Supplementary Material for
"Response Options and the Measurement of Political Knowledge"

John G. Bullock and Kelly Rader

March 1, 2021

Auxiliary Information and Related Analyses

Importance of Studying *Levels* of Knowledge

Our research speaks to characterizations of levels of knowledge in the public. Although knowledge is often a predictor or a moderating variable in studies of political behavior, a large body of research focuses instead on characterizing levels of knowledge—that is, on estimating how much or how little the public knows about politics. For example, an influential argument in Luskin (1987) is that levels of knowledge, as measured by answers to factual questions about politics, are good proxies for levels of sophistication of political thinking. Small libraries can be filled with arguments about different ways of measuring knowledge, and a question that unites many of these efforts is whether most Americans are, as many suggest, ignorant of politics. (For examples, compare Luskin 2002 to Boudreau and Lupia 2011.) And partly because knowledge predicts some forms of political behavior, Delli Carpini and Keeter (1996, chs. 2-3) take up aggregate levels of knowledge, and changes in these levels, in minute detail. We do not deny the importance of knowledge as a predictor or a moderator, but it seems plain that for as long as scholars are interested in characterizing levels of knowledge, we should scrutinize the ways in which those characterizations are made.

Actual levels of knowledge are unobservable. We observe only people's answers to questions; we infer, from their answers, their levels of knowledge. In principle, every scholar will acknowledge a potential gap between measurement of the construct and the construct itself. But in practice, almost no attention is given to the ways in which measurement of the construct—and thus characterizations of it—depends on the number and difficulty of response options. Moreover, the research that does speak to this issue is divorced from the research in which the public's levels of knowledge are characterized: we know of no published work that takes up, more than fleetingly, both the role of response options and the public's level of political knowledge.[1]

Verbatim Responses to the ANES "Chief Justice" Question

On page 4, we write that "By our reading, an additional 25% of respondents in 2000 and 18% of respondents in 2012 gave nearly correct answers" to ANES questions about the "job or political office" held by William Rehnquist (in 2000) or John Roberts (in 2012). In this section, we elaborate on our method of coding the ANES responses.

The ANES has long asked a series of open-ended "office recognition" questions. Interviewers name a public figure, and respondents are asked to name the "job or political office" that the person holds. These questions are preceded by introductory remarks, always similar to the ones that interviewers read aloud in 2000:

> Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers and the like.

These remarks are then followed by the actual questions. And one of the ANES standards is about the sitting Chief Justice of the United States. In 2000, the question was "William Rehnquist. What job or political office does he *now* hold?"

By tradition, the ANES did not release verbatim answers to these questions. It instead released coded versions of the responses. These coded versions indicated only whether the responses were correct, incorrect, or "don't know" responses.

Gibson and Caldeira (2009, 432) gained access to the verbatim responses to the Chief Justice question in the 2000 ANES. Examining those responses, they determined that many of the responses coded by the ANES as "incorrect" were in fact "nearly correct." By their measure,

[1] Perhaps Mondak (2001, 229-30) comes closest.

349 responses fit this description: 22.4% of all responses, and 71.8% of responses that the ANES had coded as incorrect. This finding is one justification for the claim that conventional ANES coding has led scholars to understate what the public knows about the Supreme Court.

The criteria that Gibson and Caldeira used to determine whether a response was "nearly correct" are not entirely clear. Of the 349 responses to the 2000 ANES that they classified as "nearly correct," 206 seemed to earn this classification because they mentioned the Supreme Court (Gibson and Caldeira 2009, 432). Reasons for coding the remaining 143 responses as "nearly correct" were not given.

We now describe a framework that one may use to categorize responses to the Chief Justice question and then to classify them as correct or nearly correct. We apply the framework to the responses to the Chief Justice question that have been made available in the 2000, 2008, and 2012 ANES time-series studies, and in the Evaluations of Government and Society (EGSS) that was fielded by the ANES in 2011.

Any scholar who examines the verbatim records will be struck by their variety. For example, of the 2,102 respondents in 2008 who were asked the Chief Justice question, fully 1,215 unique responses were recorded.[2] By leading the ANES to make these verbatim records available to the public, Gibson and Caldeira (2009) have done a service to scholars, who can now witness the variety of responses for themselves and draw their own conclusions about the degree of knowledge that they suggest.

After examining the responses in the 2008 ANES, we constructed a fourteen-category classification scheme. The categories are mutually exclusive and exhaustive. This classification scheme does not explicitly distinguish responses by degree of correctness; instead, it characterizes

---

[2] Our enumeration of unique responses does not take into account differences in capitalization. If it did, the number of unique responses would be still larger.

the content of the responses. Reasonable people may differ about whether a given response is "correct" or "nearly correct"; the point of our classification scheme is to help readers make such determinations for themselves.[3] The categories are:

1. *Mentions Chief Justice and United States.* In 2008, there were 2 such responses (0.1% of all responses).

2. *Mentions Chief Justice and Supreme Court.* These are the responses that would have been coded as correct under the conventional, strict ANES coding. In 2008, there were 69 such responses (3.3% of all responses).

3. *Mentions Chief Justice and something like Supreme Court.* There were no responses in this category.

4. *Mentions Supreme Court and something like Chief Justice.* There were 11 responses in this category (0.5%). Examples: "head supreme court justice" and "head of the supreme court."

5. *Mentions Supreme Court and "justice."* There were 78 responses in this category (3.7%). The most common was simply "Supreme Court justice."

6. *Mentions "Chief Justice" or "Chief of Justice."* There were 13 responses in this category (0.6%).

7. *Mentions Supreme Court.* There were 82 responses in this category (3.9%). Examples: "supreme court" and "supreme court judge."

8. *Mentions "justice."* There was two responses in this category: "justice area" and "he is a federal justice . . . "

[3] See DeBell (2013) for a thoughtful and more general discussion of classification schemes.

	category number	number of responses	percentage of responses
Mentions Chief Justice and United States	1	2	0.1
Mentions Chief Justice and Supreme Court	2	69	3.3
Mentions Chief Justice and something like Supreme Court	3	0	0.0
Mentions Supreme Court and something like Chief Justice	4	11	0.5
Mentions Supreme Court and “justice”	5	78	3.7
Mentions “Chief Justice” or “Chief of Justice”	6	13	0.6
Mentions Supreme Court	7	82	3.9
Mentions “justice”	8	2	0.1
Mentions “judge” or “court”	9	16	0.8
Mentions news media or journalism profession	10	15	0.7
Mentions government, politics, or policy	11	658	31.3
Mentions other job or office	12	23	1.1
Don’t know	13	1116	53.2
Uninterpretable	14	14	0.7

Table A1: Open-ended responses in the 2008 ANES. Cell entries are frequencies and percentages of responses to the “Chief Justice” office-recognition question in the 2008 ANES. N = 2099. Each response was placed in exactly one category. Categories are ordered, such that responses were placed in larger-numbered categories only if they could not be placed in smaller-numbered categories. See page A7 for details.

9. *Mentions “judge” or “court.”* There were 16 responses in this category (0.8%). The most common was simply “judge.”

10. *Mentions news media or journalism profession.* Fifteen responses (0.7%) were in this category. Respondents who gave this answer may have been thinking of John D. Roberts, who was a foreign correspondent for CNN at the time of the 2008 ANES post-election survey.

11. *Mentions government, politics, or policy.* There were 658 responses (31.3%) in this category. "Senate" and "senator" are the most common responses. Recall that the ANES Chief Justice question was asked in the middle of a long series of political questions, and in a survey that was described to respondents as political even before it began. Had the same question been asked in a less political survey, it seems unlikely that so many people would have made guesses of this sort.

12. *Mentions other job or office.* There were 23 responses in this category (1.1%). Examples include "car salesman" and "t.v. game show host."

13. *Don't know.* There were 1,116 responses (53.2%) in this category.

14. *Uninterpretable.* There were 14 responses (0.7%) in this category. Examples include "3" and "MAKE CHANGE FOR ALL." Some records of responses in this category may have been affected by transcription errors.

