Supplementary Appendix for "Information Equivalence in Survey Experiments"

Allan Dafoe, Baobao Zhang, and Devin Caughey

November 14, 2017

A Literature Review

A.1 Review of Articles in Top Journals

A.1.1 Classifying the Articles

We reviewed articles published in the *American Journal of Political Science*, the *American Political Science Review*, and *International Organization* between 2002 and 2015. Candidate articles were selected through the following search methodology: search for the terms "survey" and "experiment" in each journal issue within the aforementioned time period[1]; check the methodology section (including any appendices) of each article returned by this search to determine whether the article employed a survey experiment; if so, include the article. We used the following definition for survey experiments:

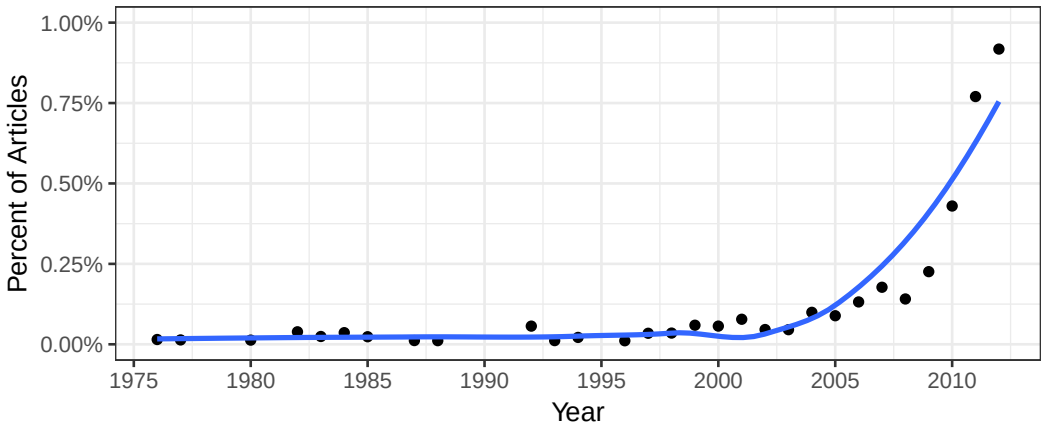
> A survey experiment systematically varies one or more elements of a survey across subjects and assesses the effect of that variation on one or more measured outcomes. Typically, subjects (respondents) are randomly assigned to either a treatment group (of which there may be more than one) or a control group. The crucial, defining element of a survey experiment is that it manipulates some aspect of the survey protocol (**marsden2010handbook**).

These could include vignette/scenario-based experiments, framing experiments, list experiments, or other types of survey experiments. The following attributes of selected articles were then recorded in a master spreadsheet: title, author(s), year of publication, journal, survey experiment type (scenario, framing, list, other), hypotheses tested, the experimental manipulation, and the sample size and subject pool (including survey service used).

We then examined all articles coded as employing scenario-based survey experiments (labeled as "Vignette"), and performed several additional coding tasks:

[1] We used JSTOR for articles published before 2014. For articles published after 2013, which are not available in JSTOR, we searched each issue of the three journals and supplemented our manual inspection with Google Scholar search.

Figure 1: Growth of Survey Experiments in Political Science



The graph above depicts the percentage of articles in political science journals that mention "survey experiment" from 1976 to 2013 (JSTOR Data for Research).

A We determined whether the authors argued that scenario-based survey experiments achieve information equivalence and allow scholars to cleanly identify causal effects (of beliefs about a feature of the scenario), and recorded relevant quotes.

B We determined whether the authors explicitly acknowledged that survey experiments could have problems with internal validity, and collected relevant quotes.

C We determined whether the authors explicitly acknowledged that survey experiments could have problems related to IE as defined in our paper.

D We then recorded three new binary variables, based on the above:

(a) Confident of Internal Validity: If the author expressed confidence that survey experiments demonstrate information equivalence as described in part A, the articled was coded 1. If not, 0.

(b) Limits to Internal Validity: If the author expressed any concerns about the internal validity of survey experiments as described in part B, the article was coded 1. If not, 0.

(c) Survey Information Equivalence Violation: If the author expressed specific concerns about the possibility of information equivalence violation in scenario-covariates[2] as described in part C, the article was coded 1. If not, 0.[3]

E Potentially Experience Information Equivalence Violation: We coded for whether survey experiments in the article might experience the problem of information equivalence

[2]The authors did not need to use the language of "information equivalence violation." All that was required was that they acknowledged that respondents' beliefs about the scenario may be affected by the manipulation in undesired ways.

[3]Articles could be coded as 1 for (a) and 1 for (c), as some authors made competing statements about the validity of their survey experiments.

violation as outlined in our paper. We considered all survey experiment types when applicable.

F For scenario-based experiments ("vignette"), we created categorical variables to classify the types of claims researchers made and what types of experiments they were conducting:

   (a) Type of Causal Claim: The author could claim that her scenario-based experiment examined the effect of X or the effect of being described as X, both, or used the experiment as a measurement tool. For instance, if the author stated that her experiment looked at the effect of regime type on support for war, we coded it as "the effect of X." If the author stated that her experiment studied the effect of a candidate being described using stereotypes on respondents' support for the candidate, we coded it as "the effect of being described as X." We created a separate category called "measurement" for when researchers used experimental vignettes to measure political attitudes.

   (b) Type of Experiment: In scenario-based experiments, the experimenter could vary some characteristics of the scenario, presentation of information, or both. For instance, if the author varied the party of candidates or the regime type of an aggressor country, she manipulated characteristics of the scenario. If the author varied the language describing a person, country, idea, or object, she manipulated the presentation of information. One example of the latter manipulation would be the researcher describing a candidate as "a shady businessman" versus "a businessman who engages in unethical practices."

A.1.2 Summary Statistics

We present some summary statistics of the articles we reviewed. Because we do not want to critique individual researchers or teams of researchers, we provide aggregate level data. The data from our literature review suggest that most researchers who used scenario-based survey experiments were not concerned about limits to internal validity. Furthermore, only five out of 35 mentioned the possibility of information equivalence violation in their survey experiments.

Table 1: Summary Statistics from Literature Review

Type of Experiment	Framing	Vignette	Framing/Vignette	Other	Other/Vignette
Number of Articles	31	32	2	13	1

	Yes	No	Not Applicable
Potentially Experience Information Equivalence Violation	16	60	2
Express Confidence of Internal Validity	16	19	43
Express Limits to Internal Validity	10	25	43
Recognize Possibility of Information Equivalence Violation	5	30	43

Type of Causal Claim	Number of Articles
Effect of X	19
Effect of Being Described as X	1
Measurement	3
Effect of X / Effect of Being Described as X	5
Effect of X / Measurement	5

Type of Experiment	Number of Articles
Characteristics of Scenarios	21
Presentation of Information	2
Both	10

A.1.3 Articles Reviewed

We reviewed survey experiments published in the *American Journal of Political Science* (AJPS), the *American Political Science Review* (APSR), and *International Organization* (IO) between 2002 and 2015. The following table contains information about the 78 articles we reviewed.

