security. Ronnie Musgrove. Failed governor, lost jobs, the beef plant scandal and now he's lying about Roger Wicker. This isn't complicated. Roger Wicker. [Roger Wicker]: "I'm Roger Wicker, and I approve this message."	1.39	Contrast (3)
[Announcer]: CEO's salaries and benefits are getting fatter and fatter... while workers face soaring gas prices, foreclosures, and rising healthcare costs. The Employee Free Choice Act gives workers the freedom to form a union so they can earn better wages, retirement security, and healthcare coverage. Call Senator Susan Collins tell her to support the Employee Free Choice Act and stop siding with wealthy CEO's over working families. American Rights At Work is responsible for the content of this advertising.	-0.64	Attack (5)

Note: When SentimentIt estimates differ from human codes, expert coding scheme's reliance on strict coding rules mischaracterizes the overall tone within the ads.

from 1999 to create a continuous scale indicating the amount of torture conveyed in the report. Hathaway (2002) codes these documents by hand on a five-point scale, with higher values indicating more entrenched, brutal, or frequent torture. Using this measure, Hathaway (2002) argues that ratifying human rights treaties is associated with greater degrees of torture. We emphasize that our aim is not to capture the actual amount or severity of torture in each country – a task well beyond the scope of this article – but only to measure the concept of torture as it is expressed in these reports.

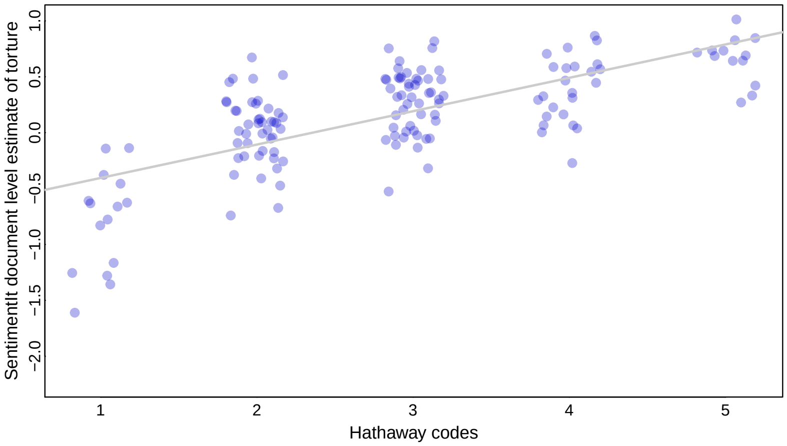
Since many of these reports are quite lengthy and would be difficult for even an expert to read and meaningfully compare, we divided the 182 documents into 1,652 paragraphs. We created 20 comparisons per paragraph. Since reading and understanding these documents is challenging, we paid workers $0.10 per task and required an extensive certification training. We asked workers to indicate which paragraph suggested more torture, which was defined in detail during the training. Our definition and coding rules were designed to approximate the Hathaway definitions (see Appendix).

We posted the tasks in batches of 1,000. The process of gathering the data took approximately three days, and 81 workers completed tasks (we banned none). To analyze the data, we adjusted our statistical model to allow for a hierarchical structure (paragraphs within documents) as described above. The resulting document-level estimates are correlated with Hathaway's coding at 0.69. Figure 4 plots the `SentimentIt` estimates of torture to the Hathaway code. In Figure 5, we show a subset of the document- and paragraph-level estimates generated by our procedure.

Broadly speaking, our measure is consonant with Hathaway's coding. For instance, all of our estimates for the countries coded as one (no torture) by Hathaway fall within the first quantile of our estimates. Yet, there are significant disagreements. One approach to arbitrating between these measures is to closely examine instances where the two codes disagree. We do this in the SI Appendix, and we believe that these examples show that the `SentimentIt` coding more faithfully reflects the level of torture in these documents than the original expert-coded measures.

However, within the space constraints of this letter, we do take two steps to assess the validity

