Pre-Analysis Plan:
Varieties of Democracy Data Update Experiments 2018

Daniel Pemstein*
Brigitte Seim†

November 12, 2018

Contents

*Associate Professor, North Dakota State University
†Assistant Professor, University of North Carolina at Chapel Hill

This protocol describes a bundled set of experiments that was deployed on Qualtrics among Turkers in September 2017 (pilot) and among V-Dem expert coders in January 2018 (full). This pre-analysis plan was written and filed before the dataset for the full experiment was provided to the co-authors.

1 Participants

The pool of potential participants in this study consists of all V-Dem country experts (CEs) participating in V-Dem's 2018 annual update. Each CE completes one or more surveys for their primary country of expertise. They are also invited to participate in a number of other tasks, such as coding anchoring vignettes, coding survey questions for bridge cases (countries other than their primary country), to complete—or update, in the case of returning coders—as post-survey questionnaire, and, in 2018, to answer a small number of prediction prompts. At the end of this process, all coders who did not drop out at an earlier stage see an invitation to participate in "…a research study to examine the conditions under which participants provide comparable data about political institutions." The sample of experts on political institutions available for this study consists of these CEs.

2 Inducement Experiment

In 2015, we embarked on a research agenda to understand expert coder incentive structure. This research was inspired by an observed puzzle: As V-Dem pay is invariate to cross-country differences in purchasing power, we should expect to see coder data quality varying with purchasing power. Yet, we don't observe this. V-Dem expert coders must be incentivized by factors other than pay (though they certainly aren't immune to monetary incentives).

To explore this issue, V-Dem interns conducted qualitative interviews with a group of V-Dem expert coders, both those who are still coding and those who have left. In this exercise, we learned that coders appear to respond to a series of incentives, including monetary, reputational, and public goods incentives.

To explore this issue further, we decided to conduct an experiment among V-Dem expert coders in the 2018 update. We conducted a pilot of this protocol among Amazon Turk users in fall of 2017.

2.1 Protocol

The inducement experiment begins on last page of the data collection tool in the v8 update. On that page, V-Dem coders were prompted with a consent form that had three randomly assigned embedded treatments:[1]

1. RANDOMLY ASSIGNED PAY TREATMENT: The first treatment randomizes compensation at the level of the coder, drawing randomly from a uniform distribution from

[1]Task type was also randomly assigned, but this is not part of the inducement experiment and we have no hypotheses regarding how the inducement treatments might interact with task type. We will nonetheless include task type as a control in all analyses.

$0.00 to $2.00 per task in $0.10 increments, for a total of 21 treatment levels.[2]

- When $0.00 is chosen as the [RANDOMIZED PAY RATE], then the text for the inducement experiment is as follows: "**We are unable to offer you payment for participating in the study.**"
- When [RANDOMIZED PAY RATE] IS > $0.00, then the text for the inducement experiment is as follows: "**You will receive [RANDOMIZED PAY RATE] for each question that you answer. At this per-question rate, the average hourly rate is [RANDOMIZED PAY RATE * 40] per hour.**"
  - Example: "You will receive $2.00 for each question that you answer. At this per-question rate, the average hourly rate is $80.00 per hour."

2. RANDOMLY ASSIGNED RECRUITMENT TREATMENT: The second treatment randomizes the recruitment message to either: 1) emphasize expertise; 2) emphasize the public good of knowledge provision; 3) emphasize compensation; 4) provide a basic "please help us" message as a placebo.

   (a) EXPERTISE: **You possess the high level of expertise necessary to answer the questions in this study.**

   (b) PUBLIC GOOD: **If you participate in this study, you will help to improve data on political institutions that are freely available to researchers, policymakers, and other interested parties.**

   (c) COMPENSATION: **If you participate in this study, you will be offered payment at the rate of [RANDOMIZED PAY RATE] for each question that you answer.**

   - If the [RANDOMIZED PAY RATE] is $0.00, and the participant is randomly assigned to the COMPENSATION recruitment message, then the [RANDOMIZED RECRUITMENT MESSAGE] says, "**You will donate your expertise to V–Dem if you participate in the survey.**"

   (d) PLACEBO: **Please consider participating in this study.**

3. RANDOMLY ASSIGNED DISTRACTOR TREATMENT: The third treatment randomizes the inclusion of distractors (no distractors or a distractor appearing with a 1/3rd probability).[3]

The distractor treatment takes the form of a simulated loading screen. The screen says "Loading…" and includes a spinning wheel. The screen remains up for between 2 seconds and 10 seconds, drawn randomly from a uniform distribution in one second increments. After the allotted time has elapsed, the screen automatically advances to the task. Figure 1 shows a screenshot of the distractor.

[2] A task is generally defined as one response to one question.

[3] This treatment was only assigned and executed once the coder opted into the study.

Figure 1: Distractor screen.

2.1.1 Consent Form

The full text of the consent form is as follows:[4]

Opportunity to Participate in Additional Study

- Thank you for your contribution to V-Dem as a Country Expert. In partnership with the V-Dem Institute, V-Dem Project Managers based at North Dakota State University and the University of North Carolina at Chapel Hill are conducting a research study to examine the conditions under which V-Dem country experts provide comparable data about political institutions.
- We would like to invite you to participate in an additional survey. Participation in this survey is entirely voluntary. **[RANDOMIZED RECRUITMENT MESSAGE]** If you agree to be in this study, you will be asked to **[RANDOMIZED TASK TREATMENT]**.
- **[RANDOMIZED PAY RATE MESSAGE]** It will likely take a few minutes to answer each question and 15-45 minutes to complete the whole survey.
- Once you start the survey, you will have 72 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. You can terminate your participation at any time. The last day to complete the survey is January 31, 2018.
- You will not be asked to provide any identifying information as part of this study, and your identity remains confidential. Your responses to this survey will be linked to V-Dem data through a randomized coder id. As with your participation as a Country Expert for V-Dem, your participation in this study will only be known to three people within the V-Dem staff.
- If you have questions, please contact Natalia Stepanova at natalia.stepanova@v-dem.net.

[4]The red and bolded font is slightly larger than the rest of the text.

In order to participate in this study, please click on this link: [CUSTOMIZED QUALTRICS LINK]

2.2 Hypotheses

In the inducement experiment, we have hypotheses regarding the effect of the treatments on response rate, attrition, coder attentiveness, and self-perception of expertise.

2.2.1 Response Rate Hypotheses

H1: *Response rate will increase with pay rate.*

H2: *Response rate will decrease with the purchasing power of the coder's country of residence.*

H3: *The effect of pay rate on response rate will decrease with purchasing power of the coder's country of residence.*

H4: *Compared to the placebo message, the expertise message will result in a higher response rate.*

H5: *Compared to the placebo message, the public goods message will result in a higher response rate.*

H6: *Compared to the placebo message, the compensation message will result in a higher response rate.*

H7: *The effect of compensation message on response rate will decrease with purchasing power of the coder's country of residence.*

H8: *Compared to the expertise message, the public goods message will result in a higher response rate.*

H9: *There will be a positive interaction effect of the compensation message and pay rate on response rate.*

2.2.2 Attrition Hypotheses

H10: *Attrition will decrease with pay rate.*

H11: *Attrition will increase with the purchasing power of the coder's country of residence.*

H12: *The effect of pay rate on attrition will decrease with purchasing power of the coder's country of residence.*

H13: *Compared to the placebo message, the expertise message will result in lower attrition.*

H14: *Compared to the placebo message, the public goods message will result in lower attrition.*

H15: *Compared to the placebo message, the compensation message will result in lower attrition.*

