V-Dem
INSTITUTE

Experts, Coders, and Crowds: An analysis of substitutability

Kyle L. Marquardt, Daniel Pemstein, Constanza Sanhueza Petrarca, Brigitte Seim, Steven Lloyd Wilson, Michael Bernhard, Michael Coppedge, Staffan I. Lindberg

October 2017

Working Paper
SERIES 2017:53
THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Varieties of Democracy (V–Dem) is a new approach to conceptualization and measurement of democracy. It is co-hosted by the University of Gothenburg and University of Notre Dame. With a V–Dem Institute at University of Gothenburg with almost ten staff, and a project team across the world with four Principal Investigators, fifteen Project Managers (PMs), 30+ Regional Managers, 170 Country Coordinators, Research Assistants, and 2,500 Country Experts, the V–Dem project is one of the largest ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V–Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

V–Dem Working Papers are available in electronic format at www.v-dem.net.

Experts, Coders, and Crowds: An analysis of substitutability*

Kyle L. Marquardt† Daniel Pemstein‡
Constanza Sanhueza Petrarca† Brigitte Seim§
Steven Lloyd Wilson¶ Michael Bernhard‖
Michael Coppedge** Staffan I. Lindberg†

*First authors are listed alphabetically and are followed by second authors, also listed alphabetically. We thank Ryan Bakker, Ken Benoit, Adam Glynn, Noah Nathan, Amy Semet and participants at the 2017 APSA Annual Meeting, EPSA General Conference, MPSA Annual Conference and V–Dem Annual Conference for comments on earlier drafts of this paper. Josefine Pernes and Natalia Stepanova provided invaluable administrative support, and Paige Ottmar provided outstanding research assistance. This research project was supported by Riksbankens Jubileumsfond, Grant M13-0559:1, PI: Staffan I. Lindberg, V–Dem Institute, University of Gothenburg, Sweden; by Knut and Alice Wallenberg Foundation to Wallenberg Academy Fellow Staffan I. Lindberg, Grant 2013.0166, V–Dem Institute, University of Gothenburg, Sweden; by the National Science Foundation under Grant No. SES-1423944, PI: Daniel Pemstein; as well as by internal grants from the Vice-Chancellor's office, the Dean of the College of Social Sciences, and the Department of Political Science at University of Gothenburg. Human subjects research approved by Regionala Etikprövningsnämnden i Göteborg, DNR 1079-16.

†V–Dem Institute, University of Gothenburg
‡North Dakota State University
§University of North Carolina, Chapel Hill
¶University of Nevada, Reno
‖University of Florida
**University of Notre Dame

Abstract

Recent work suggests that crowd workers can replace experts and trained coders in common coding tasks. However, while many political science applications require coders to both find relevant information and provide judgment, current studies focus on a limited domain in which experts provide text for crowd workers to code. To address potential over-generalization, we introduce a typology of data producing actors—experts, coders, and crowds—and hypothesize factors which affect crowd-expert substitutability. We use this typology to guide a comparison of data from crowdsourced and expert surveys. Our results provide sharp scope conditions for the substitutability of crowd workers: when coding tasks require contextual and conceptual knowledge, crowds produce substantively different data from coders and experts. We also find that crowd workers can cost more than experts in the context of cross-national panels, and that one purported advantage of crowdsourcing—replicability—is undercut by an insufficient number of crowd workers.

Political scientists often rely on experts to code data. While expertise plays an important role in a wide range of coding tasks, an increasing number of "expert surveys" (e.g., the British Election Study Expert Survey, the Chapel Hill Expert Survey, the Electoral Integrity Project, Quality of Government, Transparency International, and Varieties of Democracy) prominently tout the advantages of expert coders. Experts' purported virtues are manifold. They allow researchers to obtain information on complex topics; gather data cross-nationally and over time, even when observable indicators (e.g. roll call votes or election manifestos) are not universally available; and deductively determine the content of their measures (Hooghe, Bakker, Brigevich, de Vries, Edwards, Marks, Rovny & Steenbergen 2010). Surveys also distribute coding efforts across the research community, providing a widely accessible public good, and are inexpensive compared to fieldwork, archival research, and large-scale public and elite surveys.

Crowdsourcing—the large-scale recruitment of lay persons to code data—has emerged as a tool for data collection in traditionally expert-reliant domains (Kittur, Chi & Suh 2008, Cooper, Khatib, Treuille, Barbero, Lee, Beenen, Leaver-Fay, Baker, Popovic & Players 2010, Honaker, Berkman, Ojeda & Plutzer 2013, Benoit, Conway, Lauderdale, Laver & Mikhaylov 2016, D'Orazio, Kenwick, Lane, Palmer & Reitter 2016). The key distinction between these approaches is that expert surveys typically rely on extracting highly specific and accurate knowledge about each case from a few experts, while crowdsourced methods average across many error-prone non-experts. However, assuming a sufficiently large crowd size and no systematic bias, crowdsourcing can theoretically provide an accurate measure of any concept. Indeed, Benoit et al. (2016) argue that crowds can match or outperform experts on four dimensions:

1. **Reliability**: Experts are potentially subject to biases (e.g. ideology, socialization, education) that crowd workers are not.

2. **Validity**: A crowdsourced approach can be implemented consistently for certain exercises across contexts.

3. **Cost efficiency**: Internet-based crowdsourcing may be cheaper than expert surveys due to smaller payments to individual coders.

4. **Replicability**: Given the cost and often unclear sampling procedure for recruiting experts, expert-coded datasets are not easily reproducible. In principle, crowd-based data are replicable.

In the context of coding researcher-supplied text, Benoit et al. (2016, pp. 279) "show that properly deployed crowdsourcing generates results indistinguishable from expert ap-

proaches." In conjunction with these advantages, they argue that such substitutability serves as "as a proof of concept," with utility that "extends to all subfields of political science."

This paper demonstrates that this claim has strict scope conditions. We first develop a typology of actors—experts, trained coders, and crowd workers—who produce secondary data.[1] We then theorize about when crowds can substitute for traditional secondary data generation, in particular how the *task*, *coder* characteristics, and *incentives* affect the ability of crowds to produce data of similar quality to experts or trained coders. We also confront the questions of 1) if crowdsourcing is inherently cheaper than expert coding and 2) if there are enough crowd workers for actual replication.

Second, we combine experimental and observational evidence to examine whether crowdsourced data can substitute for expert-coded data in contexts beyond coding researcher-supplied text. Specifically, we compare crowdsourced data to data from the Varieties of Democracy (V–Dem) v7.1 data set, which provides a wide array of measures of political institutions across space and time (Coppedge, Gerring, Lindberg, Skaaning, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, Lührmann, Marquardt, Mechkova, McMann, Olin, Paxton, Pemstein, Pernes, Sanhueza Petrarca, von Römer, Saxer, Seim, Sigman, Staton, Stepanova & Wilson 2017). Two types of coders generate this V–Dem data: 1) experts, who use Likert scales to provide cross-national longitudinal measures of concepts which are difficult or impossible to directly measure (e.g. the degree to which Argentina was free from political killings in 2014); and 2) trained coders, who use a standardized set of instructions to code directly observable ("factual") questions (e.g. the minimum voting age in Argentina in 2014). We examine how incentives, task complexity, and crowd coder characteristics influence substitutability between coding approaches.

We find that crowd workers are poor substitutes for experts. The correlation between crowd and expert codings varies from weakly positive to negative across questions; there is also little evidence that crowd averages converge toward expert averages. We also find that the differences between crowds and experts increase with higher pay, perhaps because it encourages low quality crowd workers to complete all tasks. Most task and coder characteristics have no substantial influence on the difference between crowds and experts.

We further find little evidence that crowd workers can substitute for trained coders. However, some crowd worker characteristics are correlated with better performance, and crowd worker performance is sensitive to both the complexity and information availability of factual questions. These latter findings indicate that it may be possible to design a crowd survey that accurately gathers relatively accessible and simple forms of factual data.

[1]For our purposes, primary data are directly observable (e.g., public opinion survey results or roll call votes). Secondary data are observations produced in a manner requiring human judgment.

Overall, these findings constitute strong evidence that experts and trained coders remain necessary for coding enterprises where the data generating process requires contextual or conceptual knowledge. Broadly, they demonstrate that while crowd workers can provide their opinions and accurately code expert-generated text data (Benoit et al. 2016), they do not have the necessary training or incentives to provide judgment on more complicated phenomena.

1 Theorizing experts, coders, and crowds

Even in the case of directly observable data, social-scientific coding tasks generally require judgment and contextual knowledge. For example, determining the current head of government in Iran requires 1) knowledge of the respective roles of the President and Supreme Leader and 2) a conceptual basis for determining which leader has greater power. The required judgment increases for concepts that are difficult or impossible to directly observe.

The conditions necessary to generate such secondary data with high validity remains under-theorized and under-researched. While political science applications of crowdsourcing often frame crowds as an alternative to experts (Honaker et al. 2013, Benoit et al. 2016, D'Orazio et al. 2016), we argue these applications generally involve replacing "trained coders," not "experts." To clarify this argument, we make a threefold distinction between experts, trained coders, and crowds.

1.1 What is an expert?

We use Morris (1977, pp. 679) as a baseline for defining an expert: an expert is "anyone with special knowledge about an uncertain quantity or event." We consider this definition to pertain to individuals who have devoted a significant portion of their lives to developing specialized knowledge and are practiced in discovering new information, such as academics or practitioners.

Two aspects of expertise require further discussion. First, expertise is not generic. A political science professor who studies ethnic conflict in Azerbaijan is an expert on "ethnic conflict" and "Azerbaijan," but not necessarily either of these topics in other contexts (e.g. "ethnic conflict in Kenya" or "natural resource allocation in Azerbaijan"). In the context of data-collection enterprises, this qualification means that the identification of expertise is question- and case-specific.

Second, expert judgments remain subject to a variety of potential biases. As a result, scholarship has long held that gathering data from multiple experts to correct for individual

biases is a necessary practice (Tetlock 2005, pp. 31–34). Such practice is especially important when the task involves converting unobservable multidimensional data into a unidimensional scale, as is common in expert-coding enterprises.

The V–Dem project provides an illustrative example of how expert-coding enterprises operationalize this conceptualization of experts. V–Dem collects data on a variety of political attributes by country and year (Coppedge, Gerring, Lindberg, Skaaning, Teorell, Krusell, Marquardt, Mechkova, Pemstein, Pernes, Saxer, Stepanova, Tzelgov, Wang & Wilson 2017). For expert-coded questions, the project recruits approximately five country experts (CEs) for each country-year. To select CEs, V–Dem project managers based at the University of Gothenburg collaborate with regional managers (established scholars broadly aware of experts in the countries of their region) and country coordinators (scholars aware of experts on their country) to develop a list of potential recruits for each of 11 subject areas (e.g. media, elections, or political parties). The baseline for recruitment follows from the previous discussion: CEs generally hold an advanced social science degree, and coordinated validation between project managers, country coordinators, and regional managers ensures that the recruited experts have deep knowledge of the country, the specific subject area, and relevant concepts.

CEs receive monetary compensation for their service (US $1,248 for completing all eleven surveys of a single country covering the period 1900-2012, and US $25 for yearly updates of a single country/survey). The opportunity costs involved in completing a survey, which can take multiple days or weeks, often outweigh the compensation. Thus, if V–Dem experts agree to code, in many cases it is likely not solely for the money. Instead, their incentive may take the form of non-material benefits such as contributing to a public good or a sense of obligation. While this incentive structure does not necessarily lead to higher quality data, it diverges from the more material incentives of other types of coders.

1.2 What is a coder?

A common mode for secondary data generation relies on "trained coders," of which there are two types. The first uses protocols to classify provided materials, making rank-order judgments about democracy levels in different states (Honaker et al. 2013), applying a coding frame to manifesto text (Benoit et al. 2016) or coding the presence or absence of militarized interstate incidents in furnished reports (D'Orazio et al. 2016). The second conducts background research, typically using a protocol to code variables that relate to directly observable data.

The characteristics of these trained coders generally differ substantially from those of

experts. They often do not have an advanced degree in the subjects they are coding and do not typically have the information necessary for task completion prior to beginning the task (e.g., undergraduates majoring in political science who have applied for research experience credits). Instead, they are generally pre-screened for suitability for the task and trained to complete it. However, through the act of completing the task, these coders may develop a level of knowledge equivalent to expertise.

1.3 What is a crowd worker?

Individuals who constitute "crowds" are not a random selection of individuals across the world. Instead, they are individuals associated with online enterprises such as Amazon Mechanical Turk (MTurk) or Crowdflower. Research indicates many crowd workers are intelligent and trainable, like trained coders: Paolacci, Chandler & Ipeirotis (2010) found no difference between crowd workers, undergraduates, and other Internet users on a self-reported measure of numeracy that correlates highly with actual quantitative abilities. However, crowd workers often learn more slowly and have more difficulty with complex tasks than university students, perhaps reflecting differences in age and education (Crump, McDonnell & Gureckis 2013).

Unlike trained coders, crowd workers are not recruited because of their ability to acquire knowledge or technical expertise. These workers conduct a large number of different tasks, and are neither trained nor incentivized to master a specific one. They are almost wholly motivated by the financial incentive for task completion, though they may find some tasks more interesting than others.

The main advantage of crowd workers over trained coders and experts is that they are relatively cheap and numerous. Crowd workers receive only small compensation for each task, which means that a researcher can recruit hundreds or thousands of them for a relatively small cost. As of 2015, there were over half a million registered users on MTurk (Peer, Samat, Brandimarte & Acquisti 2016) providing a sufficient quantity for a range of coding tasks.

2 When can crowds replace experts and coders in practice?

The primary aim of this paper is to further establish conditions under which crowds can substitute for experts and trained coders in the generation of secondary data. The research of Benoit et al. (2016) has already demonstrated that crowd workers can substitute for

trained coders in coding data based on texts with which they are provided. As a result, we focus here on tasks that require knowledge. We first compare the resources necessary for implementing a crowdsourced project as a substitute, then theorize which task and coder characteristics should facilitate substitutability.

2.1 Resources necessary to recreate an expert-coded dataset with crowd workers

To illustrate the resources necessary to recreate an expert-coded dataset with crowdsourcing, we use the V–Dem project as a reference. We focus only on the expert-coded portion of V–Dem, since the analogous cost comparison for the trained-coder portion of the dataset is sensitive to incentive expectations (e.g. salary, course credit) of potential recruits.

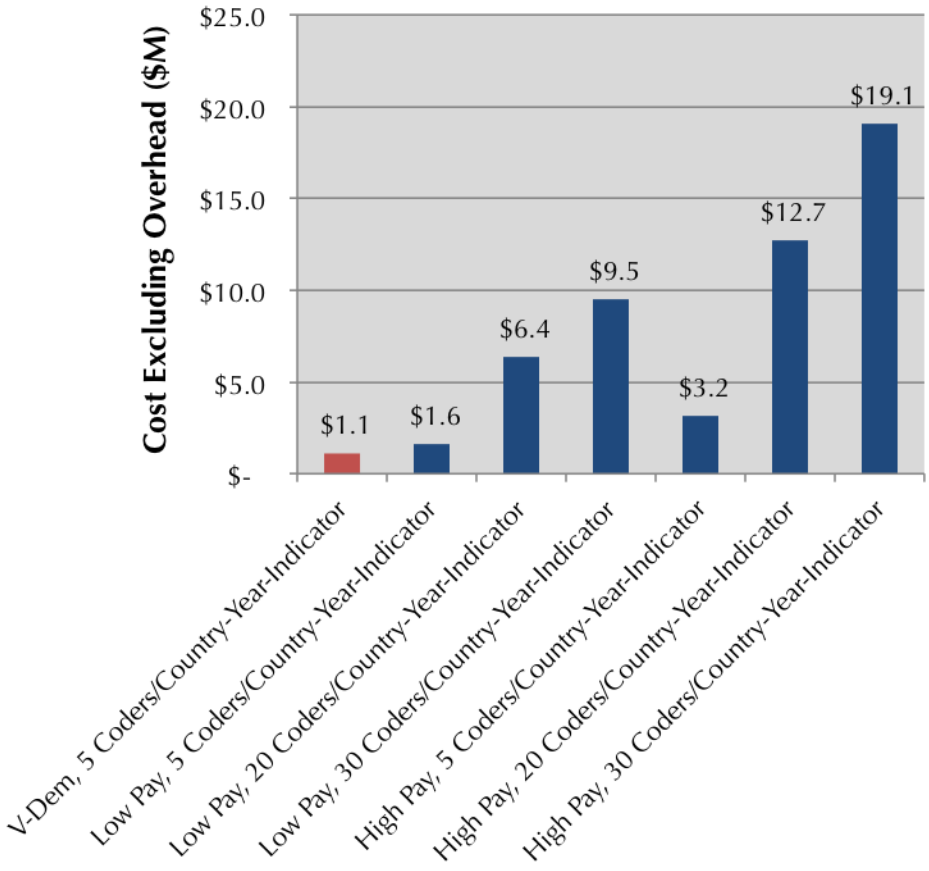
We focus on the initial V–Dem coding wave, in which 2,500 experts coded 177 countries, 113 years, and 151 indicators. Each observation generally had five or more coders. A (hypothetical) expert coding the entire set of 151 indicators (spanning eleven surveys for varying subject areas) for a given country, would receive $1,248, regardless of the number of years they coded (1900-2012). The average number of years coded was 99, which we use as an input to estimate the crowd costs. The average cost per observation is thus $0.08 per country-year-indicator.

In Figure 1, we present the total cost for producing the initial V–Dem dataset under different scenarios. The first scenario (in red) represents the cost of using expert coders. The next three scenarios represent producing the same dataset using crowd workers and paying them $0.12 to code each country-year-indicator observation, which is a typical payment on MTurk. We model the costs allowing for five coders per country-year-indicator, then 20 and 30, on the assumption that crowd coder responses will converge less quickly than expert responses. The final three scenarios represent producing the same dataset paying crowd workers $0.24 to code each observation, which is a high payment by crowdsourcing standards but perhaps reasonable given the difficulty of the coding task. Figure 1 illustrates that there is little reason to believe that crowdsourced data are cheaper than expert-coder data under these conditions.[2]

A related issue is that there may not be sufficient coders to reproduce cross-national datasets, such as V–Dem v7.1, which has over 2.5 million unique observations originating from experts' assessments. If we optimistically assume that crowd workers are willing to

[2]Please note that the scenarios depicted in Figure 1 exclude overhead costs, focusing only on coder remuneration. It is not clear which coder population—experts or crowd workers—would require more overhead: both projects would require staff and resources to develop and implement instruments, as well as oversee data collection, validation, and publication.

Figure 1: Coder Costs of Producing V–Dem Dataset with Experts and Crowds



spend an hour on a project and a minute on each task, then each worker would complete 60 observations. Replicating the V–Dem dataset with five crowd workers per observation would require over 220 thousand workers; with 20 crowd workers per observation 990 thousand; and with 30 coders per observation over 1.3 million. As of 2015, Amazon MTurk had approximately 500 thousand workers, which would only be sufficient for 11 coders per observation. Therefore, unless crowd workers are very statistically efficient, it would be difficult to replicate a dataset the size of V–Dem, given the existing worker population.

2.2 Project characteristics that affect substitutability

The above discussion assumes that, in sufficient *number*, crowd workers can produce secondary data which substitute for their expert- and trained coder equivalents. However, such substitutability may depend on the coding *task*, *crowd worker characteristics*, and *incentives* provided to coders. We discuss each of these characteristics in turn and provide observable implications for testing their effects.

Task attributes can be disaggregated into the sub-attributes *issue complexity*, *task complexity*, *information*, and *bias*. Task complexity is determined by the specialist knowledge implicit in the task itself, *issue complexity*. Tasks requiring the expert-level understanding of concepts and that are difficult to convey in a short description exhibit high issue complexity, which we expect to be more difficult for crowd workers. We hypothesize that the manner in which the researcher conveys the task to the coder, *question complexity*, is also important. While experts have experience parsing complex descriptions in their fields, and coders can be trained to handle complicated tasks, we do not expect the average crowd worker to possess such capabilities.

Information availability should also influence crowd substitutability. Experts have access to information unavailable to the general population and can draw on years of case-specific experience. Trained coders generally have the necessary time to seek out requisite sources of information, though not necessarily information unavailable to the general population. We expect that crowd workers neither have the time, nor the training, to find such information. Similarly, we predict that measuring observable concepts (e.g. *de jure* institutions like voting age) should be easier for crowd workers compared to concepts that are difficult or impossible to directly observe (e.g. *de facto* political and institutional conditions, such as the degree to which the judiciary is independent from influence by the executive).

The final *task* characteristic is relative *potential bias* across experts and crowd workers. Polarizing issues that activate implicit biases or emotional reactions are likely to disadvantage crowd workers relative to experts. While experts on a polarizing issue have an understanding

of the appropriate range that has existed across polities over time, we expect the modal crowd worker to have lower tolerance. Thus, we predict crowd workers will show explicit biases towards extreme "all-or-nothing" responses.

Crowd worker characteristics may also affect substitutability. We focus on two dimensions: *background* and *diligence*. We expect coders with a background of living or studying in a particular country, or those who speak a country's language, to better generate secondary data about the country than other workers; these characteristics are generally prerequisites for experts (though not trained coders). Similarly, those with baseline knowledge or interest in the given domain, another requirement for experts, should outperform other workers. For instance, people who follow politics in a broad sense may outperform those who find politics boring. Similarly, education level allows access to knowledge and information, and higher education should predict substitutability.

We also hypothesize that *diligence* will influence the accuracy of crowd workers' codings. While bots are the most egregious example of crowd workers who are not diligent, other workers may rush to complete the tasks without fully considering their responses. Unless crowd workers are penalized for inaccurate codings, they have every financial incentive to complete a survey as quickly as possible. Experts and trained coders, however, have strong incentives—due to a sense of obligation (experts) or material interests (trained coders)—to accurately complete tasks, regardless of time. "Screener" tasks and other measures of compliance allow researchers to assess the degree to which respondents are paying attention to the task. Similarly, measures of time investment provide researchers with a sense of whether or not coders are rushing through the survey.

Finally, because crowds are primarily pay-motivated, we hypothesize that increased payment can offset task difficulty. However, since one of the primary advantages of crowdsourcing is its (supposedly) relatively low cost, there is a clear tension between decreasing average coding error and higher payment, as Figure 1 succinctly illustrates.

Table 1 provides a reference for the categories we discuss in this section. We organize the categories into subcategories, explain how we expect variation in these characteristics to determine the relative ability of crowd workers to substitute for experts, and describe how we operationalize these hypotheses.

3 Research design

We test these hypotheses by comparing V–Dem's raw expert- and trained coder-generated secondary data to crowd responses to a March 2017 survey. We intended this survey to be a pilot for a larger endeavor, but the pilot data indicate that conducting the full-scale

Table 1: Variables and hypotheses

Category	Name	Description	Hypothesized Effect on Substitutability	Hypothesis Test
Task (Complexity)	Issue Complexity	The amount of nuance and complexity of issues considered in the task	−	Observational question trait: Average V–Dem expert coder confidence (Section A.2.1)
Task (Complexity)	Question Complexity	The complexity of the question language	−	Observational question trait: The length of the English language text of the V–Dem question plus the length of all of the choices for each question, measured in number of characters (Section A.2.1)
Task (Information)	Information Availability	The amount of information available to assist the participant in completing the task	+	Observational country-year trait: Measure based on quantity of information available online and in published texts for each country-year (Section A.2.3)
Task (Information)	Recency	Whether the task pertains to recent events or political phenomena	+	Observational year trait: Six purposively-selected five-year spans (30 years total) between 1915 and 2015 (Section A.2.2)
Task (Information)	Verifiable (vs. Perception)	Tasks that have a verifiable, correct answer (as opposed to relying on a subjective perception of a latent trait)	+	Observational question trait: Four purposively-selected factual V–Dem variables of varying findability and dimensionality (Section **??**)
Task (Bias Potential)	Issue Polarization	Polarizing issues are those that activate pre-conceived biases, emotional reactions, or personal experiences	−	Observational question trait: Purposively-selected polarizing question about political killings (Section **??**)
Coder (Background)	Case Familiarity	Whether the task pertains to a case known by the participant	+	Observational coder trait: Indicators representing whether a coder is 1) coding a country of long-term residence, and 2) fluent in an official language of the country
Coder (Background)	Baseline Knowledge	Whether the coder has a baseline level of knowledge relevant to the task	+	Observational coder trait: Indicators of whether a coder 1) discusses politics and 2) earned a degree in political science
Coder (Background)	Education	Whether the coder is educated	+	Observational coder trait: Indicator of complete or incomplete university education
Coder (Diligence)	Compliance	Whether the coder is able to pass "compliance" tests	+	Observational coder trait: Indicators of a coder's successful completion of questions that gauge 1) attention and 2) basic competency (Section **??**)
Coder (Diligence)	Time Investment	How much time is taken to complete the task	+	Observational coder trait: How long the coder took to complete each task
Incentives	Pay	The magnitude of the per-task payment received by the participant	+	Experiment: Randomly-assigned high/low payment condition

experiment is unwarranted.[3]

We ran the study on the crowdsourcing platform CrowdFlower;[4] workers self-selected into the research pool, though we randomized aspects of the task and incentives.

3.1 Task characteristics

We randomly assigned each respondent two of nine V–Dem indicators, and asked the respondents to code these indicators sequentially for six five-year periods for Argentina. Following completion of the second indicator, workers optionally coded the same (second) indicator for an additional set of six five-year periods for Senegal. We briefly discuss our rationale for each of these choices in this section; for a more in-depth discussion of some choices, see Appendix A.

We selected nine V–Dem indicators to ensure variation in information type, question and issue complexity, and polarization. Five of the nine indicators are based on perception, and CEs code them for V–Dem. The remaining four indicators pertain to verifiable, factual data; V–Dem uses trained coders to collect these data. We expect crowd workers to perform worse (i.e., be less substitutable for experts) on the perception-based indicators than on the factual indicators, as coding these indicators requires high levels of contextual and conceptual knowledge. The information and question format provided to the crowd workers corresponds to that which V–Dem CEs and trained coders see in their coding interface, making the tasks similar and thus facilitating comparison.[5]

We asked crowd workers to code a single country for two indicators to make the coding task easier, assuming that a worker could use similar data sources for both indicators; the rationale for asking her to code the same (second) indicator for the second country is the same. We chose Argentina and Senegal as countries for analysis because they are not the most internationally-prominent cases—relative to the United States, Russia or China—but are essential cases for any coding enterprise with cross-national pretensions. While the average crowd worker can likely better code indicators regarding the United States than those regarding Argentina or Senegal, evidence of substitutability in that case could exaggerate the scalability of the enterprise. Analyses in Appendix A indicate that Argentina has an intermediate level of English-language data availability in terms of Wikipedia data on poli-

[3]Since we designed this study as a pilot, we did not register a pre-analysis plan (PAP). However, we did write a PAP, describing both the pilot and the full experiment, which we distributed among co-authors prior to the pilot. Due to the more limited nature of the data in the pilot, we do not conduct every analysis in the PAP, and instead treat it as a set of pre-specified analyses. We note key areas where our analysis diverges from the PAP in the manuscript. See Appendix J for an anonymized version of the circulated PAP.

[4]For CrowdFlower details, see Peer et al. (2016) or Shapiro, Chandler & Mueller (2013).

[5]See Appendix B for example screenshots of the crowd worker interface.

tics, while Senegal has a low level, providing us with some leverage to examine the effect of information availability on substitutability: we expect the average crowd worker to better code data for Argentina relative to Senegal (Argentina is the reference level in analyses, with *Senegal* representing codings for this country).

We chose to use five year periods—not randomly-selected years or the complete period—to reduce the workload on crowd workers, while ensuring many observations per country-year-indicator. In selecting years, we considered two main criteria. First, in line with our recency hypothesis, we selected both recent and past five-year periods. Second, we selected periods that span major international political events to assess the degree to which crowd workers static-code (i.e. do not change their codings over time). The six five-year periods, which we denote by their final year, are: *2005*, *1996*, *1970*, *1950* and *1920*; we use 2011-2015 as the reference level in analyses. We expect that crowd workers will perform worse coding years farther in the past.

3.2 Incentive characteristics

We randomly assigned workers to a typical ($0.12) or high ($0.24) payment condition, determining their level of remuneration for each coding task they performed (i.e. they received a payment for each country-year-indicator they coded). We use the typical payment condition as the reference level in analyses, denoting the high payment treatment with an indicator *HighPay*. We expect crowd workers in the high payment condition to provide codings that are more substitutable for expert- and trained-coder coded data.

3.3 Coder characteristics

We also analyze the relationship between crowd worker characteristics and substitutability. To assess the effect of education on crowd performance, we use the variable *No university education*, expecting that crowd workers without university education will perform worse. We analyze baseline knowledge of political concepts with two variables: *PoliSci major*, which indicates a worker who majored in political science during their undergraduate or graduate studies; and *Does not discuss politics*, which identifies a worker who reports not discussing politics.[6] We expect crowd workers with a political science education, and who discuss politics, to provide more substitutable codings for experts and trained coders.

We also analyze case familiarity using two variables at the country-indicator level. *Resided in coded country* indicates a worker who reports having lived in the coded country for an

[6]The survey included three questions related to political awareness. We did not specify a variable preference in the PAP.

extended period. *Reads in language of coded country* represents workers who are fluent in Spanish for Argentinian cases or French for Senegalese cases. We expect workers who have greater familiarity with the cases to provide more substitutable codings.