These categories are ordered in the sense that responses were assigned to larger-numbered categories only if they could not be assigned to smaller-numbered ones. For example, responses that mentioned both "Chief Justice" and "Supreme Court" were assigned to category 2 ("mentions Chief Justice and Supreme Court") but not to category 5 (mentions Supreme Court and "justice' '). And responses such as "Supreme Court justice" were assigned to category 5 but not to category 8 (mentions "justice").

We applied this coding scheme to the four studies mentioned above. And we now use it to calculate the percentage of correct and nearly correct responses across all four studies. To do so, we code categories 1 and 2 in Table A1 as correct, categories 3-5 and 7 as "nearly correct and mentions Supreme Court," category 6 as "nearly correct and doesn't mention Supreme Court," and categories 8-9 and 11-12 as incorrect. We omit from analysis category 14 (uninterpretable). We also omit category 10 (mentions news media or journalism profession) for the reason stated above: John Roberts is the name of both a Chief Justice and a television journalist.

	ANES 2000	ANES 2008	EGSS 2011	ANES 2012
Correct	7.8	3.4	11.5	13.3
Nearly correct and mentions Supreme Court	22.9	8.3	13.6	11.8
Nearly correct and doesn't mention Supreme Court	2.5	0.6	8.7	6.3
Incorrect	9.0	33.8	10.4	10.3
Don't know	57.7	53.9	55.8	58.4
Number of observations	1555	2099	634	5510

Table A2: Open-ended responses to the Chief Justice question by degree of correctness. Cell entries are percentages of responses to the "Chief Justice" office-recognition question.

Table A2 presents the results of the analysis. Three patterns are striking. First, the percentage of "nearly correct" responses declined from 25.4% in 2000, when William Rehnquist was in his 14th year as Chief Justice, to only 8.9% in 2008, when John Roberts was in his second year. But rates of nearly-correct response rose afterward, exceeding 20% in 2011 and approaching 20% in 2012.

Second, by our coding, Gibson and Caldeira slightly understated the percentage of nearly correct responses in the 2000 ANES. (They found that 22.4% of all 2000 ANES responses were nearly correct.) Third, in every one of our four studies, most respondents said that they simply didn't know the answer to the question.

Text of Questions Listed in Table 1

Table 1 lists the eight closed-ended questions about the name of the Chief Justice, other than our own, that had been fielded with national samples as of 2017. The full text of these questions is:

1. Prior (2005, 590): Who is the Chief Justice on the U.S. Supreme Court?

2. Gibson and Caldeira (2009): [T]he former Chief Justice of the United States Supreme Court died last year and was replaced by Justice John G. Roberts, Jr. Do you recognize any of these as being the former Chief Justice of the United States Supreme Court who died last year?

3. Pew Research Center (2010): Do you happen to know who is the Chief Justice of the US (United States) Supreme Court?

4. Pew Research Center (2012): Do you happen to know who is the Chief Justice of the US (United States) Supreme Court?

5. Gibson and Caldeira (2009): I'm now going to read you three names. Do you recognize any of these as being the current Chief Justice of the United States Supreme Court?

6. Sen (2017): Please select the name of the current Chief Justice of the United States from the choices below.

7. ANES EGSS 3: Who is the Chief Justice of the U.S. Supreme Court?

8. ANES EGSS 4: Who is the Chief Justice of the U.S. Supreme Court?

"EGSS" stands for "Evaluations of Government and Society Study." See https://electionstudies.org/data-center/2010-2012-evaluations-of-government-and-society-study/.

Sample Characteristics and Sample Weighting

Survey Sampling International (SSI) maintains a large panel of people who are willing to take surveys. We contracted with SSI to provide a sample that matched the population of adults in the U.S. Census with respect to age, education, gender, income, race, and region of residence. We weighted our SSI data so that they matched the proportions of CPS respondents on all of these characteristics, save income. (We did not measure income in our study.) Available information on age, education, gender, race, and state of residence permitted us to compute weights for 1,961 subjects, and the weighted SSI data for these subjects are used in all of the analyses in our paper.

Figure A1 sheds light on the quality of the unweighted 1,961-person sample. The figure compares the proportions of various groups in the corresponding proportions of people in the weighted March 2017 Current Population Survey (Flood et al. 2018), which is run by the U.S. Census Bureau. As our sample was restricted to people who were at least 18 years old, the CPS data displayed in the figure represent CPS respondents who are at least 18 years old.

In most respects, our unweighted SSI sample seems representative. Of the twelve categories examined in Figure A1, there are nine for which the SSI and CPS samples are within 1.5 percentage points of each other. The main outliers are low-education subjects: 27% of our subjects report having no formal education beyond high school. Our sample thus compares favorably to the online sample of the 2016 ANES, 24% of which reports having no formal education beyond high school. It is comparable to the total 2016 ANES sample (26%) and the 2016 CCES sample (29%). All of these samples fall short of the benchmark set by the weighted CPS, in which 40% of respondents report having no formal education beyond high school. To adjust for the discrepancy, we follow the strategy of the ANES and the CCES (and indeed, the CPS itself): we weight our data.

Specifically, we created poststratification weights that are based on the weighted proportions in the CPS. The weights for our sample range from 0.61 to 2.35. And as one might expect, given that the survey weights are small, their use makes little difference. Table A3 is like