Table 2: Survey Experiments Published in Top Political Science Journals

Title	Authors	Year	Journal	Type
Gender Stereotypes and Vote Choice	Kira Sanbonmatsu	2002	AJPS	Vignette
Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge	Darren W. Davis and Brian D. Silver	2003	AJPS	Other
When Do Welfare Attitudes Become Racialized? The Paradoxical Effects of Education	Christopher M. Federico	2004	AJPS	Vignette
Certainty or Accessibility: Attitude Strength in Candidate Evaluations	David A. M. Peterson	2004	AJPS	Vignette
Predisposing Factors and Situational Triggers: Exclusionary Reactions to Immigrant Minorities	Paul M. Sniderman, Louk Hagendoorn, and Markus Prior	2004	APSR	Vignette
Racial Resentment and White Opposition to Race-Conscious Programs: Principles or Prejudice?	Stanley Feldman and Leonie Huddy	2005	AJPS	Vignette
The Indirect Effects of Discredited Stereotypes in Judgments of Jewish Leaders	Adam J. Berinsky and Tali Mendelberg	2005	AJPS	Vignette

The "Race Card" Revisited: Assessing Racial Priming in Policy Contests	Gregory A. Huber and John S. Lapinski	2006	AJPS	Framing
Through a Glass and Darkly: Attitudes Towards International Trade and the Curious Effects of Issue Framing	Michael J. Hiscox	2006	IO	Framing
Identity Salience, Identity Acceptance, and Racial Policy Attitudes: American National Identity as a Uniting Force	John E. Transue	2007	AJPS	Framing
Beyond Negativity: The Effects of Incivility on the Electorate	Deborah Jordan Brooks and John G. Geer	2007	AJPS	Framing
Issue Definition, Information Processing, and the Politics of Global Warming	B. Dan Wood and Arnold Vedlitz	2007	AJPS	Framing
Designing and Analyzing Randomized Experiments: Application to a Japanese Election Survey Experiment	Yusaku Horiuchi, Kosuke Imai and Naoko Taniguchi	2007	AJPS	Other
When Race Matters and When It Doesn't: Racial Group Differences in Response to Racial Cues	Ismail K. White	2007	APSR	Framing
Domestic Audience Costs in International Relations: An Experimental Approach	Michael Tomz	2007	IO	Vignette
Opinion Taking within Friendship Networks	Suzanne L. Parker, Glenn R. Parker and James A. McCann	2008	AJPS	Other
Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills	Markus Prior and Arthur Lupia	2008	AJPS Other	
Attributing Blame: The Public's Response to Hurricane Katrina	Neil Malhotra and Alexander G. Kuo	2008	AJPS	Framing
What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat	Ted Brader, Nicholas A. Valentino and Elizabeth Suhay	2008	AJPS	Vignette / Framing
Framing Public Opinion in Competitive Democracies	Dennis Chong and James N. Druckman	2008	APSR	Framing
Challenges to the Impartiality of State Supreme Courts: Legitimacy Theory and "New-Style" Judicial Campaigns	James L. Gibson	2008	APSR	Vignette
Candidate Positioning and Voter Choice	Michael Tomz and Robert P. Van Houweling	2008	APSR	Vignette
The Multiple Effects of Casualties on Public Support for War: An Experimental Approach	Scott Sigmund Gartner	2008	APSR	Vignette
How Predictive Appeals Affect Policy Opinions	Jennifer Jerit	2009	AJPS	Framing
Source Cues, Partisan Identities, and Political Value Expression	Paul Goren, Christopher M. Federico and Miki Caul Kittilson	2009	AJPS	Framing
The Electoral Implications of Candidate Ambiguity	Michael Tomz and Robert P. Van Houweling	2009	APSR	Vignette
Dynamic Public Opinion: Communication Effects Over Time	Dennis Chong and James N. Druckman	2010	APSR	Other
Are Survey Experiments Externally Valid?	Jason Barabas and Jennifer Jerit	2010	APSR	Other/Vignette

Attitudes toward Highly Skilled and Low-skilled Immigration: Evidence from a Survey Experiment	Jens Hainmueller and Michael J. Hiscox	2010	APSR	Vignette
Electoral Incentives and Partisan Conflict in Congress: Evidence from Survey Experiments	Laurel Harbridge and Neil Malhotra	2011	AJPS	Framing
The Political Costs of Crisis Bargaining: Presidential Rhetoric and the Role of Party	Robert F. Trager and Lynn Vavreck	2011	AJPS	Vignette
Elite Influence on Public Opinion in an Informed Electorate	John G. Bullock	2011	APSR	Framing
Explaining Mass Support for Agricultural Protectionism: Evidence from a Survey Experiment During the Global Recession	Megumi Naoi and Ikuo Kume	2011	IO	Other
Emotional Substrates of White Racial Attitudes	Antoine J. Banks and Nicholas A. Valentino	2012	AJPS	Framing
Cognitive Biases and the Strength of Political Arguments	Kevin Arceneaux	2012	AJPS	Framing
Polarizing Cues	Stephen P. Nicholson	2012	AJPS	Framing
Taking Sides in Other People's Elections: The Polarizing Effect of Foreign Intervention	Daniel Corstange and Nikolay Marinov	2012	AJPS	Vignette
Social Welfare as Small-Scale Help: Evolutionary Psychology and the Deservingness Heuristic	Michael Bang Petersen	2012	AJPS	Vignette
How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation	Justin Grimmer, Solomon Messing, and Sean J. Westwood	2012	APSR	Other
A Source of Bias in Public Opinion Stability	James N. Druckman, Jordan Fein, and Thomas J. Leeper	2012	APSR	Framing
Politics in the Mind's Eye: Imagination as a Link between Social and Political Cognition	Michael Bang Petersen and Lene Aarøe	2012	APSR	Framing
Economic Explanations for Opposition to Immigration: Distinguishing between Prevalence and Conditional Impact	Neil Malhotra, Yotam Margalit and Cecilia Hyunjung Mo	2013	AJPS	Vignette
Working Twice as Hard to Get Half as Far: Race, Work Ethic, and America's Deserving Poor	Christopher D. DeSante	2013	AJPS	Vignette
Poverty and Support for Militant Politics: Evidence from Pakistan	Graeme Blair, C. Christine Fair, Neil Malhotra, Jacob N. Shapiro	2013	AJPS	Other
How Elite Partisan Polarization Affects Public Opinion Formation	James N. Druckman, Erik Peterson, and Rune Slothuus	2013	APSR	Framing
Public Opinion and the Democratic Peace	Michael R. Tomz and Jessica L. P. Weeks	2013	APSR	Vignette
Atomic Aversion: Experimental Evidence on Taboos, Traditions, and the Non-use of Nuclear Weapons	Daryl G. Press, Scott D. Sagan, and Benjamin A. Valentino.	2013	APSR	Vignette
Explaining Social Policy Preferences: Evidence from the Great Recession	Yotam Margalit	2013	APSR	Framing
Explaining Support for Combatants During Wartime: A Survey Experiment in Afghanistan	Jason Lyall, Graeme Blair, and Kosuke Imai	2013	APSR	Other

Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils	Thad Dunning and Janhavi Nilekani	2013	APSR	Vignette
Sensitivity to Issue Framing on Trade Policy Preferences: Evidence from a Survey Experiment	Martin Ardanaz, M. Victoria Murillo, and Pablo M. Pinto	2013	IO	Framing
International Law and Public Attitudes Toward Torture: An Experimental Study	Geoffrey P.R. Wallace	2013	IO	Vignette
Partisans in Robes: Party Cues and Public Acceptance of Supreme Court Decisions	Stephen P. Nicholson and Thomas G. Hansford	2014	AJPS	Framing
The Power of Partisanship in Brazil: Evidence from Survey Experiments	David Samuels and Cesar Zucco Jr.	2014	AJPS	Framing
The Conditionality of Vote Buying Norms: Experimental Evidence from Latin America	Ezequiel Gonzales Ocantos, Chad Kiewiet de Jonge, and David W. Nickerson	2014	AJPS	Vignette
Informing the Electorate? How Party Cues and Policy Information Affect Public Opinion about Initiatives	Cheryl Boudreau and Scott A. MacKenzie	2014	AJPS	Framing
Substituting the End for the Whole: Why Voters Respond Primarily to the Election-Year Economy	Andrew Healy and Gabriel S. Lenz	2014	AJPS	Vignette
Distorted Communication, Unequal Representation: Constituents Communicate Less to Representatives Not of Their Race	David E. Broockman	2014	AJPS	Other
Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys	Adam J. Berinsky, Michele F. Margolis, and Michael W. Sances	2014	AJPS	Vignette
Partisanship in a Social Setting	Samara Klar	2014	AJPS	Other
Structural Topic Models for Open-Ended Survey Responses	Margaret E. Roberts et al.	2014	AJPS	Framing
Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan	Graeme Blair, Kosuke Imai, and Jason Lyall	2014	AJPS	Other
Preferences for International Redistribution: The Divide over the Eurozone Bailouts	Michael M. Bechtel, Jens Hainmueller, and Yotam Margalit	2014	AJPS	Vignette
The Political Mobilization of Ethnic and Religious Identities in Africa	John F. McCauley	2014	APSR	Framing
False Commitments: Local Misrepresentation and the International Norms Against Female Genital Mutilation and Early Marriage	Karisa Cloward	2014	IO	Other
Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions	Stephen Chaudoin	2014	IO	Vignette
Decision Maker Preferences for International Legal Cooperation	Emilie M. Hafner-Burton, Brad L. LeVeck, David G. Victor and James H. Fowler	2014	IO	Vignette
Attacks without Consequence? Candidates, Parties, Groups and the Changing Face of Negative Advertising	Conor M. Dowling and Amber Wichowsky	2015	AJPS	Framing

Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment	Edmund J. Malesky, Dimitar D. Gueorguiev, and Nathan M. Jensen	2015	AJPS	Other
Chief Justice Roberts's Health Care Decision Disrobed: The Microfoundations of the Supreme Court's Legitimacy	Dino P. Christenson and David M. Glick	2015	AJPS	Framing
Responsibility Attribution for Collective Decision Makers	Raymond Duch, Wojtek Przepiorka, and Randolph Stevenson	2015	AJPS	Vignette
Explaining Explanations: How Legislators Explain their Policy Decisions and How Citizens React	Christian R. Grose, Neil Malhotra, and Robert Parks Van Houweling	2015	AJPS	Framing
Fear and Loathing Across Party Lines: New Evidence on Group Polarization	Shanto Iyengar and Sean J. Westwood	2015	AJPS	Vignette
Xenophobic Rhetoric and its Political Effects on Immigrants and their Co-Ethnic	Efrén O. Pérez	2015	AJPS	Framing
The Hidden American Immigration Consensus: A Conjoint Analysis of Attitudes Towards Immigrants	Jens Hainmueller and Daniel J. Hopkins	2015	AJPS	Vignette
Decomposing Audience Costs: Bringing the Audience Back Into Audience Cost Theory	Joshua D. Kertzer and Ryan Brutger	2015	AJPS	Vignette
Expressive Partisanship: Campaign Involvement, Political Emotion, and Partisan Identity	Leonie Huddy, Lilliana Mason, and Lene Aarøe	2015	APSR	Framing
Human Rights Organizations as Agents of Change: An Experimental Examination of Framing and Micromobilization	Kayla Jo McEntire, Michele Leiby, and Matthew Krain	2015	APSR	Framing
Race, Paternalism, and Foreign Aid: Evidence from U.S. Public Opinion	Andy Baker	2015	APSR	Vignette
Religious Social Identity, Religious Belief, and Anti-Immigration Sentiment	Pazit Ben-Nun Bloom, Gizem Arikan, and Marie Courtemanche	2015	APSR	Vignette / Framing
International Knowledge and Domestic Evaluations in a Changing Society: The Case of China	Haifeng Huang	2015	APSR	Other

B "Democratic Peace" Survey Experiment Details – Justifications for Placebo Test Questions

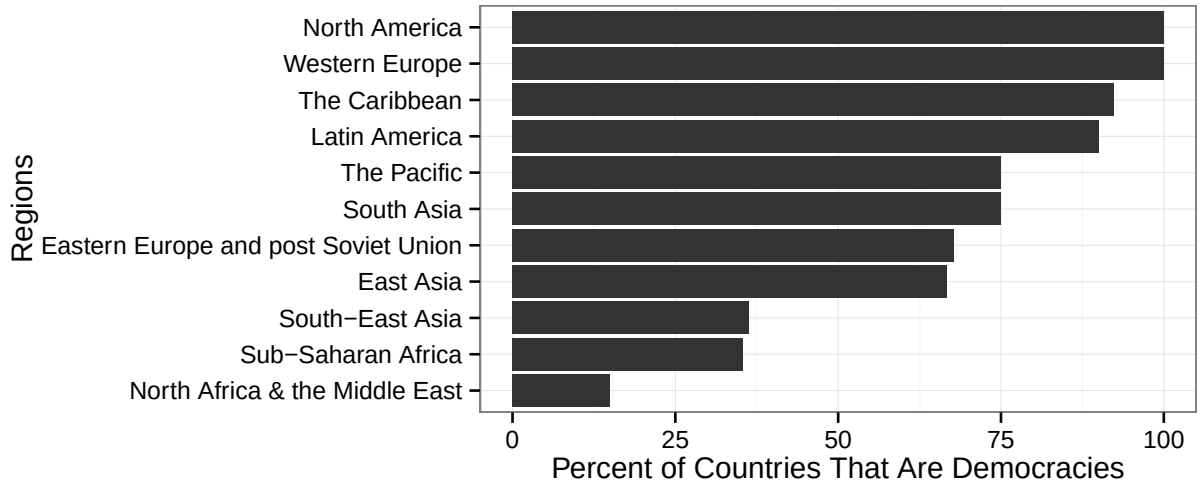
We selected placebo test variables by identifying real-world variables that show large and significant imbalances across regime types (see Table 3).[4] For our analysis, we used data from

[4]In our previous waves we selected placebo variables informally based on our intuitions. However, following the helpful comments of X on this point, for this wave we opted to select our placebos more formally by identifying real-world variables that show large and significant imbalances across regime types. This new more formal placebo selection process led us to remove placebo test questions regarding whether the country was English-speaking (insufficient imbalance) and whether the country had fought alongside the U.S. in the Iraq War, which we feared was too idiosyncratic. It also led us to include placebo test questions regarding the country's oil reserves, racial makeup, and joint military exercise with the U.S. which were sufficiently imbalanced; oil reserves and racial makeup are unlikely to be affected

the Quality of Government (GOG) Basic dataset[5], the Correlates of War (COW) formal alliance dataset[6], the COW trade dataset[7], the COW National Material Capabilities dataset[8], the CIA World Factbook Ethnic Group dataset[9], Vito D'Orazio's Joint Military Exercise dataset[10], and U.S. Department of Commerce Bureau of Economic Analysis's Foreign Direct Investment dataset[11].

First, we showed that geographic regions should be included as a placebo question because democracies and non-democracies are distributed differently across regions. Figure 2 displays the percent of countries that are democracies in the ten regions of the world. We defined democracy using the variable `chga_demo` from QOG, which is a binary coding of democracy/non-democracy from the Cheibub et al. 2010 dataset.[12]

Figure 2: Democracies in Regions of the World



by regime-type; joint military exercise is included as characteristic related but not identical to military alliance.

[5]Dahlberg, Stefan, Sören Holmberg, Bo Rothstein, Felix Hartmann & Richard Svensson. 2015. The Quality of Government Basic Dataset, version Jan15. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se.

[6]Gibler, Douglas M. 2009. International military alliances, 1648-2008. CQ Press.

[7]Barbieri, Katherine and Omar Keshk. 2012. Correlates of War Project Trade Data Set Codebook, Version 3.0. Online: http://correlatesofwar.org.