To test hypotheses regarding the relationship between coder diligence and substitutability, we analyze three variables. Two of these variables reflect basic compliance with, and understanding of, the coding task. First, prior to coding each variable, workers were presented with one of two randomly-selected hypothetical cases (corresponding to the highest and lowest level of the Likert scale for that variable), and asked to code the case using the indicator Likert scale. The text for the hypothetical cases for political killings read as follows:

1. *In Country X, political killings were practiced systematically and they were typically incited and approved by top leaders of government.* This case corresponds to the lowest level of the Political killings scale, which reads: *Not respected by public authorities. Political killings were practiced systematically and they were typically incited and approved by top leaders of government.*

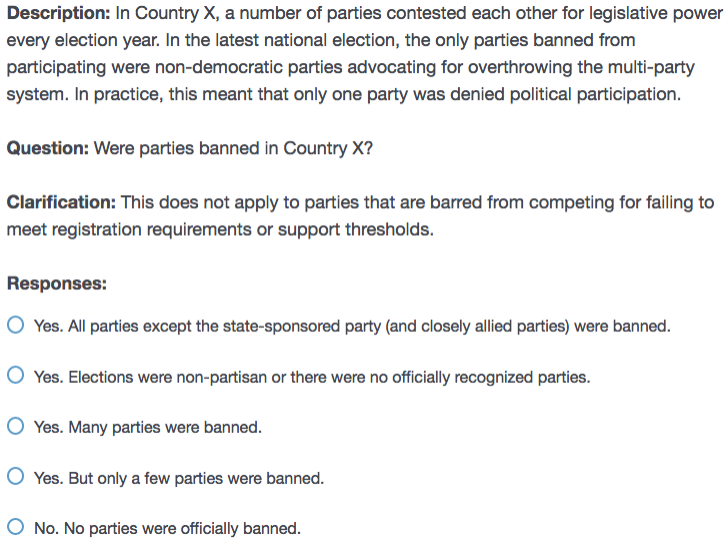
2. *In Country X, political killings were non-existent.* This case corresponds to the lowest level of the Political killings scale, which reads: *Fully respected by public authorities. Political killings were non-existent.*

These screeners serve two purposes: they 1) familiarize crowd workers with the scale, and 2) provide data on workers who are unable to perform the task. We use data on crowd responses to create a variable *All screeners correct*, which indicates a crowd worker who correctly coded both screener questions. Thirty-seven percent of workers coded both screener questions correctly; 25% of workers coded neither screener correctly. We expect workers who correctly coded both screeners to provide more substitutable data.

We also tested worker diligence with four hypothetical questions with Likert scale responses, based on V–Dem anchoring vignettes. These questions account for differential item functioning among experts, and thus have two plausible contiguous responses (out of five possible responses). Figure 2 presents an example of one of these questions, which we henceforth refer to as "Gold" questions, following Benoit et al. (2016). In the case of the example, both "4" and "5" are "correct" responses.

To correctly code these questions requires a basic ability to read, understand concepts, and provide judgment; they are thus more difficult than the screener questions but perhaps more akin to traditional crowdsourced tasks, which do not require any contextual knowledge. Accordingly, crowd workers performed worse, but not comparatively abysmally, than CEs on these tasks: "correct" response rates by question for CEs ranges from 82 to 91%, compared

Figure 2: Gold question example



to 57 to 63% for crowd workers.[7] Responses to these questions provide additional data for measuring individual-worker diligence, and we create the variable $3 \leq Gold$ to represent a worker who correctly coded three or more of the four Gold questions. We expect such workers to provide data that are more substitutable for experts.[8]

The third diligence variable, *Variable coding time*, is at the country-indicator-coder level, and represents the log-transformed time it took a crowd worker to complete a variable. Assuming that correctly answering a question requires some degree of research, crowd workers who spend more time coding should provide more substitutable codings.

Finally, we include three standard control variables: the continuous variable *Age*, a dichotomous indicator *Female*, and *Did not use V–Dem*, which controls for whether or not a worker reported using the V–Dem website in the process of coding.[9]

3.4 The sample

Our sampling strategy achieved approximately 20 observations per indicator-country-year-treatment group. We required 10,800 observations (30 years × nine variables × two treat-

[7] Thirty-one percent of workers coded all four gold questions correctly, 17% three, 24% two, 14% one and 14% zero; correct response rates ranged from 57 to 63% for individual gold questions

[8] We did not specify the operationalization of the Gold or Screener questions in the PAP. Note that Benoit et al. (2016) uses such questions to remove noncompliant crowd workers from the sample. Accordingly, we also conduct analyses of the data with samples divided by the number of screener and Gold questions the workers correctly answered (Appendix H.1).

[9] We did not include this variable in the PAP, but analyze it here because, in principle, coders could have simply accessed the V–Dem website and used online tools to replicate the dataset. Indeed, a majority of crowd workers reported using V–Dem data. Their poor performance on tasks indicates that they either erroneously reported using the data, or used them incorrectly or sporadically.

ment conditions × 20 observations), or 180 coders (two variables per coder); given the likelihood of attrition, we recruited an extra 20% above the needed sample size, for a total of 216 individuals. Given idiosyncrasies with implementation, our final sample size was 229 coders.

4 Results

We compare crowdsourced data first to expert-coded data, then trained coder-coded data. In both contexts, we examine the relationship between task characteristics and incentives and the substitutability of crowdsourced data, then turn to the effect of crowd coder characteristics.[10] Please note that these analyses concern the degree to which crowd data are equivalent to V–Dem data. Technically, they concern only substitutability, not accuracy, which is beyond the scope of this paper.

4.1 Expert-coded data

We selected perceptual indicators—indicators that V–Dem CEs code—that exhibit different combinations of issue and question complexity. We assessed all five-point Likert scale questions in the V–Dem dataset using proxies for both forms of complexity, then constructed a 2×2 table representing questions in the lower and upper quartile range on both dimensions of complexity. To proxy question complexity, we took the combined length of a variable's English language text and response-category descriptions, measured in number of characters. We proxy issue complexity by evaluating the average confidence (from 1 to 100) that V–Dem CEs self-assigned their ratings for each particular question, assuming that questions that address more nuanced issues should be answered with lower confidence, whether or not the question language is particularly complex. Finally, we randomly selected an indicator from each cell. This process yielded the following indicators:

1. *Forced labor*: The degree to which males were free from forced labor ($v2clslavem$ in V–Dem codebook). High issue and question complexity.

2. *Gender equality*: The degree to which political power was equally distributed by gender ($v2pepwrgen$). High issue complexity and low question complexity.

3. *Journalist harassment*: The degree to which journalists did not face harassment ($v2meharjrn$). Low issue complexity and high question complexity.

[10] We focus on substitutability, though Appendix D presents an item non-response analysis. See Appendix C for descriptive statistics.

4. *Judicial independence*: The degree to which a high court does not make decisions based on government positions ($v2juhcind$). Low issue and question complexity.

We expect crowd workers to provide the most substitutable data for Judicial independence, the least complex question; and least for Forced labor, the most complex question; performance on the remaining two variables should be between these two extremes (we are agnostic as to whether or not issue or question complexity makes a question more difficult). We also selected a fifth variable, *Political killings*, to represent a polarizing question on which we expect crowd workers to provide less substitutable responses.[11]

Figure 3 presents an illustrative comparison of the two types of data.[12] Specifically, it plots crowd and expert scores for Political killings for Argentina 1900-2015. A score of "4" reflects a country-year free from political killings by the government or its agents, and a "0" a country-year in which they are endemic. Different colors represent different coders, with lines representing smoothed coder-trends over time. The comparison is striking, and provides immediate grounds for skepticism regarding the substitutability of crowds for experts. While experts tend to code similar trends, it is difficult to isolate any pattern in the crowd codings. Some crowd workers may be providing valid codings, but the noise is overwhelming.

Figure 4 provides further grounds for skepticism regarding the substitutability of crowd scores, illustrating the linear relationship between expert-coded mean values and crowd scores across the five expert-coded variables, using the Argentinian data. If crowd-coded data were substitutable for expert-coded data, then the lines should be roughly on the diagonal, indicating that crowd averages correspond to expert averages. No lines come close to the diagonal, and crowd scores for Judicial Independence even show a slightly negative relationship with the expert mean.[13]

We now examine absolute differences between expert and crowd ratings, using task characteristics as explanatory factors. We take as an observation the absolute difference between the expert mean for a given country-year-variable and each crowd worker's score for that country-year-variable.[14] Given that the expert mean is often between two ordinal categories,

[11] For additional information on the indicators used, see Coppedge, Gerring, Lindberg, Skaaning, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, Marquardt, Mechkova, McMann, Paxton, Pemstein, Saxer, Staton, Seim, Sigman & Staton (2017).

[12] We create graphics using *ggplot2* (Wickham 2009) and appendix tables using *stargazer* (Hlavac 2015).

[13] Appendix E reports results from regression analyses of the correlation between expert mean and coder scores, which also indicates a weak relationship. We did not specify this analysis in the PAP.

[14] In Appendix H we provide two robustness checks of these analyses. First, to take variation in expert scores into account, we conduct bootstrap analyses in which we draw an expert and crowd worker score for each cell and take the absolute difference in codings. Second, we conduct analyses in which we control for the distance of average expert scores from the middle value (2) to ensure that findings are not a function of the placement of expert scores on the scale. Results from both analyses are congruent with those reported in the text.

Figure 3: Expert and crowd codings for Freedom from Political Killings in Argentina, 1916-2015

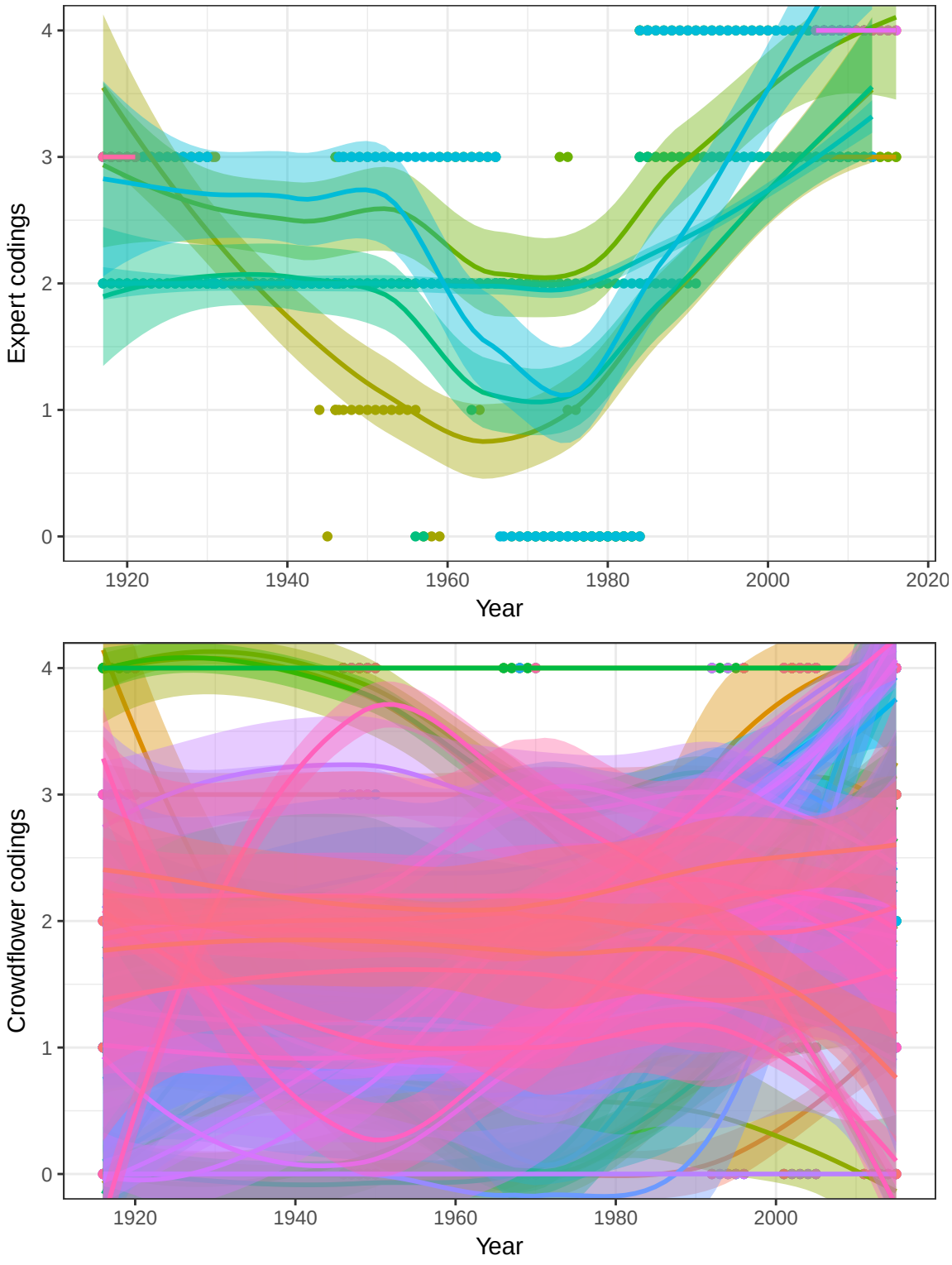
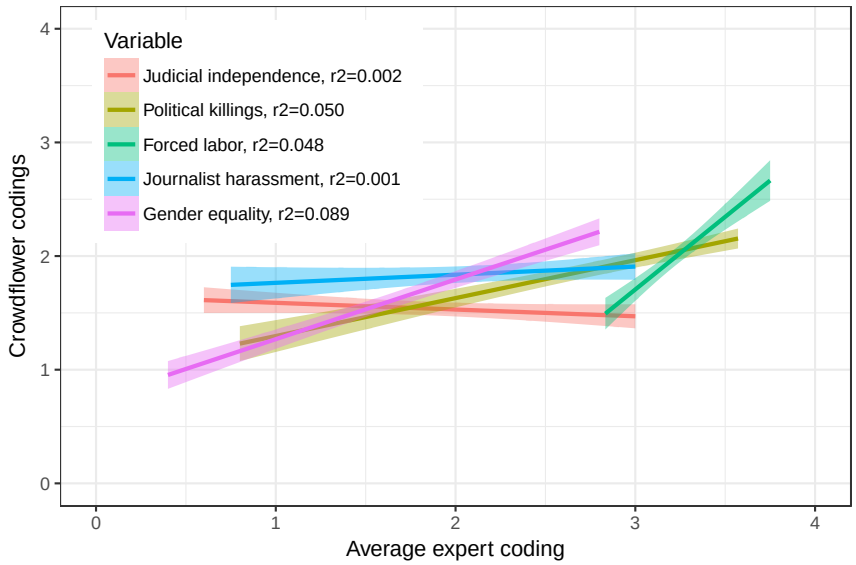


Figure 4: Linear relationship between average expert score and crowd worker scores for expert-coded questions for Argentina

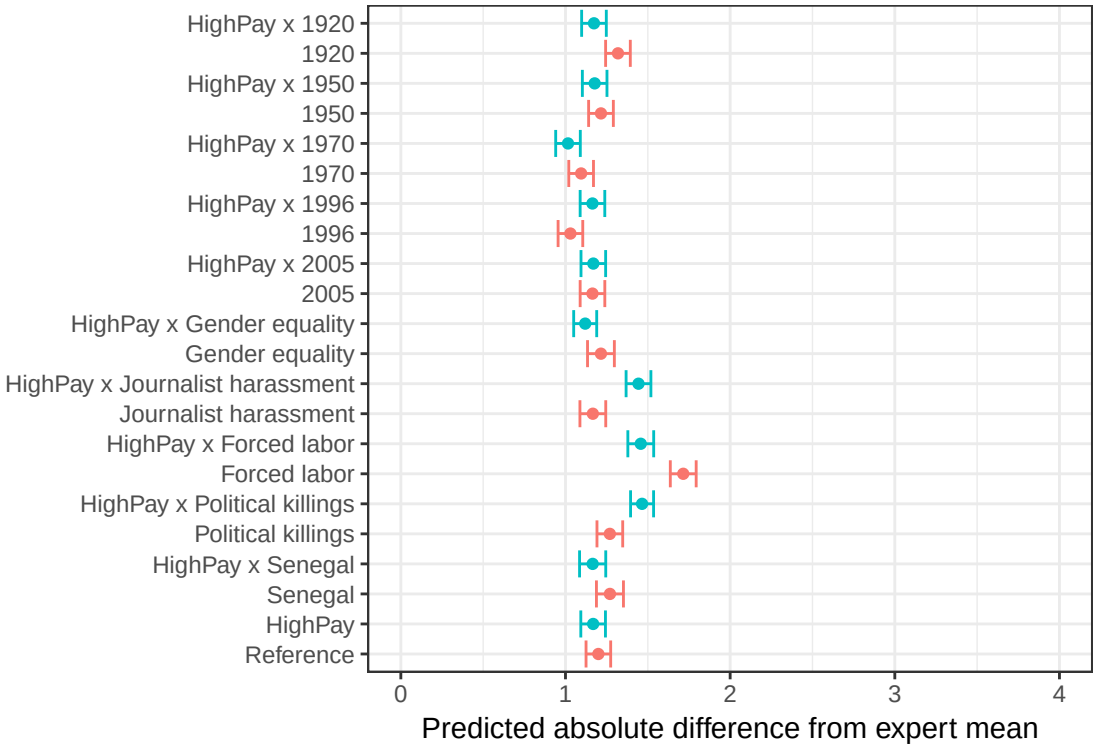


precluding crowd workers from hitting the expert score exactly, a difference of <1 should be considered "good."

Figure 5 (corresponding to Table 26 in Appendix G.1) depicts predictions of the difference between a crowd worker's score and the expert mean from a model that regresses this difference on task characteristics. Each point represents the predicted average difference between mean expert and crowd worker score for a given characteristic, along with 95 percent confidence intervals. We analyze all task characteristics, interacting all other characteristics with the payment treatment to assess heterogenous treatment effects; we use codings for the typical payment condition, of Argentina, for the period 2011-2015, and of Judicial independence as the reference group.

At the reference level, crowd workers tend to deviate by 1.3 Likert scale points from the mean expert coding. This value is slightly more than a quarter of the scale range, indicating relatively low substitutability. The high payment treatment (blue lines) causes crowd coding to generally deviate further from average expert codings, an effect that is particularly noticeable for Journalist Harassment and Political Killings (though it has the opposite effect on Forced Labor). This suggests that higher pay encourages bad coders to stay, rather than nurturing better coding. Most other task characteristics—recency, polarization, information availability, and high issue/question complexity—do not seem to have much substantive influence on absolute difference. The exception is that crowd workers who code of Forced labor tend to have greater absolute difference than workers who code other variables, which is consistent with expectations, since it is the most complex expert-coded variable.

Figure 5: Substantive effect of task characteristics over payment treatment effects on distance from expert mean



Reference level: 2015, Argentina, Judicial independence

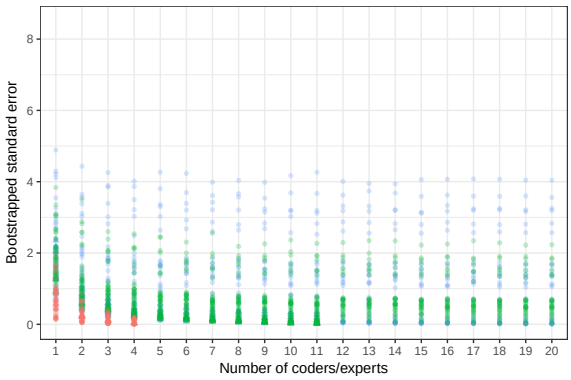
With a larger sample size, crowdsourced data might converge toward the expert mean and be more substitutable. While we did not specify this analysis in the PAP, we conducted analyses of bootstrapped standard errors across country-year-variable-treatment observations with different numbers of coders to assess whether varying the number of crowd workers would increase substitutability. More precisely, for each country-year-variable-treatment we randomly draw n coders and take their average score. We repeat this procedure 100 times with replacement, and estimate standard deviations of the difference between average scores and the expert mean across the 100 draws. If crowds are substitutable for experts, then the value of this quantity should tend toward 0 as the number of sampled coders increases. Figure 6 presents the results for the five main variables. The horizontal axis represents the number of sampled coders, and the vertical axis the bootstrapped standard error. Red points represent V–Dem experts, while blue and green points represent crowd workers in the low and high payment conditions, respectively. Triangles represent data for Senegal, circles Argentina. Experts converge relatively rapidly toward their mean: there is generally close to no bootstrapped error after the number of sampled experts is three (out of five experts). In contrast, there are a substantial number of observations for crowd workers in which bootstrapped standard errors are relatively unaffected by the number of coders sampled. Even in the case of Gender equality, where bootstrapped standard error estimates are generally below a value of one, standard error does not appear to tend toward 0, but rather a range of values between 0 and 1. There is therefore little evidence that greatly expanding the number of crowd workers would yield considerably different results than our present sample size.

We now analyze how crowd worker characteristics predict substitutability. We regressed the absolute difference between expert mean and each coder's score on worker characteristics, controlling for the task characteristic variables, as presented in Figure 7 (full results in Appendix Table 9). The results do not strongly support any of our hypotheses about crowd characteristics influencing substitutability. However, analyses with worker random effects (Appendix F) indicate that there is substantial variation in worker substitutability. The fact that our models cannot explain this variation is evidence that it is highly idiosyncratic, making it difficult for researchers to purposively recruit better crowd workers.

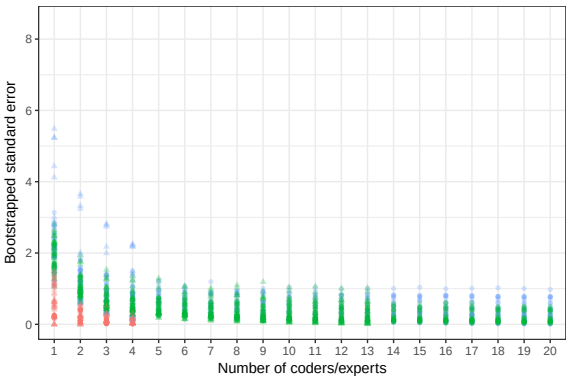
4.2 Factual questions

We purposively selected four trained-coder questions to vary along two metrics. First, we considered the dimensionality of the fact: whether coding requires an understanding of only one, or several different concepts. For example, suffrage is a unidimensional concept, whereas the requirements for referenda are complex and multidimensional (e.g., there are different

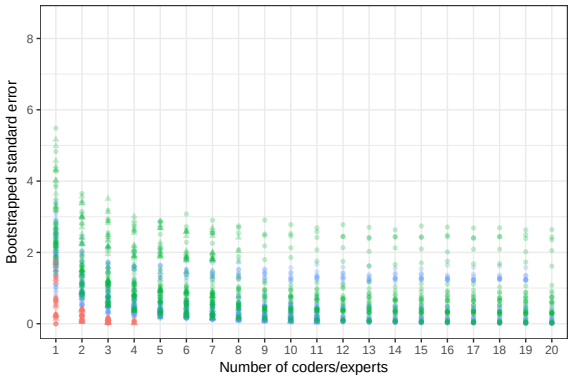
Figure 6: Bootstrapped standard errors across variables by number of crowd workers



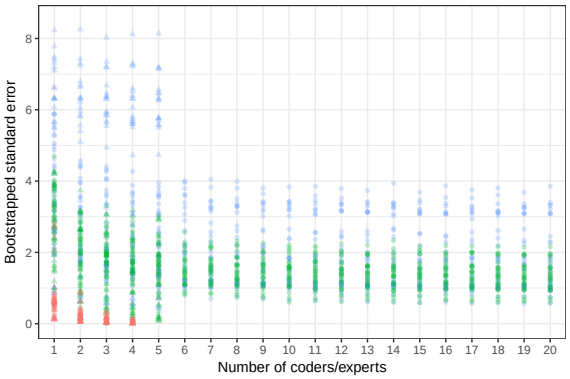
(a) Judicial independence



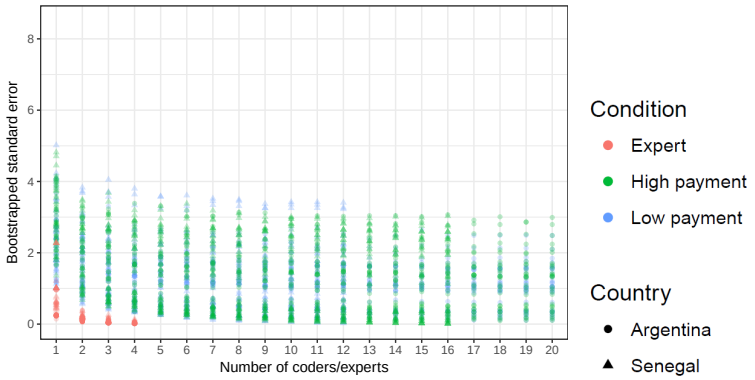
(b) Gender equality



(c) Journalist harassment

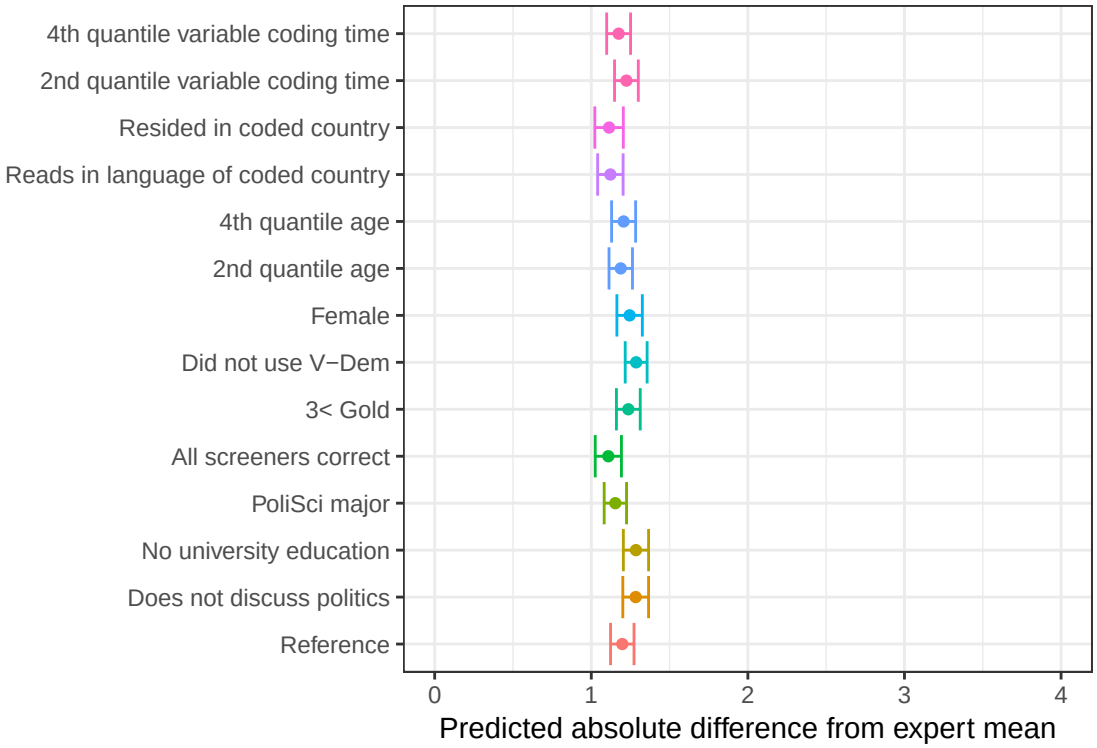


(d) Forced labor



(e) Political killings

Figure 7: Substantive effect of crowd worker characteristics on distance from expert mean, conditional on task characteristics



types of referenda, different levels, and different approval stages). Second, we consider how findable the fact was: whether crowds would know where to look to discover the fact or not. For example, the legal provisions for suffrage are located in legal documents of the country, whereas the suffrage level in practice is not always published in an official document. These criteria led us to four purposively-selected variables, again corresponding to a 2×2 table of high vs. low levels of both findability and dimensionality:

1. *Suffrage level*: Percentage of population with *de facto* suffrage ($v2elsuffrage$). Less easy to find and multidimensional.

2. *Bicameral legislature*: Number of chambers in legislature ($v2lgbicam$). Less easy to find and unidimensional.

3. *Referenda permitted*: Form of referenda permitted by law ($v2ddlegrf$). Easy to find and multidimensional.

4. *Minimum voting age*: Minimum age for voting in national elections ($v2elage$). Easy to find and unidimensional.