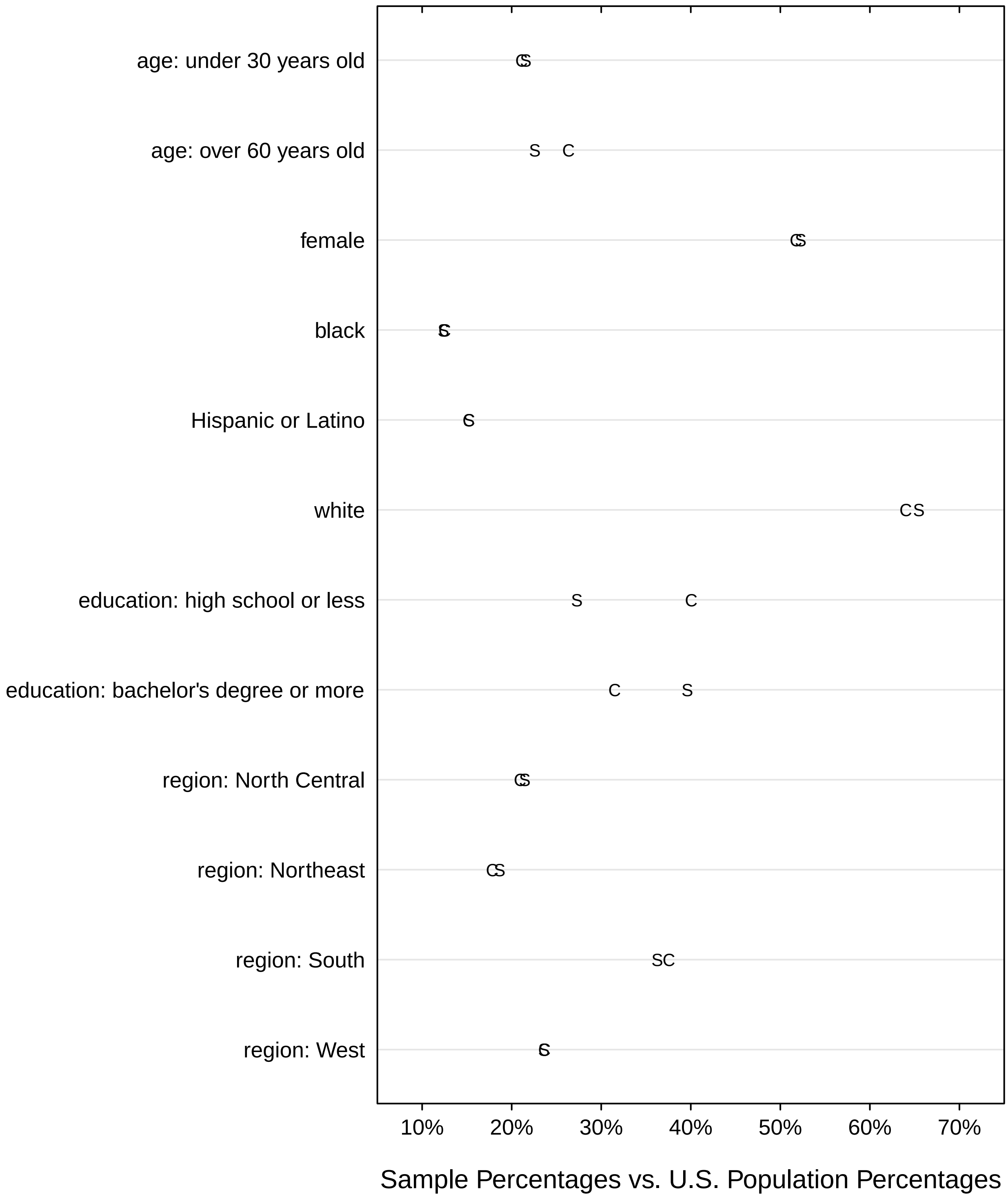

Figure A1: Sample characteristics. Each row plots the unweighted percentage of subjects in our sample (S) who share a characteristic. The corresponding weighted percentages in the March 2017 Current Population Survey (C) are also reported.

	(1)		(2)		(3)		(4)	
Intercept	.56	.01	.59	.01	.28	.01	.29	.01
five response options	−.12	.01						
difficult response options			−.24	.01				
three easy response options					.35	.01		
five difficult response options					.07	.01		
closed-ended response options							.21	.01
R^2	.13		.07		.10		.05	
Std. error of regression	.47		.48		.47		.49	
Number of questions	12		6		6		6	
Number of subjects	1957		1957		1957		1957	
Number of observations	20551		8706		9123		11739	

Table A3: Effects of number and difficulty of response options (unweighted data). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is “three response options”; for column (2), “easy response options”; and for columns (3) and (4), “open-ended questions.” Standard errors are clustered at the respondent level.

our main regression table, Table 2; it differs only in that the regressions in Table A3 are based on unweighted data. Inspection shows that only one treatment-effect estimate has changed, and it has changed by 0.01. Similarly, Table A4 shows that the estimates again change by no more than 0.01 when we both use unweighted data and remove sample restrictions, thereby accounting for subjects who did not provide complete demographic information.

Note that although we use data from 1,961 subjects to create our figures, we use data from only 1,957 subjects to create our regression tables. The difference arises because, following the “Missing Covariate Values” section of our pre-analysis plan, we estimated our regressions with data from only those subjects who answered all of our control variables. Four of our

	(1)		(2)		(3)		(4)	
Intercept	.56	.01	.59	.01	.28	.01	.28	.01
five response options	−.12	.01						
difficult response options			−.23	.01				
three easy response options					.35	.01		
five difficult response options					.07	.01		
closed-ended response options							.22	.01
R^2	.12		.06		.10		.05	
Std. error of regression	.47		.48		.47		.49	
Number of questions	12		6		6		6	
Number of subjects	2080		2080		2080		2080	
Number of observations	21834		9268		9693		12472	

Table A4: Effects of number and difficulty of response options (unweighted data, larger sample). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is "three response options"; for column (2), "easy response options"; and for columns (3) and (4), "open-ended questions." Standard errors are clustered at the respondent level.

subjects did not answer all of our political-knowledge control questions, and they were therefore excluded from our regression analyses.

Sample Sizes for Each Condition

Table A5 indicates, for each question in our experiment, the sample sizes in each condition. Our main hypothesis involved comparison of the "3 easy," "5 difficult," and open-ended conditions, and we therefore assigned more subjects to those conditions than to others.

We assigned 109 of our subjects (5.6%) to a "super-easy" condition when answering the Chief Justice question. This assignment is noted in Table A5. But as the condition plays no role in our pre-registered hypotheses or analysis strategy, we do not mention it in the paper.

	3 super-easy	3 easy	3 diff.	5 super-easy	5 easy	5 diff.	open-ended	3	5	total
Name of Chief Justice	52	494	198	57	197	492	470			1960
How many justices usually		504	214		210	525	506			1959
How many justices currently		542	207		251	507	454			1961
How many women		509	211		200	516	524			1960
Name of Senate Majority Leader		530	209		226	533	463			1961
Term length of justice		532	204		183	526	515			1960
How are justices chosen								976	985	1961
Possible to watch lawyers argue								1004	956	1960
Final say over Constitution								942	1019	1961
What happens if tie decision								950	1011	1961
Best description of Court power								970	991	1961
Possible to remove justices								948	1012	1960

Table A5: Number of subjects in each condition. For each question, subjects were assigned to a number-of-response-options condition. For each of the first six questions, subjects were also assigned to a difficulty condition.

Screening Subjects

In our pre-analysis plan, we committed to screening our subjects in several ways. (See page A43.) Following Clifford and Jerit (2016), we asked subjects to commit to not looking up answers online. We made this request at the start of our survey, before any randomization had occurred:

> It is important to us that you do not use outside sources like the Internet to search for the correct answers. We are trying to understand what people know about politics, not what they can look up. Do you agree to answer the following questions without help from outside sources?

We put this question to 2,202 SSI panelists. Fifty-seven of them answered "no," and one person skipped the question. Following our pre-analysis plan, these 58 people were excluded from our analyses.

In a pilot study, we compared the effects of this sort of commitment screener to the effects of imposing a 30-second time limit on consideration of each question. Correct-response rates were lower with the screener than with the time limit, suggesting that the screener was more effective at preventing online lookup.

Following the AAPOR Standard Definitions (2016, 14-15), we defined "break-offs" as interviews with subjects who answered fewer than half of the questions in a survey. Interviews with 64 subjects were classified as break-offs by this criterion, and we removed them from our analyses. Including these subjects makes no material difference to our results. After removing break-offs, we were left with $2{,}202 - 58 - 64 = 2{,}080$ panelists. (As described above, we had complete demographic information for 1,961 of these panelists.)

We also inserted a difficult open-ended "placebo question" (Bullock et al. 2015, esp. 529) into the middle of our survey. The question was "What job or political office did Horatio King hold?" King was a 19th-century Postmaster General, and we anticipated that few subjects would be able to provide the correct answer without looking it up. One hundred and sixty-four of our subjects (8.4%) answered correctly. We did not exclude any of these subjects from the analyses reported in the paper; in analyses not shown here, we find that excluding them makes no difference to the results. Specifically, excluding these subjects changes none of our treatment-effect estimates by more than 0.004.

In addition to the commitment screener, the screening of break-offs, and the placebo question, we included two screener questions in the style of Berinsky, Margolis, and Sances (2014). The first was about respondents' preferred news sources, and it was asked shortly after the survey had begun. The second was about justices' salaries, and it was asked shortly before the survey ended. In both cases, respondents were subtly instructed to either skip the question or to choose a particular response option. Thirty-four percent of subjects responded as instructed to the first question; 81% responded as instructed to the second question. Tables A6-A9 show that our results are slightly stronger when we exclude, from our analyses, subjects who did not follow our instructions for one or both of the questions. In particular, our main analyses—involving comparisons of estimates in columns (3) and (4)—suggest even more strongly that the difference between different kinds of closed-ended questions exceeds the difference between open- and closed-ended questions.

	(1)		(2)		(3)		(4)	
Intercept	.55	.02	.59	.02	.30	.02	.30	.02
five response options	−.11	.01						
difficult response options			−.25	.01				
three easy response options					.33	.01		
five difficult response options					.04	.01		
closed-ended response options							.20	.01
R^2	.16		.09		.12		.05	
Std. error of regression	.46		.47		.47		.49	
Number of questions	12		6		6		6	
Number of subjects	1575		1575		1575		1575	
Number of observations	16556		7020		7334		9450	

Table A6: Effects of number and difficulty of response options (excluding subjects who failed salaries screener). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is “three response options”; for column (2), “easy response options”; and for columns (3) and (4), “open-ended questions.” Standard errors are clustered at the respondent level.