[8]Singer, J. David. "Reconstructing the Correlates of War Dataset on Material Capabilities of States, 1816-1985." International Interactions 14: 115-32. Correlates of War Project National Material Capabilities Codebook, Version 4.0. http://www.correlatesofwar.org/data-sets/national-material-capabilities

[9]Ethnic Groups Dataset, CIA World Factbook, 2000. https://www.cia.gov/Library/publications/the-world-factbook/fields/2075.html

[10]http://vitodorazio.weebly.com/data.html

[11]Foreign Direct Investment in the U.S.: Balance of Payments and Direct Investment Position Data. 2015. U.S. Department of Commerce Bureau of Economic Analysis.

[12]Cheibub, José Antonio, Jennifer Gandhi, and James Raymond Vreeland. "Democracy and dictatorship revisited." *Public Choice* 143.1-2 (2010): 67-101.

In the analysis of our survey experiment, we focused on four regions that exhibit the most imbalance between regime types: Western Europe, North America, Sub-Saharan Africa, and North Africa & the Middle East. The first two have the largest percentage of countries that are democracies and the last two have the smallest percentage of countries that are democracies.[13]

To select the rest of the placebo test variables, we analyzed 114 characteristics of countries in 1998 from all the datasets mentioned in the introduction. We tried to identify variables that are the most imbalanced across regime types.[14] We selected data from 1998 so that our potential placebo variables are lagged behind the democracy variable by 10 years (the most recent year of the democracy variable `chga_demo`, which we use, is 2008).[15] Furthermore, we selected these variables because they describe characteristics that are not directly related to politics, regime type, or electoral procedure, and are thus more conceptually distinct. For each of these potential placebo variables P_k for $k \in \{1, 2, ..., 114\}$, we standardized them to create $S_{i,k}$ such that for country i:

$$S_{i,k} = \frac{P_{i,k}}{\text{Var}(P_k)} \tag{1}$$

For each country, let $D_i = 1$ if country i is a democracy in 2008 according to `chga_demo` and 0 otherwise. We estimated $\mathbb{E}(S_{i,k}|D_i = 1) - \mathbb{E}(S_{i,k}|D_i = 0)$ using $\hat{\gamma}_{i,k}$ from the following regression:

$$\mathbb{E}(S_{i,k}|D_i) = \eta_k + \gamma_k D_i \tag{2}$$

We can interpret $\hat{\gamma}_k$ as the estimated difference in means for standardized variable S_k between democracies and non-democracies. Table 3 presents the coefficient estimates and robust standard errors for the 25 variables that exhibit the greatest imbalance (in absolute value) across regime types.[16] From this list, we identified four potential placebo variables to use, in addition to the ones related to military capability, alliance, and trade (i.e., variables controlled for in the Tomz and Weeks's vignettes).

First, we constructed a placebo variable measuring how likely it is that the country in the scenario had large oil reserves. High fuel exports were highly correlated with being a non-democracy while high net energy imports were highly correlated with being a democracy. However, rather than ask about fuel exports/imports, our placebo question asked about oil reserves because it is relatively more exogenous to regime type.

[13] We also include East Asia and Central Asia among our answer choices in the placebo test question because those were popular answers in our pilot studies.

[14] The CIA World Factbook Ethnic Group dataset contains too many ethnic groups. Instead, we code the variable `majority_white` using the dataset. For each country, `majority_white` is coded 1 if the country's population is greater than 50 percent white (Causasian) and 0 otherwise. Note the data is from 2000 and not 1998; however, we think whether a country was majority white is unlikely to have changed between 1998 and 2000.

[15] Likewise, in our placebo test questions, we asked subjects to guess what the country in the scenario was like a decade ago so that their answers to the placebo questions are not affected by their beliefs about any recent change in the country's regime type, such as could be induced by the manipulation of the vignette.

[16] We also report the percentage of countries that are missing from each of the variables in the datasets.

Second, we created a placebo variable measuring how likely it is that the country in the scenario was majority Christian. As Table 3 shows, democracies had a low percentage of Muslims and a high percentage of Catholics in 1980. Since religion is slow-changing over time, we regarded it as an especially valid placebo variable (it is unlikely to be affected by regime-type on a short time scale).

Third, we created a placebo variable measuring GDP per capita. Many of the highly imbalanced variables in Table 3 are related to levels of economic development. These variables include employment in agriculture as a percentage of total employment, employment in services as a percentage of total employment, gross enrollment ratio in pre-primary schools, health expenditure as percent of GDP, and mortality rate of children under five. In selecting a placebo question we had several considerations to balance: we wanted to only ask one question to avoid burdening the respondent with multiple redundant questions; we wanted to choose a question that captures much of the common variance to these characteristics; we wanted to ask about a factor that is most likely to influence the outcome (support for using force); we wanted to ask a question that is easy to understand. These considerations led us to ask about GDP per capita. GDP per capita, itself, is 0.4 standard deviations greater for democracies than non-democracies in 1998 ($p < 0.001$).

Finally, we asked about the racial makeup of the country's population. As Table 3 shows, democracies were more likely to be majority white compared with non-democracies ($p < 0.001$).

Table 3: Top 25 Variables Most Imbalanced Across Regime Types

Variables (Standardized)	Coef	SE	% Missing
Fuel exports (% of merchandise exports)	−0.959	0.239	37
Muslims as percentage of population in 1980	−0.953	0.152	11
Employment in agriculture (% of total employment)	−0.941	0.335	56
Population ages 65 and above (% of total)	0.922	0.121	11
Heritage Foundation Economic Freedom Index: property rights	0.912	0.147	21
Heritage Foundation Economic Freedom Index	0.877	0.158	21
Employment in services (% of total employment)	0.859	0.295	56
Number of military treaties	0.822	0.109	0
Number of treaties: defense	0.810	0.109	0
Gross enrollment ratio, pre-primary schools, total	0.807	0.182	43
Number of treaties: entente	0.784	0.110	0
Population ages 0-14 (% of total)	−0.774	0.135	11
Number of treaties: non-aggression	0.758	0.111	0
Catholics as percentage of population in 1980	0.740	0.129	11
Social Globalization Index	0.732	0.142	11
Heritage Foundation Economic Freedom Index: trade freedom	0.723	0.154	21
Alternative and nuclear energy (% of total energy use)	0.698	0.149	35
Energy imports, net (% of energy use)	0.686	0.199	35
Country's population was majority white	0.675	0.120	0
Services, etc., value added (% of GDP)	0.675	0.147	16
Health expenditure, total (% of GDP)	0.666	0.135	10
Employment in industry (% of total employment)	0.655	0.336	56
Mortality rate, under-5 (per 1,000 live births)	−0.654	0.147	8
Average value of ethnolinguistic fractionalization	−0.650	0.199	47
Armed forces personnel (% of total labor force)	−0.624	0.168	17

We also examined variables that are related to military capability, alliance, and trade, three variables that were included as details in the Tomz and Weeks's survey experiment design. Potential placebo variables include those that were explicitly controlled for by the Tomz and Weeks's vignettes (i.e., non-nuclear military capability, military treaties, and volume of import/export) and those that are highly correlated with alliance and trade (i.e., iron and steel production, energy consumption, population, joint military exercises, and foreign direct investment). We estimated γ_k for these variables using the same model as described in the previous section. Table 4 contains our coefficient estimates and robust standard errors.[17] We found that none of the variables that describe military capability is statistically significant at $\alpha = 0.05$. On the other hand, variables related to trade and military alliance were all statistically different between regime types at $\alpha = 0.05$.

Based on our analysis, we asked placebo test questions regarding geographic region, GDP per capita, religion, oil reserves, race, military spending, military alliance, trade, joint military exercise, and foreign direct investment.