H16: *Compared to the expertise message, the public goods message will result in lower attrition.*

H17: *Attrition will be higher among those assigned to/with greater duration of the distractor.*

H18: *The effect of assignment to/duration of the distractor on attrition will decrease with pay rate.*

H19: *The effect of assignment to/duration of the distractor on attrition will be lower among those assigned to the public goods message.*

H20: *The effect of assignment to/duration of the distractor on attrition will be greater among those assigned to the expertise message.*

2.2.3 Attentiveness Hypotheses

H21: *Coder attentiveness will decrease with pay rate.*

H22: *Coder attentiveness will decrease with purchasing power of the coder's country of residence.*

H23: *The effect of pay rate on attrition will decrease with purchasing power of the coder's country of residence.*

H24: *Compared to the placebo message, the expertise message will result in lower attentiveness.*

H25: *Compared to the placebo message, the public goods message will result in higher attentiveness.*

H26: *Compared to the placebo message, the compensation message will result in lower attentiveness.*

H27: *Attentiveness will be lower among those assigned to/with greater duration of the distractor.*

2.2.4 Self-Perception of Expertise Hypotheses

H28: *The expertise message will result in higher ratings of self-perceived expertise.*

2.3 Sampling

All of the V-Dem CEs who observe the invitation screen participate in this portion of the study. Some of the CEs decline to click on the URL to participate in the other experimental tasks, but even those who opt out of the other experimental tasks provide data regarding response rate, which is analyzed as a separate outcome variable (see Section 2.3 and Section 2.4).

2.4 Planned Analyses

2.4.1 Outcome Variables

We have four outcome variables in our anticipated analyses: response rate; attrition; coder attentiveness; and self-perception of expertise. Response rate will be measured as binary: whether the coder completes at least one task.[5] Attrition will be measured as the number of tasks the coder completed. Attrition will take the value of zero if the coder did not respond. Coder attentiveness will be measured as the percentage of "far pairs" (in which one option is strictly dominated, see below) the coder evaluated correctly in the bottom-up task.[6] Finally, self-reported expertise will be measured as a response to the following question:

- Relative expertise (expertise)
  - **Question** Compared to others who consider similar topics, how would you rate your expertise?
  - **Response**
    1. Below others
    2. Approximately the same as others
    3. Above others

2.4.2 Explanatory Variables

The primary vector of explanatory variables for each respondent, $\mathbf{x}_i$, contains each of the three randomized treatment variables: interval-level pay rate, a vector of three dummy variables indicating assignment to the expertise, public goods, and compensation emphasis conditions (the placebo condition is the baseline), and a dummy variable indicating assignment to a distractor condition. We measure distractor duration in seconds, d_i, and this variable equals zero for respondents in the no-distractor condition. We measure the non-randomized pre-treatment variable of purchasing power using the World Bank's PPP conversion factor for private consumption;[7] this is an interval-level variable, p_i. We invert this multiplier so that p_i gets larger as the purchasing power of a US dollar in i's country gets smaller.

2.4.3 Covariates

We include a vector of basic control variables, $\mathbf{c}_i$ for each respondent, which includes task type assignment and timing of the data completion.[8]

[5] Optimally, we will measure response rate as binary whereby clicking on the experiment link counts as a response. However, we are not sure we will have access to that data. If we do, we will perform a robustness check with this alternate operationalization.

[6] Attentiveness will not be evaluated as an outcome for those individuals who completed the paired comparisons task.

[7] We will use the most recent year of available PPP conversion data for each country. We will assign country based on V-Dem's v2zzreside variable where possible. When that variable is missing we will use the coder's main country coded to proxy for residence.

[8] The number of tasks participants had the opportunity to complete was adjusted throughout the data collection period for budgetary reasons.

Questions that respondents completed in the V-Dem data collection tool as part of the "post-survey questionnaire" will also be included as covariates in some models, as these are answered pre-treatment.[9] In particular, each respondent provides a vector of PSQ responses, $\mathbf{q}_i$, that includes a gender dummy (v2zzgender), an interval-level age variable (converted from birth-dates in v2zzborn), education measured as a set of dummies (undergraduate degree or less completed (v2zzedlev 0–7), masters or professional degree completed (v2zzedlev 8, 10), PhD completed (v2zzedlev 9)), employer dummies (other (v2zzemploy 0, 7, 9, 10), university (v2zzemploy 5, 6), government job (v2zzemploy 1-4), NGO (v2zzemploy 8)), hours spent coding for V-Dem (v2zztimespent), and an ordinal 0–5 satisfaction with V-Dem coding experience (v2zzsatisf).

The additional questions we ask at the end of the experiment in Qualtrics (see Section 5) are all answered post-treatment, and therefore will not be included as covariates in the pre-registered core models, but may be included as covariates in further exploratory analyses and robustness checks.

2.4.4 Empirical Models

The primary test of each hypothesis will be a regression in which we regress the outcome variable on the explanatory variables, with a battery of controls.

First, we will model response rate and attrition jointly, using zero-inflated binomial regression, while also accounting for right censoring (CZINB). Specifically, we will use the first stage of the CZINB to model response rate[10] and the second stage will model attrition, conditional on response. We will assume that all values of the attrition variable are right-censored at 20 tasks.[11] The first stage of the zero-inflated censored negative binomial regression (CZINB) will take the following forms:

1. $y_i = \alpha + \beta \cdot f(\mathbf{x}_i^{-d}) + \gamma \mathbf{c}_i + \eta \mathbf{q}_i$ and
2. $y_i = \alpha + \beta \cdot g(\mathbf{x}_i^{-d}, p_i) + \gamma \mathbf{c}_i$,

where $\mathbf{x}_i^{-d}$ includes all elements of $\mathbf{x}$ except for the distractor task dummy. Model 1 addresses hypotheses 1, 4, 5, 6, 8, and 9. Model 2 additionally addresses hypotheses 2, 3, and 7, while serving as a robustness check for model 1 with respect to purchasing power. The function f generates all the main effects in its input variables and the interaction between compensation and pay described by hypothesis 9. The function g is similar to f, but also generates the interactions between pay and PPP, and between compensation message and PPP, described by hypotheses 3 and 7. Model 2 drops the post-survey questionnaire controls $\mathbf{q}_i$ because these are post-treatment with respect to PPP, and, because there is a plausible causal link from PPP to $\mathbf{q}_i$, their inclusion could bias estimates.

[9] We will use multiple imputation to impute missing data in these covariates, using the full sample of V-Dem coders' responses to questions included in $\mathbf{q}$, as well as the v2zzreside, v2zzbornin, and v22zzedcnt variables.

[10] If we are able to observe explicit opt-in, in addition to task count, to measure response rate, we will supplement this zero-inflated binomial with a logit regressions of this second operationalization of response.

[11] As we note below, we initially allowed participants to complete more than 20 tasks, but reduced the maximum task level to 20 during the course of data collection. We are therefore standardizing right censoring at this lower level.

The attrition stage of CZINB is similar to the response stage, but includes distractor task variables. It takes four functional forms:

1. $y_i = \alpha + \beta \cdot f'(\mathbf{x}_i) + \gamma \mathbf{c}_i + \eta \mathbf{q}_i$,
2. $y_i = \alpha + \beta \cdot f'(\mathbf{x}_i^{-d}, d_i) + \gamma \mathbf{c}_i + \eta \mathbf{q}_i$,
3. $y_i = \alpha + \beta \cdot g'(\mathbf{x}_i, p_i) + \gamma \mathbf{c}_i$, and,
4. $y_i = \alpha + \beta \cdot g'(\mathbf{x}_i^{-d}, d_i, p_i) + \gamma \mathbf{c}_i$.