	(1)		(2)		(3)		(4)	
Intercept	.59	.02	.63	.03	.31	.03	.32	.02
five response options	−.12	.01						
difficult response options			−.26	.02				
three easy response options					.33	.02		
five difficult response options					.04	.02		
closed-ended response options							.20	.02
R^2	.17		.10		.12		.06	
Std. error of regression	.46		.47		.47		.48	
Number of questions	12		6		6		6	
Number of subjects	662		662		662		662	
Number of observations	6955		2947		3116		3972	

Table A7: Effects of number and difficulty of response options (excluding subjects who failed media screener). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is "three response options"; for column (2), "easy response options"; and for columns (3) and (4), "open-ended questions." Standard errors are clustered at the respondent level.

	(1)		(2)		(3)		(4)	
Intercept	.59	.02	.63	.03	.32	.03	.33	.03
five response options	−.12	.01						
difficult response options			−.26	.02				
three easy response options					.32	.02		
five difficult response options					.03	.02		
closed-ended response options							.20	.02
R^2	.18		.10		.12		.06	
Std. error of regression	.45		.46		.47		.48	
Number of questions	12		6		6		6	
Number of subjects	635		635		635		635	
Number of observations	6675		2830		2989		3810	

Table A8: Effects of number and difficulty of response options (excluding subjects who failed either screener). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is "three response options"; for column (2), "easy response options"; and for columns (3) and (4), "open-ended questions." Standard errors are clustered at the respondent level.

	(1)		(2)		(3)		(4)	
Intercept	.55	.02	.59	.02	.29	.02	.30	.02
five response options	−.11	.01						
difficult response options			−.26	.01				
three easy response options					.34	.01		
five difficult response options					.04	.01		
closed-ended response options							.20	.01
R^2	.16		.09		.12		.05	
Std. error of regression	.46		.47		.47		.49	
Number of questions	12		6		6		6	
Number of subjects	1602		1602		1602		1602	
Number of observations	16836		7137		7461		9612	

Table A9: Effects of number and difficulty of response options (excluding subjects who failed both screeners). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is “three response options”; for column (2), “easy response options”; and for columns (3) and (4), “open-ended questions.” Standard errors are clustered at the respondent level.

	χ^2	F	p
Name of Chief Justice	474.12		.83
How many justices currently serve	361.55		.16
How many justices usually serve	359.71		.18
How many women currently serve	375.12		.07
Name of Senate Majority Leader	342.07		.40
Term length of justices	356.98		.21
How are justices chosen		.98	.53
Can one watch lawyers argue at the Court		.96	.58
Who decides constitutionality		1.08	.30
What happens when there is a tie vote		1.05	.36
When can the Court act		.95	.60
Can justices ever be removed		1.08	.29

Table A10: Randomization checks. The table reports separate randomization checks for each of our 12 questions. For each question, treatment condition was regressed on age, race, gender, education, and state of residence. The first six questions had five treatment conditions, and for these questions, we used multinomial logistic regression. The six remaining questions had only two treatment conditions ("three response options" and "five response options"), and in these cases, we used OLS.

Randomization Checks

Table A10 reports separate randomization checks for each of our 12 questions. For each question, treatment condition was regressed on age, age^2, education, gender, political knowledge (measured prior to treatment), race, and state of residence. The first six questions had five treatment conditions, and for these questions, we used multinomial logistic regression. The six remaining questions had only two treatment conditions ("three response options" and "five response options"), and in these cases, we used OLS. In our pre-analysis plan, we said that we would use F statistics for all 12 questions, but this was an error: the F statistic is appropriate only for binary treatments. See Gerber and Green (2012, 107) for a discussion.

The p values in the final column of the table suggest adequate balance. We see $p < .10$ for only one of the 12 regressions, which is about what we should expect from a randomization procedure that works as intended. (The outlying regression is for the "How many women currently serve" question; in this case, $p = .07$.) Later in the appendix, we account for imbalances in these control variables by adding them to our main regressions: see Table A12.

	minimum % correct	maximum % correct
How many women	28	74
How many justices currently	30	69
How many justices usually	42	73
Name of Senate Majority Leader	38	67
Name of Chief Justice	38	55
Term length of justice	60	73
Possible to remove justices	24	38
Final say over Constitution	63	75
Possible to watch lawyers argue	13	22
How are justices chosen	66	73
Best description of Court power	37	44
What happens if tie decision	25	29
Mean	39	68

Table A11: Ranges of percentages answering correctly across closed-ended conditions. We calculated the percentage of subjects answering each question correctly in each closed-ended condition. This table reports, for each question, the range of percentages correct, where the range is taken across closed-ended conditions.

Within each tier, questions are ordered from those with the largest range to those with the smallest. The first six questions had four closed-ended conditions: long-difficult, short-difficult, long-easy, and short-easy. The next six questions had only two conditions: long (five response options) and short (three response options).

Range of Percentages Answering Correctly for Each Question

Each of the twelve questions in our experiment was asked in multiple experimental conditions. Each question therefore has a range of percentages answering correctly, where the percentages are calculated separately for each condition. Table A11 displays this range for each question. For example, it shows that 28% of subjects answered the “How many women” question correctly in the “long-difficult” condition, but that 74% answered it correctly in the “three-easy” condition.

	(1)		(2)		(3)		(4)	
Intercept	.16	.06	.22	.08	−.14	.08	−.25	.08
five response options	−.12	.01						
difficult response options			−.24	.01				
three easy response options					.34	.01		
five difficult response options					.06	.01		
closed-ended response options							.21	.01
R^2	.23		.25		.28		.23	
Std. error of regression	.44		.43		.42		.44	
Number of questions	12		6		6		6	
Number of subjects	1957		1957		1957		1957	
Number of observations	20551		8706		9123		11739	

Table A12: Effects of number and difficulty of response options (analysis including control variables). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. All regressions include fixed effects for each question. The baseline condition for column (1) is “three response options”; for column (2), “easy response options”; and for columns (3) and (4), “open-ended questions.” Standard errors are clustered at the respondent level.

Main Regressions Estimated with Control Variables

The regressions that we report in Table 2 do not include control variables. Table A12 reports versions of those regressions that do include control variables: a quadratic function of age, education, gender, party ID, three general political knowledge questions (measured prior to treatment), race, and state of residence. The addition of control variables does not lead to any substantive changes.

Main Regressions Estimated with Data from All Conditions

The regression on which we focus most is regression 3 in Table 2. We estimated it with data from all six questions for which we had the open-ended, “three easy,” and “five difficult” treatment

	(1)		(2)	
Intercept	.25	.01	−.18	.08
three super-easy response options	.54	.06	.53	.06
three easy response options	.35	.01	.34	.01
three difficult response options	.15	.02	.15	.02
five super-easy response options	.54	.06	.50	.05
five easy response options	.29	.02	.28	.01
five difficult response options	.06	.01	.06	.01
R^2	.10		.27	
Std. error of regression	.47		.43	
Number of questions	6		6	
Number of subjects	1957		1957	
Number of observations	11739		11739	
Includes controls	no		yes	

Table A13: Effects of number and difficulty of response options (analyses including all conditions). Cell entries are OLS estimates and standard errors. The outcomes are answers to the questions: correct = 1, incorrect = 0. The listed predictors are also coded 0 or 1. Both regressions include fixed effects for each question. The baseline condition for both regressions is "open-ended questions."

conditions. Following our pre-analysis plan, we excluded data from the other conditions: "three super-easy," "three difficult," "five super-easy," and "five easy."

Because we randomly assigned subjects to different conditions, the results do not change when we include the data from the conditions that were excluded in Table 2. Table A13 illustrates the point. It reports regressions like regression 3; the sole change is that the regressions reported here do not exclude data from any condition. Comparison of the two tables reveals that this change makes no substantive difference.