[17]Again, we reported the percentage of countries that are missing in each variable.

Table 4: Variables Related to Military Capability, Alliance, and Trade with the U.S.

Variables (Standardized)	Coef	SE	% Missing
Military Capability Variables			
Iron and steel production (thousands of tons)	0.137	0.159	8
Military expenditures (thousands of $)	0.105	0.155	8
Military personnel (thousands)	−0.172	0.173	8
Energy consumption (thousands of coal-ton equivalents)	0.108	0.155	8
Total population (thousands)	−0.057	0.166	8
Urban population (thousands)	−0.089	0.181	8
Composite Index of National Capability Score	−0.010	0.172	8
Alliance Variables			
Number of treaties: defense	0.810	0.109	0
Number of treaties: non-aggression	0.758	0.111	0
Number of treaties: entente	0.784	0.110	0
Number of military treaties (all types)	0.822	0.109	0
Number of joint military exercises	0.398	0.119	0
Trade Variables			
Volume of imports	0.259	0.129	8
Volume of exports	0.323	0.122	8
Total volume of trade (imports + exports)	0.291	0.125	8
FDI: position on a historical-cost basis	0.325	0.138	46
FDI: net financial transactions	0.403	0.139	44
FDI: net income	0.386	0.128	41

B.1 Correlation Between Democracy and Percent Muslim

For each region of the world, we calculated the correlation (Pearson's r) between a country being a democracy in 2008 and the percent of its population who were Muslims in 1980. The data came from the Quality of Government Dataset. We used `chga_demo` for our binary measure of democracy and `lp_muslim80` as our measure of the percentage Muslim in each country. Note that we could not calculate the correlation for North America because all countries in North America are democracies.

Table 5: Correlation Between Democracy and Percent of Population that is Muslim by Region

Region	Pearson's r
All Countries	-0.469
Eastern Europe and post Soviet Union	-0.681
Latin America	0.087
North Africa & the Middle East	-0.656
Sub-Saharan Africa	0.048
Western Europe and North America	NA
East Asia	-0.379
South-East Asia	-0.054
South Asia	-0.537
The Pacific	-0.488
The Caribbean	-0.461

C "Democratic Peace" Survey Results – Additional Analyses

C.1 Abstract Encouragement Design

Respondents had 1/2 probability of being randomly assigned to read instructions that encouraged them to consider the vignette scenario in the abstract. They were told "For scientific validity the situation is general, and is not about a specific country in the news today." We determine those assigned to the Abstract Encouragement Design do not exhibit less imbalance in their placebo test responses. Figures 3 and 4 show that respondents in both groups exhibit similar levels of imbalance and imbalance in the same direction in almost all the placebo outcomes.

Figure 3: Effect of the Abstract Encouragement Design (Standardized)

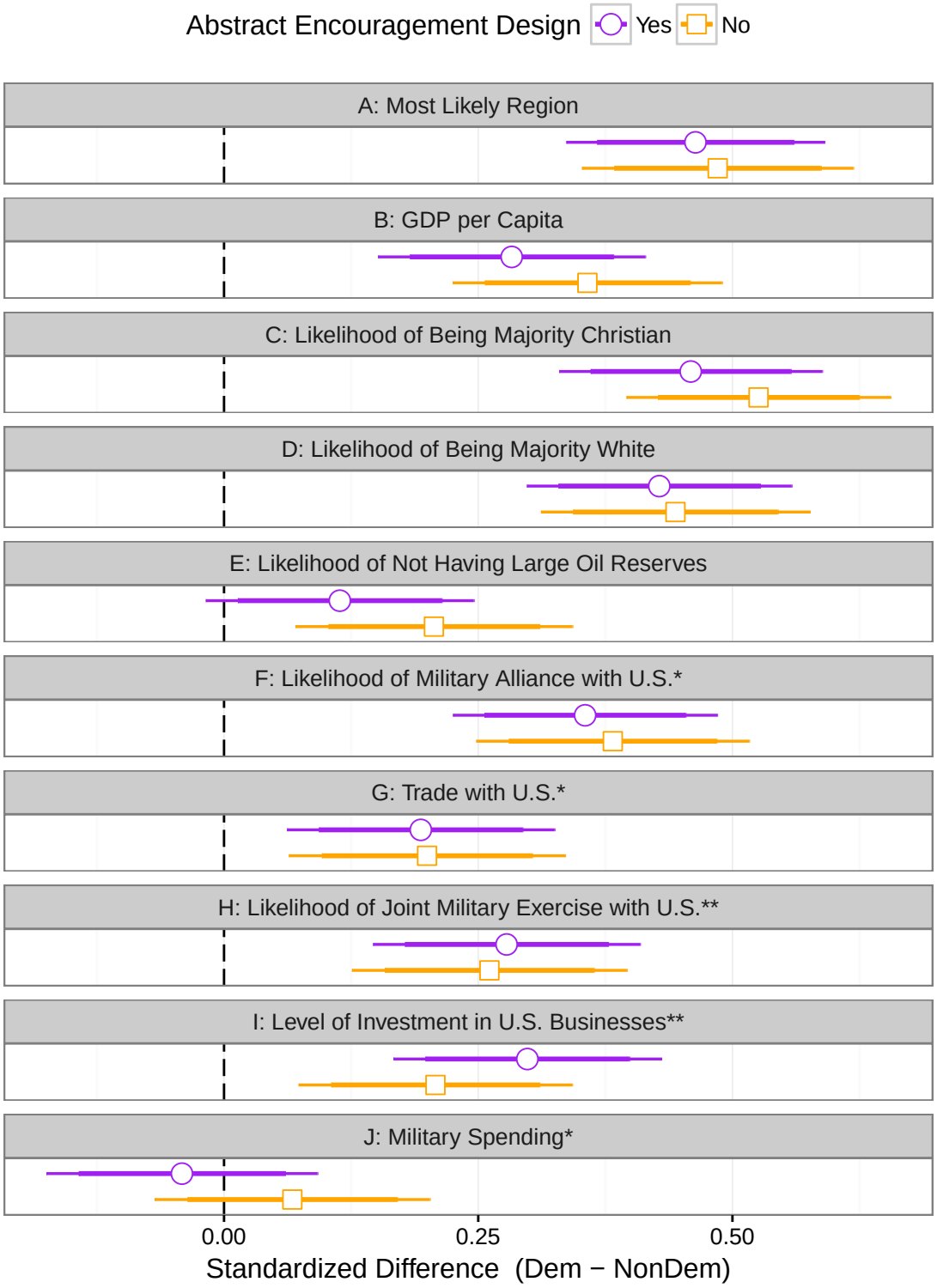
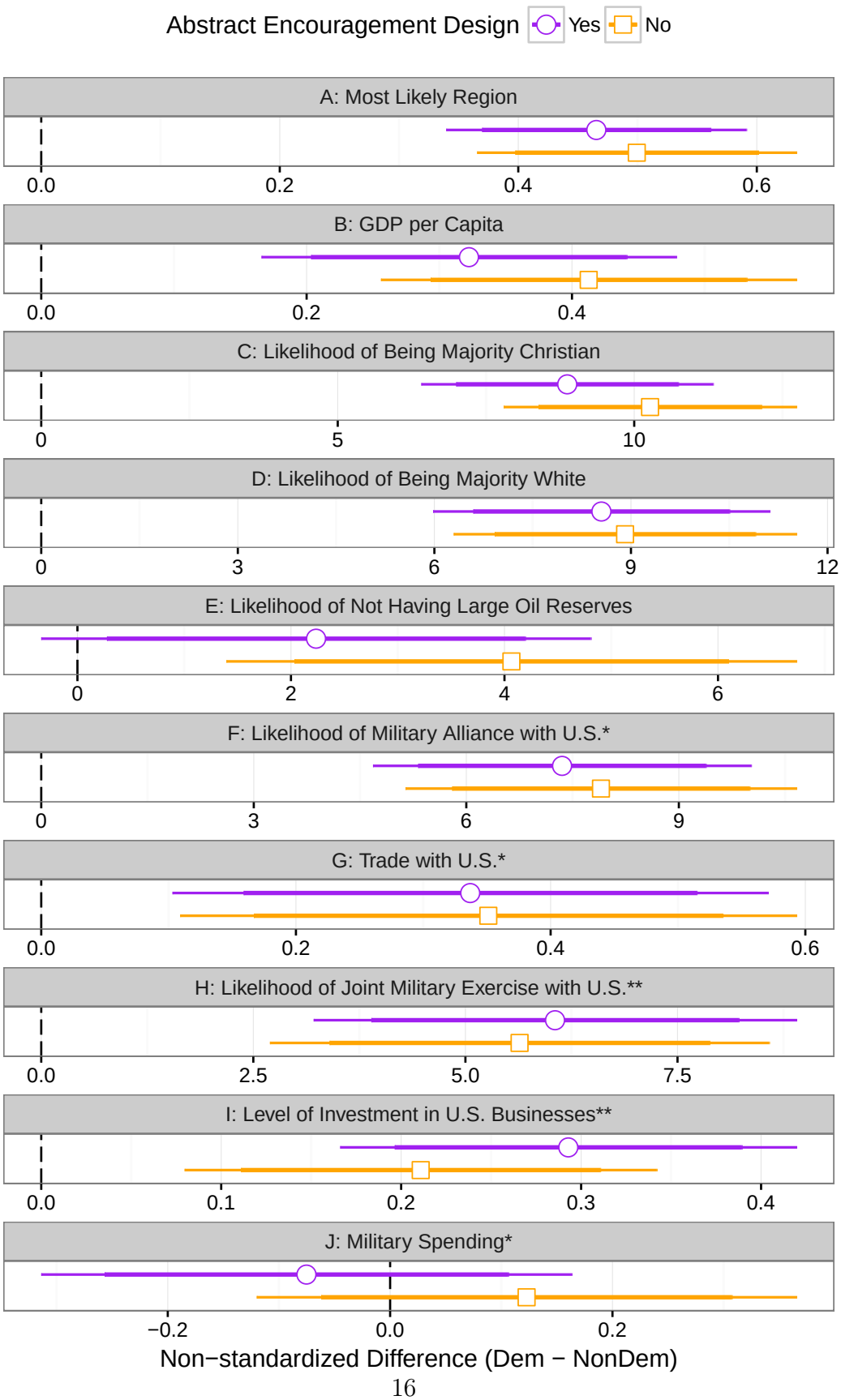