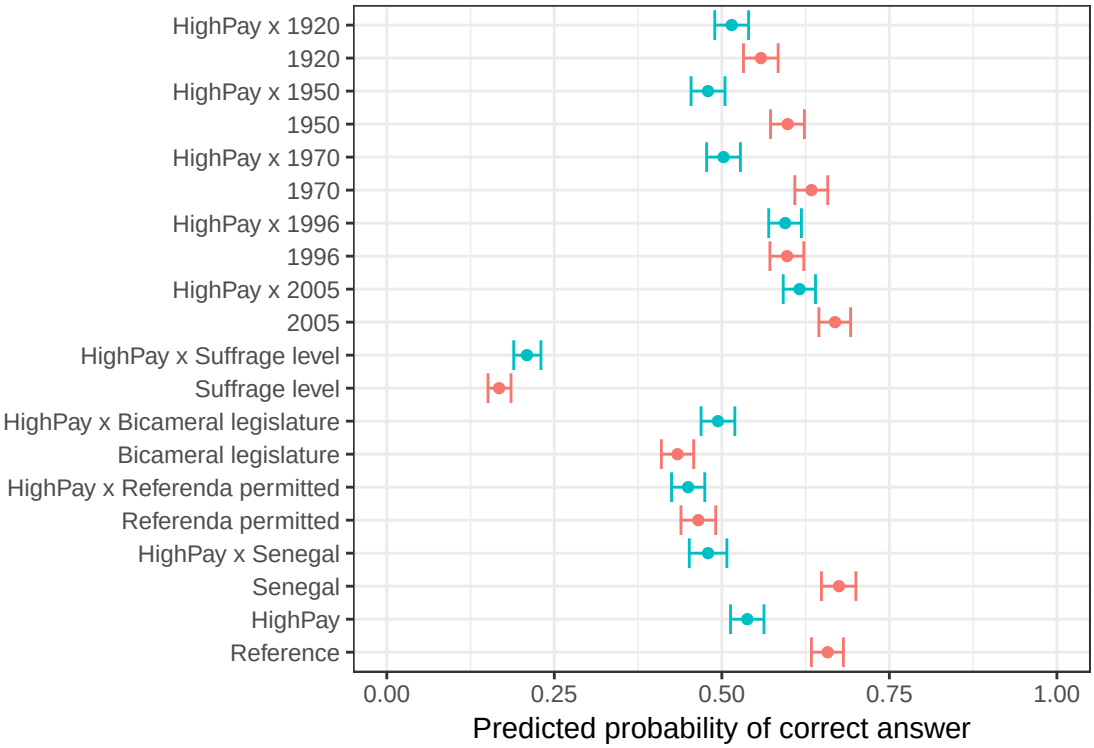
To measure substitutability, we use a dichotomous indicator of whether or not a crowd coder provided the same answer as the trained V–Dem coder, using the least complex and most findable indicator of Minimum voting age as the reference level. Figure 8 plots the predicted probability of a substitutable answer from crowd workers based on a probit model that includes task and incentive characteristics (see Table 27 in Appendix G.2 for complete results). We use the same set of variables as with the expert-coded data (i.e. year for recency, Senegal for information availability, and HighPay for incentives), substituting factual questions for the expert-coded questions examined previously.

Focusing first on the reference treatment condition (i.e. low payment), crowd workers are substantially more substitutable for trained coders in the reference indicator—minimum voting age—than for other indicators. Even in this case, however, the predicted probability that a crowd worker would produce a substitutable response is below 75 percent. Substitutability with regard to the bicameralism and referenda indicators are substantially worse (lower than 50 percent probability of substitutable answers). For suffrage, crowd workers are less than 25 percent likely to provide substitutable data. The comparative substitutability across coding tasks matches our expectations: voting age is both easy to find and unidimensional, while suffrage level is harder to find and multidimensional; the other two questions cross task complexity dimensions. As hypothesized, substitutability increases with recency, though information availability (proxied by Senegal) has little influence on substitutability. In line with the expert-coded data analyses, participants in the typical payment condition are generally more substitutable those in the high pay condition.

Figure 9 examines the relationship between worker characteristics and the substitutability of crowd workers for trained coders (full regression results in Table 10). We find no evidence that coders who regularly discuss politics are more substitutable than others. Peculiarly, coders with no university education outperformed those with education, and political science majors under-performed on this task. Less surprisingly, coders who coded screeners accurately (indication of diligence) substantially outperformed their peers and coders who looked up values on the V–Dem website did better than other participants.[15] Similarly, residing in the country in question increased substitutability, as did spending more time on the coding task.

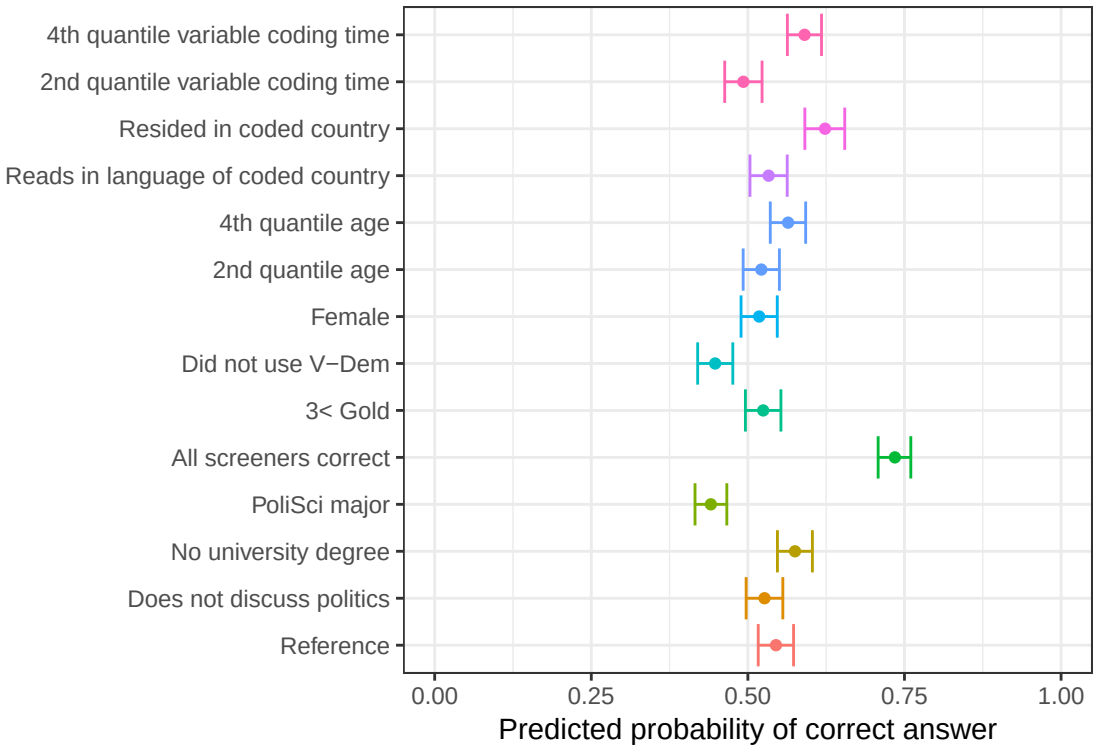
[15] A subset analysis of coders who accurately coded both screeners shows substantial differences from the analysis of the pooled data (Appendix I, Table 21). In these analyses, the high payment treatment substantially increases the probability that coders would provide substitutable responses, indicating that there may be an interactive effect between data quality and the incentives with which they are provided to perform a task, conditional on baseline coder quality.

Figure 8: Substantive effect of task characteristics on probability of correct answer to factual questions



Reference level: 2015, Argentina, Voting age

Figure 9: Substantive effect of coder characteristics on probability of correct answer to factual questions, conditional on task characteristics



5 Conclusion

Our results show that crowd workers are generally not substitutable for trained coders or experts when data production requires conceptual and contextual knowledge. In the case of trained coders, the degree of crowd substitutability has a low upper bound, and both task and worker characteristics can reduce substitutability. Results for expert-coded data are dismal: the correlation between expert-coded and crowd-coded data is minimal, and there is little evidence that crowd-coded data would converge toward the expert mean with an expanded sample size. Moreover, we show that reproducing expert-coded datasets with crowd workers would likely cost more than using experts, and may in fact be wholly infeasible given the pool of potential workers. Together, these results indicate the realm of research endeavors to which crowdsourcing may contribute is much smaller than proponents argue.

Our typology provides an explanation for these limitations: most crowd workers have neither the background nor the incentives to gather and analyze many types of data, while experts and trained coders do. However, the boundaries between trained coders and crowd workers remain somewhat unclear: our results suggest that especially diligent crowd workers may be able to perform limited tasks which require contextual and conceptual knowledge. Previous findings that crowd workers can substitute for trained coders in other domains align with this claim.

We refrain here from addressing validity and replicability, both areas of purported advantage for crowd-sourced data over other forms of coded data (Benoit et al. 2016). While we agree that producers of expert-coded datasets must address concerns regarding data validity,[16] our analyses provide no indication that crowdsourced data are more valid than other forms. Concerns regarding replicability are perhaps insurmountable given a limited pool of experts and the costs of these enterprises. However, our analyses indicate that a replicable crowdsourced dataset would look very different from a non-replicable expert-coded dataset. It is to the reader to determine whether or not issues of replicability outweigh the benefits of expertise.

[16] See McMann, Pemstein, Seim, Teorell & Lindberg (2016) and Teorell, Coppedge, Skaaning & Lindberg (2016) for examples of data validation with expert-coded data.

References

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver & Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(2):278–295.

Cooper, S., F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic & F. Players. 2010. "Predicting Protein Structures With a Multiplayer Online Game." *Nature* 466:756–760.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Anna Lührmann, Kyle L. Marquardt, Valeriya Mechkova, Kelly McMann, Moa Olin, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Constanza Sanhueza Petrarca, Johannes von Römer, Laura Saxer, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Natalia Stepanova & Steven Wilson. 2017. V–Dem Dataset v7.1. Technical report Varieties of Democracy Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle L. Marquardt, Valeriya Mechkova, Kelly McMann, Pamela Paxton, Daniel Pemstein, Laura Saxer, Jeffrey Staton, Brigitte Seim, Rachel Sigman & Jeffrey Staton. 2017. Varieties of Democracy Codebook v7. Technical report Varieties of Democracy Project: Project Documentation Paper Series.
Accessed at: https://ssrn.com/abstract=2968274

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Joshua Krusell, Kyle L. Marquardt, Valeriya Mechkova, Daniel Pemstein, Josefine Pernes, Laura Saxer, Natalia Stepanova, Eitan Tzelgov, Yi-ting Wang & Steven Wilson. 2017. Varieties of Democracy Methodology v7. Technical report Varieties of Democracy Project: Project Documentation Paper Series.
Accessed at: https://ssrn.com/abstract=2968284

Crump, Matthew J. C., John V. McDonnell & Todd M. Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." *PloS One* 8(3):e57410.

D'Orazio, Vito, Michael Kenwick, Matthew Lane, Glenn Palmer & David Reitter. 2016. "Crowdsourcing the Measurement of Interstate Conflict." *PLoS ONE* 11(6).

Hlavac, Marek. 2015. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Cambridge, USA: Harvard University. R package version 5.2.

Honaker, James, Michael Berkman, Chris Ojeda & Eric Plutzer. 2013. "Sorting Algorithms for Qualitative Data to Recover Latent Dimensions with Crowdsourced Judgments: Measuring State Policies for Welfare Eligibility under TANF.".
Accessed at: http://projects.iq.harvard.edu/files/applied_stats/files/james_honaker-_sorting_algorithms.pdf

Hooghe, Lisbet, Ryan Bakker, Anna Brigevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny & Marco Steenbergen. 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.

Kittur, A., E.H. Chi & B. Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the ACM SIGCHU Conference on Human Factors in Computing Systems*. pp. 453–456.

McMann, Kelly M, Daniel Pemstein, Brigitte Seim, Jan Teorell & Staffan I Lindberg. 2016. "Strategies of Validation: Assessing the Varieties of Democracy Corruption Data." *Varieties of Democracy Institute Working Paper* 23.
Accessed at: http://dx.doi.org/10.2139/ssrn.2727595

Morris, Peter A. 1977. "Combining expert judgments: A Bayesian approach." *Management Science* 23(7):679–693.

Paolacci, Gabriele, Jesse Chandler & Panagiotis G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5):411–419.

Peer, Eyal, Sonam Samat, Laura Brandimarte & Alessandro Acquisti. 2016. "Beyond the Turk: An empirical comparison of alternative platforms for crowdsourcing online behavioral research.".
Accessed at: http://dx.doi.org/10.2139/ssrn.2594183

Shapiro, Danielle N., Jesse Chandler & Pam A. Mueller. 2013. "Using Mechanical Turk to Study Clinical Populations." *Clinical Psychological Science* 1(2):213–220.

Teorell, Jan Teorell, Michael Coppedge, Sven-Erik Skaaning & Staffan I. Lindberg. 2016. "Measuring Electoral Democracy with V-Dem Data: Introducing a New Polyarchy Index." *Varieties of Democracy Institute Working Paper* 25.
Accessed at: http://dx.doi.org/10.2139/ssrn.2740935

Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.

Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. Springer New York.

A Operationalization

A.1 Data collection instrument

Our instrument is a Qualtrics online survey. We base the tasks (or questions) in the instrument on indicators from the V–Dem expert questionnaire, though we alter them slightly due to the different format of this study.[17]

A.2 Additional details on operationalization of hypotheses about task characteristics

A.2.1 Question and issue complexity

Question complexity: We measure question complexity across all V–Dem questions by taking the length of the English language text of the question plus the length of all of the descriptions of the responses categories for each question, both measured in number of characters. We combine the lengths of question text and response text because the two are highly negatively correlated with each other, likely because some questions load their complexity into the question text and thus have proportionately simpler responses, and vice versa. Accordingly, we posit that the total of the two best represents the overall textual complexity of the questions. In addition, textual length is highly positively correlated with the average length of time between a coder first reading a question and first entering any rating of any kind, providing some evidence that length is related to complexity.

Issue complexity: We measure issue complexity across all V–Dem questions by evaluating the average confidence (from 1 to 100) that coders self-assigned their ratings for each particular question. In theory, questions that address more nuanced issues should be answered with lower confidence, whether or not the question language is particularly complex. For example, "What is democracy?" is a very simply worded question but asks about complex issues, and thus should be answered with lower confidence, all else equal. We have reason to believe this measure is orthogonal to the question complexity proxy as the two measures have a vanishlingly low correlation. In addition, the average self-reported confidence for a question is highly correlated with the average confidence in the vignette codings for that

[17] Specifically, we specify that the coders are to code a specific country across specific years, and we also change the wording to the past tense.

same question. As these two quantities are measured independently (the coders answer vignettes separately and without indication of connection to another question elsewhere), this suggests stability in the confidence measure, as opposed to being merely noise.

A.2.2 Recency

The V–Dem project covers most countries from 1900 to 2016. We have selected six five-year periods ranging from distant years to more recent years. In selecting years, we considered three criteria. First, in line with our recency hypothesis, we deliberately selected some five-year periods that are recent and some that are farther in the past, expecting to find lower substitutability going back in time. Second, we carefully selected times covering major political events to assess the degree to which coders static-code (i.e., do not change their codings over time). Finally, we deliberately chose some five-year periods prior to which the crowd workers are likely to have been born.

- 2011-2015: The most recent time period, and thus the period with which coders are most likely familiar.
- 2001-2005: A more distant time period in which crowds were nonetheless alive and aware.
- 1992-1996: Third wave of democracy, still in many coders' lifetimes.
- 1966-1970: Turbulent period prior to birth of coders.
- 1946-1950: Turbulent post-war period, far in past.
- 1916-1920: Turbulent post-war period, very far in past.

A.2.3 Information availability

We quantify information availability at the country level and year level using custom measures detailed below.

Information availability by country: Our approach in developing a measure of information availability at the country level was to evaluate the "googleability" of different countries. To do this, we wrote software to measure the amount of data on each country in the world available on Wikipedia, as a proxy for this general notion of easily available information. Wikipedia has a number of standardized pages and hierarchies for organizing pages of a similar nature. In addition, it has a number of "hidden" (in the sense that they

aren't linked directly from substantive pages, yet they are still open and available for public viewing) pages that provide meta-directories of such pages, to an arbitrary depth of hierarchy. Our approach was to select one of these meta-directories as appropriate to the country level unit of analysis and then recursively count and download the pages therein for each country.

We downloaded the full text of all pages on Wikipedia, three levels deep in the organizational tree for "Politics by Country." It was important to get more than just the top level pages because those by design are truncated at a certain length and then subdivided. For example, the "Politics of (Country Name)" pages are almost all of about the same length, even though there is vastly more information on certain countries. So all countries look the same at the first level in terms of length. On the other hand, too deep of a recursion into the hierarchy only increases skew. That is, past three levels, most countries have only a few brief pages, and all the additional data is simply adding to the countries that already have the highest counts. In addition, we also screen out pages Wikipedia has labeled "stubs" (i.e., placeholder pages with at most a sentence or two of descriptive content) so as to ensure measurement of actual usable information. We downloaded the full contents of 187,319 Wikipedia pages, and aggregated the data into counts of the total numbers of characters on the pages associated with every country in the world. This provides our measure of the amount of information generally available about each country.

As MTurk users are typically native English-speakers, we also consider the official language of the country as a proxy indicator for whether easily-accessed online material is readable for the average crowd coder.

Finally, as another—highly V–Dem specific—operationalization of information availability, we also consider whether the V–Dem data had been publicly released as of April 2017, deliberately including one country in the study for which data had not yet been made available on the V–Dem website.

With this in mind, we selected four countries for our intended full experiment—United States, Russia, Singapore, Benin—and two countries—Argentina (primary country) and Senegal (optional) for the pilot reported here. As can be seen in Table 2, the selected countries vary in terms of their political characteristics that might affect information availability: electoral democracy in the form of the V–Dem Polyarchy Index over time, information availability score, likely case familiarity for crowd workers (MTurk), and official languages. Argentina is in the middle range of information availability and case familiarity, while Senegal is on the low end of the range.

Table 2: Countries

Name Name	Political Characteristics	Polyarchy Index (CI)	Information Availabilty	Case Familiarity for Turkers	English Official Language?	V–Dem Data Released?
United States	One of the most stable advanced democracies, democratic since 1776	0.86 (0.81; 0.89)	10,889 (859 million)	High	Yes	Yes
Russia	Revolution, disintegration of the Soviet Union	0.06 (0.02; 0.08)	2,019 (138 million)	Intermediate	No	Yes
Singapore	Authoritarian but low corruption, repression, conflict	N/A	638 (53 million)	Intermediate	Yes	No
Benin	Recent colonial past; country changed names	0.46 (0.35; 0.56)	472 (27 million)	Low	No	Yes
Argentina	History of both dictatorship and democracy	0.81 (0.76; 0.86)	1,143 (71 million)	Intermediate	No	Yes
Senegal	Relatively stable and consolidated democracy	0.74 (0.69; 0.79)	548 (39 million)	Low	No	Yes

B Qualtrix screenshots

Figure 10: Screener coding example for expert-coded indicators

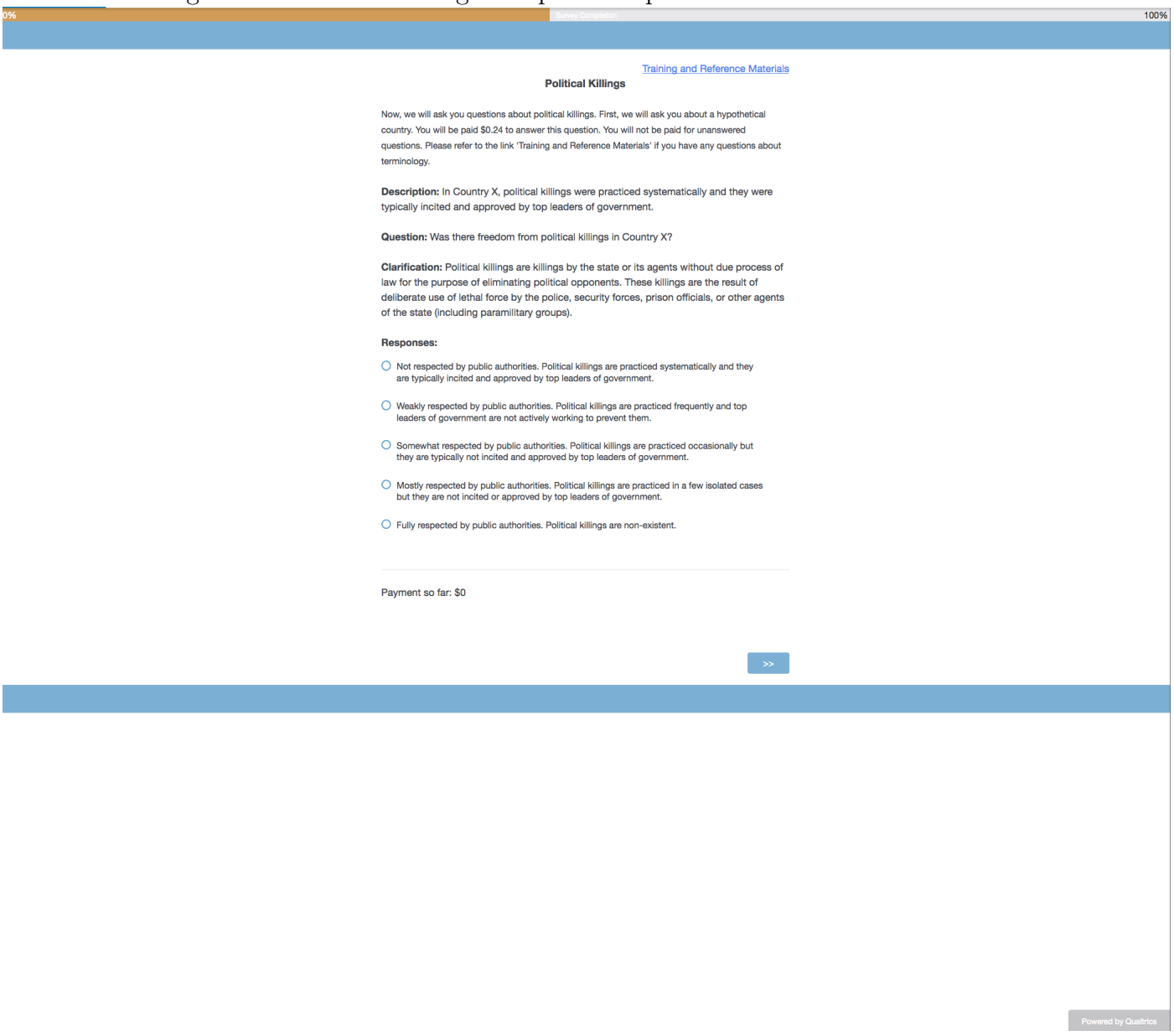


Figure 11: Coding example for expert-coded indicators

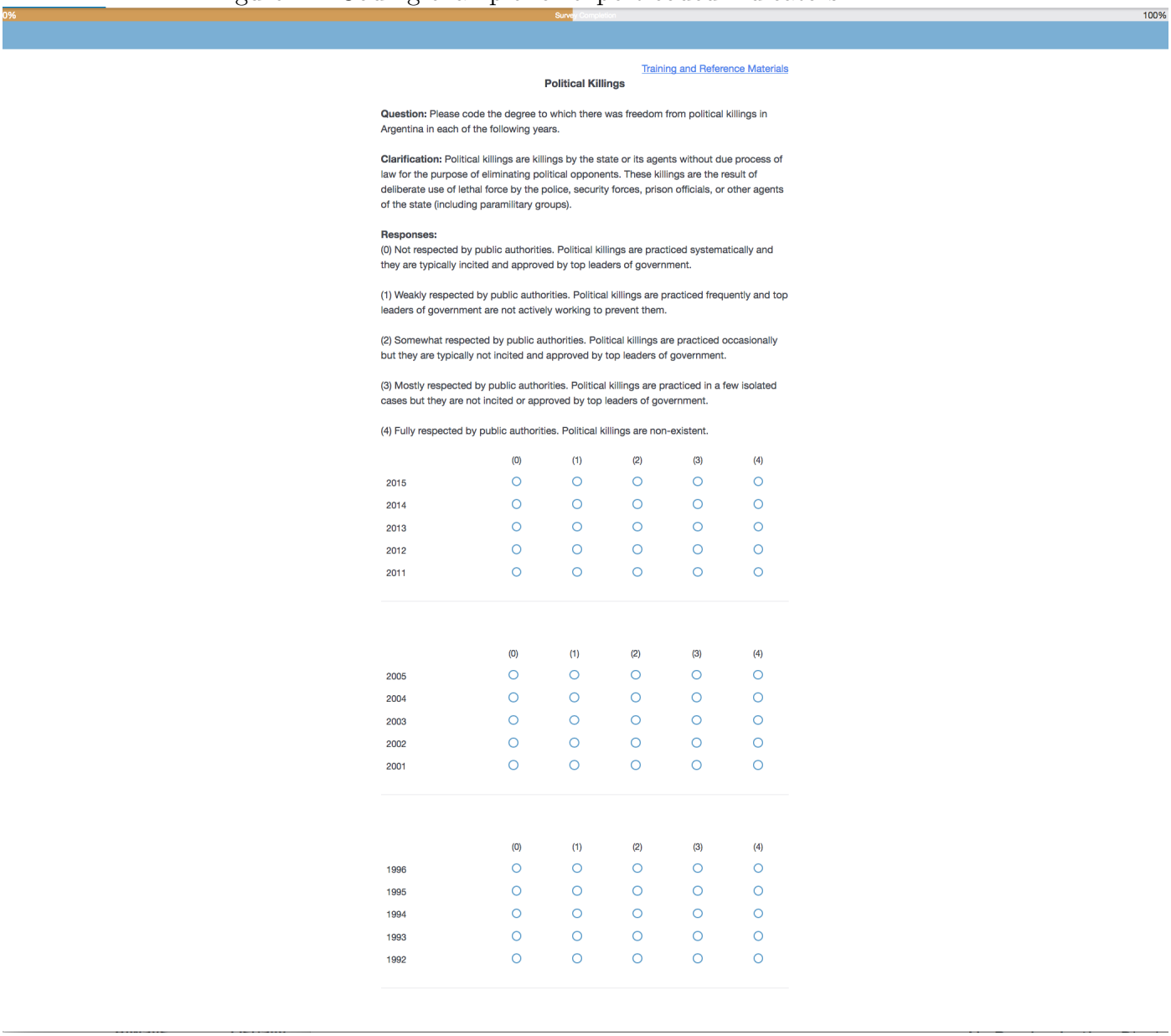


Figure 12: Screener coding example for coder-coded indicators

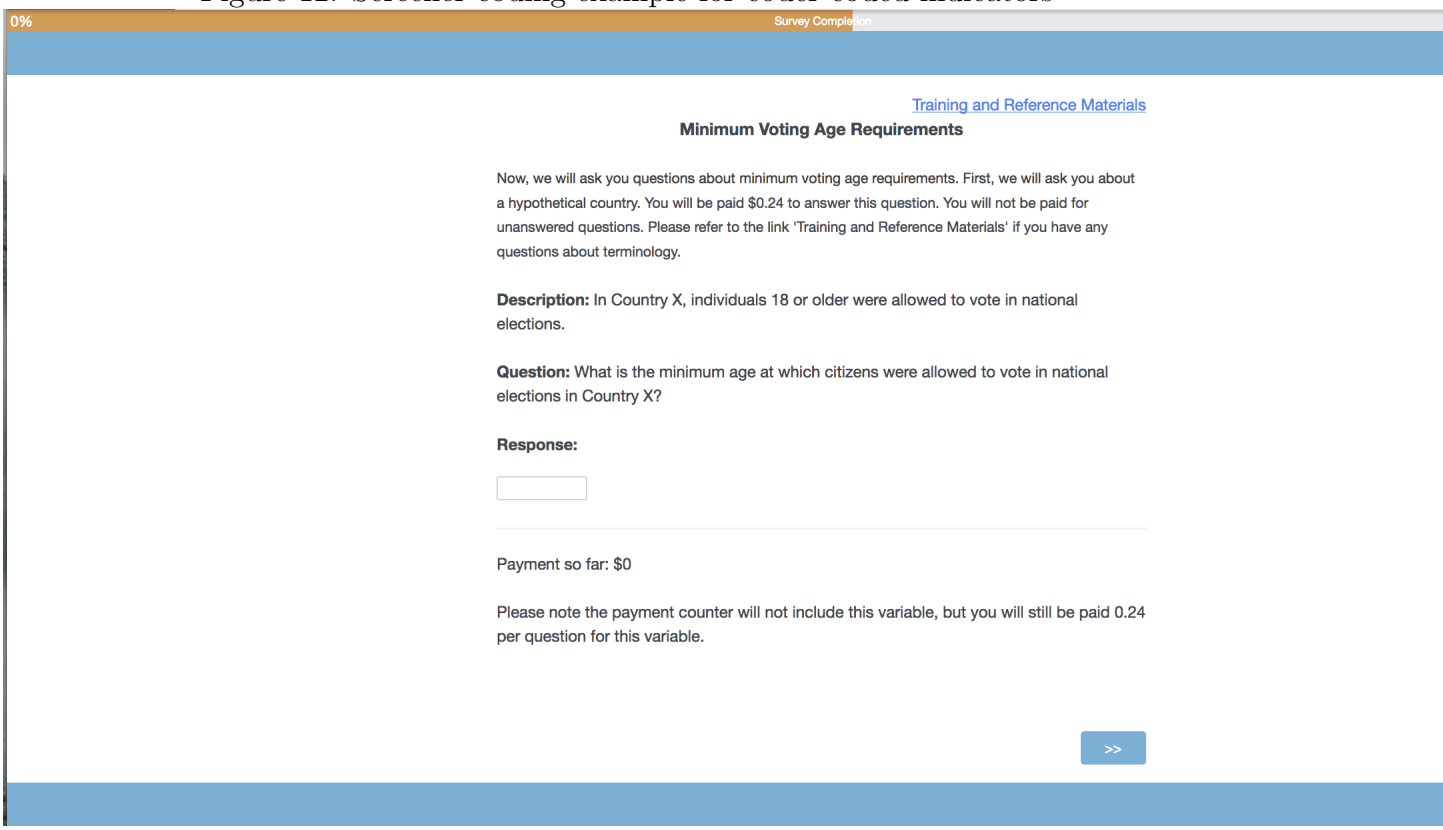
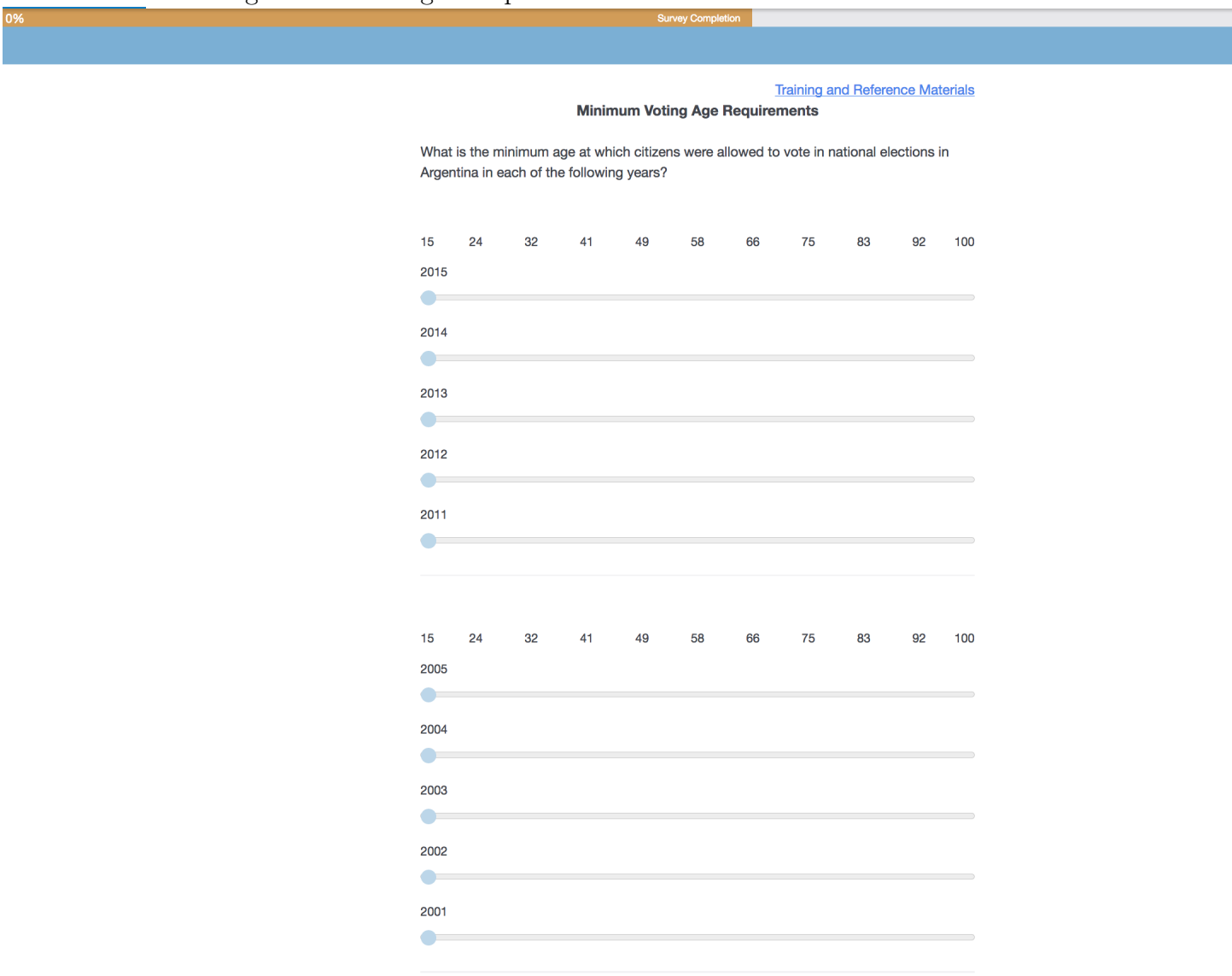