Here, $f'(\cdot)$ and $g'(\cdot)$ are similar to f and g. Both generate all the main effects of their input variables. Additionally, f' generates the interactions described by hypotheses 18–20 and g' additionally generates the interaction described by hypothesis 12. Models 1 and 2 address hypotheses 10, 13–20. Models 3 and 4 additionally address hypotheses 11 and 12 while serving as robustness checks for models 1 and 2 with respect to purchasing power. Models 1 and 3 use an either/or operationalization of the distractor task, while models 2 and 4 measures distractor in seconds, capturing both operationalizations of hypotheses 17-20. Overall, we will fit four CZINB regressions, using the following first stage-second stage combinations: 1, 1; 1, 2; 2, 3; 2, 4. The first two CZINB models exclude the observational variable purchasing power, while the second two include it.

We will model attentiveness using OLS, and the following specifications:

1. $y_i = \alpha + \beta \mathbf{x}_i + \gamma \mathbf{c}_i + \eta \mathbf{q}_i$,
2. $y_i = \alpha + \beta (\mathbf{x}_i^{-d}, d_i) + \gamma \mathbf{c}_i + \eta \mathbf{q}_i$,
3. $y_i = \alpha + \beta \cdot h(\mathbf{x}_i, p_i) + \gamma \mathbf{c}_i$, and,
4. $y_i = \alpha + \beta \cdot h(\mathbf{x}_i^{-d}, d_i, p_i) + \gamma \mathbf{c}_i$.

Here, models 1 and 2 investigate hypotheses 21, 24–27 while models 3 and 4 additionally test hypotheses 22–23, while serving as robustness checks for models 1 and 2 with respect to purchasing power. Models 1 and 3 operationalize the distractor task as binary, while models 2 and 4 use the interval operationalization. The function h generates main effects and the interaction term described by hypothesis 23. Since attentiveness is only defined for those participating in the bottom-up task, $\mathbf{c}_i$ will contain only timing of data completion in this instance.

Finally, we model self-perception of expertise (hypothesis 28) using the same four regression equations we use to model attentiveness, but now including the entire sample of participants. We include four models to investigate the robustness of effects to operationalization of distractor and inclusion of purchasing power.

3 Bottom-Up Aggregation

The second part of the study examines if we can leverage experts' shared notions about democratic institutions to aggregate low-level information about democratic institutions

into higher-level measures of democracy, potentially side-stepping the difficult process of designing universal aggregation rules. It also attempts to unpack how scholars of democratic institutions understand a country's level of electoral democracy, as a function of its constituent institutional parts. We have two primary research questions:

1. Can we use an inductive process of pairwise comparisons to generate a electoral democracy scale that has a rank ordering with high face validity? Specifically, does this process produce a scale that is similar to V-Dem's current electoral democracy scale?

2. What lower-level institutions most influence how experts rank order cases with respect to electoral democracy?

3.1 Protocol

We perform a conjoint experiment where we ask respondents to compare pairs of country descriptions and to report their opinions on both which country is more democratic (rank order) and on the level of democracy for each country. We present respondents with a repeated conjoint task.[12]

Initially, at each round, we present respondents with the following instructions:

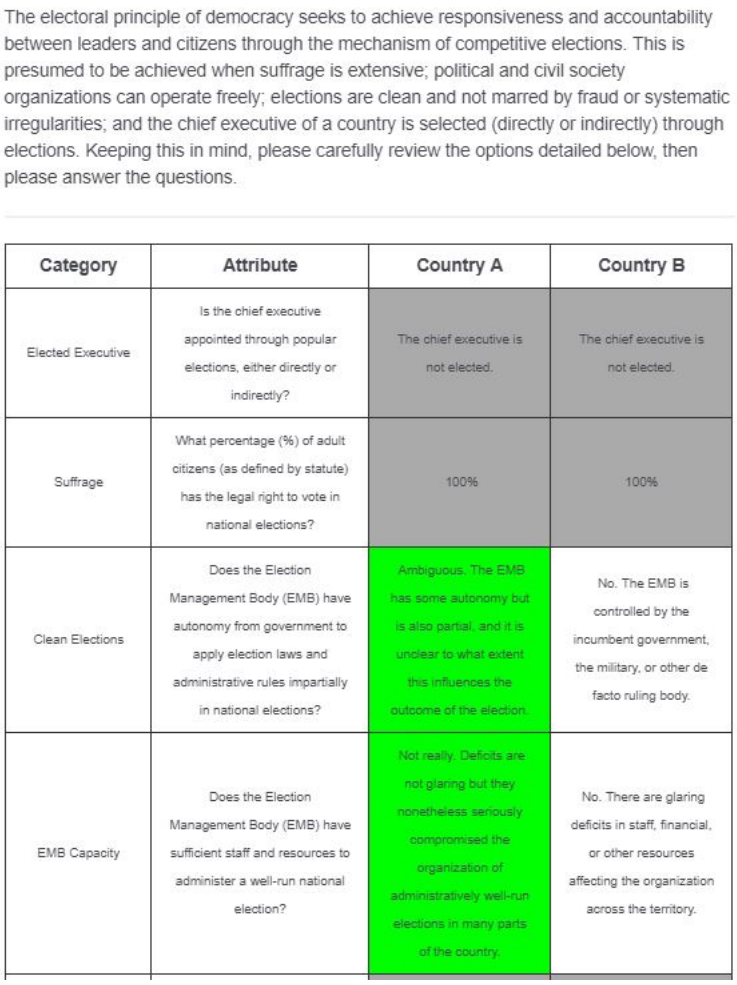
> The table on the next page describes political conditions in two countries. Entries in the column labeled "Category" identify the group of political indicators. Entries in the column labeled "Attribute" list the specific political indicators in that category, and entries in the remaining two columns provide information about how the two countries rate for each indicator. For each indicator, the country that is more democratic is highlighted in **GREEN**. For indicators where the countries have the same level of democracy, both countries will be highlighted in **GRAY**.
>
> After reviewing the information about these countries, you will be asked to evaluate **which country has more electoral democracy overall**. This is a question about the opinion you formed based on the information in the table, not a question with a right or wrong answer. If you are unsure or if you do not believe there is a difference between the two countries, select one of those choices. Finally, if you believe one country is more democratic than the other, use the slide bar that appears at the bottom of this page to estimate **how much more electoral democracy** that country has.

After these initial instructions, respondents are presented with the following question, and the table of categories, attributes, and descriptions of grades, all based on electoral democracy sub-components measured as part of the V-Dem survey. Table 1 provides the

[12] Initially, we allowed respondents to complete at most 35 tasks, but we reduced this to 30 tasks after two days, and to 20 tasks about halfway through the course of the experiment. These reductions were based on RA reports of the gross number of tasks completed by participants–without accompanying information about treatment conditions—and were adjusted to keep the overall spending on participant compensation within budget.

categories and attributes presented on the survey and provides a mockup of the interface.[13] Figure 2 shows a mock-up of part of the table, as viewed by respondents.

> The electoral principle of democracy seeks to achieve responsiveness and accountability between leaders and citizens through the mechanism of competitive elections. This is presumed to be achieved when suffrage is extensive; political and civil society organizations can operate freely; elections are clean and not marred by fraud or systematic irregularities; and the chief executive of a country is selected (directly or indirectly) through elections. Keeping this in mind, please carefully review the options detailed below, then please answer the questions.



The electoral principle of democracy seeks to achieve responsiveness and accountability between leaders and citizens through the mechanism of competitive elections. This is presumed to be achieved when suffrage is extensive; political and civil society organizations can operate freely; elections are clean and not marred by fraud or systematic irregularities; and the chief executive of a country is selected (directly or indirectly) through elections. Keeping this in mind, please carefully review the options detailed below, then please answer the questions.