Figure 4: Effect of the Abstract Encouragement Design (Non-standardized)

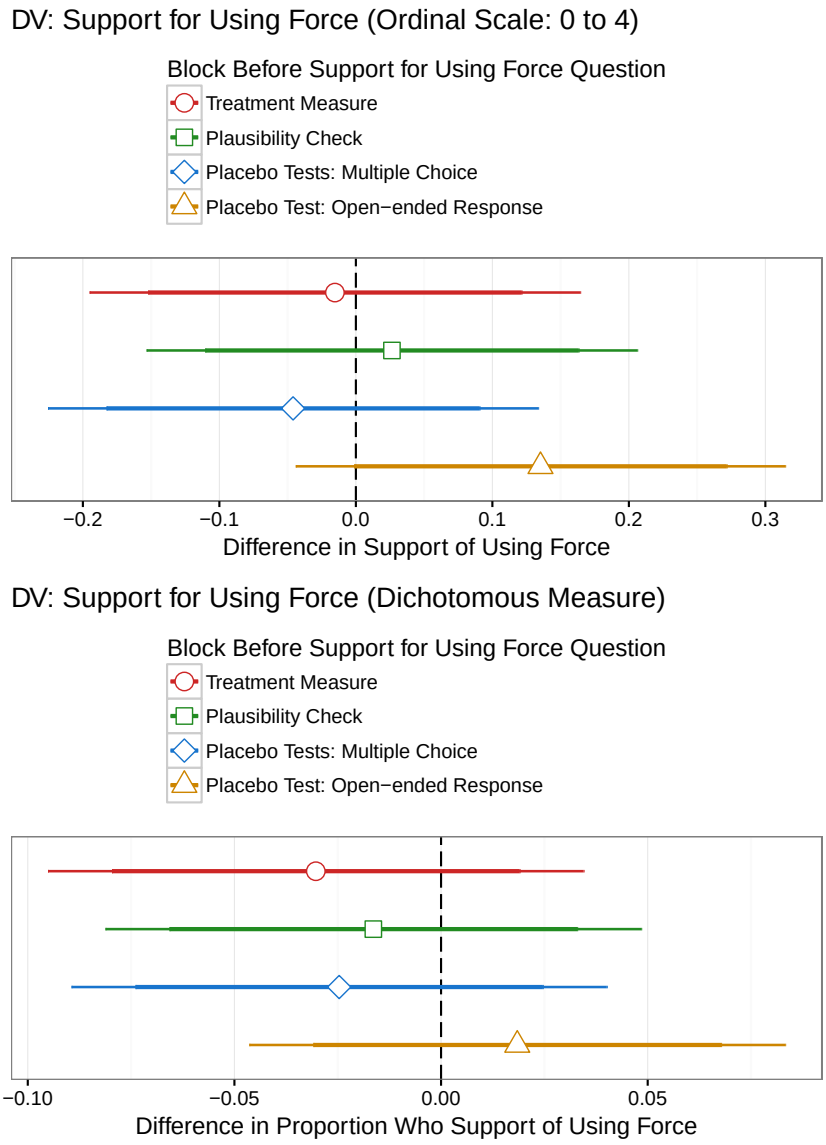


C.2 Question Block Order Does Not Affect Substantive Outcome Measure

C.2.1 Question Block Order in Together-Placebos Design

We examine whether the order of the question blocks changes how respondents answer the substantive outcome question (i.e., support for using force). In the Together-Placebos, we randomized the order of four question blocks relative to the substantive outcome block. In Figure 5, for each of the four blocks, we estimate the difference-in-mean in support for using force between respondents who saw the block before the substantive outcome block and those who saw it after. We find that the order of the question blocks mostly do not affect respondents' support for using force.