Figure 13: Coding example for coder-coded indicators



C Descriptive Statistics

Table 3: Descriptive Statistics

Variable	Population Mean (SD)
Age	33 years (9.4)
Female	30.0%
Spent Time in Argentina	13.7%
Spent Time in Senegal	5.8%
Coursework on Argentina	13.6%
Coursework on Senegal	10.1%
Speaks Spanish	32.1%
Speaks French	20.8%
Lives in India	14.2%
Lives in Spain	9.7%
Lives in Venezuela	8.0%
University Degree	59.9%
Political Science Major	49.4%
Political Science Courses (Non-Major)	32.2%
Discusses Politics with Friends and Family	73.1%
Interested in Public Affairs	77.5%
Follows Politics	78.3%
Voted in the Last Election	90.1%
Believes Democratic Government is Important	93.4%
Used V–Dem Data in Completing Survey	51.0%
Randomly Assigned: High Pay Condition	52%

D Analysis of item non-response

In this analysis, we consider item non-response (a missing value by a crowd worker at the country-year-indicator level) as the outcome variable. We regress item non-response on the task complexity variables and coder characteristics considered in the main text analyses, pooling perceptions and factual questions in the same regression (since the item non-response outcome variable is the same for both).

Table 4: Relationship between item non-response and task characteristics

	Dependent variable: Item non-response (0 or 1)				
	(1)	(2)	(3)	(4)	(5)
Reference	0.12*** (0.002)	0.12*** (0.003)	0.04*** (0.01)	0.04*** (0.01)	−0.03** (0.01)
1920			0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
1950			0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
1970			0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
1996			0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
2005			0.01 (0.01)	0.01 (0.01)	0.01 (0.01)
Senegal			0.14*** (0.005)	0.14*** (0.005)	0.14*** (0.01)
Political killings			−0.02** (0.01)	−0.02** (0.01)	0.07*** (0.01)
Forced labor			0.08*** (0.01)	0.08*** (0.01)	0.10*** (0.01)
Journalist harassment			0.08*** (0.01)	0.08*** (0.01)	0.06*** (0.01)
Gender equality			0.001 (0.01)	0.001 (0.01)	0.16*** (0.01)
Referenda permitted			0.09*** (0.01)	0.09*** (0.01)	0.15*** (0.01)
Minimum voting age			0.01 (0.01)	0.01 (0.01)	0.10*** (0.01)
Bicameral legislature			−0.03*** (0.01)	−0.03*** (0.01)	0.05*** (0.01)
Suffrage level			0.10*** (0.01)	0.10*** (0.01)	0.21*** (0.01)
High payment		−0.001 (0.005)		0.002 (0.004)	0.13*** (0.02)
High payment × 1920					−0.005 (0.02)
High payment × 1950					−0.001 (0.02)
High payment × 1970					−0.01 (0.02)
High payment × 1996					−0.001 (0.02)
High payment × 2005					−0.005 (0.02)
High payment × Senegal					−0.01 (0.01)
High payment × Political killings					−0.17*** (0.02)
High payment × Forced labor					−0.04** (0.02)
High payment × Journalist harassment					0.04** (0.02)
High payment × Gender equality					−0.29*** (0.02)
High payment × Referenda permitted					−0.12*** (0.02)
High payment × Minimum voting age					−0.16*** (0.02)
High payment × Bicameral legislature					−0.17*** (0.02)
High payment × Suffrage level					−0.22*** (0.02)
Observations	20,610	20,610	20,610	20,610	20,610
R^2	0.00	0.0000	0.06	0.06	0.08
Adjusted R^2	0.00	−0.0000	0.06	0.06	0.08
Note:					*p<0.1; **p<0.05; ***p<0.01
					Reference level: 2015, Argentina, Judicial independence

Table 5: Relationship between item non-response and coder characteristics

	Dependent variable: Item non-response (0 or 1)	
	(1)	(2)
Reference	−0.004 (0.01)	0.01 (0.01)
Does not discuss politics	0.10*** (0.01)	0.10*** (0.01)
No university education	0.01 (0.01)	0.01 (0.01)
Political science major	−0.06*** (0.01)	−0.07*** (0.01)
All screeners correct	0.01 (0.005)	0.01* (0.005)
2 ¡ gold standard correct	−0.01*** (0.005)	−0.02*** (0.005)
Did not use V–Dem	0.07*** (0.005)	0.07*** (0.005)
Female	0.01*** (0.005)	0.01*** (0.005)
log(Age)	−0.02*** (0.01)	−0.02*** (0.01)
Reads coded-country language	−0.06*** (0.01)	−0.06*** (0.01)
Resided in country	−0.004 (0.01)	−0.003 (0.01)
log(Variable coding time)	−0.04*** (0.002)	−0.04*** (0.002)
High payment		−0.03*** (0.004)
1920	0.01 (0.01)	0.01 (0.01)
1950	0.01 (0.01)	0.01 (0.01)
1970	0.01 (0.01)	0.01 (0.01)
1996	0.01 (0.01)	0.01 (0.01)
2005	0.01 (0.01)	0.01 (0.01)
Senegal	0.11*** (0.005)	0.11*** (0.005)
Political killings	−0.02** (0.01)	−0.01 (0.01)
Forced labor	0.11*** (0.01)	0.11*** (0.01)
Journalist harassment	0.12*** (0.01)	0.13*** (0.01)
Gender equality	0.03*** (0.01)	0.04*** (0.01)
Referenda permitted	0.04*** (0.01)	0.04*** (0.01)
Minimum voting age	0.07*** (0.01)	0.07*** (0.01)
Bicameral legislature	−0.01 (0.01)	−0.01 (0.01)
Suffrage level	0.08*** (0.01)	0.08*** (0.01)
Observations	17,430	17,430
R^2	0.14	0.15
Adjusted R^2	0.14	0.15
Note:		*p<0.1; **p<0.05; ***p<0.01
	Reference level: 2015, Argentina, Judicial independence	

E Correlation regression analyses

Table 25 presents regression results in which we examine the relationship between crowd and expert data more concretely, using a standard OLS regression of each crowd coder's rating on 1) the expert mean and 2) task characteristics that could condition substitutability. First, note that all models have very low R^2 values, indicating that neither the expert mean nor task characteristics explain a great deal of the variation in crowd codings. For example, Model 1 presents results from a model in which crowd scores are regressed on the expert mean for each country-year-variable. While the correlation between the expert mean and crowd scores is positive and significant, indicating the expected positive correlation, the R^2 is .03, indicating that the expert mean explains little variation in coder scores. Model 2 includes a variety of controls that may influence expert scores. After controlling for these variables, the relationship between the expert mean and crowd scores decreases: perhaps most interestingly, all the recency variables have a negative correlation with coder scores, which indicates coders tended to code earlier years as having lower values. Since lower scores reflect normatively worse values, it is possible that this pattern reflects crowd workers as generally coding earlier years as being "worse" for each variable.

Model 3 interacts the expert mean with the experimental payment condition, which shows little interactive relationship with coder scores. This finding demonstrates that crowd workers with higher payment did not tend to have a higher correlation with expert mean, which indicates that the treatment did not greatly incentivize them to provide more substitutable data.

Model 4 presents results from a model in which we interact the expert mean with question variables. The results indicate that the relationship between expert codings and crowd-provided values is not consistent: while the expert mean for reference variable (v2juhcind) has a *negative* relationship with the crowd codings, the expert mean for other variables appears to have a weakly positive relationship with crowd scores. Equally importantly, Model 3 illustrates that being randomly assigned to the high (double) pay condition does not interact with the average expert score to significantly predict the crowd member's rating. In other words, higher pay does not improve the correlation between crowd and expert ratings.

Table 6: Relationship between coder scores and average expert score

	(1)	(2)	(3)	(4)
Expert mean	0.25^{***} (0.01)	0.16^{***} (0.03)	0.14^{***} (0.03)	-0.22^{***} (0.05)
1920		-0.52^{***} (0.05)	-0.51^{***} (0.05)	-0.37^{***} (0.05)
1950		-0.35^{***} (0.05)	-0.35^{***} (0.05)	-0.39^{***} (0.05)
1970		-0.26^{***} (0.05)	-0.26^{***} (0.05)	-0.29^{***} (0.05)
1996		-0.22^{***} (0.05)	-0.22^{***} (0.05)	-0.29^{***} (0.05)
2005		-0.13^{***} (0.04)	-0.13^{***} (0.04)	-0.13^{***} (0.04)
Senegal		-0.19^{***} (0.03)	-0.19^{***} (0.03)	-0.16^{***} (0.03)
Political killings		0.04 (0.04)	0.03 (0.04)	-0.72^{***} (0.13)
Forced labor		-0.01 (0.05)	-0.01 (0.05)	-1.96^{***} (0.23)
Journalist harassment		0.12^{***} (0.04)	0.12^{***} (0.04)	-0.18 (0.14)
Gender equality		-0.01 (0.04)	-0.02 (0.04)	-0.98^{***} (0.11)
Expert mean × Political killings				0.41^{***} (0.06)
Expert mean × Forced labor				0.79^{***} (0.08)
Expert mean × Journalist harassment				0.20^{***} (0.06)
Expert mean × Gender equality				0.57^{***} (0.06)
High payment			-0.004 (0.07)	
High payment × Expert mean			0.03 (0.03)	
Constant	1.15^{***} (0.03)	1.62^{***} (0.08)	1.63^{***} (0.09)	2.28^{***} (0.10)
Observations	10,403	10,403	10,403	10,403
R^2	0.03	0.04	0.04	0.06
Adjusted R^2	0.03	0.04	0.04	0.05

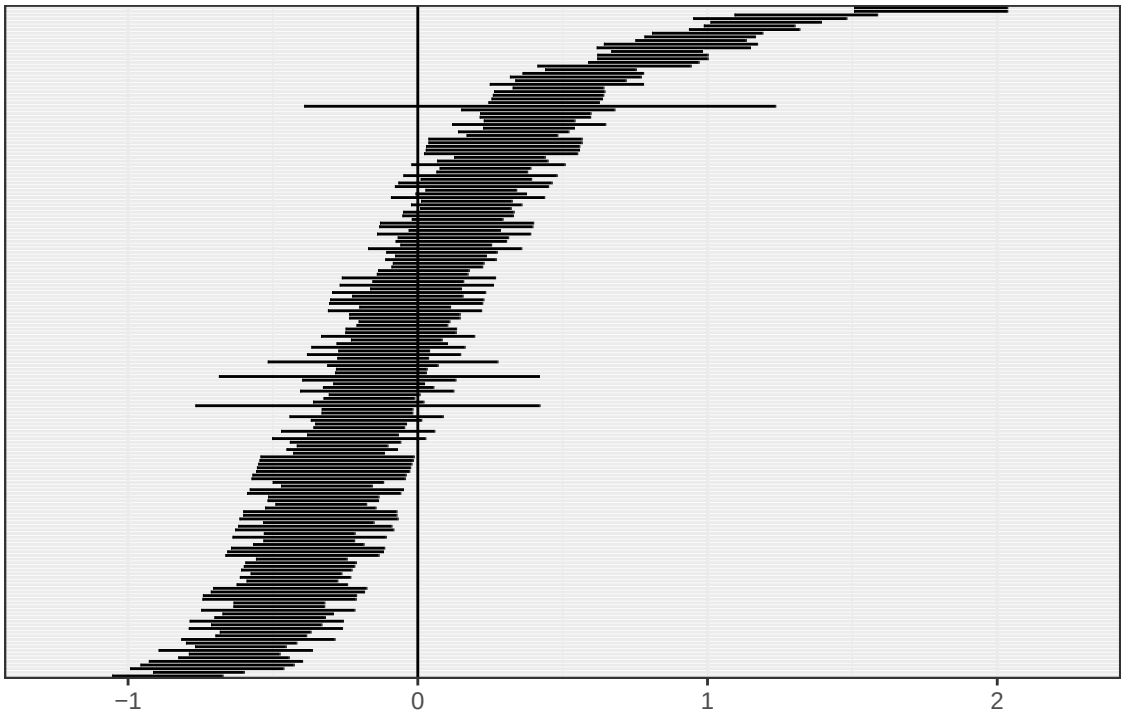
Note: *p<0.1; **p<0.05; ***p<0.01

Reference level: 2015, Argentina, Judicial independence

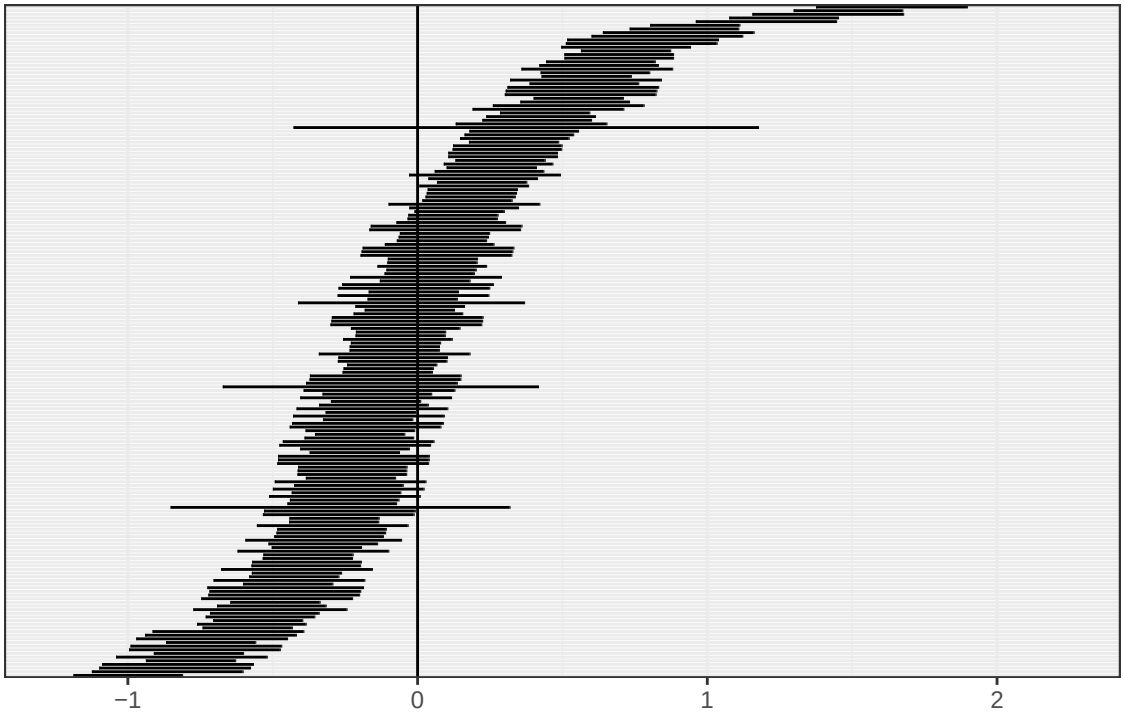
F Random effects for expert-coded variables

Figure 14 presents coder random effects models, without variable or treatment fixed effects (panel a) and with interacted variable and treatment fixed effects (panel b). These models indicate high variation in absolute differences across coders. Finally, Figure 7 reports the effect of coder characteristics on the difference between expert and crowd ratings. None of the covariates significantly shrink or grow the difference between expert and crowd ratings.

Figure 14: Coder random intercepts from mixed models



(a) Model without variable or treatment fixed effects



(b) Model with interacted variable and treatment fixed effects

G Regression tables

G.1 Expert-coded question regression tables

Table 7: Distance between coder scores and average expert score

	(1)	(2)	(3)	(4)	(5)
Reference	1.30*** (0.01)	1.30*** (0.01)	1.19*** (0.03)	1.18*** (0.03)	1.20*** (0.04)
High payment		−0.001 (0.02)		0.01 (0.02)	−0.03 (0.05)
1920			0.06** (0.03)	0.06** (0.03)	0.12*** (0.04)
1950			0.01 (0.03)	0.01 (0.03)	0.02 (0.04)
1970			−0.13*** (0.03)	−0.13*** (0.03)	−0.10** (0.04)
1996			−0.08*** (0.03)	−0.08*** (0.03)	−0.17*** (0.04)
2005			−0.02 (0.03)	−0.02 (0.03)	−0.04 (0.04)
Senegal			0.02 (0.02)	0.02 (0.02)	0.07** (0.03)
Political killings			0.20*** (0.03)	0.20*** (0.03)	0.07* (0.04)
Forced labor			0.41*** (0.03)	0.41*** (0.03)	0.52*** (0.04)
Journalist harassment			0.12*** (0.03)	0.12*** (0.03)	−0.03 (0.04)
Gender equality			−0.03 (0.03)	−0.04 (0.03)	0.02 (0.04)
High payment × 1920					−0.11* (0.06)
High payment × 1950					−0.01 (0.06)
High payment × 1970					−0.05 (0.06)
High payment × 1996					0.17*** (0.06)
High payment × 2005					0.04 (0.06)
High payment × Senegal					−0.07** (0.04)
High payment × Political killings					0.23*** (0.05)
High payment × Forced labor					−0.23*** (0.06)
High payment × Journalist harassment					0.31*** (0.05)
High payment × Gender equality					−0.06 (0.05)
Observations	10,403	10,403	10,403	10,403	10,403
R^2	0.00	0.0000	0.03	0.03	0.05
Adjusted R^2	0.00	−0.0001	0.03	0.03	0.05
Note:					*p<0.1; **p<0.05; ***p<0.01
					Reference: 2015, Argentina, Judicial independence

Table 8: Coder demographics and distance from expert mean

Reference	1.19*** (0.04)	1.20*** (0.04)
Does not discuss politics	0.09*** (0.02)	0.09*** (0.02)
No university education	0.09*** (0.03)	0.09*** (0.03)
Political science major	−0.04* (0.02)	−0.04* (0.02)
All screeners correct	−0.09*** (0.02)	−0.09*** (0.02)
2 < gold correct	0.04* (0.02)	0.04* (0.02)
Did not use V–Dem	0.09*** (0.02)	0.09*** (0.02)
Female	0.05** (0.02)	0.05** (0.02)
log(Age)	0.05 (0.04)	0.05 (0.04)
Reads coded-country language	−0.07*** (0.02)	−0.08*** (0.02)
Resided in country coded	−0.08*** (0.03)	−0.08*** (0.03)
log(Variable coding time)	−0.04*** (0.01)	−0.04*** (0.01)
High payment		−0.01 (0.02)
1920	0.08*** (0.03)	0.08*** (0.03)
1950	0.02 (0.03)	0.02 (0.03)
1970	−0.15*** (0.03)	−0.15*** (0.03)
1996	−0.10*** (0.03)	−0.10*** (0.03)
2005	−0.03 (0.03)	−0.03 (0.03)
Senegal	−0.003 (0.02)	−0.003 (0.02)
Political killings	0.14*** (0.03)	0.14*** (0.03)
Forced labor	0.33*** (0.03)	0.33*** (0.03)
Journalist harassment	0.09*** (0.03)	0.09*** (0.03)
Gender equality	−0.06** (0.03)	−0.06** (0.03)
Observations	8,799	8,799
R^2	0.05	0.05
Adjusted R^2	0.04	0.04

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

G.2 Trained coder questions regression tables

Table 9: Task characteristics and correct answer to factual questions

	(1)	(2)	(3)	(4)	(5)
Reference	−0.23*** (0.01)	−0.27*** (0.02)	0.24*** (0.05)	0.22*** (0.05)	0.41*** (0.06)
1920			−0.16*** (0.05)	−0.16*** (0.05)	−0.26*** (0.07)
1950			−0.15*** (0.05)	−0.15*** (0.05)	−0.16** (0.07)
1970			−0.08 (0.05)	−0.08 (0.05)	−0.06 (0.07)
1996			−0.005 (0.05)	−0.005 (0.05)	−0.16** (0.07)
2005			0.12** (0.05)	0.12** (0.05)	0.03 (0.07)
Senegal			−0.05 (0.03)	−0.05 (0.03)	0.05 (0.05)
Referenda permitted			−0.35*** (0.04)	−0.35*** (0.04)	−0.49*** (0.06)
Bicameral legislature			−0.34*** (0.04)	−0.34*** (0.04)	−0.57*** (0.06)
Suffrage level			−1.14*** (0.05)	−1.14*** (0.05)	−1.37*** (0.06)
High payment		0.07** (0.03)		0.03 (0.03)	−0.31*** (0.09)
High payment × 1920					0.20* (0.10)
High payment × 1950					0.01 (0.10)
High payment × 1970					−0.02 (0.10)
High payment × 1996					0.30*** (0.10)
High payment × 2005					0.17* (0.10)
High payment × Senegal					−0.19*** (0.07)
High payment × Referenda permitted					0.27*** (0.08)
High payment × Bicameral legislature					0.46*** (0.08)
High payment × Suffrage level					0.47*** (0.09)
Observations	7,648	7,648	7,648	7,648	7,648
Log Likelihood	−5,168.97	−5,165.81	−4,804.14	−4,803.47	−4,770.76
Akaike Inf. Crit.	10,339.94	10,335.62	9,628.28	9,628.94	9,581.52

Note: *p<0.1; **p<0.05; ***p<0.01
Reference: Voting age, 2015, Argentina

Table 10: Coder characteristics and correct answer to factual questions

	(1)	(2)
Reference	0.12* (0.07)	0.11 (0.07)
Does not discuss politics	−0.05 (0.04)	−0.05 (0.04)
No university education	0.08 (0.05)	0.08 (0.05)
Political science major	−0.26*** (0.04)	−0.26*** (0.04)
All screeners correct	0.51*** (0.04)	0.51*** (0.04)
2 < gold correct	−0.05 (0.04)	−0.05 (0.04)
Did not use V–Dem	−0.24*** (0.04)	−0.24*** (0.04)
Female	−0.07* (0.04)	−0.07* (0.04)
log(Age)	0.30*** (0.07)	0.30*** (0.07)
Reads coded-country language	−0.03 (0.04)	−0.03 (0.04)
Resided in coded country	0.20*** (0.05)	0.20*** (0.05)
log(Variable coding time)	0.20*** (0.02)	0.20*** (0.02)
High payment		0.02 (0.03)
1920	−0.11** (0.06)	−0.11** (0.06)
1950	−0.12** (0.06)	−0.12** (0.06)
1970	−0.02 (0.06)	−0.02 (0.06)
1996	0.01 (0.06)	0.01 (0.06)
2005	0.13** (0.06)	0.13** (0.06)
Senegal	0.06 (0.04)	0.06 (0.04)
Referenda permitted	−0.20*** (0.05)	−0.20*** (0.05)
Bicameral legislature	−0.29*** (0.05)	−0.30*** (0.05)
Suffrage level	−1.06*** (0.05)	−1.06*** (0.05)
Observations	6,842	6,842
Log Likelihood	−4,072.61	−4,072.48
Akaike Inf. Crit.	8,187.22	8,188.96

Note: *p<0.1; **p<0.05; ***p<0.01
Reference: Voting age, 2015, Argentina

H Additional Likert-scale robustness checks

Table 11: Distance between coder scores and average expert score, bootstrapped

	(1)	(2)	(3)	(4)	(5)
Intercept	1.41*** (0.04)	1.41*** (0.06)	1.20*** (0.14)	1.20*** (0.14)	1.15*** (0.19)
1920			0.04 (0.15)	0.04 (0.15)	0.07 (0.21)
1950			0.08 (0.15)	0.08 (0.15)	0.10 (0.21)
1970			-0.03 (0.14)	-0.03 (0.14)	0.05 (0.19)
1996			-0.01 (0.14)	-0.01 (0.14)	-0.02 (0.19)
2005			0.04 (0.14)	0.04 (0.14)	0.05 (0.20)
Senegal			0.07 (0.08)	0.07 (0.08)	0.10 (0.12)
Political killings			0.22* (0.13)	0.22* (0.13)	0.16 (0.18)
Forced labor			0.43*** (0.13)	0.43*** (0.13)	0.61*** (0.19)
Journalist harassment			0.15 (0.12)	0.15 (0.12)	0.04 (0.17)
Gender equality			-0.02 (0.13)	-0.02 (0.13)	0.03 (0.18)
High Payment		-0.00 (0.08)		-0.00 (0.08)	0.10 (0.27)
High Payment × 1920					-0.06 (0.30)
High Payment × 1950					-0.04 (0.30)
High Payment × 1970					-0.16 (0.28)
High Payment × 1996					0.02 (0.28)
High Payment × 2005					-0.02 (0.29)
High Payment × Senegal					-0.07 (0.17)
High Payment × Political killings					0.11 (0.26)
High Payment × Forced labor					-0.37 (0.26)
High Payment × Journalist harassment					0.22 (0.25)
High Payment × Gender equality					-0.10 (0.25)
Observations	600	600	600	600	600
Simulations	2000	2000	2000	2000	2000

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

Table 12: Distance between coder scores and average expert score, controlling for expert coding characteristics

	(1)	(2)	(3)	(4)	(5)
Reference	1.33*** (0.01)	1.33*** (0.02)	1.23*** (0.03)	1.22*** (0.03)	1.23*** (0.04)
1.5 < Expert mean < 2.5	−0.17*** (0.02)	−0.17*** (0.02)	−0.15*** (0.02)	−0.15*** (0.02)	−0.16*** (0.02)
3 < Expert mean < 1	0.06*** (0.02)	0.06*** (0.02)	−0.04* (0.02)	−0.04* (0.02)	−0.04 (0.02)
1920			0.05 (0.03)	0.05 (0.03)	0.11*** (0.04)
1950			0.05 (0.03)	0.05 (0.03)	0.06 (0.04)
1970			−0.13*** (0.03)	−0.13*** (0.03)	−0.10** (0.04)
1996			−0.04 (0.03)	−0.04 (0.03)	−0.12*** (0.04)
2005			0.03 (0.03)	0.03 (0.03)	0.02 (0.04)
Senegal			0.05*** (0.02)	0.05*** (0.02)	0.11*** (0.03)
Political killings			0.18*** (0.03)	0.18*** (0.03)	0.06 (0.04)
Forced labor			0.37*** (0.03)	0.37*** (0.03)	0.48*** (0.04)
Harassment of journalists			0.12*** (0.03)	0.12*** (0.03)	−0.03 (0.04)
Gender equality			−0.03 (0.03)	−0.03 (0.03)	0.02 (0.04)
High Payment		−0.0001 (0.02)		0.01 (0.02)	−0.02 (0.05)
High Payment × 1920					−0.12** (0.06)
High Payment × 1950					−0.01 (0.06)
High Payment × 1970					−0.05 (0.06)
High Payment × 1996					0.17*** (0.06)
High Payment × 2005					0.04 (0.06)
High Payment × Senegal					−0.08** (0.04)
High Payment × Political killings					0.23*** (0.05)
High Payment × Forced labor					−0.23*** (0.06)
High Payment × Journalist harassment					0.31*** (0.05)
High Payment × Gender equality					−0.07 (0.05)
Observations	10,403	10,403	10,403	10,403	10,403
R^2	0.01	0.01	0.04	0.04	0.05
Adjusted R^2	0.01	0.01	0.04	0.04	0.05