Category	Attribute	Country A	Country B
Elected Executive	Is the chief executive appointed through popular elections, either directly or indirectly?	The chief executive is not elected.	The chief executive is not elected.
Suffrage	What percentage (%) of adult citizens (as defined by statute) has the legal right to vote in national elections?	100%	100%
Clean Elections	Does the Election Management Body (EMB) have autonomy from government to apply election laws and administrative rules impartially in national elections?	Ambiguous. The EMB has some autonomy but is also partial, and it is unclear to what extent this influences the outcome of the election.	No. The EMB is controlled by the incumbent government, the military, or other de facto ruling body.
EMB Capacity	Does the Election Management Body (EMB) have sufficient staff and resources to administer a well-run national election?	Not really. Deficits are not glaring but they nonetheless seriously compromised the organization of administratively well-run elections in many parts of the country.	No. There are glaring deficits in staff, financial, or other resources affecting the organization across the territory.

Figure 2: Partial view of the bottom-up table.

There are 16 variables in the V-Dem dataset that map into the concept of electoral democracy. As described above, these variables fall in four categories: elected executive,

[13]The mockup includes example country values for the first two attributes. In practice, all values are filled in. All these values are drawn from the wording of ordinal response levels in the V-Dem codebook.

Table 1: V-Dem Indicators of Electoral Democracy

Category	Attribute	Country A	Country B
Elected Executive	Is the chief executive appointed through popular elections, either directly or indirectly?	The chief executive is not appointed through popular elections, either directly or indirectly	The chief executive is not appointed through popular elections, either directly or indirectly
Suffrage	What percentage (%) of adult citizens (as defined by statute) has the legal right to vote in national elections?	80%	20%
Clean Elections	Does the Election Management Body (EMB) have autonomy from government to apply election laws and administrative rules impartially in national elections?	...	...
EMB Capacity	Does the Election Management Body (EMB) have sufficient staff and resources to administer a well-run national election?		
Election Voter Registry	In this national election, was there a reasonably accurate voter registry in place and was it used?		
Election Vote Buying	In this national election, was there evidence of vote and/or turnout buying?		
Election Other Voting Regularities	In this national election, was there evidence of other intentional irregularities by incumbent and/or opposition parties, and/or vote fraud?		
Election Government Intimidation	In this national election, were opposition candidates/parties/campaign workers subjected to repression, intimidation, violence, or harassment by the government, the ruling party, or their agents?		
Election Other Voting Irregularities	In this national election, was the campaign period, election day, and post-election process free from other types (not by the government, the ruling party, or their agents) of violence related to the conduct of the election and the campaigns (but not conducted by the government and its agents)?		
Election Free and Fair	Taking all aspects of the pre-election period, election day, and the post-election process into account, would you consider this national election to be free and fair?		
Party Ban	Are any parties banned?		
Barriers to Parties	How restrictive are the barriers to forming a party?		
Opposition Parties Autonomous	Are opposition parties independent and autonomous of the ruling regime?		
Elections Multiparty	Was this national election multiparty?		
CSO Entry and Exit	To what extent does the government achieve control over entry and exit by civil society organizations (CSOs) into public life?		
CSO Repression	Does the government attempt to repress civil society organizations (CSOs)?	...	...

suffrage, clean elections, and freedom of association. These indicators were measured through the main V-Dem survey of country and regional experts. The questions associated with these indicators are presented in the "attribute" column of Table 1. The elected executive question has two possible answers,[14] and all other questions have five possible answers,[15] so we cannot practically randomize over all combinations of all variables. We therefore limit the possible treatments to a subset of the combinations of these variables that we observe in the V-Dem data.[16] For questions other than suffrage, the V-Dem codebook provides the English descriptions of ordinal answer categories that we use to populate the interface. The relevant variables are (in the order presented in table 1):

- Executive Elected
  - v2x_accex
- Suffrage
  - v2elsuffrage
- Clean Elections
  - v2elembaut
  - v2elembcap
  - v2elrgstry
  - v2elvotbuy
  - v2elirreg
  - v2elintim
  - v2elpeace
  - v2elfrfair
- Freedom of Association
  - v2psparban
  - v2psbars
  - v2psoppaut
  - v2elmulpar
  - v2cseeorgs
  - v2csreprss

[14]The elected executive question has possible answers between 0 (no) and 1 (yes). However, the vast majority of country years produces a yes or no answer, so we limit the analysis these types of country years.

[15]Suffrage is given as the percentage value in the dataset, such as 20% or 98%.

[16]To be precise, for a given country year in the data set, we use the maximum posterior probability answer for each question. See the sampling section for details on which countries-years we used.

At the bottom of the screen presenting the conjoint table we provide two feedback mechanisms. First, respondents answer the question: "Which of these countries is more democratic?", choosing from the following responses:

- Country A
- Country B
- Neither

Second, they are presented with the prompt: "Please rank each country's level of electoral democracy on a scale of 1 (least democratic) to 10 (most democratic)." This prompt is followed by slider bars for both Country A and Country B. See figure 3 for a screenshot of these questions.

Figure 3: Conjoint table questions.

At the bottom of each task page an option states, "To skip the remaining expert tasks and go to the final demographic questions and receive payment, click here." The link allows respondents to end the task questions and redirects them to the covariates questions. Additionally, respondents are provided with a link at the bottom of each page that allows them to exit and return to the survey at any point during the 72-hour window.

3.1.1 Conjoint Variables

For the variables with ordinal values, the following descriptors are used to populate the conjoint table, with the exception of suffrage, for which we provide percentages. The variables are taken from version 7 of the V–Dem dataset.

- v2x_accex_ord
  - 0: The chief executive is not appointed through popular elections (either directly or indirectly).

  - 1: The chief executive is appointed through popular elections (either directly or indirectly).

- v2elembaut_ord
  - 0: No. The EMB is controlled by the incumbent government, the military, or other de facto ruling body.
  - 1: Somewhat. The EMB has some autonomy on some issues but on critical issues that influence the outcome of elections, the EMB is partial to the de facto ruling body.
  - 2: Ambiguous. The EMB has some autonomy but is also partial, and it is unclear to what extent this influences the outcome of the election.
  - 3: Almost. The EMB has autonomy and acts impartially almost all the time. It may be influenced by the de facto ruling body in some minor ways that do not influence the outcome of elections.
  - 4: Yes. The EMB is autonomous and impartially applies elections laws and administrative rules.

- v2elembcap_ord
  - 0: No. There are glaring deficits in staff, financial, or other resources affecting the organization across the territory.
  - 1: Not really. Deficits are not glaring but they nonetheless seriously compromised the organization of administratively well-run elections in many parts of the country.
  - 2: Ambiguous. There might be serious deficiencies compromising the organization of the election but it could also be a product of human errors and co-incidence or other factors outside the control of the EMB.
  - 3: Mostly. There are partial deficits in resources but these are neither serious nor widespread.
  - 4: Yes. The EMB has adequate staff and other resources to administer a well-run election.

- v2elrgstry_ord
  - 0: No. There was no registry, or the registry was not used.
  - 1: No. There was a registry but it was fundamentally flawed (meaning 20% or more of eligible voters could have been disenfranchised or the outcome could have been affected significantly by double-voting and impersonation).
  - 2: Uncertain. There was a registry but it is unclear whether potential flaws in the registry had much impact on electoral outcomes.
  - 3: Yes, somewhat. The registry was imperfect but less than 10% of eligible voters may have been disenfranchised, and double-voting and impersonation could not have affected the results significantly.