Guessing-Corrected Results

It is relatively easy for completely unknowledgeable respondents to guess the correct answers to closed-ended questions with only a few response options. While there is typically no way to know whether any particular correct response reflects knowledge or lucky guessing, we can estimate the extent to which lucky guessing affects the correct-response *rate* by using a model developed by Zinger (1972) and Luskin (2002, 287-89). The model makes three assumptions: all incorrect answers are guesses, all correct answers are guesses or answers from actually knowledgeable respondents, and the number of correct guesses is inversely proportionate to the number of response options, such that there will be fewer correct guesses if the number of response options increases. Under these assumptions, we can purge correct-response rates of the effects of guessing, thereby estimating a "guessing-corrected" correct-response rate. For example, if a question has two response options, and 90% of responses are correct while 10% are incorrect, we infer that 10% of the correct responses are due to lucky guessing. The percentage of responses that are both correct and truly knowledgeable is thus $90\% - 10\% = 80\%$. More generally, given correct-response rate C, incorrect-response rate I, and J response options (only one of which is correct), the guessing-corrected correct-response rate is $G = C - I\left(\frac{1}{J-1}\right)$.[4]

In the paper, we dicuss the variation of correct-response rates for eight national-sample closed-ended Chief Justice questions. The guessing-corrected response rates for these questions are shown in Table A14. They range from 22% in the 2010 Pew survey to 66% in Gibson and

[4] If subjects don't know the correct answer but can nevertheless eliminate one or more incorrect response options from consideration, this model understates the chance of making a lucky guess. In turn, it overstates G, the proportion of subjects who are knowledgeable. In other words, if unknowledgeable subjects can eliminate one or more response options as incorrect, this model will lead to an overestimate of the proportion who actually know the answer. See Ahler and Goggin (2017, esp. 20) for a related discussion.

	% correct (raw)	% correct (adj.)	response options	mode	year
Prior (2005)	63	55	Rehnquist, Thomas, Scalia, Kennedy	internet	2003
Gibson and Caldeira (2009)	71	66	Rehnquist, Powell, White	phone	2006
Pew Research Center	28	22	Roberts, Stevens, Marshall, Reid	phone	2010
Pew Research Center	34	29	Roberts, Breyer, Rehnquist, Reid	phone	2012
Gibson and Caldeira (2009)	46	41	Roberts; J. Harvie Wilkinson, III; Theodore Olson	phone	2006
Sen (2017)	52	40	Roberts, Rehnquist, Breyer, Scalia, Kennedy	internet	2013
ANES EGSS 3	69	61	Roberts, David Cole, Kennedy, Larry Thompson	internet	2011
ANES EGSS 4	69	62	Roberts, David Cole, Kennedy, Larry Thompson	internet	2012

Table A14: Variation in responses to closed-ended "Chief Justice" questions. This table is very similar to Table 1. The sole difference is that this table also reports guessing-corrected correct-response rates, per the formula on page A26. These percentages appear in the "adjusted" column. "EGSS" stands for "Evaluations of Government and Society Study." The correct answer to the questions reported in the first two rows is "William Rehnquist"; for the questions reported in the other rows, it is "John Roberts."

Caldeira (2009). These percentages are lower than the raw percentages in the previous column, with an average drop across the eight studies of 6.6 percentage points.

Figure A2 depicts the results of our experiment with and without guessing-corrected response rates.[5] In each row of the figure, the left-hand panel depicts the same data that we show in Figure 1, and the right-hand panel shows how the patterns change after we account

[5]Open-ended response rates cannot be corrected for guessing, but it's unlikely that guessing plays a significant role in determining the correct-response rates for these questions.

for guessing. Together, they show that correcting for guessing amplifies the main pattern that we observe in the "raw" data. To see that correcting for guessing has this effect, recall that our most important contrast is a difference of differences: the difference between the "three easy" and "five difficult" conditions on the one hand, and the difference between the "five difficult" and open-ended conditions on the other hand. In the raw data, this difference of differences is 28% − 6% = 22 percentage points. In the guessing-corrected data, it is 32% − −4% = 36 percentage points.

After we correct for guessing, the correct-response rate in the "five difficult" condition is actually lower than the correct-response rate in the open-ended condition. This result runs contrary to the general expectation that closed-ended response options will yield higher estimates of political knowledge. And it further emphasizes the importance of considering response-option characteristics when estimating political knowledge.

The top panels of Figure A2 also shed some light on the mechanisms through which increasing the number of response options has its effects. We see in Panel 1 that, on average, increasing the number of response options reduced the correct-response rate by 12.4 percentage points. If this reduction in the correct-response rate were due entirely to the increased difficulty of blind guessing—you are less likely to make a lucky guess with five options than with three—we should see evidence of it in Panel 2, which corrects for guessing. Specifically, we should see no decline at all in the guessing-corrected response rate when we switch from three response options to five. The black line in the panel should be flat.

It is not flat. To be sure, its slope is not as steep: when we correct for guessing, the decline in the correct-response rate is only 8.7 percentage points. This 12.4 − 8.7 = 3.7 percentage-point decline in the slope suggests that 3.7 percentage points of the decline that we observe in Panel 1, or 3.7/12.4 ≈ 30% of the decline, is consistent with an increase in difficulty due to blind guessing: the "purely mechanical" mechanism that we describe on page 3. But that calculation leaves the bulk of the decline, 8.7 percentage points, intact. It seems that most of the