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

I Subset analyses

I.1 Subset by screener questions

I.1.1 No screeners correct

Table 13: Relationship between coder scores and average expert score

	(1)	(2)	(3)	(4)
Expert mean	0.05** (0.03)	−0.01 (0.07)	0.02 (0.07)	−0.17** (0.08)
1920		−0.41*** (0.11)	−0.41*** (0.11)	−0.41*** (0.11)
1950		−0.09 (0.12)	−0.09 (0.12)	−0.09 (0.12)
1970		−0.09 (0.16)	−0.09 (0.16)	−0.09 (0.16)
1996		−0.09 (0.08)	−0.09 (0.08)	−0.09 (0.08)
2005		−0.04 (0.08)	−0.04 (0.08)	−0.04 (0.08)
Senegal		−0.11** (0.05)	−0.12** (0.05)	−0.11** (0.05)
Political killings		0.03 (0.06)	0.07 (0.06)	−0.73*** (0.22)
Forced labor		0.37*** (0.07)	0.40*** (0.07)	−0.54** (0.24)
Journalist harassment		0.69*** (0.07)	0.69*** (0.07)	0.66*** (0.23)
Gender equality		0.17*** (0.06)	0.19*** (0.06)	−0.47** (0.22)
Expert mean × Political killings				0.29*** (0.08)
Expert mean × Forced labor				0.35*** (0.09)
Expert mean × Journalist harassment				0.01 (0.08)
Expert mean × Gender equality				0.24*** (0.08)
High payment			0.04 (0.15)	
Expert mean × High payment			−0.08 (0.05)	
Constant	1.53*** (0.08)	1.64*** (0.25)	1.60*** (0.26)	2.06*** (0.28)
Observations	3,007	3,007	3,007	3,007
R^2	0.001	0.06	0.06	0.07

Note: *p<0.1; **p<0.05; ***p<0.01
Reference: 2015, Argentina, Judicial independence

Table 14: Distance between coder scores and average expert score

	(1)	(2)	(3)	(4)	(5)
Reference	1.25*** (0.02)	1.25*** (0.02)	1.16*** (0.05)	1.18*** (0.05)	1.06*** (0.06)
1920			−0.07 (0.05)	−0.07 (0.05)	0.005 (0.07)
1950			−0.12** (0.05)	−0.12** (0.05)	0.10 (0.07)
1970			−0.26*** (0.05)	−0.26*** (0.05)	−0.07 (0.07)
1996			−0.30*** (0.05)	−0.30*** (0.05)	−0.31*** (0.07)
2005			−0.13** (0.05)	−0.13** (0.05)	−0.14** (0.07)
Senegal			−0.02 (0.03)	−0.02 (0.03)	0.06 (0.05)
Political killings			0.53*** (0.04)	0.54*** (0.04)	0.47*** (0.06)
Forced labor			0.56*** (0.05)	0.57*** (0.05)	0.74*** (0.07)
Journalist harassment			0.13*** (0.05)	0.13*** (0.05)	0.06 (0.06)
Gender equality			0.16*** (0.04)	0.16*** (0.04)	0.22*** (0.06)
High payment		0.01 (0.03)		−0.05* (0.03)	0.19** (0.09)
High payment × 1920					−0.17 (0.10)
High payment × 1950					−0.50*** (0.10)
High payment × 1970					−0.44*** (0.10)
High payment × 1996					0.03 (0.10)
High payment × 2005					0.01 (0.10)
High payment × Senegal					−0.15** (0.07)
High payment × Political killings					0.11 (0.09)
High payment × Forced labor					−0.35*** (0.10)
High payment × Journalist harassment					0.14 (0.10)
High payment × Gender equality					−0.12 (0.09)
Observations	3,007	3,007	3,007	3,007	3,007
R^2	0.00	0.0000	0.08	0.09	0.11
Adjusted R^2	0.00	−0.0003	0.08	0.08	0.10
Note:					*p<0.1; **p<0.05; ***p<0.01
					Reference: 2015, Argentina, Judicial independence

Table 15: Task characteristics and correct answer to factual questions

	(1)	(2)	(3)	(4)	(5)
Reference	−0.71*** (0.03)	−0.62*** (0.04)	−0.32*** (0.11)	−0.19 (0.12)	0.05 (0.16)
1920			0.16 (0.13)	0.17 (0.13)	0.18 (0.17)
1950			0.14 (0.13)	0.14 (0.13)	−0.003 (0.17)
1970			0.18 (0.13)	0.18 (0.13)	0.09 (0.17)
1996			0.22* (0.13)	0.23* (0.13)	0.06 (0.17)
2005			0.17 (0.13)	0.17 (0.13)	0.02 (0.17)
Senegal			0.06 (0.08)	0.03 (0.08)	−0.11 (0.10)
Referenda permitted			−0.28*** (0.09)	−0.34*** (0.09)	−0.44*** (0.14)
Bicameral legislature			−0.39*** (0.10)	−0.44*** (0.10)	−0.60*** (0.15)
Suffrage level			−2.14*** (0.17)	−2.18*** (0.17)	−2.17*** (0.21)
High payment		−0.18*** (0.07)		−0.20*** (0.08)	−0.74*** (0.23)
High payment × 1920					−0.003 (0.27)
High payment × 1950					0.38 (0.26)
High payment × 1970					0.25 (0.26)
High payment × 1996					0.41 (0.26)
High payment × 2005					0.39 (0.27)
High payment × Senegal					0.44*** (0.17)
High payment × Referenda permitted					0.24 (0.19)
High payment × Bicameral legislature					0.33 (0.21)
High payment × Suffrage level					−0.33 (0.41)
Observations	1,737	1,737	1,737	1,737	1,737
Log Likelihood	−957.36	−953.64	−780.91	−777.44	−768.53
Akaike Inf. Crit.	1,916.73	1,911.28	1,581.81	1,576.88	1,577.07

I.1.2 0 < screeners correct

Table 16: Relationship between coder scores and average expert score

	(1)	(2)	(3)	(4)
Expert mean	0.27^{***} (0.02)	0.02 (0.05)	−0.002 (0.05)	-0.29^{***} (0.06)
1920		-0.74^{***} (0.08)	-0.74^{***} (0.08)	-0.75^{***} (0.08)
1950		-0.69^{***} (0.08)	-0.69^{***} (0.08)	-0.70^{***} (0.08)
1970		-0.57^{***} (0.11)	-0.57^{***} (0.11)	-0.57^{***} (0.11)
1996		-0.38^{***} (0.05)	-0.37^{***} (0.05)	-0.38^{***} (0.05)
2005		-0.21^{***} (0.05)	-0.21^{***} (0.05)	-0.21^{***} (0.05)
Senegal		-0.25^{***} (0.04)	-0.26^{***} (0.04)	-0.26^{***} (0.04)
Political killings		0.19^{***} (0.05)	0.18^{***} (0.05)	-0.81^{***} (0.16)
Forced labor		0.09^{*} (0.05)	0.10^{*} (0.05)	-0.92^{***} (0.17)
Journalist harassment		−0.05 (0.05)	−0.05 (0.05)	-0.76^{***} (0.17)
Gender equality		-0.16^{***} (0.05)	-0.19^{***} (0.05)	-1.46^{***} (0.17)
Expert mean × Political killings				0.38^{***} (0.06)
Expert mean × Forced labor				0.38^{***} (0.06)
Expert mean × Journalist harassment				0.27^{***} (0.06)
Expert mean × Gender equality				0.49^{***} (0.06)
High payment			0.05 (0.11)	
Expert mean × High payment			0.04 (0.04)	
Constant	0.97^{***} (0.05)	2.13^{***} (0.18)	2.11^{***} (0.19)	2.96^{***} (0.21)
Observations	7,396	7,396	7,396	7,396
R^2	0.03	0.06	0.06	0.07
Adjusted R^2	0.03	0.06	0.06	0.06

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

Table 17: Distance between coder scores and average expert score

	(1)	(2)	(3)	(4)	(5)
Reference	1.32*** (0.01)	1.32*** (0.02)	1.23*** (0.03)	1.22*** (0.04)	1.32*** (0.05)
1920			0.11*** (0.04)	0.11*** (0.04)	0.18*** (0.05)
1950			0.07* (0.04)	0.07* (0.04)	−0.03 (0.05)
1970			−0.08** (0.04)	−0.08** (0.04)	−0.12** (0.05)
1996			0.01 (0.04)	0.01 (0.04)	−0.10* (0.05)
2005			0.03 (0.04)	0.03 (0.04)	0.02 (0.05)
Senegal			0.02 (0.02)	0.02 (0.02)	0.05 (0.03)
Political killings			0.04 (0.03)	0.04 (0.03)	−0.16*** (0.05)
Forced labor			0.32*** (0.03)	0.32*** (0.03)	0.36*** (0.05)
Journalist harassment			0.09*** (0.03)	0.09*** (0.03)	−0.13*** (0.05)
Gender equality			−0.14*** (0.03)	−0.14*** (0.03)	−0.14*** (0.05)
High payment		−0.02 (0.02)		0.03 (0.02)	−0.16** (0.07)
High payment × 1920					−0.12* (0.07)
High payment × 1950					0.17** (0.07)
High payment × 1970					0.08 (0.07)
High payment × 1996					0.18** (0.07)
High payment × 2005					0.02 (0.07)
High payment × Senegal					−0.05 (0.04)
High payment × Political killings					0.36*** (0.06)
High payment × Forced labor					−0.11* (0.07)
High payment × Journalist harassment					0.40*** (0.07)
High payment × Gender equality					0.04 (0.07)
Observations	7,396	7,396	7,396	7,396	7,396
R^2	0.00	0.0001	0.03	0.03	0.05
Adjusted R^2	0.00	−0.0001	0.03	0.03	0.04

Note: *p<0.1; **p<0.05; ***p<0.01
Reference: 2015, Argentina, Judicial independence

Table 18: Task characteristics and correct answer to factual questions

	(1)	(2)	(3)	(4)	(5)
Reference	−0.11*** (0.02)	−0.17*** (0.02)	0.35*** (0.05)	0.31*** (0.05)	0.47*** (0.07)
1920			−0.22*** (0.06)	−0.22*** (0.06)	−0.37*** (0.08)
1950			−0.21*** (0.06)	−0.21*** (0.06)	−0.20** (0.08)
1970			−0.13** (0.06)	−0.13** (0.06)	−0.11 (0.08)
1996			−0.05 (0.06)	−0.05 (0.06)	−0.22*** (0.08)
2005			0.11* (0.06)	0.11* (0.06)	0.03 (0.08)
Senegal			−0.05 (0.04)	−0.05 (0.04)	0.12** (0.05)
Referenda permitted			−0.32*** (0.05)	−0.33*** (0.05)	−0.39*** (0.07)
Bicameral legislature			−0.35*** (0.04)	−0.35*** (0.04)	−0.55*** (0.06)
Suffrage level			−0.95*** (0.05)	−0.95*** (0.05)	−1.23*** (0.07)
High payment		0.12*** (0.03)		0.08** (0.03)	−0.22** (0.10)
High payment × 1920					0.29** (0.12)
High payment × 1950					−0.02 (0.12)
High payment × 1970					−0.04 (0.12)
High payment × 1996					0.33*** (0.12)
High payment × 2005					0.16 (0.12)
High payment × Senegal					−0.34*** (0.07)
High payment × Referenda permitted					0.15 (0.10)
High payment × Bicameral legislature					0.42*** (0.09)
High payment × Suffrage level					0.60*** (0.10)
Observations	5,911	5,911	5,911	5,911	5,911
Log Likelihood	−4,074.73	−4,068.12	−3,865.94	−3,863.12	−3,822.20
Akaike Inf. Crit.	8,151.46	8,140.23	7,751.87	7,748.23	7,684.40

I.1.3 All screeners correct

Table 19: Relationship between coder scores and average expert score

	(1)	(2)	(3)	(4)
Expert mean	0.36^{***} (0.03)	-0.01 (0.07)	0.03 (0.08)	-0.33^{***} (0.10)
1920		-0.98^{***} (0.11)	-0.98^{***} (0.11)	-0.97^{***} (0.11)
1950		-0.96^{***} (0.12)	-0.97^{***} (0.12)	-0.96^{***} (0.12)
1970		-0.83^{***} (0.16)	-0.83^{***} (0.16)	-0.83^{***} (0.16)
1996		-0.48^{***} (0.08)	-0.48^{***} (0.08)	-0.48^{***} (0.08)
2005		-0.29^{***} (0.08)	-0.29^{***} (0.08)	-0.29^{***} (0.08)
Senegal		-0.45^{***} (0.05)	-0.45^{***} (0.05)	-0.44^{***} (0.05)
Political killings		0.61^{***} (0.08)	0.59^{***} (0.08)	-0.54^{**} (0.27)
Forced labor		0.52^{***} (0.08)	0.49^{***} (0.08)	-0.25 (0.28)
Journalist harassment		0.36^{***} (0.08)	0.27^{***} (0.08)	-0.34 (0.27)
Gender equality		-0.29^{***} (0.08)	-0.40^{***} (0.08)	-1.60^{***} (0.28)
Expert mean × Political killings				0.43^{***} (0.10)
Expert mean × Forced labor				0.29^{***} (0.10)
Expert mean × Journalist harassment				0.26^{***} (0.10)
Expert mean × Gender equality				0.49^{***} (0.10)
High payment			0.60^{***} (0.15)	
Expert mean × High payment			-0.07 (0.05)	
Constant	0.76^{***} (0.08)	2.18^{***} (0.27)	1.90^{***} (0.28)	3.05^{***} (0.33)
Observations	3,601	3,601	3,601	3,601
R^2	0.04	0.14	0.16	0.15
Adjusted R^2	0.04	0.14	0.16	0.14

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

Table 20: Distance between coder scores and average expert score

	(1)	(2)	(3)	(4)	(5)
Reference	1.28*** (0.01)	1.33*** (0.02)	1.20*** (0.05)	1.24*** (0.05)	1.12*** (0.07)
1920			0.04 (0.05)	0.04 (0.05)	0.24*** (0.07)
1950			0.01 (0.05)	0.01 (0.05)	−0.04 (0.07)
1970			−0.09* (0.05)	−0.09* (0.05)	−0.08 (0.07)
1996			−0.02 (0.05)	−0.01 (0.05)	−0.005 (0.07)
2005			0.06 (0.05)	0.06 (0.05)	0.12* (0.07)
Senegal			−0.03 (0.03)	−0.03 (0.03)	0.05 (0.05)
Political killings			0.002 (0.05)	0.01 (0.05)	−0.05 (0.07)
Forced labor			0.30*** (0.05)	0.31*** (0.05)	0.64*** (0.07)
Journalist harassment			0.18*** (0.05)	0.19*** (0.05)	0.09 (0.07)
Gender equality			−0.05 (0.05)	−0.03 (0.05)	0.18** (0.08)
High payment		−0.09*** (0.03)		−0.09*** (0.03)	0.16 (0.11)
High payment × 1920					−0.35*** (0.10)
High payment × 1950					0.09 (0.10)
High payment × 1970					−0.02 (0.10)
High payment × 1996					−0.02 (0.10)
High payment × 2005					−0.12 (0.10)
High payment × Senegal					−0.12* (0.06)
High payment × Political killings					0.09 (0.10)
High payment × Forced labor					−0.67*** (0.11)
High payment × Journalist harassment					0.11 (0.10)
High payment × Gender equality					−0.35*** (0.11)
Observations	3,601	3,601	3,601	3,601	3,601
R^2	0.00	0.003	0.03	0.03	0.06
Adjusted R^2	0.00	0.002	0.02	0.03	0.06
Note:					*p<0.1; **p<0.05; ***p<0.01
					Reference: 2015, Argentina, Judicial independence

Table 21: Task characteristics and correct answer to factual questions

	(1)	(2)	(3)	(4)	(5)
Reference	0.10*** (0.02)	0.004 (0.03)	0.80*** (0.08)	0.70*** (0.08)	0.68*** (0.10)
1920			−0.32*** (0.08)	−0.32*** (0.08)	−0.48*** (0.11)
1950			−0.28*** (0.08)	−0.28*** (0.08)	−0.24** (0.11)
1970			−0.14* (0.08)	−0.14* (0.08)	−0.03 (0.11)
1996			−0.10 (0.08)	−0.10 (0.08)	−0.26** (0.11)
2005			0.08 (0.08)	0.08 (0.08)	0.07 (0.11)
Senegal			−0.01 (0.05)	−0.02 (0.05)	0.04 (0.07)
Referenda permitted			−0.80*** (0.07)	−0.92*** (0.07)	−0.91*** (0.11)
Bicameral legislature			−0.66*** (0.06)	−0.70*** (0.06)	−0.65*** (0.08)
Suffrage level			−0.77*** (0.07)	−0.81*** (0.07)	−0.78*** (0.09)
High payment		0.21*** (0.05)		0.34*** (0.05)	0.38** (0.16)
High payment × 1920					0.35** (0.16)
High payment × 1950					−0.10 (0.16)
High payment × 1970					−0.22 (0.16)
High payment × 1996					0.36** (0.16)
High payment × 2005					0.04 (0.16)
High payment × Senegal					−0.12 (0.10)
High payment × Referenda permitted					−0.05 (0.16)
High payment × Bicameral legislature					−0.12 (0.14)
High payment × Suffrage level					−0.09 (0.16)
Observations	3,027	3,027	3,027	3,027	3,027
Log Likelihood	−2,088.39	−2,077.85	−1,987.09	−1,963.01	−1,951.34
Akaike Inf. Crit.	4,178.78	4,159.70	3,994.18	3,948.03	3,942.68

I.2 Subset by gold questions

I.2.1 3 > gold correct

Table 22: Relationship between coder scores and average expert score

	(1)	(2)	(3)	(4)
Expert mean	0.14^{***} (0.02)	-0.01 (0.05)	-0.04 (0.06)	-0.23^{***} (0.07)
1920		-0.52^{***} (0.08)	-0.52^{***} (0.08)	-0.52^{***} (0.08)
1950		-0.34^{***} (0.09)	-0.34^{***} (0.09)	-0.33^{***} (0.09)
1970		-0.28^{**} (0.12)	-0.28^{**} (0.12)	-0.28^{**} (0.12)
1996		-0.20^{***} (0.06)	-0.20^{***} (0.06)	-0.19^{***} (0.06)
2005		-0.12^{**} (0.06)	-0.12^{**} (0.06)	-0.12^{**} (0.06)
Senegal		-0.24^{***} (0.04)	-0.24^{***} (0.04)	-0.24^{***} (0.04)
Political killings		-0.16^{***} (0.05)	-0.15^{***} (0.05)	-1.11^{***} (0.18)
Forced labor		0.16^{***} (0.06)	0.16^{***} (0.06)	-0.59^{***} (0.19)
Journalist harassment		0.04 (0.05)	0.02 (0.06)	-0.53^{***} (0.19)
Gender equality		0.09^{*} (0.05)	0.09^{*} (0.05)	-0.60^{***} (0.17)
Expert mean × Political killings				0.36^{***} (0.06)
Expert mean × Forced labor				0.28^{***} (0.07)
Expert mean × Journalist harassment				0.22^{***} (0.07)
Expert mean × Gender equality				0.26^{***} (0.06)
High payment			-0.24^{**} (0.12)	
Expert mean × High payment			0.07 (0.04)	
Constant	1.35^{***} (0.06)	2.03^{***} (0.19)	2.16^{***} (0.20)	2.61^{***} (0.22)
Observations	5,404	5,404	5,404	5,404
R^2	0.01	0.03	0.03	0.04
Adjusted R^2	0.01	0.03	0.03	0.04

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

Table 23: Distance between coder scores and average expert score

	(1)	(2)	(3)	(4)	(5)
Reference	1.29*** (0.01)	1.26*** (0.02)	1.17*** (0.04)	1.15*** (0.04)	1.10*** (0.05)
1920			0.0004 (0.04)	0.0003 (0.04)	−0.03 (0.06)
1950			−0.09** (0.04)	−0.09** (0.04)	−0.09 (0.06)
1970			−0.20*** (0.04)	−0.20*** (0.04)	−0.18*** (0.06)
1996			−0.14*** (0.04)	−0.14*** (0.04)	−0.28*** (0.06)
2005			−0.09** (0.04)	−0.09** (0.04)	−0.10* (0.06)
Senegal			0.01 (0.03)	0.01 (0.03)	0.07* (0.04)
Political killings			0.39*** (0.04)	0.38*** (0.04)	0.31*** (0.06)
Forced labor			0.39*** (0.04)	0.39*** (0.04)	0.58*** (0.06)
Journalist harassment			0.11*** (0.04)	0.12*** (0.04)	0.16*** (0.05)
Gender equality			0.19*** (0.04)	0.19*** (0.04)	0.33*** (0.05)
High payment		0.05** (0.02)		0.03 (0.02)	0.13* (0.07)
High payment × 1920					0.06 (0.08)
High payment × 1950					−0.01 (0.08)
High payment × 1970					−0.04 (0.08)
High payment × 1996					0.26*** (0.08)
High payment × 2005					0.02 (0.08)
High payment × Senegal					−0.11** (0.05)
High payment × Political killings					0.10 (0.07)
High payment × Forced labor					−0.38*** (0.08)
High payment × Journalist harassment					−0.10 (0.08)
High payment × Gender equality					−0.26*** (0.07)
Observations	5,404	5,404	5,404	5,404	5,404
R^2	0.00	0.001	0.04	0.04	0.05
Adjusted R^2	0.00	0.001	0.03	0.03	0.05
Note:					*p<0.1; **p<0.05; ***p<0.01
					Reference: 2015, Argentina, Judicial independence

Table 24: Task characteristics and correct answer to factual questions

	(1)	(2)	(3)	(4)	(5)
Constant	−0.32*** (0.02)	−0.34*** (0.03)	−0.002 (0.06)	0.002 (0.07)	0.03 (0.10)
1920			0.01 (0.07)	0.01 (0.07)	−0.08 (0.11)
1950			0.14* (0.07)	0.14* (0.07)	0.17 (0.11)
1970			0.15** (0.07)	0.15** (0.07)	0.23** (0.11)
1996			0.18** (0.07)	0.18** (0.07)	0.09 (0.11)
2005			0.19** (0.07)	0.19** (0.07)	0.08 (0.11)
Senegal			−0.13*** (0.05)	−0.14*** (0.05)	−0.11 (0.07)
Referenda permitted			−0.29*** (0.06)	−0.29*** (0.06)	−0.11 (0.09)
Bicameral legislature			−0.31*** (0.06)	−0.31*** (0.06)	−0.35*** (0.09)
Suffrage level			−1.16*** (0.07)	−1.16*** (0.07)	−1.51*** (0.11)
High payment		0.04 (0.04)		−0.01 (0.04)	−0.05 (0.13)
High payment × 1920					0.18 (0.15)
High payment × 1950					−0.04 (0.15)
High payment × 1970					−0.14 (0.15)
High payment × 1996					0.17 (0.15)
High payment × 2005					0.19 (0.15)
High payment × Senegal					−0.13 (0.10)
High payment × Referenda permitted					−0.36*** (0.12)
High payment × Bicameral legislature					0.07 (0.12)
High payment × Suffrage level					0.65*** (0.14)
Observations	3,817	3,817	3,817	3,817	3,817
Log Likelihood	−2,527.03	−2,526.59	−2,344.98	−2,344.97	−2,311.45
Akaike Inf. Crit.	5,056.06	5,057.17	4,709.96	4,711.93	4,662.90

I.2.2 2 < gold correct

Table 25: Relationship between coder scores and average expert score

	(1)	(2)	(3)	(4)
Expert mean	0.29*** (0.02)	0.03 (0.06)	0.04 (0.06)	−0.31*** (0.07)
1920		−0.78*** (0.09)	−0.78*** (0.09)	−0.80*** (0.09)
1950		−0.72*** (0.10)	−0.72*** (0.10)	−0.72*** (0.10)
1970		−0.61*** (0.14)	−0.61*** (0.14)	−0.61*** (0.13)
1996		−0.40*** (0.07)	−0.40*** (0.07)	−0.40*** (0.07)
2005		−0.20*** (0.07)	−0.20*** (0.07)	−0.20*** (0.06)
Senegal		−0.22*** (0.04)	−0.23*** (0.04)	−0.23*** (0.04)
Political killings		0.50*** (0.06)	0.47*** (0.06)	−0.47** (0.19)
Forced labor		0.22*** (0.06)	0.21*** (0.06)	−1.03*** (0.21)
Journalist harassment		0.31*** (0.06)	0.27*** (0.06)	−0.20 (0.20)
Gender equality		−0.31*** (0.06)	−0.38*** (0.06)	−2.16*** (0.21)
Expert mean × Political killings				0.36*** (0.07)
Expert mean × Forced labor				0.47*** (0.07)
Expert mean × Journalist harassment				0.19*** (0.07)
Expert mean × Gender equality				0.69*** (0.08)
High payment			0.26** (0.13)	
Expert mean × High payment			−0.02 (0.05)	
Constant	0.90*** (0.07)	1.94*** (0.22)	1.83*** (0.23)	2.84*** (0.25)
Observations	4,999	4,999	4,999	4,999
R^2	0.03	0.09	0.09	0.11
Adjusted R^2	0.03	0.09	0.09	0.10

Note: *p<0.1; **p<0.05; ***p<0.01

Reference: 2015, Argentina, Judicial independence

Table 26: Distance between coder scores and average expert score

	(1)	(2)	(3)	(4)	(5)
Reference	1.30*** (0.01)	1.33*** (0.02)	1.21*** (0.04)	1.21*** (0.04)	1.27*** (0.05)
1920			0.12*** (0.04)	0.12*** (0.04)	0.28*** (0.06)
1950			0.13*** (0.04)	0.13*** (0.04)	0.13** (0.06)
1970			−0.05 (0.04)	−0.05 (0.04)	−0.02 (0.06)
1996			−0.02 (0.04)	−0.02 (0.04)	−0.05 (0.06)
2005			0.07 (0.04)	0.07 (0.04)	0.04 (0.06)
Senegal			0.03 (0.03)	0.03 (0.03)	0.06 (0.04)
Political killings			0.01 (0.04)	0.01 (0.04)	−0.13*** (0.05)
Forced labor			0.41*** (0.04)	0.41*** (0.04)	0.46*** (0.05)
Journalist harassment			0.12*** (0.04)	0.11*** (0.04)	−0.21*** (0.05)
Gender equality			−0.34*** (0.04)	−0.34*** (0.04)	−0.44*** (0.06)
High payment		−0.05** (0.02)		0.01 (0.02)	−0.17** (0.08)
High payment × 1920					−0.30*** (0.08)
High payment × 1950					−0.01 (0.08)
High payment × 1970					−0.06 (0.08)
High payment × 1996					0.07 (0.08)
High payment × 2005					0.05 (0.08)
High payment × Senegal					−0.04 (0.05)
High payment × Political killings					0.31*** (0.07)
High payment × Forced labor					−0.07 (0.08)
High payment × Journalist harassment					0.62*** (0.08)
High payment × Gender equality					0.24*** (0.08)
Observations	4,999	4,999	4,999	4,999	4,999
R^2	0.00	0.001	0.07	0.07	0.09
Adjusted R^2	0.00	0.001	0.07	0.07	0.09
Note:					*p<0.1; **p<0.05; ***p<0.01
					Reference: 2015, Argentina, Judicial independence

Table 27: Task characteristics and correct answer to factual questions

	(1)	(2)	(3)	(4)	(5)
Reference	−0.15*** (0.02)	−0.21*** (0.03)	0.47*** (0.06)	0.43*** (0.07)	0.73*** (0.09)
1920			−0.32*** (0.07)	−0.32*** (0.07)	−0.41*** (0.10)
1950			−0.44*** (0.07)	−0.44*** (0.07)	−0.43*** (0.10)
1970			−0.29*** (0.07)	−0.29*** (0.07)	−0.30*** (0.10)
1996			−0.18** (0.07)	−0.18** (0.07)	−0.37*** (0.10)
2005			0.05 (0.07)	0.06 (0.07)	−0.01 (0.10)
Senegal			0.01 (0.05)	0.01 (0.05)	0.09 (0.06)
Referenda permitted			−0.38*** (0.06)	−0.39*** (0.06)	−0.87*** (0.09)
Bicameral legislature			−0.37*** (0.06)	−0.37*** (0.06)	−0.75*** (0.08)
Suffrage level			−1.13*** (0.06)	−1.13*** (0.06)	−1.34*** (0.08)
High payment		0.13*** (0.04)		0.10** (0.04)	−0.50*** (0.13)
High payment × 1920					0.18 (0.15)
High payment × 1950					−0.03 (0.15)
High payment × 1970					0.02 (0.15)
High payment × 1996					0.40*** (0.15)
High payment × 2005					0.14 (0.15)
High payment × Senegal					−0.22** (0.09)
High payment × Referenda permitted					0.94*** (0.12)
High payment × Bicameral legislature					0.85*** (0.11)
High payment × Suffrage level					0.39*** (0.13)
Observations	3,831	3,831	3,831	3,831	3,831
Log Likelihood	−2,626.42	−2,621.54	−2,416.81	−2,414.09	−2,364.24
Akaike Inf. Crit.	5,254.84	5,247.08	4,853.62	4,850.19	4,768.47

J Circulated Pre-Analysis Plan

Pre-analysis plan: Experts vs. crowds pilot study

Michael Bernhard*
Michael Coppedge†
Adam Glynn‡
Staffan I. Lindberg§
Kyle L. Marquardt§
Daniel Pemstein¶
Constanza Sanhueza Petrarca§
Brigitte Seim‖
Steven Wilson§

March 29, 2017

Abstract

Scholars increasingly use expert-coded data in their statistical analyses of political phenomena. Benoit, Conway, Lauderdale, Laver & Mikhaylov (2016) argue that crowd-sourced data can substitute for expert-coded data. However, Benoit et al. base their claims on a very specific type of data - coding issue positions of party manifestos. This project examines the circumstances under which crowd-sourced data can substitute for expert-coded data. Specifically, it asks how characteristics of the *task*, *coder*, and *incentives* provided can condition the ability of crowds to produce data of similar quality to expert coders.