Figure 5: Question Block Order Does Not Affect Support for Using Force



C.2.2 Placebo Test Questions Do Not Affect Substantive Outcome Measure

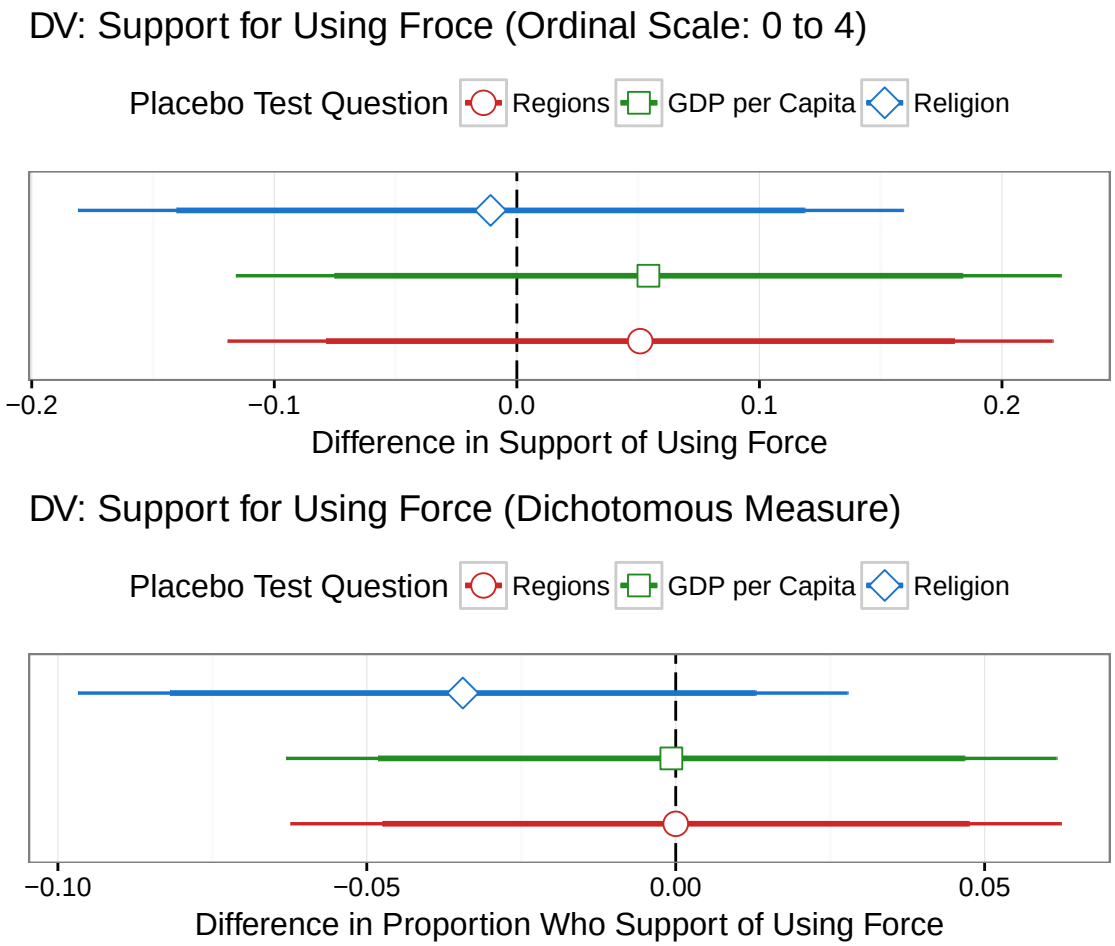
In the Separated-Placebos Design, we randomize whether zero, one, two, or three placebo test question(s) appear(s) before the support for using force question. The three placebo test questions we use are:

- Regions of the world
- GDP per capita
- Likelihood of being majority Christian

When placebo test question(s) appear(s) before the support for using force question, which questions are asked and the order of those questions (when there are two or more such questions) are randomized. Let X_1 be a dummy for whether the GDP per capita question appeared before the outcome question; let X_2 be a dummy for whether the religion question appeared before the outcome question; let X_2 be a dummy for whether the regions question appeared before the outcome question; and let Y be responses to the outcome measure.

First, we test if there is a significant difference-in-means in Y depending on X_j for $j \in \{1, 2, 3\}$. The results are reported in Figure 6. Whether each of the placebo test question appeared before the support for using force question did not affect respondents' support for using force.

Figure 6: Separated-Placebos Design Results



We also estimate the effect of the placebo test questions appearing before the support for using force question using a fully interacted regression:

$$\begin{aligned}\mathbb{E}(Y_i|X_{1,i}, X_{2,i}, X_{3,i}) &= \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i} \\ &+ \beta_4 X_{1,i} X_{2,i} + \beta_5 X_{1,i} X_{3,i} + \beta_6 X_{2,i} X_{3,i} + \beta_7 X_{1,i} X_{2,i} X_{3,i}\end{aligned} \tag{3}$$

As Table 6 shows, none of the coefficients are statistically significant. Respondents' support for using force is unchanged by all combinations of placebo test questions appearing before it.

Finally, we consider whether the placebo test questions appearing before the support for using force question affects our ITT estimates (estimates of the effect of treatment assignment Z on Y). For $j \in \{1, 2, 3\}$, we estimate $\beta_{3,j}$ from the following regression:

$\mathbb{E}(Y_i|X_{i,j}, Z_i) = \beta_{0,j} + \beta_{1,j}X_{i,j} + \beta_{2,j}Z_i + \beta_{3,j}X_{i,j}Z_i$

In Tables 7 and 8, we show that $\hat{\beta}_{3,j}$ for all j are not statistically significant at $p = 0.05$. This means that each of the placebo test questions appearing before the support for war question does not seem to affect our ITT estimates.

Table 6: Results from the Fully Interacted Regressions

	DV 1: Support for Using Force (Ordinal Scale) DV 2: Support for Using Force (Dichotomous Measure)	
	(1)	(2)
GDP per Capita	0.058	−0.031
	(0.129)	(0.049)
Religion	0.042	−0.039
	(0.131)	(0.050)
Regions	0.005	−0.056
	(0.130)	(0.047)
GDP per Capita × Religion	−0.183	−0.042
	(0.204)	(0.075)
GDP per Capita × Regions	0.063	0.091
	(0.204)	(0.075)
Religion × Regions	−0.083	0.014
	(0.204)	(0.075)
GDP per Capita × Religion × Regions	0.175	0.049
	(0.289)	(0.106)
Constant	1.727***	0.370***
	(0.067)	(0.025)
Note:		*p<0.05; **p<0.01; ***p<0.001

Table 7: Effect of Placebo Tests on ITT Estimates

	DV: Support for Using Force (Ordinal Scale)		
	(1)	(2)	(3)
Democracy	−0.595***	−0.455	−0.436***
	(0.090)	(0.363)	(0.102)
Regions	−0.053		
	(0.093)		
Democracy × Regions	0.208		
	(0.130)		
GDP per Capita		−0.031	
		(0.028)	
Democracy × GDP per Capita		−0.003	
		(0.041)	
Religion			−0.008***
			(0.002)
Democracy × Religion			0.0005
			(0.003)
Constant	2.029***	2.290***	2.257***
	(0.066)	(0.242)	(0.065)
Note:			*p<0.05; **p<0.01; ***p<0.001

Table 8: Effect of Placebo Tests on ITT Estimates

	DV 2: Support for Using Force (Dichotomous Measure)		
	(1)	(2)	(3)
Democracy	-0.173^{***}	-0.044	-0.127^{***}
	(0.034)	(0.133)	(0.037)
Regions	-0.031		
	(0.036)		
Democracy × Regions	0.062		
	(0.048)		
GDP per Capita		-0.002	
		(0.011)	
Democracy × Regions		-0.010	
		(0.015)	
Religion			-0.003^{***}
			(0.001)
Democracy × Religion			0.0004
			(0.001)
Constant	0.426^{***}	0.432^{***}	0.490^{***}
	(0.025)	(0.094)	(0.024)
Note:			*p<0.05; **p<0.01; ***p<0.001

C.3 Substantive Outcome Question Does Not Affect Severity of Imbalance in Placebo Outcomes

In the Together-Placebos design, we randomize whether the substantive outcome question (i.e., support for using force) comes before or after the placebo test question block. In this subsection, we analyze whether the support for using force question impacts the severity of imbalance in the placebo outcomes using a difference-in-difference approach.