  - 4: Yes. The voter registry was reasonably accurate (less than 1% of voters were affected by any flaws) and it was applied in a reasonable fashion.

- v2elvotbuy_ord
  - 0: Yes. There was systematic, widespread, and almost nationwide vote/turnout buying by almost all parties and candidates.
  - 1: Yes, some. There were non-systematic but rather common vote-buying efforts, even if only in some parts of the country or by one or a few parties.
  - 2: Restricted. Money and/or personal gifts were distributed by parties or candidates but these offerings were more about meeting an 'entry-ticket' expectation and less about actual vote choice or turnout, even if a smaller number of individuals may also be persuaded.
  - 3: Almost none. There was limited use of money and personal gifts, or these attempts were limited to a few small areas of the country. In all, they probably affected less than a few percent of voters.
  - 4: None. There was no evidence of vote/turnout buying.

- v2elirreg_ord
  - 0: Yes. There were systematic and almost nationwide other irregularities.
  - 1: Yes, some. There were non-systematic, but rather common other irregularities, even if only in some parts of the country.
  - 2: Sporadic. There were a limited number of sporadic other irregularities, and it is not clear whether they were intentional or disfavored particular groups.
  - 3: Almost none. There were only a limited number of irregularities, and many were probably unintentional or did not disfavor particular groups' access to participation.
  - 4: None. There was no evidence of intentional other irregularities. Unintentional irregularities resulting from human error and/or natural conditions may still have occurred.

- v2elintim_ord
  - 0: Yes. The repression and intimidation by the government or its agents was so strong that the entire period was quiet.
  - 1: Yes, frequent: There was systematic, frequent and violent harassment and intimidation of the opposition by the government or its agents during the election period.
  - 2: Yes, some. There was periodic, not systematic, but possibly centrally coordinated – harassment and intimidation of the opposition by the government or its agents.

  - 3: Restrained. There were sporadic instances of violent harassment and intimidation by the government or its agents, in at least one part of the country, and directed at only one or two local branches of opposition groups.
  - 4: None. There was no harassment or intimidation of opposition by the government or its agents, during the election campaign period and polling day.

- v2elpeace_ord
  - 0: No. There was widespread violence between civilians occurring throughout the election period, or in an intense period of more than a week and in large swaths of the country. It resulted in a large number of deaths or displaced refugees.
  - 1: Not really. There were significant levels of violence but not throughout the election period or beyond limited parts of the country. A few people may have died as a result, and some people may have been forced to move temporarily.
  - 2: Somewhat. There were some outbursts of limited violence for a day or two, and only in a small part of the country. The number of injured and otherwise affected was relatively small.
  - 3: Almost. There were only a few instances of isolated violent acts, involving only a few people; no one died and very few were injured.
  - 4: Peaceful. No election-related violence between civilians occurred.

- v2elfrfair_ord
  - 0: No, not at all. The elections were fundamentally flawed and the official results had little if anything to do with the 'will of the people' (i.e., who became president; or who won the legislative majority).
  - 1: Not really. While the elections allowed for some competition, the irregularities in the end affected the outcome of the election (i.e., who became president; or who won the legislative majority).
  - 2: Ambiguous. There was substantial competition and freedom of participation but there were also significant irregularities. It is hard to determine whether the irregularities affected the outcome or not (as defined above).
  - 3: Yes, somewhat. There were deficiencies and some degree of fraud and irregularities but these did not in the end affect the outcome (as defined above).
  - 4: Yes. There was some amount of human error and logistical restrictions but these were largely unintentional and without significant consequences.

- v2psparban_ord
  - 0: Yes. All parties except the state-sponsored party (and closely allied parties) are banned.
  - 1: Yes. Elections are non-partisan or there are no officially recognized parties.
  - 2: Yes. Many parties are banned.

  - 3: Yes. But only a few parties are banned.
  - 4: No. No parties are officially banned.

- v2psbars_ord
  - 0: Parties are not allowed.
  - 1: It is impossible, or virtually impossible, for parties not affiliated with the government to form (legally).
  - 2: There are significant obstacles (e.g. party leaders face high levels of regular political harassment by authorities).
  - 3: There are modest barriers (e.g. party leaders face occasional political harassment by authorities).
  - 4: There are no substantial barriers.

- v2psoppaut_ord
  - 0: Opposition parties are not allowed.
  - 1: There are no autonomous, independent opposition parties. Opposition parties are either selected or co-opted by the ruling regime.
  - 2: At least some opposition parties are autonomous and independent of the ruling regime.
  - 3: Most significant opposition parties are autonomous and independent of the ruling regime.
  - 4: All opposition parties are autonomous and independent of the ruling regime.

- v2elmulpar_ord
  - 0: No. No-party or single-party and there is no meaningful competition (includes situations where a few parties are legal but they are all de facto controlled by the dominant party).
  - 1: Not really. No-party or single-party (defined as above) but multiple candidates from the same party and/or independents contest legislative seats or the presidency.
  - 2: Constrained. At least one real opposition party is allowed to contest but competition is highly constrained – legally or informally.
  - 3: Almost. Elections are multiparty in principle but either one main opposition party is prevented (de jure or de facto) from contesting, or conditions such as civil unrest (excluding natural disasters) prevent competition in a portion of the territory.
  - 4: Yes. Elections are multiparty, even though a few marginal parties may not be permitted to contest (e.g. far-right/left extremist parties, anti-democratic religious or ethnic parties).

- v2cseeorgs_ord
  - 0: Monopolistic control. The government exercises an explicit monopoly over CSOs. The only organizations allowed to engage in political activity such as endorsing parties or politicians, sponsoring public issues forums, organizing rallies or demonstrations, engaging in strikes, or publicly commenting on public officials and policies are government-sponsored organizations. The government actively represses those who attempt to defy its monopoly on political activity.
  - 1: Substantial control. The government licenses all CSOs and uses political criteria to bar organizations that are likely to oppose the government. There are at least some citizen-based organizations that play a limited role in politics independent of the government. The government actively represses those who attempt to flout its political criteria and bars them from any political activity.
  - 2: Moderate control. Whether the government ban on independent CSOs is partial or full, some prohibited organizations manage to play an active political role. Despite its ban on organizations of this sort, the government does not or cannot repress them, due to either its weakness or political expedience.
  - 3: Minimal control. Whether or not the government licenses CSOs, there exist constitutional provisions that allow the government to ban organizations or movements that have a history of anti-democratic action in the past (e.g. the banning of neo-fascist or communist organizations in the Federal Republic of Germany). Such banning takes place under strict rule of law and conditions of judicial independence.
  - 4: Unconstrained. Whether or not the government licenses CSOs, the government does not impede their formation and operation unless they are engaged in activities to violently overthrow the government.
- v2csreprss_ord
  - 0: Severely. The government violently and actively pursues all real and even some imagined members of CSOs. They seek not only to deter the activity of such groups but to effectively liquidate them. Examples include Stalinist Russia, Nazi Germany, and Maoist China.
  - 1: Substantially. In addition to the kinds of harassment outlined in responses 2 and 3 below, the government also arrests, tries, and imprisons leaders of and participants in oppositional CSOs who have acted lawfully. Other sanctions include disruption of public gatherings and violent sanctions of activists (beatings, threats to families, destruction of valuable property). Examples include Mugabe's Zimbabwe, Poland under Martial Law, Serbia under Milosevic.
  - 2: Moderately. In addition to material sanctions outlined in response 3 below, the government also engages in minor legal harassment (detentions, short-term incarceration) to dissuade CSOs from acting or expressing themselves. The government may also restrict the scope of their actions through measures that restrict association of civil society organizations with each other or political parties, bar

civil society organizations from taking certain actions, or block international contacts. Examples include post-Martial Law Poland, Brazil in the early 1980s, the late Franco period in Spain.