*University of Florida
†University of Notre Dame
‡Emory University
§V–Dem Institute, University of Gothenburg
¶North Dakota State University
‖University of North Carolina

Contents

1 Introduction

1.1 Core research question

Under what circumstances can crowd-sourced data substitute for expert-coded data? Specifically, how can characteristics of the *task*, *coder*, and *incentives* affect the ability of crowds to produce data of similar quality to that produced by expert coders?[1]

1.2 Motivation

Political science research increasingly uses expert indicators. Examples of large-scale expert-coding enterprises include the British Election Study Expert Survey, the Chapel Hill Expert Survey, the Electoral Integrity Project, Quality of Government, Transparency International and Varieties of Democracy (V–Dem). The virtues of expert surveys are manifold. In particular: "...they do not require that specific sources of information (e.g., roll call votes, opinion surveys of elite position, election manifestos or elite surveys) be accessible in all cases, and they are relatively inexpensive to administer. In addition, expert surveys allow the researcher to use a single format to ask a common set of questions. Whereas roll call votes, surveys and manifesto tabulations provide data that the researcher interprets after simplifying the data (e.g., using factor analysis or scaling techniques), expert surveys allow the researcher to design dimensions deductively" (Hooghe, Bakker, Brigevich, de Vries, Edwards, Marks, Rovny & Steenbergen 2010).

However, scholars have criticized expert surveys with regard to a) data and measurement and b) the data generation process. Benoit et al. (2016) compare crowd-sourced to expert coder-sourced data of party manifestos and show that crowd-sourced estimates of party policy positions can be used as substitutes for the trained coder estimates in this context. More generally, Benoit et al. (2016) argue that crowds can match or outperform experts on four dimensions:

1. **Reliability**: Expert surveys present challenges related to the subjectivity inherent in how experts interpret questions, which results from inter-personal differences in the properties that experts attribute to specific concepts. Scholars question the overall consistency of expert-coded measurements, since experts are subject to several biases (e.g., ideology, socialization, education).

2. **Validity**: Given the limitations of expert-coded data, scholars have questioned their validity. Benoit et al. (2016) find that a crowd-sourced approach is potentially more valid in certain contexts (e.g., coding text), as it can be implemented consistently for certain exercises across contexts.

3. **Cost efficiency**: The costs of running expert surveys are large, in particular in comparison to quick and low-cost methods like internet-based crowd-sourcing, which distributes small tasks to a large number of online workers in exchange for a small financial reward.

4. **Replicability**: Given the cost and often unclear sampling procedure for experts, expert-based datasets are generally not reproducible. They therefore do not satisfy common transparency and reproducibility standards for data generation.

In this study, we evaluate these arguments in a new context, comparing expert-coded data on key political indicators from the V–Dem data set to crowd-sourced data covering the same topics.

[1] In this project, we are agnostic about what constitutes high-quality data. We focus on identifying significant differences across expert-coded and crowd-sourced data.

Our study is thus a direct extension of Benoit et al. (2016) and others who explicitly or implicitly suggest that crowd-sourced data substitute for all forms of expert-coded data, including those which rely on substantial expertise in a field or case. Combining observational data with an experimental setup, we examine the scope conditions that determine when crowd-sourced data can substitute for expert-coded data. We ask how characteristics of the *task*, *coder*, and *incentives* provided affect the ability of crowds to produce data of similar quality to expert coders.

1.3 Definitions

1.3.1 What is an expert?

Benoit et al. (2016) and others explicitly or implicitly suggest that crowd-sourced data can be used as a substitute for expert-coded data. While acknowledging that this may be true for some types of tasks, we assert that there is a spectrum of tasks involved in coding enterprises, ranging from those that crowds can certainly complete to those which experts unambiguously must complete. As a deliberately trivial example, crowds could almost certainly compile a dataset of presidents' birthdates: the question is straightforward, the construct of "birth date" is fixed and known across different demographics, and the data are readily available. In contrast, crowds would be less able to code the degree to which social identities are relevant for political participation in a given country-year, since coding these data requires both knowledge of concepts (i.e., the definition of "social identities" as well as "political participation") and cases (i.e., relevant contextual information in the country and year that affected identity and politics), as well as several cognitive steps (i.e., a process to determine which identities are "relevant" to political participation).

As we think about this spectrum of tasks with its two extremes, we believe that there is a point at which crowds can no longer substitute for experts because either: 1) the task is too complex; or 2) it relies too heavily on information (either conceptual or case-related) that crowds do not have and cannot or will not obtain. In the following section we develop a theory of how these factors collectively determine that point on the spectrum.

Before we continue, it is necessary to clarify what constitutes an "expert." While there is no consensus in the literature, we argue that the following characteristics provide a baseline for a definition:

- An expert is "anyone with special knowledge about an uncertain quantity or event" (Morris 1977, pp. 679). By definition, they must invest a great deal of resources over a long time period to acquire their knowledge.
- With regard to politics and economics, an expert is "a professional who makes his or her livelihood by commenting or offering advice on political and economic trends of significance to the well-being of particular states, regional clusters of states, or the international systems a whole" (Tetlock 2005, pp. 239).
- Given that they hold a great deal of baseline knowledge regarding their tasks, and that they are practiced in discovering new information within their area of expertise, the quality of the information experts provide is most often solely dependent on their investment in the task.
- In the particular context of the V-Dem project, the country experts (CEs) who code tasks fulfill all of the above criteria. To select CEs, V-Dem project managers based at the University of Gothenburg collaborate with regional managers (established scholars who are broadly aware of experts in the countries in their region) and country coordinators (scholars aware of experts in their country) to develop a list of potential recruits. CEs are typically scholars

or professionals with deep knowledge of a country and of a particular area of politics in that country. V-Dem CEs generally hold a PhD, or in some developing nations an MA, a strong signal of investment in a topic. They also have specialized knowledge in at least one country and one of the conceptual sections of the V-Dem survey (e.g., civil society, party politics). A majority of experts are generally citizens or residents of the country being coded, meaning that they have first-hand experience in their country of interest. CEs receive monetary compensation for their service (US $25/survey).

1.3.2 What is a coder?

Many of the articles written about "experts" are not always discussing "experts" as we conceive of them above. They actually discuss a third, middle category of individuals that produce data. In this study, we refer to such individuals as "coders." We conceive of coders as follows:

- Coders often do not have an advanced degree in the subjects they are coding, and they do not typically have the information they need before they collect it to complete the coding task. Perhaps after they complete the coding task they could be considered an "expert" in the area, but not before. In other words, they engage in task-specific expertise acquisition.
- Coders are often well-educated, intelligent, and competent. Some coders are able to appropriately complete a coding task in the face of situations that might not have been anticipated ex ante, unlike a coder who follows a pre-specified coding protocol. But many coders are simply trained to apply a well-specified protocol to data (e.g., classifying Manifesto sentence fragments).
- As with experts, coders' success in completing a given task depends largely on how invested they are in the task.
- Coders often receive monetary compensation from the data project that employs them.
- A data-gathering enterprise employs coders to code information and includes it in the dataset. Many projects (e.g., Polity, the Comparative Manifesto Project) make use of such coders. In fact, V-Dem uses such "coders" to code "A" variables, or variables that relate to factual data and do not require deep expertise to understand, research, or apply a concept.

1.3.3 What is a crowd worker?

A final important definition regards those individuals whom scholars such as Benoit et al. (2016) consider members of "crowds." Such individuals are not a random selection of individuals across the world; they are individuals associated with an online survey system such as Amazon Mechanical Turk. There has been significant research on this group of individuals and we draw on this research to make several points:

- Mechanical Turk (MTurk) is an online labor market created by Amazon, and has become a popular tool among social scientists as a pool for survey and experimental data. As of 2014, the MTurk workforce is composed of more than 500,000 individuals from 190 countries. CrowdFlower provides a similar online market as well as other services, from its base in San

Fransisco, United States. We will use CrowdFlower to run a pilot, and MTurk for the main study.[2]

- Surveys consistently show that the MTurk workforce is dominated by workers residing in the United States and India, with less than a quarter of workers residing elsewhere (Paolacci, Chandler & Ipeirotis 2010, Ross, Irani, Silberman, Zaldivar & Tomlinson 2010).
- MTurk workers are diverse, but not necessarily representative of the populations they are drawn from, reflecting the fact that Internet users differ systematically from non-Internet users.
- Workers tend to be younger (about 30 years old), overeducated, underemployed, less religious, and more liberal than the general population (Berinsky, Huber & Lenz 2012, Paolacci, Chandler & Ipeirotis 2010, Shapiro, Chandler & Mueller 2013).
- Within the United States, Asians are overrepresented and Blacks and Hispanics are underrepresented, relative to the population as a whole (Berinsky, Huber & Lenz 2012).
- Less is known about workers' cognitive abilities, and this remains a fruitful area of investigation. Paolacci, Chandler & Ipeirotis (2010) found no difference between workers, undergraduates, and other Internet users on a self-report measure of numeracy that correlates highly with actual quantitative abilities. However, workers may learn more slowly and have more difficulty with complex tasks than university students, perhaps reflecting differences in age and education (Crump, McDonnell & Gureckis 2013).
- Their success in completing a given task depends largely on how invested they are in the task.
- Crowd workers receive a small monetary compensation for completing tasks online.
- While huge in principle, the pool of crowd workers available for coding tasks is limited in practice. While there are approximately half a million registered users on MTurk, there are only 10,000-20,000 on each of the primary alternative platforms (Peer et al. 2016).

2 Hypotheses

The overarching research question of this project is: Under which conditions and in which domains can crowd-sourced workers substitute for expert-coded data? In considering what might condition the degree of substitutability of crowd-sourced data for expert-coded data, we consider characteristics of the *task*, *coder*, and *incentives*. A delineation of the variables we consider, with related our hypotheses and tests, appears in Table 1.

3 Research design

3.1 Research participants

We will use the crowd-sourcing platform Amazon Mechanical Turk to recruit crowd coders.[3] Coders will self-select into the sample. We will determine the degree to which these coders differ from the

[2]The participants in MTurk and CrowdFlower tend to be distinct and the response rate at CrowdFlower tends to be faster. However, the users on CrowdFlower tend to be less attentive and less familiar with coding tasks than those on MTurk. Therefore, using CrowdFlower for the pilot allows us to focus on the most attentive respondent pool of crowd workers in the full study (Peer, Samat, Brandimarte & Acquisti 2016).

[3]After running a pilot on CrowdFlower.

Table 1: Variables and hypotheses

Category	Name	Description	Hypothesized Effect on Substitutability	Hypothesis Test
Task	Question Complexity	The complexity of the question language	−	Observational question trait: The length of the English language text of the V-Dem question plus the length of all of the choices for each question, both measured in number of characters (Section 3.5.1)
Task	Issue Complexity	The amount of nuance and complexity of issues considered in the task	−	Observational question trait: Average V-Dem expert coder confidence (Section 3.5.1)
Task	Issue Polarization	Polarizing issues are those that activate pre-conceived biases, emotional reactions, or personal experiences	−	Observational question trait: We purposively select a polarizing question about political killings (Section 3.5.2)
Task	Verifiable (vs. Perception)	Tasks that have a verifiable, correct answer (as opposed to relying on a subjective perception of a latent trait)	+	Observational question trait: We purposively select four fact-based V-Dem "A" variables of varying obscurity and dimensionality (Section 3.5.3)
Task	Recency	Whether the task pertains to recent events or political phenomena	+	Observational year trait: We purposively select six five-year spans (30 years total) between 1915 and 2015 (Section 3.5.4)
Task	Information Availability	The amount of information available to assist the participant in completing the task	+	Observational country-year trait: We create a measure based on the information available online and in published texts for each country-year (Section 3.5.5)
Task	Paired (vs. Likert)	Participants either rank paired cases or to place a single case on a Likert scale	+	Experiment: We randomly assign participants to complete paired or Likert tasks (Section 3.5.6)
Coder	Training	Whether the participant receives training in preparation for completing the task	+	Observational coder trait: We record whether the coder opens a training glossary we have available to all coders
Coder	Case Familiarity	Whether the task pertains to a case known by the participant	+	Observational coder trait: We record whether the coder coded a country of past or present long-term residence, and whether the coder is fluent in the official language(s) of the country
Coder	Baseline Knowledge	Whether the coder has a baseline level of knowledge relevant to the task	+	Observational coder trait: We record whether the coder follows politics, discusses politics, is interested in public affairs, and/or earned an advanced degree in political science
Coder	Education	Whether the coder is educated	+	Observational coder trait: We record the education level of the coder
Coder	Compliance	Whether the coder is able to pass "compliance" tests	+	Observational coder trait: We track the coder's ability to answer questions that gauge attention and basic competency (Section 3.6.5)
Coder	Time Investment	How much time is taken to complete the task	+	Observational coder trait: We record how long the coder took to complete each task
Incentives	Pay	The magnitude of the per-task payment received by the participant	+	Experiment: We randomly assign participants to a high/low payment condition

universe of MTurk coders (if at all) using demographic information provided either by MTurk or demographic questions on the pre- and post-survey questionnaires and then compared to past research documenting the characteristics of the MTurk user pool (see, for example, Peer et al. (2016) or Shapiro, Chandler & Mueller (2013)).

3.2 Sample size

In order to run the crowd-sourced data through the V–Dem measurement model (see Section 4), we aim to have approximately 40 observations per indicator-country-year-treatment group. We are collecting crowd-sourced data on 30 years (six 5-year blocks), four countries, nine V-Dem indicators, and four experimental treatment combinations. As a result, we require 172,800 observations ($30 \times 4 \times 9 \times 4 \times 40$). If each coder in the full experiment completes 60 tasks, 3,000 coders will result in 180,000 observations.

In the pilot, we will primarily collect data on one country (and respondents who complete their answers are asked to volunteer to code one question for an additional country to gauge their willingness to do more than one country), and we will only run the experiment on pay, meaning there will only be two experimental treatment combinations. This implies we require 21,600 observations ($30 \times 1 \times 9 \times 2 \times 40$), or 360 coders.

Given the high likelihood of attrition, we plan to recruit an extra 10% above the needed sample size in both the pilot and full study, for a total of 400 coders in the pilot and 3,300 coders in the full study.

3.3 Statistical power

Separate from the requirements of the measurement model, discussed in Section 3.2, we consider the sample size necessary to identify a statistically significant and substantively significant difference in the mean score crowds assign to a country-year-indicator compared to expert coders. While what constitutes a substantively significant effect size is not set, we assert that a difference in means larger than one-tenth of a standard deviation—based on expert coder responses—would be a substantively significant difference between crowds and experts. In Figure 1 and Table 2, we depict the relationship between the anticipated effect size and the number of crowd coders we need to complete a given country-year-question coding task to achieve statistical significance. Here, the effect size has been standardized to be in units of average expert case-level standard deviations, and we are assuming $\alpha = 0.05$ and $power = .80$. For example, we can see that a sample of 785 crowd coders would ensure that differences across experts and crowds greater than one-tenth a standard deviation would result in the rejection of the null hypothesis at the 95% confidence level 80% of the time. As our 2x2 factorial design has four experimental treatment groups, we need a sample of 3140 individuals (784×4). Even though each individual is expected to do 20 coding tasks, we conservatively estimate sample size without dividing by 20, in case there are unprecedented levels of attrition or there are strong learning or fatigue trends across the tasks.

3.4 Data collection instrument

Our instrument is a Qualtrics online survey. We base the tasks (or questions) in the instrument on indicators from the V–Dem expert questionnaire, though we alter them slightly due to the different format of this study.[4]

[4]Specifically, we specify that the coders are to code a specific country across specific years, and we also change the wording to the past tense.

Figure 1: Relationship between effect size and sample size

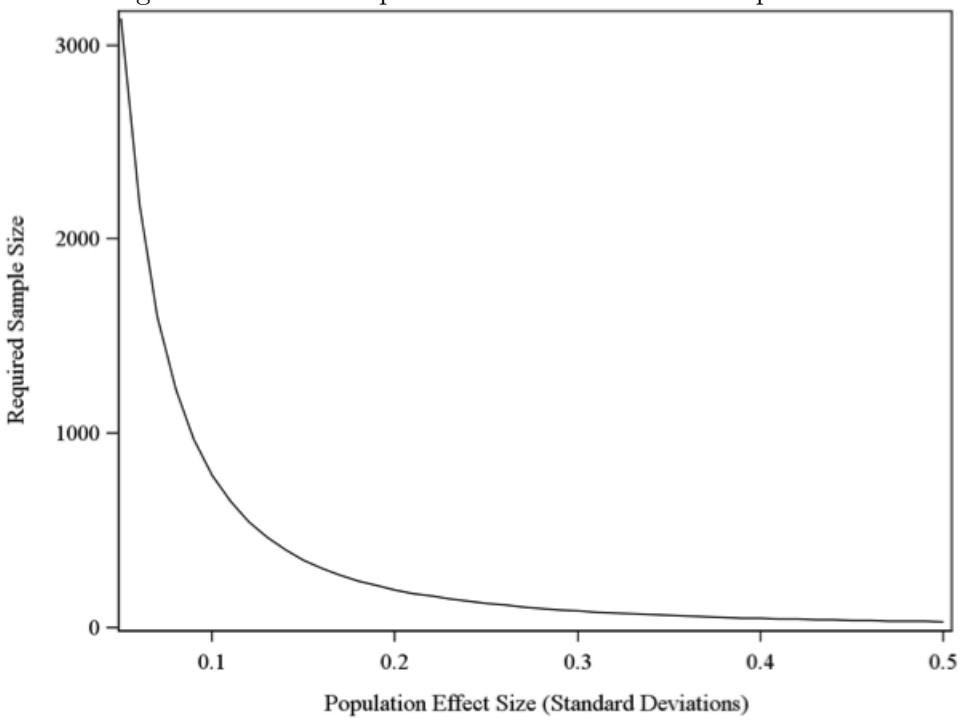


Table 2: Relationship between effect size and sample size

Effect Size (Standard Deviations)	**Sample Size**
0.05	3140
0.10	785
0.20	197
0.30	88
0.40	50
0.50	32

In both the pilot and the full study, out of a set of nine selected V-Dem indicators (see sections 3.5.1, 3.5.2 and 3.5.3 for explanation on selection), two indicators will be randomly selected for each crowd coder. In the full study, out of set of four countries (see Section 3.5.5 for explanation on selection), one country will be randomly selected for each crowd coder.[5] We chose to assign each coder to one country (as opposed to randomly selecting multiple countries) to ease the coding task and more closely replicate the expert-survey design. The coder will code these two variables for this one country for 30 years (see Section 3.5.4 for explanation on selection). To allow for some bridging (as exists in the expert-coded V-Dem data), we will give coders the option to code an additional country, for the second randomly selected indicator that they coded. If the coder takes this option, then we will obtain 90 country-year-indicator observations per coder. If not, we will obtain 60 country-year-indicator observations per coder.

3.5 Operationalizing hypotheses about task characteristics

3.5.1 Question and issue complexity

Four V-Dem indicators have been selected based on measurements of question and issue complexity. We developed measures for question and issue complexity (detailed below) and identified V-Dem indicators that fall into the lowest or highest quartile for each measure. We created a two-by-two table mapping indicators with low issue complexity and high question complexity, indicators with high issue complexity and low question complexity, indicators with high issue complexity and high question complexity, and indicators with low issue complexity and low question complexity. We removed indicators with low coverage across years from this table, and randomly selected an indicator from each cell in the table. The randomly selected indicators are as follows.[6]

- **Freedom from forced labor for men**: High issue and question complexity.
- **Harassment of journalists**: Low issue complexity and high question complexity.
- **Power distributed by gender**: High issue complexity and low question complexity.
- **High court independence**: Low issue and question complexity.

Question complexity: We measure question complexity across all V-Dem questions by taking the length of the English language text of the question plus the length of all of the descriptions of the responses categories for each question, both measured in number of characters. We combine the lengths of question text and response text because the two are highly negatively correlated with each other, likely because some questions load their complexity into the question text and thus have proportionately simpler responses, and vice versa. Accordingly, we posit that the total of the two best represents the overall textual complexity of the questions. In addition, textual length is highly positively correlated with the average length of time between a coder first reading a question and first entering any rating of any kind, providing some evidence that length is related to complexity.

Issue complexity: We measure issue complexity across all V-Dem questions by evaluating the average confidence (from 1 to 100) that coders self-assigned their ratings for each particular question. In theory, questions that address more nuanced issues should be answered with lower confidence, whether or not the question language is particularly complex. For example, "What is democracy?"

[5] In the pilot, all coders are assigned to one and the same country.
[6] See Appendix F for the full text of each question.

is a very simply worded question but asks about complex issues, and thus should be answered with lower confidence, all else equal. We have reason to believe this measure is orthogonal to the question complexity proxy as the two measures have essentially zero correlation. In addition, the average self-reported confidence for a question is highly correlated with the average confidence in the vignette codings for that same question. As these two quantities are measured independently (the coders answer vignettes separately and without indication of connection to another question elsewhere), this suggests stability in the confidence measure, as opposed to being merely noise.

3.5.2 Issue polarization

Some issues may evoke an emotional, biased, or simply highly volatile reaction when posed to coders. As stated above in Table 1, it may be more challenging for non-experts to set this kind of visceral reaction aside. To test this expectation, we consider this V-Dem indicator:

- **Freedom from political killings**: Purposively selected because we expect it to be highly polarizing, in that it is likely to activate pre-conceived biases, emotional reactions, or personal experiences.

The Likert scale response categories for this indicator include options where there is a moderate level of politically-motivated murder in the country for that year.[7] While the V-Dem coders are able to evaluate the level of political killings in a country and select these mid-level answer categories where appropriate, we anticipate that crowd coders will be less able to implicitly tolerate any level of political killing, and will be more likely to select the answer categories at the extremes of the spectrum, implicitly judging that either there are no political killings or there are far too many. We will use this indicator to test our expectation that crowd-sourced estimates of polarizing topics such as political killings have greater variance.

3.5.3 Verifiable (vs. perception)

To test our hypothesis that crowds can substitute for experts more effectively when completing verifiable, fact-based tasks, we purposively selected four factual V-Dem "A" indicators (i.e., variables that are not expert-coded in the V–Dem dataset, but rather pertain to objective responses) based on two criteria. First, we consider the dimensionality of the fact; whether only one concept is employed to search for and code the fact or whether finding the fact implicitly requires an understanding of several different concepts. For example, suffrage is a unidimensional concept, whereas the requirements for referendum are complex and multidimensional (e.g., different types of referenda, different levels, different approvals stages, etc.). Second, we consider how obscure the fact was: whether crowds (or experts) would know where to look to discover the fact or not. For example, the legal provisions for suffrage are located in legal documents of the country, whereas the suffrage level in practice is not always published in an official document. With this in mind, the variables we chose are as follows.[8]

- **Minimum voting age**: Easy to find and unidimensional.
- **Bicameral legislature**: Less easy to find and unidimensional.
- **Referenda permitted**: Easy to find and multidimensional.
- **Suffrage level**: Less easy to find and multidimensional.

[7] See Appendix F for the full text of this question.
[8] See Appendix F for the full text of the questions.

3.5.4 Recency

The V-Dem project covers most countries from 1900 to 2016. We have selected six five-year periods ranging from distant years to more recent years. We chose to use five year periods to reduce the workload on coders, while ensuring a relatively large number of observations per country-year-variable. Randomly selecting years from the entire period would risk losing coverage, while asking coders to code the entire time period would not be feasible. In selecting years, we considered three criteria. First, in line with our recency hypothesis, we deliberately selected some five-year periods that are recent and some that are far in the past, expecting to find lower substitutability going back in time. Second, we carefully selected times covering major political events to assess the degree to which coders static-code (i.e., do not change their codings over time). Finally, we deliberately chose some five-year periods prior to which the crowd coders are likely to have been born.

- 2011-2015: The most recent time period, and thus the period with which coders are most likely familiar.
- 2001-2005: A more distant time period in which crowds were nonetheless alive and aware.
- 1992-1996: Third wave of democracy, still in many coders' lifetimes.
- 1966-1970: Revolutions, coups, protests, far in past.
- 1946-1950: Post-war, far in past.
- 1916-1920: Post-war, very far in past.

3.5.5 Information availability

We quantify information availability at the country level and year level using custom measures detailed below.

Information availability by country: Our approach in developing a measure of information availability at the country level was to evaluate the "googleability" of different countries. To do this, we wrote software to measure the amount of data on each country in the world available on Wikipedia, as a proxy for this general notion of easily available information. Wikipedia has a number of standardized pages and hierarchies for organizing pages of a similar nature. In addition, it has a number of "hidden" (in the sense that they aren't linked directly from substantive pages, yet they are still open and available for public viewing) pages that provide meta-directories of such pages, to an arbitrary depth of hierarchy. Our approach was to select one of these meta-directories as appropriate to the country level unit of analysis and then recursively count and download the pages therein for each country.

We downloaded the full text of all pages on Wikipedia, three levels deep in the organizational tree for "Politics by Country." It was important to get more than just the top level pages because those by design are truncated at a certain length and then subdivided. For example, the "Politics of (Country Name)" pages are almost all of about the same length, even though there is vastly more information on certain countries. So all countries look the same at the first level in terms of length. On the other hand, too deep of a recursion into the hierarchy only increases skew. That is, past three levels, most countries have only a few brief pages, and all the additional data is simply adding to the countries that already have the highest counts. In addition, we also screen out pages Wikipedia has labeled "stubs" (i.e., placeholder pages with at most a sentence or two of

Table 3: Countries

Name Name	Political Characteristics	Polyarchy Index (CI)	Information Availabilty	Case Familiarity for Turkers	English Official Language?	V-Dem Data Released?
United States	One of the most stable advanced democracies, democratic since 1776	0.86 (0.81; 0.89)	10,889 (859 million)	High	Yes	Yes
Russia	Revolution, disintegration of the Soviet Union	0.06 (0.02; 0.08)	2,019 (138 million)	Intermediate	No	Yes
Singapore	Authoritarian but low corruption, repression, conflict	N/A	638 (53 million)	Intermediate	Yes	No
Benin	Recent colonial past; country changed names	0.46 (0.35; 0.56)	472 (27 million)	Low	No	Yes
Argentina	History of both dictatorship and democracy	0.81 (0.76; 0.86)	1,143 (71 million)	Intermediate	No	Yes
Senegal	Relatively stable and consolidated democracy	0.74 (0.69; 0.79)	548 (39 million)	Low	No	Yes

descriptive content) so as to ensure measurement of actual usable information. We downloaded the full contents of 187,319 Wikipedia pages, and aggregated the data into counts of the total numbers of characters on the pages associated with every country in the world. This provides our measure of the amount of information generally available about each country.

As MTurk users are typically native English-speakers, we also consider the official language of the country as a proxy indicator for whether easily-accessed online material will be readable for the average crowd coder.

Finally, as another—highly V-Dem specific—operationalization of information availability, we consider whether the V-Dem data have been publicly released, deliberately including one country in the study for which data have not yet been made available on the V-Dem website.

With this in mind, we selected four countries for the full experiment—United States, Russia, Singapore, Benin[9]—and two for the pilot—Argentina (main country) and Senegal (optional).

As it can be seen in Table 3, the selected countries vary in terms of their political characteristics that might affect information availability, polyarchy index, information availability score, likely case familiarity, and official languages.

Information availability by year: In order to quantify the variation over time in the amount of information available, we made use of Google Ngram database, which contains the number of instances of every word in every book scanned and analyzed by the Google Books project. In terms of time frame, the project has scanned books back to the 16th century, and so can serve as a reasonable proxy for the amount of published material available in general, year by year. We downloaded the summary file for the project, which contains, by publication year, the total number of scanned volumes, pages, and words. The latter gives us a relative measure of the amount of information available in a general sense as we move backwards in time.

We also want to measure how much information is available over time about specific countries over time, as we expect that certain countries will have more available information than others,

[9]We may swap Benin for Kuwait, another country for which V-Dem data have not been publicly released, depending on the results of the pilot and whether coders seem to be using the V-Dem data as a resource.

with variance over different time periods. We also leveraged the Google Ngram database for this effort. The database tracks n-grams, which are the occurrence of words next to each other. So "one word" is a 2-gram of the words "one" and "word". This is essential because about a quarter of countries in the world have names longer than one word. We wrote software to download the appropriate data files for the n-grams that capture the names of all the countries in the world, some 500 gigabytes of data. We then searched the appropriate files for occurrences of the full names of countries. The result of this is a count of the number of times each country is mentioned in any published text, for every year going back to 1900 (our chosen start year). This serves as a measure for how much information is generally available about every country on a temporal basis.