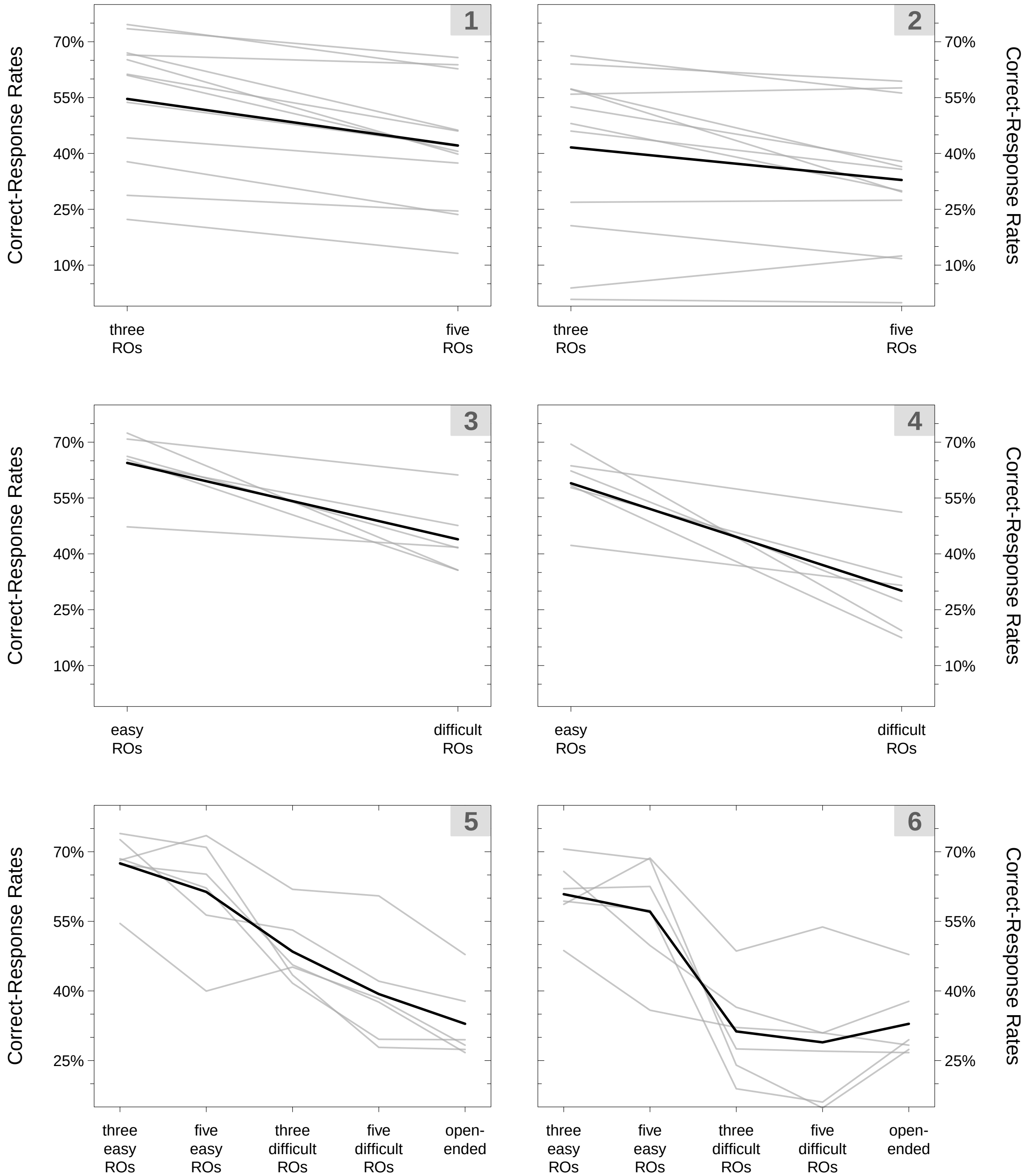

Figure A2: Correct-response rates by response-option set, with corrections for guessing. Panels 1, 3, and 5 report the same data that we reported in Figure 1. The other panels show how the patterns change when we account for guessing. "ROs" are "response options." Grey lines indicate correct-response rates for individual questions, and the black line in each panel indicates the average correct-response rate. Respondents are from SSI; N = 1,961.

decline in the correct-response rate that is brought about by increasing the number of response options may be due to non-blind guessing, satisficing, or other mechanisms.

Survey Text

It is important to us that you do not use outside sources like the Internet to search for the correct answers. We are trying to understand what people know about politics, not what they can look up. Do you agree to answer the following questions without help from outside sources? [Yes, No.]

What job or political office does Mike Pence hold? [Open-ended.]

What job or political office does Janet Yellen hold? [*Chair of the Federal Reserve System,* U.S. Attorney General, Secretary of Labor, don't know.]

How long is the term of a U.S. Senator? [2 years, 4 years, *6 years,* 8 years, 10 years, don't know.]

People have different trusted news sources. People also differ in how carefully they read survey questions. To show that you've read carefully, please ignore the question below and do not select any of the choices. Just move to the next page.

What is your most trusted source of news? [New York Times, Huffington Post, CNN, Fox News, MSNBC, ABC News, CBS News, Other.]

How many justices *currently* serve on the Supreme Court? [Difficult response options: 7, *8*, 9, 10, 12, don't know. Easy response options: 4, *8,* 16, 21, 50, don't know. Some subjects were assigned to an open-ended version of the question.]

How are Supreme Court justices chosen? [*Nominated by the President, then confirmed by the Senate;* Elected by the American people; Nominated by the President, then confirmed by the justices already on the Court; Appointed by a bipartisan Congressional committee; Recommended by the Department of Justice, then approved by the Senate; Don't know.]

Is it possible to watch lawyers argue before the Supreme Court? [No. Lawyers do not argue before the Supreme Court in person. They only submit written arguments; No. Lawyers do argue before the Supreme Court in person, but those arguments are private; Yes, except for highly sensitive national security cases. In those cases, all proceedings are private; Yes, the hearings are televised on C-SPAN; *Yes, if one goes to Washington, D.C. to see the Court in person*; Don't know.]

When there is a conflict over the meaning of the U.S. Constitution, who has the final say? [*Supreme Court;* Congress; President; Secretary of State; Department of Justice; don't know.]

Who is the Chief Justice of the Supreme Court? [Difficult response options: William Rehnquist; *John Roberts*; Clarence Thomas; Antonin Scalia; Earl Warren; don't know. Easy response options: Theodore Olson; *John Roberts*; J. Harvie Wilkinson, III; Mark Rockefeller; Homer Stille Cummings; don't know. Some subjects were assigned to an open-ended version of the question.]

Who is the Majority Leader of the U.S. Senate? [Difficult response options: John Boehner; Harry Reid; Chuck Schumer; Paul Ryan; *Mitch McConnell*; don't know. Easy response options: Elliot Anderson; Philip Baruth; James Ohrenschall; Brian Collamore; *Mitch McConnell*; don't know. Some subjects were assigned to an open-ended version of the question.]

How many justices *usually* serve on the Supreme Court? [Difficult response options: 7, 8, *9,* 10, 12, don't know. Easy response options: 4, *9,* 16, 21, 50, don't know. Some subjects were assigned to an open-ended version of the question.]

What job or political office did Horatio King hold? [Open-ended.]

There are currently eight justices on the Supreme Court. If they split 4 to 4 on a ruling in a case, what happens? [*The decision of a lower court stands;* The case is sent to Congress, which must decide it; The case is sent to the Senate Judiciary Committee, which must decide it; The Supreme Court must reconsider the case after hearing new arguments from both sides; The Attorney General must break the tie; don't know.]

How many of the current Supreme Court justices are women? [Difficult response options: 1, 2, *3,* 4, 5, don't know. Easy response options: 0, *3,* 8, 9, 12, don't know. Some subjects were assigned to an open-ended version of the question.]

Supreme Court Justices' salaries are determined by Congress. The Constitution prohibits Congress from reducing Justices' salaries while the Justices are currently serving. Please choose the "Don't know" option below. [$233,000; $300,000; $412,000; Don't know.]

How long is the term of a Supreme Court justice? [Difficult response options: 4 years, 6 years, 8 years, 10 years, *Life term*, don't know. Easy response options: 1 year, 2 years, 3 years, Until age 60, *Life term*, don't know. Some subjects were assigned to an open-ended version of the question.]

Which of these statements best describes the Supreme Court's power? [*It can interpret laws only to settle disputes between two parties in a legal case;* It can give advice on the interpretation of a law at any time, even when no legal case is involved; It can advise the president on the meaning of a law when he asks for advice; It can advise the Senate Judiciary Committee on the meaning of a law when it asks for advice; It can give advice on the interpretation of a law only if the law was passed within the last year; Don't know.]

Can Supreme Court justices ever be removed from office? [Yes, if their fellow justices unanimously agree to impeach them; *Yes, if they are impeached by the House of Representatives and convicted by the Senate;* Yes, if they are too sick or too old to do the job; Yes, if they are prosecuted by the Attorney General; No, they can never be removed; Don't know.]

Generally speaking, do think of yourself as a . . . [Democrat, Republican, Member of another party, Independent or unaffiliated.]

[Democrats only.] Would you call yourself . . . [A strong Democrat, A Democrat.]

[Republicans only.] Would you call yourself . . . [A strong Republican, A Republican.]

[Only independents, unaffiliated subjects, or members of minor parties.] Do you think of yourself as closer to the Democratic Party of the Republican Party? [Democratic Party, Republican Party, Equally close to both.]

In what year were you born? [Open-ended.]

What is your race? [Open-ended.]

What is your gender? [Open-ended.]

What is the highest level of schooling that you have completed? [No formal schooling; 1st grade; 2nd grade; 3rd grade; 4th grade; 5th grade; 6th grade; 7th grade; 8th grade; 9th grade; 10th grade; 11th grade; 12th grade, no GED or diploma; 12th grade, diploma; Graduate equivalence degree; Some college, no degree; Associate's degree; Bachelor's degree; Master's degree; Other post-college degree.]

Where do you live now? [Options for 50 states, American Samoa, District of Columbia, Guam, Northern Mariana Islands, Virgin Islands.]