Figure 4: `SentimentIt` document-level estimates of torture on Hathaway codes



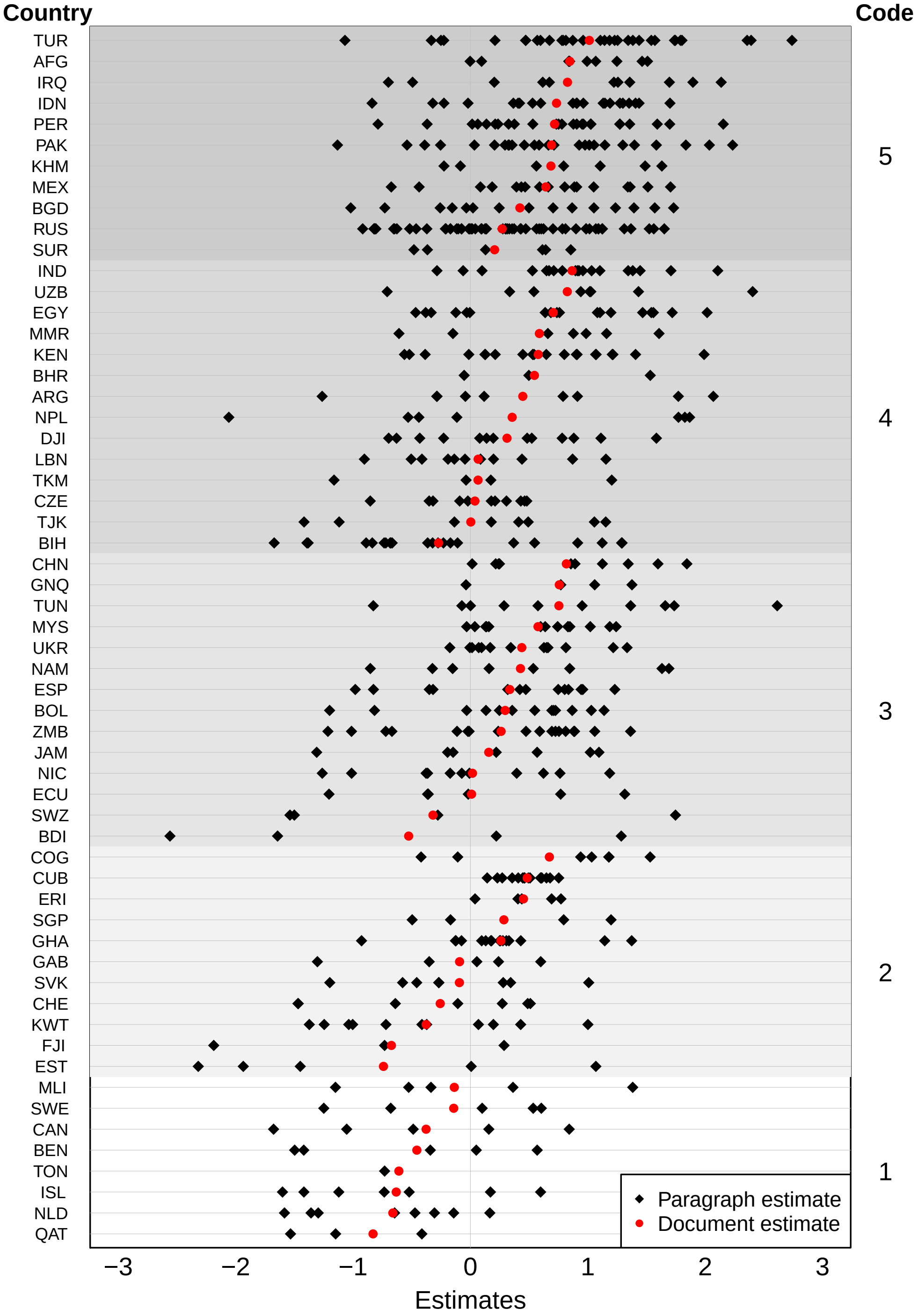
of these estimates. First, we compare the measures with three other well-known measures of torture drawn (in part) from this same document set. Specifically, we analyze the State Department variable of the Political Terror Scale (PTS) data on human rights violations, the Ill-Treatment and Torture (ITT) data, and the torture variable from The CIRI Human Rights Dataset.[4]

Second, we test the degree to which Hathaway's primary substantive claim – that signing anti-torture conventions leads to more abuse (see also Neumayer 2005) – is supported using our measure. Hathaway (2002) finds that signing an anti-torture convention is positively but unreliably associated with torture. We argue that if our measure is valid, this relationship should hold with our measure as well.

For each variable, we conduct separate regressions and calculate standardized regression coefficients. We maintain all controls of the original analysis. Our only deviation is that we look at only one year while the original analysis included 15 years of reports. The results are shown in Table 2. In almost all cases, we find that both the Hathaway and `SentimentIt` scores are positively related to the other measures ($p \leq 0.05$). The only exception is that the relationship between level of torture and ratifying an anti-torture convention is estimated unreliably using Hathaway's coding. However, although both measures are correlated with the alternative torture measures, both the standardized coefficients and R^2 values indicate that the `SentimentIt` measure is more highly related to these other variables. The only exception is the CIRI measure, where the two approaches

[4] Additional information on these measures is provided in the SI Appendix.

Figure 5: `SentimentIt` document- and paragraph-level estimates of randomly selected countries and countries estimated differently than Hathaway



Note: Document-level estimates are shown in red, and paragraph-level estimates are shown in black. There is a clear trend that as Hathaway's codes increase, so do ours. The notable exceptions are explicitly included and discussed in the SI Appendix. Where our estimates deviate from Hathaway's codes, there is strong reason to believe estimates from `SentimentIt` are outperforming expert codings.