3.5.6 Likert Scales vs. Paired Comparisons

NOTE: *Given the already complex nature of this project, and the coding burden involved in creating paired tasks, we are considering separating the likert vs. paired experiment in a separate study. We plan to discuss further after the pilot and MPSA.*

We will randomly assign coders to one of two tasks. The first set of coders will complete tasks identical to those assigned to V-Dem experts. Specifically, they will code country-years for our chosen questions on the Likert scales provided by the question response categories. These tasks require coders to evaluate country-years according to the given scale and then to place cases on an ordinal scale, forcing them to make fine-grained and comparable judgments across cases.

The second type of task requires coders to rank order pairs of cases on a given question. Specifically, we will present coders with a question—and sets of Likert-scale response categories,[10] since we know that questions often pack a lot of information into their response categories—and two country-years; we will ask coders to indicate which country-year ranks higher on the given scale, and will provide the option to say that the two cases are equal. To reduce the complexity of the task, we will limit each coder to comparisons between two countries, and cases within countries will be drawn from the same set that we present to Likert raters. While we could use adaptive algorithms to optimize the presentation of paired comparisons, we will satisfice in order to reduce implementation complexity. Specifically, using expert-coded V-Dem data, we will calculate the distribution of pairwise differences in point estimates across cases within each potential country-year pairing for each indicator. Then, we will split this distribution into quartiles. Next, to ensure that coders provide disproportionately more ratings for "closer" cases, we will adjust pair sampling probabilities as follows:

1. If the proportion of pairs in the lowest quartile—the toughest cases to compare—is greater than 0.5, simply draw pairings at random.

2. If the proportion of pairs in the lowest quartile is less than 0.5, oversample from the lowest quartile so as to set that proportion at 0.5, proportionally reducing the probability of sampling from the upper 3 quartiles.

Once we have determined the sample of pairs to present to the coders, the coder will see one pair at a time, holding the indicator constant within and across the set of pairs that they see. For example, the coder might be asked: "Considering Argentina in 2012 and Chile in 2014, which had greater political killings?" The response categories will simply be "Argentina in 2012" and "Chile in 2014" and "These two cases had the same level of political killings." Across the sample of coders

[10] In other words, the full V-Dem question text.

and the number of observations, this approach will generate a dataset of rankings for an expansive dataset of country-year pairs.

Once we obtain the crowd-sourced rankings, we will use contest scoring techniques (Schnakenberg & Penn 2014, Coppedge, Glynn, Lindberg, Pemstein & Seim 2015) to estimate latent scales from the participant-provided rankings and compare country-year scores on the crowd-sourced scales to those produced by applying IRT models to expert-produced Likert scores.

We expect that latent scales extracted from the paired comparisons task will better approximate expert-based estimates than will estimates drawn from crowd-sourced Likert ratings. Fundamentally, we expect crowd-sourced coders to have trouble matching experts in fine-grained, complex, tasks, like Likert scale placement, requiring consistency in scale application over space and time. On the other hand, crowd-coders may be able to carry out the comparatively simpler task of rank-ordering with reasonable accuracy. Moreover, given sufficient crowd-coders of sufficient accuracy, it is theoretically possible to extract valid measures from error-prone paired comparisons (Honaker, Berkman, Ojeda & Plutzer 2013). An earlier experiment (Coppedge et al. 2015) leads us to expect that the paired comparison task may provide the best way to leverage the power of crowds to replace the sorts of expert judgments that we examine here.

3.6 Operationalizing hypotheses about coder characteristics

3.6.1 Training

To operationalize training, coders will have access to a "training manual" that compiles all of the clarifications and definitions provided to V-Dem expert coders in an indexed and searchable document. We will track whether or not coders open this document. We expect to see greater substitutability across crowds and experts among those crowd coders who open this document.

3.6.2 Case familiarity

Since "expertise" is closely related to experience and familiarity, in the pre- and post-survey questionnaires, we will document whether the coder coded a country of past or present long-term residence and whether the coder is fluent in the official language(s) of the country. See sections D and G for exact question wording.

3.6.3 Baseline knowledge

Similarly, "expertise" is also related to the level of formal and informal exposure to the topic, in the pre- and post-survey questionnaires, we will document whether the coder follows politics, discusses politics, is interested in public affairs, or earned an advanced degree in political science. We expect these variables to be positively correlated with substitutability. See sections D and G for exact question wording.

3.6.4 Education

Since higher education may provide general expertise and research skills, we will document the education level of the coder, expecting to observe that more education leads to greater substitutability for expert coders. See sections D and G for exact question wording.

Table 4: Vignettes Selected as Gold Standard Questions

Vignette	% Expert Answers within Expected Answer Categories	n
v2psparban	91.2%	194
v2eldonate	86.7%	135
v2pepwrort	84.9%	251
v2juaccnt	83.4%	163

3.6.5 Compliance

We will ask the crowd coders to complete two sets of tasks that are designed to weed out "non-compliers," or those who are simply not attentive or competent.

The first kind of compliance questions uses anchoring vignettes written for V-Dem to measure basic competency. During the administration of the V-Dem surveys, we ask experts to code hypothetical country vignettes, which are questions about hypothetical cases written in order to evaluate the coder's threshold between different answer categories on existing non-hypothetical questions. As such, each question has several corresponding vignettes, designed to straddle the threshold between two of the original question's Likert-scale answer categories. In an ideal sense, perfectly written vignettes coded by perfectly performing coders should never have any submitted answers other than the two choices designed to be tested. Therefore, the percentage of answers submitted other than those two choices for a given vignette can be used as a measure of how well-performing a particular vignette is. By extension, this makes verifiably well-performing vignettes a good instrument for measuring coder competence; they require no country-specific or time-specific knowledge, but do require an ability to understand the political science concepts being examined in the question. We have therefore selected four of the best performing vignettes to serve as "gold standard" questions. They are listed in Table 4 below. We will create an additive index of these gold standard questions (number of vignettes answered correctly) as a covariate in analyses that captures coder competence.

The second kind of compliance task is a "screener," a task primarily designed to weed out fraudulent (automated) coders. It will also serve as a training that all crowd coders will receive. In this task, a hypothetical country vignette (which we used as anchoring vignettes within V-Dem) will appear with instructions on how it should be coded. For example, the coder will read a vignette and then be directed to assign it a four. There are two screener vignette tasks for each of the nine indicators (representing the extremes of the Likert scale), and one will be randomly selected for the coder to complete before they complete the 30 years of coding for that indicator.[11] See Section F for the text of these screener questions.

3.6.6 Time investment

We will collect data about the time to complete each task and move on to the next one. While we cannot directly measure "attention" or "effort," the time spent on the task is a proxy indicator for the time investment, which we believe is positively related to the substitutability between crowd-sourced and expert-coded data.

[11] Ideally, we would randomly intersperse additional screeners between five-year periods as both a reliability check and to refocus coders. However, this would likely be onerous for coders and disrupt the flow of coding.

3.7 Incentives

Coders will be randomly assigned to one of two treatment conditions - low pay ($0.12 per task) and high pay ($0.24 per task) - with equal probability (0.5). We expect better-paid coders to outperform lower paid coders. We will strengthen the nature of this treatment by displaying a "Payment Earned so Far" tracker at the bottom of the screen to the coder throughout the survey.

We also expect higher levels of attrition from the low-payment sample. However, we do not have any information that would allow us to assign numeric values to the rate of attrition; indeed, gathering information about this rate is one of the goals of this pilot.

3.8 Data collection and processing

We will field the pilot in late March 2017 and field the full study in mid-April the same year. Data will be stored and processed on CrowdFlower and Amazon Mechanical Turk servers according to their confidentiality procedures. We will then transfer data to a secure server at University of Gothenburg to which only three co-authors working there have access, and randomly assign coders new IDs to replace their CrowdFlower/Mturk coder IDs. IP addresses will be replaced by country of location and deleted. All data will be deidentified prior to storage on a SVN server available to the rest of the team, and prior to analysis. The deidentified dataset will be made publicly available following publication of an article related to this project. Until that point, data will be collectively owned by the authors and not circulated further without unanimous consent.

3.9 Ethics

The *Regionala Etikprövningsnämnden i Göteborg* (a regional ethics review board to which all projects at the University of Gothenburg must send their applications) has approved the study. It has been deemed exempt from human subjects review at UNC and NDSU. See the consent language in Appendix C.

4 Empirical analysis

4.1 Independent Variables

We operationalize our research design in terms of the following independent variables. Bold variables provide information about research questions, while non-bold variables are pure controls.

1. **Pay treatment** (t_r^{pay}): A coder-level indicator for high pay.
2. **Task treatment** (t_r^{task}): A coder-level indicator for paired task.
3. Question-level covariates ($\mathbf{s}_q$):
   (a) **Question complexity**: The natural logarithm of the continuous measure we describe in section 3.5.1.
   (b) **Issue complexity**: The continuous measure described in section 3.5.1.
   (c) **Issue polarization**: A dummy variable for the freedom from political killings question.
   (d) **Question type**: A dummy variable for verifiable, as opposed to subjective, questions.
4. Country-level covariates ($\mathbf{t}_c$):

(a) **Wikipedia words**: The natural logarithm of the number of Wikipedia words dedicated to the country, as we describe in section 3.5.5.

(b) **Official English**: A dummy indicating whether or not English is the country's official language.

(c) **V-Dem public**: A dummy indicating whether or not V-Dem data are publicly available for the country.

5. Year-level covariates ($\mathbf{u}_y$):

   (a) **Recency**: A categorical variable for the six time-periods described in section 3.5.4.

6. Country-year-level covariates ($\mathbf{v}_{cy}$):

   (a) **Ngram mentions**: The logarithm of the number of times the country was mentioned in the given year in the Google Ngram database, as we describe in section 3.5.5.

7. Coder-level covariates ($\mathbf{w}_r$):

   (a) **Training**: A dummy variable indicating whether or not the coder opened the provided manual.

   (b) **Follows politics**: A dummy variable for whether or not the coder follows politics.

   (c) **Discusses politics**: An ordinal variable for whether how much the coder discusses politics.

   (d) **Interested in public affairs**: An ordinal variable for how interested the coder is in public affairs.

   (e) **Political science education**: A dummy variable representing a coder who focused on political science at the bachelors level or higher.

   (f) **Education**: A categorical variable representing 1) high school or less education and 2) Graduate-level education (reference is college-level education, either at present or completed).

   (g) **Gold standard score**: An additive index of vignettes answered correctly, ranging from zero to four.

   (h) **Screener**: A dummy variable indicating if the coder answered all screener question correctly.

   (i) V–Dem use: An indicator of whether or not coder reports using V–Dem data in coding.[12]

   (j) Female: Indicator for female coders.

   (k) Age: Mean-centered natural logarithm of coder age.

8. Coder-case-question[13] level covariates $\mathbf{x}_{cyqr}$

   (a) **Country residence**: A dummy for countries where the coder has been a long-term resident.

   (b) **Language skill**: A dummy indicating whether or not the coder is fluent in the coded country's official language(s).

   (c) Time spent: A continuous variable representing log-scale time spent on variable.

[12] We can validate these responses with IP checks.

[13] Some of these are coder-country, others coder-question.

4.2 Likert Response Analysis

Here we compare crowd-sourced averages to expert averages, focusing only on crowd coders for whom $t_r^{\text{task}} = 0$. The unit of observation is therefore the country-year-question *cyq*, disaggregated, by $t = t_r^{\text{pay}}$. The response variable, d^a_{cyqt}, is 1) the absolute difference between the average crowd response and the average expert response for Likert scale questions (from the V–Dem v7 data set) and 2) the percentage of crowd coders who correctly code a factual question (for interval-level factual data, we will treat responses ± 1 on a 0-100 scale as being correct).[14] Given the difference in scale, we will conduct separate analyses for factual and Likert-scale questions.

4.2.1 Average Response

We will fit a series of regression models. First, we can fit the basic model

$$d^a_{cyqt} = \alpha, \tag{1}$$

where our null hypothesis is $\alpha = 0$, or more conservatively that $\alpha < c$, where c is one tenth of one standard deviation in expert responses on the given scale, as described above.

Next, we can test for a basic treatment effect, fitting the model

$$d^a_{cyqt} = \alpha + \beta t \tag{2}$$

where our null hypothesis is $\beta = 0$. We expect a negative relationship between the treatment indicator and the outcome.

We can then examine observation-level covariates, allowing us to test task effects.[15] Specifically, we can conduct these tests by fitting the models

$$\begin{aligned} d^a_{cyqt} &= \alpha + \boldsymbol{\gamma}\mathbf{z}_{cyq} \\ d^a_{cyqt} &= \alpha + \beta t + \boldsymbol{\gamma}\mathbf{z}_{cyq} \end{aligned} \tag{3}$$

where $\mathbf{z}_{cyq} = [\mathbf{s}_q \ \mathbf{t}_c \ \mathbf{u}_y \ \mathbf{v}_{cy}]$, and each element of $\boldsymbol{\gamma}$ tests the effect of a case-level covariate, and our null hypotheses are that $\boldsymbol{\gamma} = 0$. Refer to table 1 for hypothesized effect directions. We will also fit an interactive model,

$$d^a_{cyqt} = \alpha + \beta t + \boldsymbol{\gamma}\mathbf{z}_{cyq} + \boldsymbol{\delta}(t \cdot \mathbf{z}_{cyq}), \tag{4}$$

where each element of $\boldsymbol{\delta}$ captures the interaction between the payment treatment and given task characteristic. While we do not have a priori expectations about how the treatment might interact with task-level covariates, we will fit this model for descriptive purposes, and to test the core task hypothesis using a flexible model. We may fit higher order interactions, or consider interactions between task characteristics from a purely exploratory perspective.[16]

The averages used to compute each d^a_{cyqt} are based on small samples, especially on the expert side. Therefore, we will use parametric bootstrapping, computing average crowd and expert responses, based on sample-sized draws, with replacement. Specifically, each country-question-year-treatment cell is represented by n^{crowd}_{cyqt} crowd and n^{expert}_{cyqt} expert ratings. We will resample 1000 times, randomly drawing a bootstrap sample of size $n^{\text{crowd}}_{cyqt}(n^{\text{expert}}_{cyqt})$, with replacement, from the set of crowd(expert)-raters who provided a rating for cell *cyqt*, calculating averages for resampled

[14] In other words, we calculate country-year-question averages/percentages correct within pay conditions, yielding two scores per country-year-question.

[15] Excepting the paired/likert treatment, of course.

[16] In general, findings from the pilot may guide our design strategy in the final study.

crowd and expert rating sets, and computing d^a_{cyqt} for each resampling draw. We will fit each of the above models using this bootstrapping procedure.

In a second approach, we will bootstrap crowd averages, as described above, but replace resampled draws from expert averages with random draws from the posterior ordinal response distributions produced by the V-Dem measurement model. The measurement model factors in expert reliability and differential item functioning, and thus represents a harder test for crowd-sourced data.

4.2.2 Individual Responses

Here our response variable is the absolute difference between the response for coder r on country-year-item cyq and the average expert response.[17] Call this d_{cyqrt}. Adding coder-level covariates, we will refit equation 3, which becomes

$$\begin{aligned} d^a_{cyqrt} &= \alpha + \boldsymbol{\gamma}\mathbf{z}_{cyq} + \boldsymbol{\lambda}\mathbf{w}_r + \boldsymbol{\xi}\mathbf{x}_{cr} \\ d^a_{cyqrt} &= \alpha + \beta t + \boldsymbol{\gamma}\mathbf{z}_{cyq} + \boldsymbol{\lambda}\mathbf{w}_r + \boldsymbol{\xi}\mathbf{x}_{cr}. \end{aligned} \tag{5}$$

Interpreting α is fraught at the individual level, since the average absolute difference between crowd and expert averages (α in section 4.2.1) is not the same as the average absolute differences between individual crowd sourced estimates and expert averages (α here), so we will treat the α parameters in these regressions as tangential to our core questions. Nonetheless, β and $\boldsymbol{\gamma}$ provide alternative tests of our pay and task characteristic hypothesis, controlling for individual characteristics. Furthermore, $\boldsymbol{\lambda}$ and $\boldsymbol{\xi}$ provide information about how individual characteristics of crowd-sourced raters affect their propensity to generate scores close to expert-based point estimates. Again, table 1 provides hypothesized effect directions.

We will use bootstrapping to account for small expert response samples. The procedure is identical to that in section 4.2.1 except that we need only resample expert responses when calculating averages. Again, we will use random draws from the ordinal posteriors produced by the V-Dem model as an alternative to expert coder averages.

4.2.3 Variation in Response

For Likert response variables We will run the models described in section 4.2.1 and 4.2.2 with an additional dependent variable, namely the difference in standard deviations between experts and crowds d^v_{cyqt}. Specifically, we will subtract crowd standard deviations from expert deviations; thus negative scores will indicate that crowds are providing more variable responses than experts. Given this coding, table 1 continues to describe our hypothesized directions for the effects of the pay treatment and included covariates.

4.2.4 IRT Model Fits

Finally, we will fit ordinal IRT models to the crowd data, just as V-Dem does with expert data. We will use the same prior specification strategy that V-Dem uses for expert-based data. We will fit two models, one to the low-pay and one to the high-pay group. Because the crowd and expert IRT fits will produce non-comparable scales, we will evaluate ranks across models. First, we will examine rank correlations between crowd and expert-based model estimates, across treatment conditions. We expect higher rank correlations between crowd and expert-based estimates in the high pay condition.

[17] We will list-wise delete cases with missing covariates.

We will also fit logit models analogous to the linear models described in section 4.2.1. For each dyad-treatment, we will code the dependent variable as 1 if the ranking produced by the crowd IRT model matches that produced by the expert IRT model, and zero otherwise. Because question characteristics vary across pairs, we need to create dyadic versions of $\mathbf{t}_c, \mathbf{u}_y$, and, $\mathbf{v}_{cy}$. Specifically, we will average continuous variables, convert *Official English* and *V-Dem public* into ordinal—neither, one, both—variables, and treat recency as a series of dummy variables for each period, where up to two periods might both be set to one for a given pair. We will use method of composition to account for posterior uncertainty in the measurement model estimates. If computationally necessary, we will randomly subsample no fewer than 10,000 dyads.

4.3 Paired Responses

We need not conduct separate analyses of the verifiable and subjective questions for these analyses.[18] The text below refers to the subjective questions. When considering verifiable questions, we will use true answers rather than measurement model estimates and we can omit the method of composition since these scores are not measure with error.

4.3.1 Average Response

Here our observation level is paired country-years, by item. Our dependent variable is the proportion of crowd-sourced responses that match the rank-ordering produced by the measurement model for the paired country-years. Using GLM to deal with the fact that our response variable is proportions, we will fit the models described by section 4.2.1. Because question characteristics vary across pairs, we need to create dyadic versions of $\mathbf{t}_c, \mathbf{u}_y$, and, $\mathbf{v}_{cy}$. Specifically, we will average continuous variables, convert *Official English* and *V-Dem public* into ordinal—neither, one, both—variables, and treat recency as a series of dummy variables for each period, where up to two periods might both be set to one for a given pair. We will use method of composition to account for posterior uncertainty in the measurement model estimates.

4.3.2 Individual Response

Here we replicate section 4.2.2, using logit instead of OLS. The observation level is rater-diad and the response variable is a dummy indicating whether or not the provided rank ordering matches the MM ranking. Again, we can use method of composition to account for uncertainty in measurement model estimates. We now need to adjust coder-country level variables in $\mathbf{x}_{cr}$, transforming *Country residence* and *Language skill* into neither/one/both ordinal variables.

4.3.3 Comparing Model Based Ranks

We start by splitting the paired data by payment treatment. Then, once for each pay group, we will fit a contest scoring model (Schnakenberg & Penn 2014) to the paired rank data, bootstrapping the procedure 1000 times to produce estimates of uncertainty. For each question this will produce a rank ordering of the cases considered by crowd coders. Then, once for each of the 1000 bootstrapped scoring model fits, we will compare the ordering of pairs produced by crowd coders to the ranks produced by the V-Dem measurement model, randomly selecting a posterior draw to compare with the bootstrap draw at each iteration.[19] This will produce our dependent variable, y_{ddqt}, which is a

[18] We must treat the binary variables as ordinal; that is more or less bicameral and referenda more or less permitted.

[19] We may have to sample a subset of pairs to make this tractable. Again, we will subsample no fewer than 10,000 dyads.

distribution over a dyad-question-treatment level dummy variable, indicating whether or not the dyad is ranked "correctly," based on the V-Dem measurement model. Our analysis of these data will follow the same pattern as section 4.3.1, except the model is a simple logit, and we will use the method of composition to reflect estimation uncertainty.

4.4 Likert vs Paired

We expect crowds to perform better at paired comparisons than Likert scale responses. To evaluate this prediction we will compare the rank orderings produced by IRT model fits to crowd-based likert responses (see section 4.2.4) to those produced when we apply contest scoring models to paried comparisons (section 4.3.3). Here we will compare rank correlations with expert-based IRT-based estimates, using the procedure described in section 4.2.4.

4.5 Item Non-Response and Attrition

We will fit count models of item non-response, following the basic structure of sections 4.2.1 and 4.2.2. In other words we can fit those models described in section 4.2.1, but our response variable will be aggregate (at the country-year-item level for Likert responses and at the dyad-item level for paired responses) and individual non-response counts. We expect the same relationship between variables of interest and the outcome variables.

5 Research team

The research team consists of the authors of this pre-analysis plan. All co-authors are Principal Investigators, Project Managers, and Post-Doctoral Research Fellows of V-Dem.

6 Deliverables

There will be several products from this study. We will present some of the results from the pilot at MPSA 2017.[20] The results of the full study will be presented at V-Dem 2017, EPSA 2017, and APSA 2017, and will be prepared for journal publication.

7 Budget

For the pilot, we plan to recruit 400 individuals, half of which will be randomly assigned to the high-pay condition ($0.24) and half of which will be assigned to the low-pay condition ($0.12). Each participant will complete 66 tasks at this pay rate, with an option to complete an additional 31. All participants, regardless of pay treatment status, will receive $0.10 per answered question on the pre- and post-survey questionnaires. This results in a budget of $7,840 in participant payments and a 20% CrowdFlower transaction fee of $1,568, for a total of $9,408.

In the full study, we plan to recruit 3,300 individuals. The pay rates and number of tasks are otherwise the same, for an anticipated budget in the full study of $63,888 in participant payments and $12,778 in a 20% transaction fee, for a total of $76,666.

[20] We will consider whether these results in and of themselves are suitable for publication, either in combination with the results of the full study or as a stand-alone set of results. However, the primary goal of the pilot is to refine our approach, not academic publication.

A Notes for Coder

- Section headers should be displayed to respondents, with the following modifications:
  - Pre-Survey Questionnaire: "Pre-Survey Questionnaire"
  - Gold Standard Questions: "Hypothetical Questions"
  - Main Coding Task: *See note at the beginning of Section F.*
  - Post-Survey Questionnaire: "Post-Survey Questionnaire"
- Paging design: Display one question per page unless specified otherwise, and display the introduction to each section on its own screen.
- Page-specific or question-specific notes for coding are displayed between brackets [ ]. These should not be displayed to the respondent.
- Variable names are displayed between parentheses ( ). These should not be displayed to the respondent.
- Questions are specified after **"Question"**. **"Question"** should not be displayed to the respondent.
- Responses are specified after: **"Response"** **"Response"** should not be displayed to the respondent.
- Codes for responses are specified before ":" next to the responses. These should not be displayed to the respondent.
- Items are single-punch radio buttons, unless specified otherwise.
- Nonresponse prompting: Unless specified otherwise, for every question that the respondent fails to respond, please display the following text in a pop-up.
  - Text: "There is 1 unanswered question on this page. Would you like to continue? You will not be paid for questions you do not answer." Buttons: "Continue Without Answering" and "Answer the Question"
    - **NOTE:** We will not pay for unanswered questions.
- Recording nonresponse prompting: Please record whether the nonresponse pop up is shown.
- Missing data and nonresponse: Please assign a numeric code to all variables as follows:
  - 88 - Don't know: Note that this option should not be displayed to the respondent unless otherwise stated below.
  - 99 - No answer: The question was displayed to the respondent but the respondent clicked "Next" without answering the question and then clicked "Continue Without Answering" on the nonresponse pop up.
  - 98 - Other/error: Data is missing due to technical problem
  - 97 - Breakoff: The questionnaire was terminated before reaching this question
  - 96 - Not applicable: Note that this option should be displayed to the respondent for certain questions, as noted below.

- Timing: Record all items' timing
- Time: Record local time when survey was completed.
- Errors: Please display error text as follows:
  - "Sorry, there was an unexpected error in processing this question."

B Recruitment

[FOR DISPLAY IN AMAZON MECHANICAL TURK/E-SERVICE]
[START SCREEN]
[DISPLAY ONLY]
Dear Contributor, We are a group of researchers at the University of Gothenburg and we are interested in measuring some indicators of democracy across the world. We would like to invite you to participate in a short survey, which will take 30-60 minutes of your time. If you choose to participate in the survey, renumeration will be offered up to a maximum of USD 25.42.
Please click this link to begin the survey: [DISPLAY LINK TO SURVEY]
Regards,
Professor Staffan I. Lindberg
xlista@gu.se

C Consent

[IN QUALTRICS]
[START SCREEN]
Thank you for taking the time to participate in this survey. Below is a consent form that describes the nature of the experiment, what to expect, and what we will do with the data you provide. Please take the time to read this over before deciding to participate in this study.

University of Gothenburg
Consent to Act as a Research Subject
Democracy Study

1. **Will you be compensated for participating in this study?**
   - Yes. We will ask you questions about politics in different contexts, both real and hypothetical. You will receive [RANDOMIZED PAY RATE] for each question that you answer. In addition to these questions about politics, we will ask you a few questions about your background, and you will be paid $0.10 for each of these questions. Finally, the very last question of the survey will ask you for your feedback, and you will be paid $0.24 to answer this question. Once you agree to participate and start the survey, you will have 12 hours to complete it. During this time you can close and open the survey as many times as you need. If you do not complete the survey within this time frame, you will not receive any payment.
2. **Who is conducting the study, why have you been asked to participate, how were you selected, and what is the approximate number of participants in the study?**

- A team of researchers based at the University of Gothenburg is conducting a research study to measure several indicators of politics around the world. You have been asked to participate in this study because you are a registered user on this survey platform. There will be approximately 5,000 participants in this study in total.

3. **Why is this study being done?**

   - The purpose of this study is to gain a better understanding of how to measure elements of politics around the world.

4. **What will happen to you in this study and which procedures are standard of care and which are experimental?**

   - If you agree to be in this study, the following will happen to you:
     (a) You will be asked to evaluate certain aspects of a country (e.g., elections, parties, civil society)
     (b) You will be asked to answer a few questions about yourself.
     (c) You do not have to answer any question you do not wish to answer.
     (d) Your participation in this study will take 30-60 minutes.

5. **What risks are associated with this study?**

   - Responses with identifying information will only be shared among collaborators, and will be stripped of identifying information prior to circulation. Upon publishing the results of the research, the anonymized dataset will be released publicly on the authors' websites
   - Participation in this study may involve some added risks or discomforts. These include the following:
     (a) There is a risk that the study may induce boredom or fatigue. You may opt out of the study at any time, which will minimize this risk.
     (b) Because this is a research study, there may also be some unknown risks that are currently unforeseeable.

6. **What are the alternatives to participating in this study?**

   - The alternative to participation in this study is to not participate.

7. **What benefits can be reasonably expected?**

   - There may or may not be any direct benefit to you from participating this study. You may find that information included in the study is useful to you. You will also be helping the team of investigators learn more about how to measure democracy across the world and how political data can be generated. You will receive monetary compensation for your participation.

8. **Can you choose to not participate or withdraw from the study without penalty?**

   - Participation in this research is entirely voluntary. If you decide that you no longer wish to continue in this study, there will be no further requirements of you. You may terminate your participation at any time. Please note that if you do not complete the survey, you will not be compensated (even for the questions you may have answered at that point).

9. **Can you be withdrawn from the study without your consent?**
    - No.
10. **Are there any costs associated with participating in this study?**
    - There will be no cost to you for participating in this study.
11. **Who can you contact if you have questions?**
    - If you have other questions beyond those explained in this information sheet, you may reach the contact researcher, Professor Staffan I. Lindberg, via email at xlista@gu.se.

1. Consent (consent)
    - **Question** You have read the above information. You are at least 18 years old. By selecting "I agree to participate," you consent to participate in the study.
    - **Response**
        (a) I agree to participate.
        (b) No, I do not wish to participate.

D Pre-Survey Questionnaire

First, we have a few questions about you. Remember that you will be paid $0.10 for each answer that you provide in this section.

1. Year of birth (v2zzborn)
    - **Question** In what year were you born?
    - **Response** Year [DROP DOWN MENU, STARTING AT 2000 AND GOING BACKWARDS IN TIME]
2. Gender (v2zzgender)
    - **Question** What is your gender?
    - **Response**
        (a) Male
        (b) Female
        (c) Other/Prefer not to answer
3. Time in Argentina (ctrArg)
    - **Question** Have you ever lived in Argentina for a period greater than three months?
    - **Response**
        (a) No
        (b) Yes
4. Time in Senegal (ctrSen)
    - **Question** Have you ever lived in Senegal for a period greater than three months?