- 3: Weakly. The government uses material sanctions (fines, firings, denial of social services) to deter oppositional CSOs from acting or expressing themselves. They may also use burdensome registration or incorporation procedures to slow the formation of new civil society organizations and sidetrack them from engagement. The government may also organize Government Organized Movements or NGOs (GONGOs) to crowd out independent organizations. One example would be Singapore in the post-Yew phase or Putin's Russia.
- 4: No. Civil society organizations are free to organize, associate, strike, express themselves, and to criticize the government without fear of government sanctions or harassment.

3.2 Hypotheses

We hypothesize, broadly, that we will be able to use information from CE's paired rankings of cases to construct an inductive electoral democracy index that closely mirrors the index in the V-Dem data (v2x_EDcomp_thick). We have two strong, and one weak version of this hypothesis.

H29 (Strong): *Our inductive electoral democracy index will rank-correlate with V-Dem's electoral democracy index at Kendall's* $\tau > 0.9$.

H30 (Strong): *Our inductive electoral democracy index will exhibit a rank-correlation (measured by Kendall's* τ*) similar to that between draws from the posterior and the point estimates of V-Dem's electoral democracy index.*

H31 (Weak): *Our inductive electoral democracy index will rank-correlate with V-Dem's electoral democracy index at Kendall's* $\tau > 0.8$.

We have weak expectations about which variables will drive experts' ranking decisions. Nonetheless, we posit two hypotheses:

H32: *Whether or not the executive is elected will have a particularly large influence on ranking decisions.*

H33: *Whether elections are free and fair will have a particularly large influence on ranking decisions.*

Finally, we hypothesize that we can use machine learning methods to take information from expert rankings on a limited number of cases, and extrapolate it to new ones, producing out-of-sample classifiers that generate rankings indistinguishable from V-Dem's bespoke index:

H34: *We expect H29, H30, and H31 to hold for an index produced by learning from the expert rankings and using a learned function to classify cases—specifically, all countries for which there exists a electoral democracy scores, for the year 2000, in V-Dem v8—not directly rated by the experts.*

H35: *We expect H29, H30, and H31 to hold for an index produced by learning from the expert scores and using a learned function to classify cases—specifically, all countries for which there exists a electoral democracy scores in V-Dem v8—not directly scored by the experts.*

3.3 Sampling

3.3.1 Participants

V-Dem CEs who consent to participate in the study enter this condition with a probability of 0.5.

3.3.2 Country Pairs

We draw countries from version 7 of the V-Dem dataset, and use point estimates from the public data to determine ordinal levels of country characteristics. For most variables this means using the ordinal scale versions of the variables with the "_ord" suffix. The two exceptions, as noted above, are the elected executive and suffrage variables. We draw countries to compare, and populate the table, as follows:

1. We limit the potential sample to observations in the year 1966.
2. We drop cases with missing data for the relevant variables.
3. We drop cases with v2x_accex values other than 0 or 1.
4. We draw a random subsample of 85 countries.
5. We include all of the close (v2x_polyarchy scores within 0.2 of one another) country pairs.
6. We add randomly selected, without replacement, far (v2x_polyarchy scores greater than 0.2 from one another) pairs until the total sample contains 70% close cases.

In practice, this procedure produced a sample of 1177 pairs, covering 84 countries, from which we randomly sample during the task. The countries include: Afghanistan, Albania, Algeria, Angola, Argentina, Australia, Barbados, Belgium, Benin, Bhutan, Bolivia, Brazil, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Chad, Chile, China, Colombia, Republic of the Congo, Costa Rica, Cyprus, Czech Republic, Denmark, Ecuador, Egypt, Eritrea, Ethiopia, Finland, France, German Democratic Republic, Germany, Ghana, Greece, Guinea, Guinea-Bissau, Guyana, Honduras, Hungary, Indonesia, Israel, Italy, Ivory Coast, Jamaica, Japan, Jordan, Korea, North, Lesotho, Liberia, Libya, Madagascar, Mexico, Mongolia, Morocco, Mozambique, Namibia, Nepal, Netherlands, New Zealand, Norway,

Pakistan, Palestine/Gaza, Papua New Guinea, Paraguay, Philippines, Russia, Rwanda, Somalia, South Africa, South Yemen, Swaziland, Switzerland, Syria, Taiwan, Togo, Turkey, Uganda, United Kingdom, United States, Democratic Republic of Vietnam, Republic of Vietnam, and Zanzibar.

3.4 Planned Analyses

3.4.1 Scale Retrieval

To score the countries, we will use the method developed in Schnakenberg & Penn (2014), creating an inductive democracy index for the 84 countries in our sample. The method is a generalization of Bradley-Terry models for paired comparisons that is grounded in social choice theory.

We will use simple rank correlations to investigate the ability of our inductive electoral democracy index to mimic a more traditional approach, computing Kendall's τ statistics for V-Dem's electoral democracy (v2x_EDcomp_thick), V-Dem's polyarchy (v2x_polyarchy), Polity (and sub-components), Freedom House (and sub-components), and the Unified Democracy scores. We will use these correlation coefficients to descriptively analyze the quality of our inductive index. We will also plot our inductive index against these established indeces, and visually identify outliers and notably misclassified cases.

We conduct the following hypothesis tests of scale retrieval:

- For H29 and H31, we will compute Kendall's τ, for the 84 cases in our sample, between our inductive democracy index and the posterior draws for V-Dem v7's v2x_EDcomp_thick index, and test whether or not 95% HPD intervals for rank correlations exceed 0.9 and 0.8, respectively.

- For H30 we will compute Kendall's τ between each draw from the posterior and point estimates (posterior medians) of v2x_EDcomp_thick. We will then compute the 95% HPD for this distribution of τ statistics, and determine if the τ computed above falls within this HPD interval. This tests whether or not our inductively contstructed ranking is plausibly drawn from the same distribution as V-Dem's measure.

3.4.2 Conjoint Analysis

We will conduct two analyses of the conjoint experiment to determine which attributes drive CEs' democracy rankings. All analyses will omit coders who indicated that they consulted the V-Dem dataset during the experiment.

First, using both rank orderings and relative scores, we will conduct regression analyses to identify key determinants of rankings. To the extent possible, we will use Hainmueller, Hopkins & Yamamoto's (2014) technique for estimating average marginal component effects (AMCEs) under conditionally independent randomization. Nonetheless, because we designed primarily to recover an inductive electoral democracy index for true cases, and wanted to maximize data comparing these cases, we only truly randomize at the pair level, and all components are highly conditional, because they reflect only externally valid combinations. Thus we expect that we will need to satisfice and assume away higher order conditional

relationships between components in order to produce estimates, trading some potential bias for tractability.

Second, we will fit random forests (Breiman 2001) to both the pairwise ranks and 1–10 democracy scores that participants provide. We will include a vector of component dummies—for both cases in the ranking model, for the case in question in the democracy score model—as the only predictors in the random forests. We will construct 70% training, 20% testing, and 10% tuning datasets, drawn randomly from the available data, for both models.

We will conduct the following hypothesis tests:

- For H32, we will test the null hypothesis that the AMCE for elected executive is greater than all of the other case attribute AMCEs, save that for clean elections.
- For H32 we will test the expectation that dropping elected executive from the forests increases in-sample mean squared error more than all other variables save clean elections.
- For H33, we will test the null hypothesis that the AMCE for clean elections is greater than all of the other case attribute AMCEs, save that for elected executive.
- For H33 we will test the expectation that dropping clean elections from the forests increases in-sample mean squared error more than all other variables save elected executive.