Table 2: Standardized OLS coefficients for SentimentIt and Hathaway measures regressed on torture measures and treaty ratification

	Dependent variable:							
	ITT		PTS		CIRI		Torture Convention	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Hathaway	**0.724**		**0.263**		**0.275**		0.252	
	(0.210)		(0.079)		(0.064)		(0.523)	
SentimentIt		**0.964**		**0.298**		**0.275**		**1.296**
		(0.205)		(0.078)		(0.064)		(0.517)
Observations	93	94	106	108	103	105	106	108
R^2	0.382	0.442	0.654	0.663	0.511	0.510	0.343	0.367

Note: Bolded coefficients are significant at the $p \leq 0.05$ level. Additional controls for per capita GDP, population, population growth, trade, foreign aid, GDP growth, civil war, level of democracy, and nation durability not shown. SentimentIt and Hathaway codes are standardized for comparability. CIRI increases as the degree of torture decreases, and is reverse coded in the analyses.

are essentially indistinguishable in terms of their predictive validity.

Finally, to test for reliability of our measures we separately estimate the (roughly) first 10 comparisons to the last 10 comparisons for each paragraph. The correlation between paragraph estimates is 0.77. The correlation between document estimates is 0.88. Considering we are only comparing the estimates between two rounds of 10 comparisons, we consider this to be strong evidence indicating the SentimentIt platform is generating highly reliable measures.[5]

CONCLUSION

In this paper, we propose a novel framework approach to analyzing political texts and measuring latent traits that provides the reliability of automated methods but leverages the superior abilities of humans to read and understand natural language. Specifically, we rely on having an online workforce complete pairwise comparisons of texts. By having the workers indicate which of two documents is more extreme along a dimension of interest (e.g., positivity), including documents in multiple pairwise comparisons, training and monitoring workers, and statistically post-processing

[5] Hathaway randomly sampled 20% of the documents and had a second researcher code the documents, and reports a Cohen's κ of 0.8, suggesting fairly strong agreement between coders (Hathaway 2002, p. 1971). Because we do not have these secondary codes, we cannot directly compare our level of agreement with those reported in Hathaway, and our method does not lend well to calculating Cohen's κ. Thus, although the very high level of correlation we achieve using our system suggests a high level of reliability, we cannot directly compare it to Cohen's encoding.

the worker evaluations, we are able to reliably produce valid estimates of latent traits within texts. Further, we provide software that can automate as little or as much of the process as the researcher desires and provide details about the workflow and steps needed to replicate our approach. In the SI Appendix, we provide extensive documentation, diagnostics, and discussions of practical issues.[6]

We close by noting that, although the method we propose is powerful, it is not entirely unproblematic. Although the examples above show the versatility of the method, there may be unknown limits. Asking coders to evaluate the "tone" of a political ad may differ in kind from asking them to evaluate the quality of the legal reasoning in court cases. Moreover, the tasks we designed were specifically aimed at evaluating aspects of a text that are somewhat objective. Asking workers to draw more deeply on their own judgment to evaluate, say, the "persuasiveness" of a text, may require a more representative workforce. Further, although the method is scalable relative to having trained experts coding data, there is still a per-document cost. Yet, even these may be reduced by coding subsets of larger document sets and using supervised learning methods to create estimates for the remaining set.

[6] It is worth noting that the `SentimentIt` system is a webservice, which requires ongoing support and maintenance. While not unprecedented (e.g, King 2007), providing a webservice as a method rather than a standard piece of software is somewhat unusual in the field. However, we note that, first, even supposedly "stand alone" software tools require regular maintenance and updating to adjust to, for instance, changes in operating systems. Second, the webservice is itself a piece of open-source software that could in principal be hosted by another researcher or university. Finally, given the direction of technology towards cloud computing and distributed systems, we believe that creating and hosting online apps like `SentimentIt` will be increasingly common in the field of political methodology.

References

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: Reproducible and agile production of political data." *American Political Science Review* 110(2): 278–295.

Berinsky, Adam J., Gergory A. Huber, and Gabriel S. Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3): 329–50.

Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael A. Betancourt, Michael Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2016. "Stan: A probabilistic programming language." *Journal of Statistical Software* .

Hathaway, Oona A. 2002. "Do human rights treaties make a difference?" *The Yale Law Journal* 111(8): 1935–2042.

Henderson, John A. 2015. "Using experiments to improve ideal point estimation in text with an application to political ads." Unpublished manuscript.

Hitlin, Paul. 2016. *Research in the crowdsourcing age, a case study*. `www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/`: Pew Research Center.

King, Gary. 2007. "An introduction to the Dataverse Network as an infrastructure for data sharing." *Sociological Methods and Research* 36(2): 173–199.

King, Gary, Christopher JL Murray, Joshua A Salomon, and Ajay Tandon. 2004. "Enhancing the validity and cross-cultural comparability of measurement in survey research." *American Political Science Review* 98(01): 191–207.

Neumayer, Eric. 2005. "Do international human rights treaties improve respect for human rights?" *Journal of Conflict Resolution* 49(6): 925–953.

Oishi, Shigehiro, Jungwon Hahn, Ulrich Schimmack, Phanikiran Radhakrishan, Vivian Dzokoto, and Stephen Ahadi. 2005. "The measurement of values across cultures: A pairwise comparison approach." *Journal of Research in Personality* 39(2): 299–305.

Sheng, Victor S, Foster Provost, and Panagiotis G Ipeirotis. 2008. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM pp. 614–622.