Coding Race

We measured our subjects' races via an open-ended question: "What is your race?" SSI had previously used a different question—the U.S. Census Bureau's question—to measure the races of the same subjects. Comparing the two sets of answers, we found that approximately 7% of the sample had identified as Hispanic in the SSI data while identifying as white, or giving an uncodable answer, in our data. In the analyses of ours that include control variables, we rely mainly on the SSI race variable. It is missing values for 120 cases; we fill in these cases with data from our open-ended variable. Unsurprisingly, given that our data are experimental, it makes no difference to our analyses if we instead use the open-ended variable alone, use the SSI variable alone, or simply drop the race variable from our analyses.

Construction of Open-Ended Items

One major point of this study is to compare open-ended items, as commonly constructed, to closed-ended items, as commonly constructed. Thus, like every other open-ended knowledge question of which we are aware (for example, those in the ANES), the open-ended questions asked here were not accompanied by a separate "don't know" option. In addition to endangering the applicability of the results presented here to ordinary research, the use of a separate "don't know" option would raise doubts about whether the questions are really open-ended at all.

One might object that the absence of a separate "don't know" option—for example, a separate "don't know" checkbox—threatens the internal validity of the experiment, given that our closed-ended items did include "don't know" options. But this objection is not plausible. The outcomes of interest in our study are the percentages of respondents answering questions correctly. To affect our results, the inclusion of a separate "don't know" option would need to cause subjects to answer correctly at meaningfully different rates. And this outcome is not plausible.

Perhaps the best evidence on this point comes from an experiment on open-ended questions in the 2000 ANES, in which some subjects were asked open-ended questions in a neutral format, while others who initially answered "don't know" were probed with a follow-up question ("Well, what's your best guess?"). The results show that a relatively strong discouragement of "don't know" responses—probing—had only a minimal effect on the overall percentages answering correctly. For none of the four questions in the experiment did the percentage answering correctly increase by as much as three percentage points; the average increase was 1.6 percentage points (Luskin and Bullock 2011, 552-53). Not including a "don't know" response option for open-ended questions may be considered a milder discouragement of "don't know" responses, and in turn, its effects on the percentages answering correctly are likely to have been milder as well. In any case, even changes larger than 1.6 percentage points would not affect any of the substantive conclusions that can be drawn from our study.

Coding of Open-Ended Items

For six questions in our experiment, some subjects were randomly assigned to an open-ended question format: they were presented with the question and given a blank text box in which to enter their answers.

For some questions, the coding of open-ended responses is notoriously difficult: see page A3 for a discussion. The difficulties were somewhat muted in our case because the open-ended questions in our experiment were relatively simple, and the correct answers relatively unambiguous:

- How many justices *currently* serve on the Supreme Court?
- Who is the Chief Justice of the Supreme Court?

- Who is the Majority Leader of the U.S. Senate?
- How many justices *usually* serve on the Supreme Court?
- How many of the current Supreme Court justices are women?
- How long is the term of a Supreme Court justice?

For each of these questions, coding of answers as correct or incorrect proceeded in three stages. First, we converted all responses to lowercase. Second, we conducted a preliminary coding by stipulating a set of patterns (that is, a regular expression), at least one of which we expected to find in any correct response. Third, we read the open-ended responses and reviewed the preliminary coding, making any necessary adjustments.

An example will illustrate the procedure. Consider the question in our experiment whose responses are arguably most difficult to code: “How long is the term of a Supreme Court justice?” Article III of the U.S. Constitution says that justices “shall hold their Offices during good Behaviour,” and this clause has been taken to mean that justices serve for as long as they please or unless they are impeached. Colloquially, they have “life terms.” After converting all responses to lowercase, we began by preliminarily coding as correct responses that contained any of the following words or patterns: “death”; “forever”; “indefinite” (including “indefinitely”); “life” (including “lifetime”); “long” followed by “want” or “choose” (as in “as long as they choose”); “no limit”; “resign” (as in “until they resign”); and “retire” (as in “until they retire”) or “retirement.”

Obviously, this set of patterns does not speak to all possible correct answers. We therefore read every open-ended response and adjusted the coding as needed. This review led us to mark as correct three responses that we had preliminarily coded as incorrect:

1. “as long as they can serve,”
2. “Indefintely” (note the misspelling), and

3. "For ever" (with a space between the two words).

The same procedure also led us to change the coding for three responses that were given in answer to the other five questions.

Pre-Analysis Plan

Our pre-analysis plan was pre-registered at the Political Science Registered Studies Dataverse: see https://doi.org/10.7910/DVN/HU04WI. The remainder of this section includes the content of the plan.

Recent research suggests that political scientists have understated popular knowledge of the U.S. Supreme Court. This research implies that popular knowledge of other aspects of government has been understated as well. The implication may be correct. But these revisionist studies of political knowledge may err in the other direction, by overstating popular knowledge of politics. Focusing on the Supreme Court, we will use an experiment conducted with a representative national sample to test the idea that inferences about popular knowledge of politics depend heavily on little-appreciated aspects of survey design and analysis.

Experimental Manipulations

We will field a survey of a national sample of U.S. adults. Most of the questions in the survey will be factual questions about the U.S. Supreme Court.

The study will include two main types of manipulation. One is a number-of-response options manipulation: for a given question, subjects will be randomly assigned to see three or five response options (in addition to "don't know"). When subjects are assigned to see only three options, the two incorrect options that they see will be drawn at random from the complete set of four incorrect options. Subjects who are instead assigned to the five-response-option condition will simply see the correct response and the four alternatives, presented in random order.

For some questions, some subjects will be assigned to an open-ended condition. These subjects will see no response options at all; instead, they will see an open-ended version of the question.

The second type of manipulation involves the difficulty of response options. For a given question, subjects will be randomly assigned to see a "difficult" or an "easy" set of response options; a "difficult" set is one that we expect will make it more difficult to identify the correct option.

Twelve of our questions involve a number-of-response-options manipulation. Six involve a difficulty-of-response-options manipulation. All six questions that involve a difficulty manipulation also involve a number-of-response-options manipulation and an open-ended condition. There are thus five conditions for each of these six questions: three easy options, five easy options, three hard options, five hard options, and an open-ended condition.[6]

Random assignments will be done at the question-subject level. For example, a subject who is assigned to the easy three-option condition for one question may be assigned to the difficult five-option condition for another question, and to the open-ended condition for a third question.

Hypotheses

1. Correct-response rates are decreasing in the number of response options.
2. Correct-response rates are decreasing in the difficulty of response options.
3. The difference between correct-response rates to questions that have a short-easy set of response options and questions that have a long-difficult set of response options will be greater than, or equal to, the difference between correct-response rates to questions that have a long-difficult set of response options and the corresponding open-ended questions.

We refer to Hypothesis 3 as our "main hypothesis."

[6] We will also include a set of "extremely easy" response options to one question. See the questionnaire at the end of this document for details.

Estimation

To test Hypothesis 1, we will pool responses to all of the questions in which we manipulate the number of response options. We will use the pooled data to estimate two models:

$$correct_{ij} = \alpha + \beta_1 (\textit{five response options})_{ij} + \lambda_j + \epsilon_{it}, \text{ and} \tag{A1a}$$

$$correct_{ij} = \alpha + \beta_1 (\textit{five response options})_{ij} + \lambda_j + \gamma \mathbf{X_{it}} + \epsilon_{it}, \tag{A1b}$$

where i indexes subjects, j indexes questions, and $correct_{ij}$ equals 1 if subject i answered question j correctly, 0 otherwise. λ_j is a vector of fixed effects for each question, and $\mathbf{X}_{it}$ is a vector of control variables: age, age^2, educational attainment, gender, an index of general political knowledge (measured prior to treatment), party ID, race, and state of residence. We will not causally interpret these covariates; we will include them only to improve the precision of the estimates of β_1.

Our procedure for the testing of Hypothesis 2 is much the same. We will pool responses to the all of the questions in which we manipulate response-option difficulty. We will use the pooled data to estimate two models:

$$correct_{ij} = \alpha + \beta_1 difficult_{ij} + \lambda_j + \epsilon_{it}, \text{ and} \tag{A2a}$$

$$correct_{ij} = \alpha + \beta_1 difficult_{ij} + \lambda_j + \gamma \mathbf{X_{it}} + \epsilon_{it}, \tag{A2b}$$

where $difficult_{ij}$ indicates whether subject i was assigned to the “difficult response options” condition for question j. The other variables are as described above.