We will conduct a sensitivity analysis, using the random forests, to ensure that we have sufficient variation in our data to effectively identify influential case characteristics. In particular, we will plot percent increase in MSE, when each variable is dropped, against a measure of dispersion for each of the ordinal variables in the analysis. Specifically, the consensus level of an ordinal variable X is

$$\text{Cns}(X) = 1 + \sum_{i=1}^{k} p_i \log_2 \left(\frac{|X_i - \mu_X|}{d_X} \right) \tag{1}$$

where p_i is the proportion of cases in category i, μ_X is the mean of X, and $d_X = X_{\max} - X_{\min}$ (Tastle & Wierman 2007). This measure takes a value of zero when X exhibits even dispersion across ordinal categories and a value of one when observations cluster at a single level. We will use this plot to evaluate the extent to which each variable's (lack of) influence on the MSE of the model is being driven by relative variability.

3.4.3 Aggregation Rule Learning/Extrapolation

We will attempt to use the random forests to develop an automated approach to electoral democracy scale assignment. We will first heuristically[17] evaluate the out-of-sample prediction accuracy of the two random forests, using the test sets. Next, we will apply the fitted pairwise-scoring forest to the V-Dem v8 dataset, scoring every pairwise country comparison

[17] We do not pre-specify what "good" is here, but we do pre-specify how we will evaluate prediction quality.

for the year 2000. We will feed these ranks into Schnakenberg & Penn's (2014) ranking model, generating an index. We will test H34 by replicating the analysis described in section 3.4.1, substituting our forest-produced index for our expert-produced index, and focusing on cases from 2000. This is a hard test, because cases in 2000 look quite different, on average, than cases in 1966.

Extrapolating from the pairwise model is computationally costly because the number of potential comparisons is large. We will therefore replicate this procedure using the random regression forest—which is fitted to CE's 1–10 scores, rather than to their pairwise ranks—to produce scores directly for each case, obviating the need to evaluate pairs, and omitting the Schnakenberg & Penn (2014) step, to see if we can achieve a reasonable index through this simpler technique. Because this approach obviates the need to generate pairs, we can generate scores for every case in the v8 dataset that contains an electoral demorcacy score (and therefore, the necessary sub-components) in a computationally tractable fashion. Thus we can do a full-scale comparison between these inductively generated scores and the current index across the entire 1900–2017 period. We will test H35 by replicating the analysis described in section 3.4.1, substituting our forest-produced index for our expert-produced index, and focusing on the entire contemporary (post-1900) V-Dem dataset. Note that, while we state H35 as an affirmative hypothesis, we nonetheless would not be surprised to reject it, because assigning democracy scores on a near-interval scale is a more cognitively demanding task than making rank-order decisions.

3.4.4 Compliance Checks

A subset of comparisons feature pairs where one country weakly dominates the other, in the sense that all of its components are greater than or equal to its comparison case. We will run all analyses including all coders, and also dropping coders who miscode more than 5% of these "easy" cases with respect to ordering.[18]

4 Pairwise Comparisons

The third part of the study asks if we can use experts' pairwise rankings of cases with respect to specific V-Dem C indicators to recover information consistent with that produced by the standard process for elicitating ratings for C indicators.

Normally, 5+ V-Dem experts rate country-years on these indicators, using Likert responses. CEs primarily rate cases in their country of expertise, although they sometimes rate "bridge" cases for which they feel they possess sufficient expertise. CEs also only answer questions on surveys for which they feel they have sufficient expertise, although, in practice, CEs take 6 surveys, on average.[19] The V-Dem measurement model (MM) then processes these responses, and responses to anchoring vignettes, to aggregate ratings, while adjusting for differential item functioning (DIF), and random error (Pemstein, Marquardt, Tzelgov, Wang, , Krusell & Miri 2018).

[18] This includes pairs that are equal, which coders should indicate are equal.

[19] In other words, they are reasonably optimistic about their domain expertise.

Here we assign experts to cases and questions at random; while some experts may therefore consider country-years and questions within their core areas of expertise, most will not. But we ask CEs to do a less cognitively demanding task—pairwise ranking instead of Likert scale assignment—and ask many more coders to rank each pair than we can recruit to code cases on Likert scales. Thus we trade expertise for rating density, although our "crowds" is special population particularly well-suited to the task. We have two overarching research questions:

1. Can we use crowd-sourced pairwise comparisons from CEs to recover rank orderings of cases that are similar to those produced by the standard V-Dem rating process?

2. How do question, case, and coder characteristics affect coders' ability to rank-order cases accurately?

4.1 Protocol

If randomly assigned to receive this task, coders rank order pairs of cases on a given question. Specifically, we present coders with a question—and sets of Likert-scale response categories,[20] since we know that questions often pack a lot of information into their response categories—and two country-year cases. We then ask coders to indicate which country-year ranks higher on the given scale, along with the option to say that the two cases are equal.

Rather than making it possible for coders to see every possible country-year pair for every possible indicator, we restrict the set as follows:

- Years: 1916, 1946, 1966, 1992, 2001, 2011
- Countries: Senegal and Argentina
- Indicators: v2clkill, v2clslavem, v2juhcind, v2meharjrn, v2pepwrgen

The coder sees 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:

- **Question** Considering Argentina in 2012 and Chile in 2014, which had greater political killings?
- **Response**
  - Argentina in 2012
  - Chile in 2014
  - These two cases had the same level of political killings.

We present respondents with a repeated paired comparison task.[21] At the bottom of each task an option states, "To skip the remaining expert tasks and go to the final demographic

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

[21] As in the Bottom-Up task, we allowed respondents to complete at most 35 tasks, but we reduced this to 30 tasks after two days, and to 20 tasks about halfway through the course of the experiment.

questions and receive payment, click here." The link allows respondents to end the task questions and redirects them to the covariates questions. Additionally, respondents are provided with a link at the bottom of each page that allows them to exit and return to the survey at any point during the 72-hour window.

4.2 Hypotheses

As in the Bottom-Up study, we posit two strong hypotheses, and one weak hypothesis, about our ability to recover V-Dem scores using our alternative method:

H36 (Strong): *Our crowd-sourced indicators will rank-correlate with V-Dem's indicators at Kendall's* $\tau > 0.9$.

H37 (Strong): *Our crowd-sourced indicators will exhibit a rank-correlation (measured by Kendall's* τ*) similar to that between draws from the posterior and the point estimates of V-Dem's indicators.*

H38 (Weak): *Our crowd-sourced indicators will rank-correlate with V-Dem's indicators at Kendall's* $\tau > 0.8$.

Table 2: Selected V–Dem expert-coded variables

		Question complexity	
		Low	High
Issue complexity	Low	*Judicial independence* (v2juhcind)	*Journalist harassment* (v2meharjrn)
	High	*Gender equality* (v2pepwrgen)	*Forced labor* (v2clslavem)

We also expect characteristics of the task to influence expert performance. Each of these hypotheses relates to *accuracy*, defined as the extent to which rank orders provided by experts correspond with rank orders implied by the V-Dem measures. These hypotheses are drawn from (Marquardt, Pemstein, Sanhueza Petrarca, Seim, Wilson, Bernhard, Coppedge & Lindberg 2017), which provides theory and defines terms. Table 2 breaks questions down by issue and question complexity.