Let F_i be an indicator variable for whether subject i answered the support for using force question before the placebo test questions. Adopting notation from previous subsections, let $Y_{i,j}$ be subject i's standardized response to placebo test question j and Z_i be her treatment assignment. We estimate

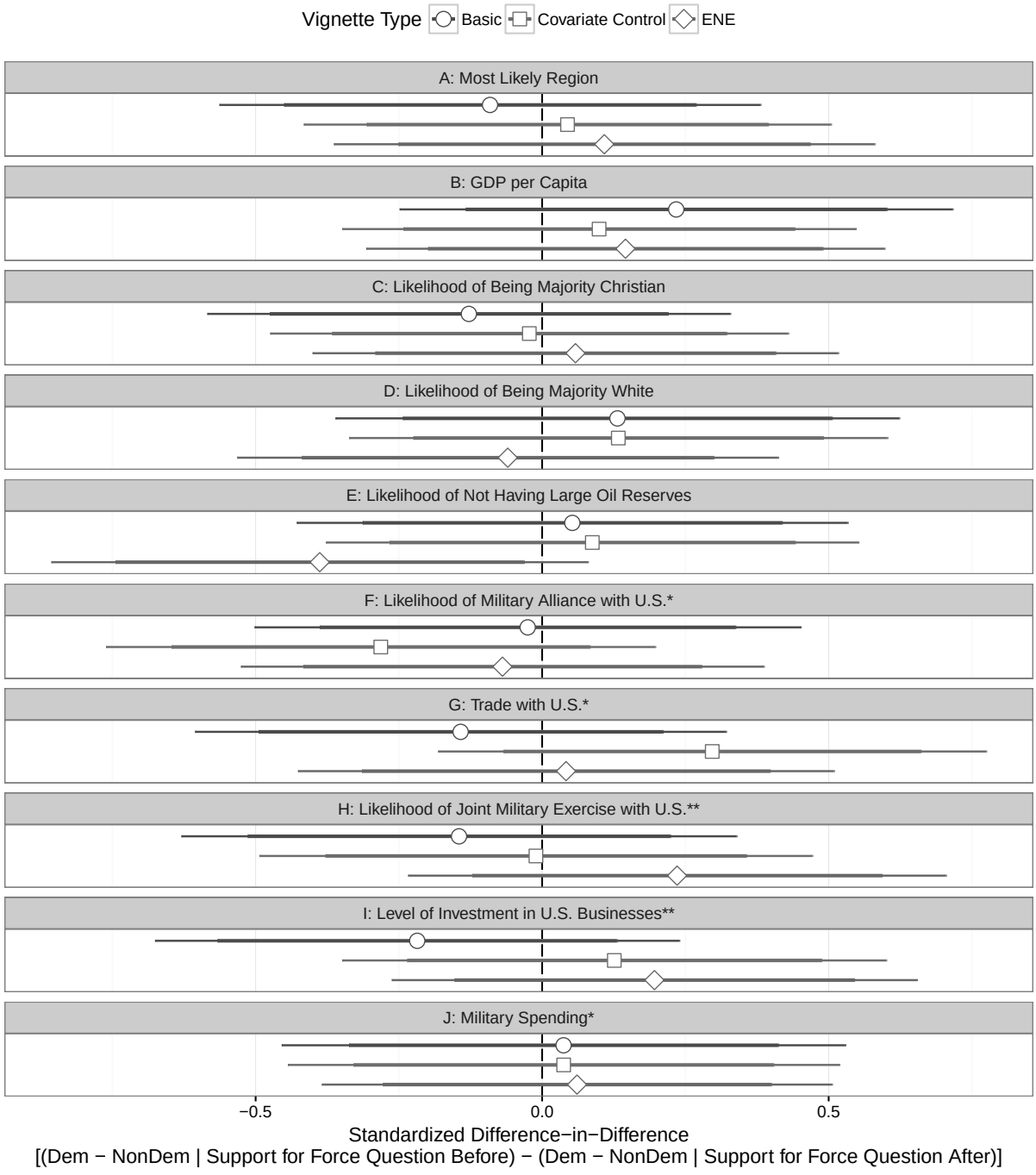
$$\mathbb{E}\left\{[Y_{i,j}(Z_i = 1, F_i = 1) - Y_{i,j}(Z_i = 0, F_i = 1)] - [Y_{i,j}(Z_i = 1, F_i = 0) - Y_{i,j}(Z_i = 0, F_i = 0)]\right\}$$

using $\hat{\beta}_{3,j}$ from the regression:

$$\mathbb{E}(Y_{i,j}|Z_i, F_i) = \beta_{0,j} + \beta_{1,j}Z_i + \beta_{2,j}F_i + \beta_{3,j}Z_iF_i \tag{4}$$

We report our estimates for $\beta_{3,j}$ in Figure 7 for each vignette type. The results demonstrate that answering the substantive outcome question does not affect the severity of imbalance respondents exhibit in their placebo outcomes.

Figure 7: Substantive Question Order and Severity of Imbalance in Placebo Outcomes



C.4 Attention Checks

We use three measures to check how much respondents are paying attention to our survey. First, we analyze respondents' answers to the attention check question. Thirty-eight respondents, or 1.31 percent of respondents, failed the attention check.

Second, we examine whether respondents took too little or too much time to complete the survey. Those who spent too little time likely rushed through the questions; those who took too much time might have been pre-occupied with other activities. The average amount of time respondents' took to complete the survey is 12.91 minutes and the median is 10.70 minutes. We code respondents who were in the bottom and top five percentile of time spent (less than 4.65 minutes or more than 25.83 minutes) as inattentive.

Finally, we look at how frequently respondents chose the first or last answers in the multiple-choice placebo test questions; respondents who are rushing through the survey are likely to simply click on the first or last answer choice. Thirty-three percent of respondents did not select the first/last answer choices for any of the 10 questions; only six respondents exclusively selected the first/last answer choices.

For dimension reduction, we using principal component analysis (PCA) to combine the three measures into a single principal component that measures attentiveness. We use the PCA score to test whether respondents paid attention produced greater imbalance in their placebo tests. Note that a higher PCA score means that the respondent is more attentive.

In Table 9, the interaction effect between treatment assignment Z and the PCA score is statistically significant at $\alpha = 0.05$ for seven out of 10 placebo outcomes. Furthermore, the positive signs suggests that the more attentive respondents are, the more likely they think the country described as a democracy in the scenario has real-world characteristics of democracies.

Table 9: Attentiveness and Responses to Placebo Test Questions

	DV: Responses to Placebo Test Questions A through E				
	(A)	(B)	(C)	(D)	(E)
Democracy	0.455^{***}	0.293^{***}	0.403^{***}	0.334^{***}	0.140^{***}
	(0.037)	(0.037)	(0.035)	(0.035)	(0.037)
PCA Score	-0.115^{***}	-0.163^{***}	-0.410^{***}	-0.477^{***}	-0.040
	(0.029)	(0.036)	(0.030)	(0.030)	(0.032)
Democracy × PCA Score	-0.042	0.102^{*}	0.119^{*}	0.171^{***}	-0.037
	(0.046)	(0.048)	(0.048)	(0.047)	(0.046)
Constant	-0.218^{***}	-0.132^{***}	-0.164^{***}	-0.121^{***}	-0.068^{**}
	(0.022)	(0.027)	(0.024)	(0.024)	(0.025)

	DV: Responses to Placebo Test Questions F through J				
	(F)	(G)	(H)	(I)	(J)
Democracy	0.284^{***}	0.117^{**}	0.184^{***}	0.172^{***}	0.006
	(0.036)	(0.036)	(0.037)	(0.036)	(0.037)
PCA Score	-0.358^{***}	-0.337^{***}	-0.374^{***}	-0.369^{***}	-0.025
	(0.031)	(0.032)	(0.031)	(0.032)	(0.035)
Democracy × PCA Score	0.113^{*}	0.162^{***}	0.175^{***}	0.158^{**}	-0.023
	(0.044)	(0.046)	(0.045)	(0.049)	(0.049)
Constant	-0.107^{***}	-0.027	-0.053^{*}	-0.050^{*}	-0.002
	(0.025)	(0.025)	(0.026)	(0.026)	(0.026)

Note: *p<0.05; **p<0.01; ***p<0.001