Our main hypothesis is a difference of differences. To estimate it, we will pool data from the six questions for which we manipulate both the number and the difficulty of response options. As the hypothesis is about correct-response rates in three conditions—“three easy response options," "five hard response options," and the open-ended condition—we will exclude data from other conditions. We will estimate two models:

$$correct_{ij} = \alpha + \beta_1(three\ easy)_{ij} + \beta_2(five\ difficult)_{ij} + \lambda_j + \epsilon_{it}, \text{ and} \tag{A3a}$$

$$correct_{ij} = \alpha + \beta_1(three\ easy)_{ij} + \beta_2(five\ difficult)_{ij} + \lambda_j + \gamma\mathbf{X_{it}} + \epsilon_{it}, \tag{A3b}$$

where $(three\ easy)_{ij}$ + $(five\ difficult)_{ij}$ indicate the condition to which subject i was assigned for question j. The reference category is the open-ended condition. The test of the hypothesis is $(\beta_1 - \beta_2) > \beta_2 \Leftrightarrow \beta_1 - 2\beta_2 > 0$.

Statistical Power and Allocation of Subjects

In an analysis that follows the form of Equation A3a, pilot-test data suggest that $\alpha = .50$ and $\beta_2 = .17$. Suppose that the difference between correct-response rates to questions that have a short-easy set of response options and questions that have a long-difficult set of response options is 30% greater than the difference between correct-response rates to questions that have a long-difficult set of response options and the corresponding open-ended questions. We therefore have $\beta_1 = 2.3(\beta_2) = .391$. Our simulations suggest that, under these conditions, we will detect $\beta_1 - 2\beta_2 > 0$ at $p < .05$ (one-tailed) in approximately 80% of experiments that have 467 subjects in each of the "three easy," "five difficult," and open-ended conditions.

We will collect data from approximately 2,000 subjects. For most questions in which we manipulate both the difficulty and the number of response options, we will use this allocation schedule: approximately 536 subjects in the "three-easy" and "five-difficult" conditions, 500 subjects in the open-ended condition, and 214 in the "three-difficult" and "five-easy" conditions.

One of our questions includes an "extremely easy" closed-ended condition. For this question, we will use a slightly different allocation schedule: approximately 500 subjects in the "three-easy," "five-difficult," and open-ended conditions; approximately 200 in the “three-difficult” and “five-easy” conditions; and approximately 100 in the extremely easy condition.

Measurement and Coding of Covariates

The last pages of the appended questionnaire include the questions that we will use to measure each covariate.

Education. We will treat the educational attainment of our subjects as a categorical variable. That is, the regressions that include covariates will include indicators for every level of education (“12th grade, diploma,” “12th grade, no GED or diploma,” and so on), save for a reference category.

Gender. Our question about the gender of our subjects is open-ended. We will use the responses to this question to create an indicator for “female”; this is the gender covariate that we will include when we control for covariates.

Political knowledge (general). Our survey will include three general political knowledge questions, all asked before any treatment assignments are conducted. We will create one indicator for each variable (1 = correct response; 0 = incorrect response, “don’t know,” or skipped question) and will control for all of these variables when we estimate models 1b, 2b, and 3b. We will not control for an index of political knowledge when we estimate those models.

Race. Our question about the race of our subjects is open-ended. We will use the responses to this question to create six race indicators: “Asian,” “black,” “Hispanic or Latino,” “multiracial,” “white,” and “other race.” We will control for five of these six indicators; the sixth will be our reference category.

Screening and Elimination of Subjects

Commitment screener. Following Clifford and Jerit (2016), we will ask subjects to commit to not looking up answers online. We will make this request at the start of the survey, and we will omit from our analyses all subjects who do not make the commitment. In our previous research, fewer than 1% of subjects refused to make the commitment.

Placebo question. As in Bullock et al. (2015), our survey will contain a "placebo question." The question is about an obscure historical figure; the answer is known to almost no one, but it is trivial to discover if one uses Google or other online search tools. We will use this question as a measure of the extent to which subjects use online tools to look up the answers to our other questions. But we will not exclude subjects who answer the question correctly from our analyses.

Attention screeners. Following Berinsky, Margolis, and Sances (2014), we will use two "attention screeners" to gauge whether respondents are attending to our questions. We will not exclude subjects who fail these screeners from our analyses.

Break-offs. Following the AAPOR Standard Definitions (AAPOR 2016, 14-15), we define break-offs as interviews with subjects who answered fewer than half of our questions. We will eliminate all break-offs from our analyses.

Missing Covariate Values

Some of our models call for use of the battery of covariates described in the "Estimation" section. To ensure that all models are estimated with data from the same set of subjects, we will estimate models with data from only those subjects who provide information (e.g., demographic information) for all of the covariates listed above.

Implausible age data. Age will be computed from answers to our question about year of birth. Answers indicating age less than 18 or greater than 100 (e.g., year of birth of 1874 or 22222) will be treated as missing data.

Checks for Covariate Imbalance

We will regress an indicator for each treatment on the covariates described in the "Estimation" section. For each regression, the F statistic will be our gauge of covariate imbalance (Gerber and Green 2012, 107). F-statistic p values of less than 0.01 will prompt a review of the random assignment procedures and possible data-handling mistakes. If we find no errors, we will report the imbalance, proceed on the assumption that it is due to chance, and report estimates with and without covariate adjustment.

Appendix References

Ahler DJ and Goggin SN (2017) Assessing Political Knowledge: Problems and Solutions in Online Surveys. Working paper, 2017 May 02. http://www.sgoggin.org/papers/AhlerGoggin.KnowledgePaper.pdf.

American Association for Public Opinion Research (2016) Standard Definitions. https://perma.cc/6V72-MDN5.

Berinsky AJ, Margolis MF and Sances MW (2014) Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys. *American Journal of Political Science* **58** (3), 739–53.

Boudreau C and Lupia A (2011) Political Knowledge. In Druckman JN et al. (eds.), *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press.

Bullock JG et al. (2015) Partisan Bias in Factual Beliefs about Politics. *Quarterly Journal of Political Science* **10**, 519–78.

Clifford S and Jerit J (2016) Cheating in Political Knowledge Questions in Online Surveys. *Public Opinion Quarterly* **80** (4), 858–87.

DeBell M (2013) Harder than It Looks: Coding Political Knowledge on the ANES. *Political Analysis* **21**, 393–406.

Delli Carpini MX and Keeter S (1996) *What Americans Know about Politics and Why It Matters*. New Haven, CT: Yale University Press.

Flood S et al. (2018) Integrated Public Use Microdata Series, Current Population Survey: Version 6.0 [dataset]. Minneapolis, MN: IPUMS, 2018. https://doi.org/10.18128/D030.V6.0.

Gerber AS and Green DP (2012) *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton.

Gibson JL and Caldeira GA (2009) Knowing the Supreme Court? A Reconsideration of Public

Ignorance of the High Court. *Journal of Politics* **71** (2), 429–41.

Luskin RC (1987) Measuring Political Sophistication. *American Journal of Political Science* **31** (4), 856–99.

Luskin RC (2002) From Denial to Extenuation (and Finally Beyond): Political Sophistication and Citizen Performance. In Kuklinski JH (ed.), *Thinking about Political Psychology*. New York: Cambridge University Press.

Mondak JJ (2001) Developing Valid Knowledge Scales. *American Journal of Political Science* **45** (1), 224–38.

Prior M (2005) News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout. *American Journal of Political Science* **49** (3), 577–92.

Sen M (2017) How Political Signals Affect Public Support for Judicial Nominations: Evidence from a Conjoint Experiment. *Political Research Quarterly* **70** (2), 374–93.

Zinger A (1972) A Note on Multiple-Choice Items. *Journal of the American Statistical Association* **67** (338), 340–41.