H39 (Recency): *We expect experts to be more accurate when rating more recent cases.*

H40 (Information Availability): *On average, we expect experts to be more accurate when rating Argentina (high information availability) than when rating Senegal (low information availability).*

H41 (Issue Complexity): *We expect experts to be more accurate when evaluating lower issue complexity questions. Specifically, we expect experts to be more accurate evaluating judicial independence (journalist harassment) than gender equality (forced labor).*

H42 (Question Complexity): *We expect experts to be more accurate when evaluating questions with lower question complexity. Specifically, we expect experts to be more accurate evaluating judicial independence (gender equality that journalist harassment (forced labor).*

H43 (Expertise): *We expect experts to be more accurate when rating cases in their region of expertise.*

4.3 Sampling

4.3.1 Participants

V-Dem CEs who consent to participate in the study enter this condition with a probability of 0.5.

4.3.2 Pairs

We select indicators, years, and countries fully at random (with replacement), excluding same-case pairs. In practice we eliminate same-case pairs by resampling the second country-year in the pair, up to two more times, given a match with the first country-year. Thus, it is technically possible for a respondent to rate a same-case pair, although highly unlikely.

4.4 Planned Analyses

4.4.1 Scale Retrieval

As in the Bottom-Up study, our first goal is scale retrieval. We will again use Schnakenberg & Penn's (2014) method to aggregate rankings for each indicator. Then, for each indicator included in the analysis, we will conduct the following hypothesis tests of scale retrieval:

- For H36 and H38, seperately for each indicator, we will compute Kendall's τ for the cases in our sample, between our crowd-sourced and the posterior draws for V-Dem's version of the indicator, and test whether or not 95% HPD intervals for rank correlations exceed 0.9 and 0.8, respectively.
- For H37, separately for each indicator, we will compute Kendall's τ between each draw from the posterior and point estimates (posterior medians) of V-Dem's version of the indicator. We will then compute the 95% HPD for this distribution of τ statistics, and determine if the τ computed above falls within this HPD interval. This tests whether or not our inductively contstructed ranking is plausibly drawn from the same distribution as V-Dem's measures.

We will conduct these tests both for the full set of cases, and strictly within countries, because the extant V-Dem measures provide a more plausible gold standard within countries than across. The key test is, therefore, within country.

4.4.2 Task & Accuracy

We will construct two measures of accuracy, at the rating level, one binary and one interval-valued:

1. A rating is accurate if its rank order matches that of the point estimates in the V-Dem dataset.

2. If a rating's rank-order does not match that of the point estimates in the V-Dem dataset $a_i = -1/d_i$, where d_i is the difference between the V-Dem point estimates for rating i.[22] Otherwise, $a_i = 1/d_i$.

We will run logit and OLS analyses for the two indicators of accuracy. Each analysis will take the forms:

$$a_i = f\left(\alpha_r + \beta c_i + \delta\frac{t_1 + t_2}{2} + \gamma x_i + \eta e_i\right) \tag{2}$$

and

$$a_i = f\left(\alpha + \beta c_i + \delta\frac{t_1 + t_2}{2} + \gamma x_i + \eta e_i + \lambda z_i\right), \tag{3}$$

where c_i contains two dummies for whether each country in the pair is Argentina, x_i is a vector of dummies, one each for the V-Dem variables, e_i is a count of the cases within the rating expert's region,[23] and z_i is a vector of rater-specific variables drawn from the post-experiment questionairre, specifically questions: 1–3, 8–10, and 11–12 (correct dummies). The first regression model includes expert-specific fixed effects a_r, while the second treats experts as exchangeable, but includes expert characteristics.

We expect every coefficient in β will be positive (H40), $\delta < 0$ (H39), the coefficients in x_i allow us to test H41 and H42, and we expect $e_i > 0$. We include z_i for descriptive purposes.

5 Post-Experiment Survey Questions

The following questions are asked of all CEs, regardless of treatment assignments. These questions were always asked after the bottom-up aggregation or pairwise comparisons tasks were completed. As the inducement experiment begins on the closing screen of the V–Dem update with the opt-in link, these questions are also all post-treatment covariates for the inducement experiment.

Before being asked these questions, there is an introductory screen that states, "Finally, we will ask you to answer a few questions about yourself. You will be paid $0.10 for each response in this section."

1. Spanish language (Spanish)

[22] These are country-time-pair-indicator-rater observations.

[23] Ranging from 0–2: 0 if the expert's region is outside Latin America and Africa, 1 if the two cases represent different countries and the expert specializes in Latin America and Africa, and 2 if the two cases represent the same country and the expert specializes in that region. Specialization is based on the expert's primary country of expertise in the V-Dem data.

- **Question** Are you able to read written materials (e.g., newspaper articles) in Spanish?
- **Response**
  (a) No
  (b) Yes
  (c) Don't know

2. French language (French)

   - **Question** Are you able to read written materials (e.g., newspaper articles) in French?
   - **Response**
     (a) No
     (b) Yes
     (c) Don't know

3. 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)

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 (v2zzpolcourses) [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. 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

7. 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

8. Data entry non-profit (non-profit)

   - **Question** What is the minimum amount you would accept to do data entry for a local for-profit company for one 8 hour day?
   - **Clarification** Please enter the amount you would accept in US dollars.
   - **Response** [INTEGER]

9. Data entry charity (charity)

   - **Question** What is the minimum amount you would accept to do data entry for a local charity for one 8 hour day?
   - **Clarification** Please enter the amount you would accept in US dollars.
   - **Response** [INTEGER]

10. Relative expertise (expertise)

    - **Question** Compared to others who consider similar topics, how would you rate your expertise?
    - **Response**
      (a) Above others

(b) Approximately the same as others
(c) Below others

For the next two questions, we would like to test their technological competence by asking them to copy and paste the URLs for two easily found websites and recording the time it takes to copy and paste each URL. Before being asked to copy and paste the URLs, there is an introductory screen that states, "On the following pages, you will be asked to find a website and copy and paste its URL. Please copy and paste the entire URL of the site into the text box provided."

11. URL 1 (url1)
    - **Question**Please copy and paste the URL for the homepage of Ambrose Alli University:
    - **Response** [TEXT BOX]
12. URL 2 (url2)
    - **Question** Please copy and paste the URL for the homepage of New Zealand's Ministry for Pacific Peoples:
    - **Response** [TEXT BOX]
13. 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
      (c) Don't know
14. Comments (comment)
15. **Question** Add here any comments you have about any of the previous questions.
    - **Response** [TEXT BOX]

References

Breiman, Leo. 2001. "Random forests." *Machine Learning* 45(1):5–32.

Hainmueller, Jens, Daniel J. Hopkins & Teppei Yamamoto. 2014. "Causal inference in conjoint analysis: understanding multi-dimensional preferences via stated preference experiments." *Political Analysis* 22(1):1–30.

Marquardt, Kyle L., Daniel Pemstein, Constanza Sanhueza Petrarca, Brigitte Seim, Steven Lloyd Wilson, Michael Bernhard, Michael Coppedge & Staffan I. Lindberg. 2017. "Experts, Coders, and Crowds: An Analysis of Substitutability." *V-Dem Institute Working Paper* 53.
URL: https://www.v-dem.net/media/filer_public/9e/81/9e81b209-4b20-4188-878b-9ed92781ff56/v-dem_working_paper_2017_53.pdf

Pemstein, Daniel, Kyle L Marquardt, Eitan Tzelgov, Yi-ting Wang, , Joshua Krusell & Farhad Miri. 2018. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." *Varieties of Democracy Institute Working Paper* 21(3rd ed).

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

Tastle, William J. & Mark J. Wierman. 2007. "Consensus and dissention: A measure of ordinal dispersion." *International Journal of Approximate Reasoning* 45:531–545.