- **Response**
  (a) No
  (b) Yes

5. Spanish language (Spanish)

   - **Question** Are you able to read written materials (e.g., newspaper articles) in Spanish?
   - **Response**
     (a) No
     (b) Yes

6. French language (French)

   - **Question** Are you able to read written materials (e.g., newspaper articles) in French?
   - **Response**
     (a) No
     (b) Yes

7. Education level (v2zzedlev)

   - **Question** What is the highest level of education you have completed?
   - **Response**
     (a) Incomplete primary.or left before eighth grade
     (b) Primary completed or completed eighth grade
     (c) Incomplete secondary or left in grades 9-12
     (d) Secondary or high school completed, or obtained GED
     (e) Post-secondary trade/vocational school
     (f) University undergraduate degree incomplete
     (g) University undergraduate degree completed
     (h) Masters degree (MA)
     (i) Ph.D
     (j) Juris Doctor or other professional degree (medicine, business)

E Gold Standard Questions

[RANDOMIZE THE ORDER OF QUESTIONS IN THIS SECTION]

Now, we will shift to asking you political questions about hypothetical countries. You will be paid [RANDOMIZED PAY RATE] per answer in this section.

1. Party ban (v2psparban34)

   - **Description** In Country X, a number of parties contested each other for legislative power every election year. In the latest national election, the only parties banned from participating were non-democratic parties advocating for overthrowing the multi-party system. In practice, this meant that only one party was denied political participation.

- **Question** Were parties banned in Country X?
- **Clarification** This does not apply to parties that were barred from competing for failing to meet registration requirements or support thresholds.
- **Response**
  (a) Yes. All parties except the state-sponsored party (and closely allied parties) were banned.
  (b) Yes. Elections were non-partisan or there were no officially recognized parties.
  (c) Yes. Many parties were banned.
  (d) Yes. But only a few parties were banned.
  (e) No. No parties were officially banned.

2. Disclosure of campaign donations (v2eldonate12)

- **Description** National elections in Country X were often criticized for lack of transparency. Although several laws guided how electoral parties must disclose information such as campaign spending, donations and distribution of spending per region, several NGOs found that these laws were not being followed. When they were followed, it was only by smaller parties that were not been able to make their way into the mainstream. The NGO findings were refuted by the government.
- **Question** Were there disclosure requirements for donations to national election campaigns in Country X?
- **Response**
  (a) No. There were no disclosure requirements.
  (b) Not really. There were some, possibly partial, disclosure requirements in place but they were not observed or enforced most of the time.
  (c) Ambiguous. There were disclosure requirements in place, but it is unclear to what extent they were observed or enforced.
  (d) Mostly. The disclosure requirements may not be fully comprehensive (some donations not covered), but most existing arrangements were observed and enforced.
  (e) Yes. There were comprehensive requirements and they were observed and enforced almost all the time.

3. State ownership of economy (v2clstown23)

- **Description** In Country X, many industries remained subject to regulation by the national government. However, in recent years economic hardship forced the regime to loosen its controls of several small industries. In addition private schools became commonplace due to an overall increase in wealth caused by less restrictive government controls. These schools, although not initially deemed valuable by the country's leadership, resulted in an influx of academics from surrounding countries, and lead to a higher level of education. This manifested itself in the establishment and growth of industrial sectors that the government no longer had direct control over, making them relatively more meaningful than those controlled by the state.
- **Question** What was the level at which the state (the government of Country X) owned or directly controlled important sectors of the economy?

- **Clarification** This question gauges the degree to which the state owns and controls capital (including land) in the industrial, agricultural, and service sectors. It does not measure the extent of government revenue and expenditure as a share of total output; indeed, it is quite common for states with expansive fiscal policies to exercise little direct control (and virtually no ownership) over the economy.
- **Response**
  (a) Virtually all valuable capital belonged to the state or was directly controlled by the state. Private property may be officially prohibited.
  (b) Most valuable capital either belonged to the state or was directly controlled by the state.
  (c) Many sectors of the economy either belonged to the state or were directly controlled by the state, but others remained relatively free of direct state control.
  (d) Some valuable capital either belonged to the state or was directly controlled by the state, but most remained free of direct state control.
  (e) Very little valuable capital belonged to the state or was directly controlled by the state.

4. Judicial accountability (v2juaccnt01)

- **Description** In Country X, there were a number of judges who came under scrutiny for questionable conduct. The country's executive had the power to punish judges for not fulfilling their duties adequately, however as the judges were usually appointed by the executive through personal connections, this was a rare occurrence. Thorough investigations were rarely conducted as any transgressions were usually handled between the executive and the judiciary..
- **Question** When judges were found responsible for serious misconduct, how often were they removed from their posts or otherwise disciplined in Country X?
- **Response**
  (a) Never
  (b) Seldom
  (c) About half of the time
  (d) Usually
  (e) Always

F Main Coding Task

[Coders should be randomly assigned two of the nine variables listed in the following nine subsections. The coder will be randomly assigned a variable, and then will be randomly assigned to see ONE of the two screener questions for that variable, which will appear on its own screen. Then, upon clicking through to the next screen, the coder will see the question text for the variable at the top, and then a matrix of 30 years for one country. The coder will code 30 years for that country for that variable, and then will be randomly assigned to a second variable. After the second variable's screener question and 30-year table is complete, the coder should be asked the question that appears in the 10th and final sub-section of this "Main Coding Task" section. Coders who answer "Yes" to this question will then code 30 years for the second randomly selected variable

out of this section but for a different country. Before each screener should appear the following statement:]

Now, we will ask you questions about [VARIABLE]. First, we will ask you about a hypothetical country. You will be paid [RANDOMIZED PAY RATE] to answer this question. Please refer to the link 'Training and Reference Materials' if you have any questions about terminology.

[Then, the following statement should appear before each matrix of 30 years:]

Now, you will be asked to evaluate six five-year periods for the country [COUNTRY] about [VARIABLE], gathering information and conducting research as necessary. You will be paid [pay rate] for each year you evaluate. Please refer to the link 'Training and Reference Materials' if you have any questions about terminology.

F.1 Political killings

1. Screener 1

   - **Description** In Country X, political killings were practiced systematically and they were typically incited and approved by top leaders of government.
   - **Question** Was there freedom from political killings in Country X?
   - **Clarification** Political killings are killings by the state or its agents without due process of law for the purpose of eliminating political opponents. These killings are the result of deliberate use of lethal force by the police, security forces, prison officials, or other agents of the state (including paramilitary groups).
   - **Response**
     (a) Not respected by public authorities. Political killings were practiced systematically and they were typically incited and approved by top leaders of government.
     (b) Weakly respected by public authorities. Political killings were practiced frequently and top leaders of government were not actively working to prevent them.
     (c) Somewhat respected by public authorities. Political killings were practiced occasionally but they were typically not incited and approved by top leaders of government.
     (d) Mostly respected by public authorities. Political killings were practiced in a few isolated cases but they were not incited or approved by top leaders of government.
     (e) Fully respected by public authorities. Political killings were non-existent.

2. Screener 2

   - **Description** In Country X, political killings were non-existent.
   - **Question** Was there freedom from political killings in Country X?
   - **Clarification** Political killings are killings by the state or its agents without due process of law for the purpose of eliminating political opponents. These killings are the result of deliberate use of lethal force by the police, security forces, prison officials, or other agents of the state (including paramilitary groups).
   - **Response**
     (a) Not respected by public authorities. Political killings were practiced systematically and they were typically incited and approved by top leaders of government.
     (b) Weakly respected by public authorities. Political killings were practiced frequently and top leaders of government were not actively working to prevent them.

(c) Somewhat respected by public authorities. Political killings were practiced occasionally but they were typically not incited and approved by top leaders of government.

(d) Mostly respected by public authorities. Political killings were practiced in a few isolated cases but they were not incited or approved by top leaders of government.

(e) Fully respected by public authorities. Political killings were non-existent.

3. Variable

- **Question** Please code the degree to which there was freedom from political killings in [COUNTRY] in each of the following years.
- **Clarification** Political killings are killings by the state or its agents without due process of law for the purpose of eliminating political opponents. These killings are the result of deliberate use of lethal force by the police, security forces, prison officials, or other agents of the state (including paramilitary groups).
- **Response**

(a) Not respected by public authorities. Political killings were practiced systematically and they were typically incited and approved by top leaders of government.

(b) Weakly respected by public authorities. Political killings were practiced frequently and top leaders of government were not actively working to prevent them.

(c) Somewhat respected by public authorities. Political killings were practiced occasionally but they were typically not incited and approved by top leaders of government.

(d) Mostly respected by public authorities. Political killings were practiced in a few isolated cases but they were not incited or approved by top leaders of government.

(e) Fully respected by public authorities. Political killings were non-existent.

F.2 Harassment of journalists

1. Screener 1

- **Description** In Country X, no journalists dared to engage in journalistic activities that would offend powerful actors because harassment or worse would be certain to occur.
- **Question** Were individual journalists harassed—i.e., threatened with libel, arrested, imprisoned, beaten, or killed—by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities in Country X?
- **Response**

(a) No journalists dared to engage in journalistic activities that would offend powerful actors because harassment or worse would be certain to occur.

(b) Some journalists occasionally offended powerful actors but they were almost always harassed or worse and eventually were forced to stop.

(c) Some journalists who offended powerful actors are forced to stop but others managed to continue practicing journalism freely for long periods of time.

(d) It was rare for any journalist to be harassed for offending powerful actors, and if this were to happen, those responsible for the harassment would be identified and punished.

(e) Journalists were never harassed by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities.

2. Screener 2

   - **Description** In Country X, journalists were never harassed by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities.
   - **Question** Were individual journalists harassed—i.e., threatened with libel, arrested, imprisoned, beaten, or killed—by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities in Country X?
   - **Response**
     (a) No journalists dared to engage in journalistic activities that would offend powerful actors because harassment or worse would be certain to occur.
     (b) Some journalists occasionally offended powerful actors but they were almost always harassed or worse and eventually were forced to stop.
     (c) Some journalists who offended powerful actors are forced to stop but others managed to continue practicing journalism freely for long periods of time.
     (d) It was rare for any journalist to be harassed for offending powerful actors, and if this were to happen, those responsible for the harassment would be identified and punished.
     (e) Journalists were never harassed by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities.

3. Variable

   - **Question** Please code the degree to which individual journalists were harassed - i.e., threatened with libel, arrested, imprisoned, beaten, or killed – by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities in [COUNTRY] in each of the following years.
   - **Response**
     (a) No journalists dared to engage in journalistic activities that would offend powerful actors because harassment or worse would be certain to occur.
     (b) Some journalists occasionally offended powerful actors but they were almost always harassed or worse and eventually were forced to stop.
     (c) Some journalists who offended powerful actors are forced to stop but others managed to continue practicing journalism freely for long periods of time.
     (d) It was rare for any journalist to be harassed for offending powerful actors, and if this were to happen, those responsible for the harassment would be identified and punished.
     (e) Journalists were never harassed by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities.

F.3 Freedom from forced labor for men

1. Screener 1

   - **Description** In Country X, male servitude or other kinds of forced labor was widespread and accepted (perhaps even organized) by the state.
   - **Question** Were adult men free from servitude and other kinds of forced labor in Country X?

- **Clarification** Involuntary servitude occurs when an adult is unable to quit a job s/he desires to leave not by reason of economic necessity but rather by reason of employer's coercion. This includes labor camps but not work or service which forms part of normal civic obligations such as conscription or employment in command economies.
- **Response**
  (a) Male servitude or other kinds of forced labor was widespread and accepted (perhaps even organized) by the state.
  (b) Male servitude or other kinds of forced labor was substantial. Although officially opposed by the public authorities, the state was unwilling or unable to effectively contain the practice.
  (c) Male servitude or other kinds of forced labor exists but was not widespread and usually actively opposed by public authorities, or only tolerated in some particular areas or among particular social groups.
  (d) Male servitude or other kinds of forced labor was infrequent and only found in the criminal underground. It was actively and sincerely opposed by the public authorities.
  (e) Male servitude or other kinds of forced labor was virtually non-existent.

2. Screener 2

- **Description** In Country X, male servitude or other kinds of forced labor was virtually non-existent.
- **Question** Were adult men free from servitude and other kinds of forced labor in Country X?
- **Clarification** Involuntary servitude occurs when an adult is unable to quit a job s/he desires to leave not by reason of economic necessity but rather by reason of employer's coercion. This includes labor camps but not work or service which forms part of normal civic obligations such as conscription or employment in command economies.
- **Response**
  (a) Male servitude or other kinds of forced labor was widespread and accepted (perhaps even organized) by the state.
  (b) Male servitude or other kinds of forced labor was substantial. Although officially opposed by the public authorities, the state was unwilling or unable to effectively contain the practice.
  (c) Male servitude or other kinds of forced labor exists but was not widespread and usually actively opposed by public authorities, or only tolerated in some particular areas or among particular social groups.
  (d) Male servitude or other kinds of forced labor was infrequent and only found in the criminal underground. It was actively and sincerely opposed by the public authorities.
  (e) Male servitude or other kinds of forced labor was virtually non-existent.

3. Variable

- **Question** Please code the degree to which adult men were free from servitude and other kinds of forced labor in [COUNTRY] in each of the following years.

- **Clarification** Involuntary servitude occurs when an adult is unable to quit a job s/he desires to leave not by reason of economic necessity but rather by reason of employer's coercion. This includes labor camps but not work or service which forms part of normal civic obligations such as conscription or employment in command economies.
- **Response**
  (a) Male servitude or other kinds of forced labor was widespread and accepted (perhaps even organized) by the state.
  (b) Male servitude or other kinds of forced labor was substantial. Although officially opposed by the public authorities, the state was unwilling or unable to effectively contain the practice.
  (c) Male servitude or other kinds of forced labor exists but was not widespread and usually actively opposed by public authorities, or only tolerated in some particular areas or among particular social groups.
  (d) Male servitude or other kinds of forced labor was infrequent and only found in the criminal underground. It was actively and sincerely opposed by the public authorities.
  (e) Male servitude or other kinds of forced labor was virtually non-existent.

F.4 Power distribution by gender

1. Screener 1
   - **Description** In Country X, men had a near-monopoly on political power.
   - **Question** Was political power distributed according to gender in Country X?
   - **Response**
     (a) Men had a near-monopoly on political power.
     (b) Men had a dominant hold on political power. Women had only marginal influence.
     (c) Men had much more political power but women had some areas of influence.
     (d) Men had somewhat more political power than women.
     (e) Men and women had roughly equal political power.
2. Screener 2
   - **Description** In Country X, men and women had roughly equal political power.
   - **Question** Was political power distributed according to gender in Country X?
   - **Response**
     (a) Men had a near-monopoly on political power.
     (b) Men had a dominant hold on political power. Women had only marginal influence.
     (c) Men had much more political power but women had some areas of influence.
     (d) Men had somewhat more political power than women.
     (e) Men and women had roughly equal political power.
3. Variable
   - **Question** Please code the degree to which political power was distributed according to gender in [COUNTRY] in each of the following years.

- **Response**
  (a) Men had a near-monopoly on political power.
  (b) Men had a dominant hold on political power. Women had only marginal influence.
  (c) Men had much more political power but women had some areas of influence.
  (d) Men had somewhat more political power than women.
  (e) Men and women had roughly equal political power.

F.5 High court independence

1. Screener 1

   - **Description** When the high court in the judicial system of Country X was ruling in cases that are salient to the government, it always made decisions that merely reflect government wishes regardless of its sincere view of the legal record.
   - **Question** When the high court in the judicial system of Country X was ruling in cases that are salient to the government, how often would you say that it made decisions that merely reflect government wishes regardless of its sincere view of the legal record?
   - **Clarification** We are seeking to identify autonomous judicial decision-making and its absence. Decisions certainly can reflect government wishes without "merely reflecting" those wishes, i.e. a court can be autonomous when its decisions support the government's position. This is because a court can be fairly persuaded that the government's position is meritorious. By "merely reflect the wishes of the government" we mean that the courts own view of the record, its sincere evaluation of the record, is irrelevant to the outcome. The court simply adopts the government's position regardless of its sincere view of the record.
   - **Response**
     (a) Always
     (b) Usually
     (c) About half of the time
     (d) Seldom
     (e) Never

2. Screener 2

   - **Description** When the high court in the judicial system of Country X was ruling in cases that are salient to the government, it never made decisions that merely reflect government wishes regardless of its sincere view of the legal record.
   - **Question** When the high court in the judicial system of Country X was ruling in cases that are salient to the government, how often would you say that it made decisions that merely reflect government wishes regardless of its sincere view of the legal record?
   - **Clarification** We are seeking to identify autonomous judicial decision-making and its absence. Decisions certainly can reflect government wishes without "merely reflecting" those wishes, i.e. a court can be autonomous when its decisions support the government's position. This is because a court can be fairly persuaded that the government's position is meritorious. By "merely reflect the wishes of the government" we mean that the courts own view of the record, its sincere evaluation of the record, is irrelevant to the outcome.

The court simply adopts the government's position regardless of its sincere view of the record.

- **Response**
  (a) Always
  (b) Usually
  (c) About half of the time
  (d) Seldom
  (e) Never

3. Variable

- **Question** When the high court in the judicial system of [COUNTRY] was ruling in cases that are salient to the government, how often would you say that it made decisions that merely reflect government wishes regardless of its sincere view of the legal record in each of the following years?
- **Clarification** We are seeking to identify autonomous judicial decision-making and its absence. Decisions certainly can reflect government wishes without "merely reflecting" those wishes, i.e. a court can be autonomous when its decisions support the government's position. This is because a court can be fairly persuaded that the government's position is meritorious. By "merely reflect the wishes of the government" we mean that the courts own view of the record, its sincere evaluation of the record, is irrelevant to the outcome. The court simply adopts the government's position regardless of its sincere view of the record.
- **Response**
  (a) Always
  (b) Usually
  (c) About half of the time
  (d) Seldom
  (e) Never

F.6 Minimum voting age requirements

1. Screener 1

- **Description** In Country X, individuals 18 or older were allowed to vote in national elections.
- **Question** What was the minimum age at which citizens were allowed to vote in national elections in Country X?
- **Response**
  (a) NA
  (b) Scale [CONSTRAIN TO TWO DIGIT INTEGER]

2. Screener 2

- **Description** In Country X, individuals 60 or older were allowed to vote in national elections.

- **Question** What was the minimum age at which citizens were allowed to vote in national elections in Country X?
- **Response**
  (a) NA
  (b) Scale [CONSTRAIN TO TWO DIGIT INTEGER]

3. Variable

- **Question** What was the minimum age at which citizens were allowed to vote in national elections in [COUNTRY] in each of the following years?
- **Response**
  (a) NA
  (b) Scale [CONSTRAIN TO TWO DIGIT INTEGER]

F.7 Bicameral legislatures

1. Screener 1

- **Description** In Country X, the legislature contained 2 chambers.
- **Question** How many chambers did the legislature of Country X contain?
- **Response**
  (a) 0 chambers (NA)
  (b) 1 chamber
  (c) 2 or more chambers

2. Screener 2

- **Description** In Country X, the legislature contained 0 chambers.
- **Question** How many chambers did the legislature of Country X contain?
- **Response**
  (a) 0 chambers (NA)
  (b) 1 chamber
  (c) 2 or more chambers

3. Variable

- **Question** How many chambers did the legislature of [COUNTRY] contain in each of the following years?
- **Response**
  (a) 0 chambers (NA)
  (b) 1 chamber
  (c) 2 or more chambers

F.8 Referendums

1. Screener 1

   - **Description** In Country X, referendums were not allowed.
   - **Question** Was there a legal provision for referendums in Country X?
   - **Clarification** These are measures placed on the ballot through a citizen petition process, not by the legislature or the executive. They may concern either the rejection of a recently approved law or a bill discussed in parliament. (They do not include recall elections.)
   - **Response**
     (a) Not allowed
     (b) Allowed but non-binding (or with an intervening institutional veto)
     (c) Allowed and binding

2. Screener 2

   - **Description** In Country X, referendums were allowed and binding.
   - **Question** Was there a legal provision for referendums in Country X?
   - **Clarification** These are measures placed on the ballot through a citizen petition process, not by the legislature or the executive. They may concern either the rejection of a recently approved law or a bill discussed in parliament. (They do not include recall elections.)
   - **Response**
     (a) Not allowed
     (b) Allowed but non-binding (or with an intervening institutional veto)
     (c) Allowed and binding

3. Variable

   - **Question** Was there a legal provision for referendums in [COUNTRY] in each of the following years?
   - **Clarification** These are measures placed on the ballot through a citizen petition process, not by the legislature or the executive. They may concern either the rejection of a recently approved law or a bill discussed in parliament. (They do not include recall elections.)
   - **Response**
     (a) Not allowed
     (b) Allowed but non-binding (or with an intervening institutional veto)
     (c) Allowed and binding

F.9 Suffrage rates

1. Screener 1

   - **Description** In Country X, there had never been national elections.

- **Question** What percentage (%) of adult citizens (as defined by statute) had the legal right to vote in national elections in Country X?
- **Clarification** This question does not take into consideration restrictions based on age, residence, having been convicted for crime, or being legally incompetent. It covers legal (de jure) restrictions, not restrictions that may be operative in practice (de facto). The adult population (as defined by statute) is defined by citizens in the case of independent countries or the people living in the territorial entity in the case of colonies. Universal suffrage is coded as 100%. Universal male suffrage only is coded as 50%. Years before electoral provisions are introduced are scored 0%. The scores do not reflect whether an electoral regime was interrupted or not. Only if new constitutions, electoral laws, or the like explicitly introduce new regulations of suffrage, the scores were adjusted accordingly if the changes suggested doing so. If qualifying criteria other than gender apply (such as property, tax payments, income, literacy, region, race, ethnicity, religion, and/or "economic independence"), estimates have been calculated by combining information on the restrictions with different kinds of statistical information (on population size, age distribution, wealth distribution, literacy rates, size of ethnic groups, etc.), secondary country-specific sources, and—in the case of very poor information—the conditions in similar countries or colonies. The scores reflect de jure provisions of suffrage extension in percentage of the adult population. If the suffrage law is revised in a way that affects the extension, the scores reflect this change as of the calendar year the law was enacted.
- **Response** Scale (0-100)

2. Screener 1

- **Description** In Country X, all adult citizens (as defined by statute) had the legal right to vote in national elections.
- **Question** What percentage (%) of adult citizens (as defined by statute) had the legal right to vote in national elections in Country X?
- **Clarification** This question does not take into consideration restrictions based on age, residence, having been convicted for crime, or being legally incompetent. It covers legal (de jure) restrictions, not restrictions that may be operative in practice (de facto). The adult population (as defined by statute) is defined by citizens in the case of independent countries or the people living in the territorial entity in the case of colonies. Universal suffrage is coded as 100%. Universal male suffrage only is coded as 50%. Years before electoral provisions are introduced are scored 0%. The scores do not reflect whether an electoral regime was interrupted or not. Only if new constitutions, electoral laws, or the like explicitly introduce new regulations of suffrage, the scores were adjusted accordingly if the changes suggested doing so. If qualifying criteria other than gender apply (such as property, tax payments, income, literacy, region, race, ethnicity, religion, and/or "economic independence"), estimates have been calculated by combining information on the restrictions with different kinds of statistical information (on population size, age distribution, wealth distribution, literacy rates, size of ethnic groups, etc.), secondary country-specific sources, and—in the case of very poor information—the conditions in similar countries or colonies. The scores reflect de jure provisions of suffrage extension in percentage of the adult population. If the suffrage law is revised in a way that affects the extension, the scores reflect this change as of the calendar year the law was enacted.
- **Response** Scale (0-100)

3. Variable

   - **Question** What percentage (%) of adult citizens (as defined by statute) had the legal right to vote in national elections in [COUNTRY] in each of the following years?
   - **Clarification** This question does not take into consideration restrictions based on age, residence, having been convicted for crime, or being legally incompetent. It covers legal (de jure) restrictions, not restrictions that may be operative in practice (de facto). The adult population (as defined by statute) is defined by citizens in the case of independent countries or the people living in the territorial entity in the case of colonies. Universal suffrage is coded as 100%. Universal male suffrage only is coded as 50%. Years before electoral provisions are introduced are scored 0%. The scores do not reflect whether an electoral regime was interrupted or not. Only if new constitutions, electoral laws, or the like explicitly introduce new regulations of suffrage, the scores were adjusted accordingly if the changes suggested doing so. If qualifying criteria other than gender apply (such as property, tax payments, income, literacy, region, race, ethnicity, religion, and/or "economic independence"), estimates have been calculated by combining information on the restrictions with different kinds of statistical information (on population size, age distribution, wealth distribution, literacy rates, size of ethnic groups, etc.), secondary country-specific sources, and—in the case of very poor information—the conditions in similar countries or colonies. The scores reflect de jure provisions of suffrage extension in percentage of the adult population. If the suffrage law is revised in a way that affects the extension, the scores reflect this change as of the calendar year the law was enacted.
   - **Response** Scale (0-100)

F.10 Additional country

1. Additional Country

   - **Question** Would you be willing to code another 30-year period for another country? You will be paid the same rate of [RANDOMIZED PAY RATE] for each year that you code for this country.
   - **Response**
     (a) Yes, I will code another country.
     (b) No, I want to end my participation.

G Post-Survey Questionnaire

Thank you for your assistance with this study. We will now conclude with a few more questions about you and your experiences. You will be paid $0.10 per answer in this section.

1. V-Dem use (v2vd)

   - **Question** Did you refer to data from the Varieties of Democracy (V-Dem) Project (either from the Project website, v-dem.net, or otherwise) to assist in your coding?
   - **Response**
     (a) No
     (b) Yes

2. Coursework on Argentina (crsArg)

   - **Question** Have you taken a college/university course that focused on Argentina?
   - **Response**
     (a) No
     (b) Yes

3. Coursework on Senegal (crsSen)

   - **Question** Have you taken a college/university course that focused on Senegal?
   - **Response**
     (a) No
     (b) Yes

4. Polisci major (v2zzpolmaj) [SHOW ONLY IF v2zzedlev IS F OR HIGHER]

   - **Question** Was political science or a related field (e.g., public policy, public affairs) your major or focus in your post-secondary education at any level (i.e., undergraduate or graduate)?
   - **Response**
     (a) No
     (b) Yes

5. Polisci major (v2zzpolmaj) [SHOW ONLY IF v2zzpolmaj IS A]

   - **Question** Did you take one or more political science courses during the course of your post-secondary education?
   - **Response**
     (a) No
     (b) Yes

6. Follow politics (v2folpol)

   - **Question** Do you follow international or national politics on a weekly basis (or more frequently)?
   - **Response**
     (a) No
     (b) Yes

7. Interest in Public Affairs (pubaffairsinterest)

   - **Question** How interested would you say you are in public affairs?
   - **Response**
     (a) Not very interested
     (b) Very interested

8. Discuss politics (dpol)

- **Question** When you get together with your friends or family, how often do you discuss political matters?
- **Response**
  (a) Rarely
  (b) Frequently

9. Democracy important (v2demi)

   - **Question** Is it important for you to live in a country that is governed democratically?
   - **Response**
     (a) No
     (b) Yes

10. Vote Past National Election (votpastnat)

    - **Question** Did you vote in the last national election in the country in which you are eligible to vote?
    - **Response**
      (a) No
      (b) Yes
      (c) Not eligible to vote

11. Country of birth (v2zzbornin)

    - **Question** In which country were you born?
    - **Response** Country (chosen from menu, include "Other" as an option)

12. Country of residence (v2zzreside)

    - **Question** In what country do you live today?
    - **Clarification** If your time is split between several countries, list that country where you spend the most time or that which constitutes your official residence.
    - **Response** Country (chosen from menu)

- **Question** Add here any comments you have about any of the previous questions.
- **Response** [TEXT BOX]

Thank you for participating in this study. We will now issue your payment. Please click the next button to submit your responses and receive your validation code.

References

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver & Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110(02):278–295.

Berinsky, Adam J., Gregory A. Huber & Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–368.

Coppedge, Michael, Adam Glynn, Staffan Lindberg, Daniel Pemstein & Brigitte Seim. 2015. "Conceptualizing Democracy: A Survey Experiment Using Paired Country Comparisons." *Working paper* .

Crump, Matthew J. C., John V. McDonnell & Todd M. Gureckis. 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." *PloS One* 8(3):e57410.

Honaker, James, Michael Berkman, Chris Ojeda & Eric Plutzer. 2013. "Sorting Algorithms for Qualitative Data to Recover Latent Dimensions with Crowdsourced Judgments: Measuring State Policies for Welfare Eligibility under TANF.".

Hooghe, Lisbet, Ryan Bakker, Anna Brigevich, Catherine de Vries, Erica Edwards, Gary Marks, Jan Rovny & Marco Steenbergen. 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.

Morris, Peter A. 1977. "Combining Expert Judgments: A Bayesian Approach." *Management Science* 23(7):679–693.

Paolacci, Gabriele, Jesse Chandler & Panagiotis G. Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5(5):411–419.

Peer, Eyal, Sonam Samat, Laura Brandimarte & Alessandro Acquisti. 2016. "Mechanical Turk vs Prolific Academic vs CrowdFlower." *Working paper* .

Ross, Joel, Lilly Irani, M. Six Silberman, Andrew Zaldivar & Bill Tomlinson. 2010. "Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk.".

Schnakenberg, Keith & Elizabeth Maggie Penn. 2014. "Scoring from Contests." *Political Analysis* 22(1):86–114.
URL: http://pan.oxfordjournals.org/content/22/1/86.short

Shapiro, Danielle N., Jesse Chandler & Pam A. Mueller. 2013. "Using Mechanical Turk to Study Clinical Populations." *Clinical Psychological Science* 1(2):213–220.

Tetlock, Philip. 2005. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton, NJ: Princeton University Press.