Public Opinion Quarterly (2025) **Vol 00 No 0,** 1–20

Asking about Complex Policies

Stephen Jessee[1], Neil Malhotra[2,*], Maya Sen[3]

[1]Professor, Department of Government, University of Texas, Austin, TX, US
[2]Edith M. Cornell Professor of Political Economy, Graduate School of Business, Stanford University, Stanford, CA, US
[3]Professor, John F. Kennedy School of Government, Harvard University, Cambridge, MA, US

Abstract As political issues have increased in complexity, public opinion researchers increasingly ask respondents about sophisticated political topics that may require substantive knowledge and analytic skills, raising concerns about survey satisficing. In this note, we examine the correlates of one manifestation of this kind of satisficing—response order effects, or when respondents are more likely to pick the top response options from a list. We do this by analyzing randomized response order experiments embedded in surveys conducted across three years, where 6,291 respondents provided their opinion on 40 complex issues before the Supreme Court. We find an overall response order effect of 2.8 percentage points with substantial heterogeneity related to both question length and respondent knowledge. These factors also interact with one another; among the most sensitive subpopulation—low-knowledge respondents answering long questions—the predicted response order effect was 17.4 percentage points. On the other hand, question complexity and respondent education did not moderate response order effects. Our practical advice for researchers asking about complex issues is that they should focus on being brief, rather than using extra words to make the language simpler.

Survey researchers have worried about the responses they obtain from asking about complex policies, concerned that respondents are providing nonattitudes or data measured with error (e.g., Converse 1964; Alvarez and Brehm 2002). These concerns have only increased as the world—and political issues—have increased in complexity. For instance, pollsters asked about President Joseph Biden's Build Back Better legislation, a complex piece of policy with many components (e.g., Prokop 2021). More generally, every two years, the Cooperative Election Study (CES) asks people how they

*Corresponding author: Neil Malhotra, Graduate School of Business, Stanford University, 655 Knight Way, Stanford, CA 94305, US; email: neilm@stanford.edu.

https://doi.org/10.1093/poq/nfae050

would vote on various pieces of legislation in an attempt to jointly scale the ideological locations of the public and members of Congress (e.g., Ansolabehere and Kuriwaki 2022). Public opinion researchers also have asked about complex legal issues concerning the role of the Supreme Court in US politics (e.g., Jessee, Malhotra, and Sen 2022). These issues are compounded by worries that inattentive survey respondents populate internet survey panels, which are an increasingly used data collection methodology (e.g., Berinsky, Margolis, and Sances 2014).

The literature has generally argued that task difficulty and respondent ability can compromise data quality because they are related to survey satisficing (e.g., Krosnick 1991, 1999). One sign of satisficing is response order effects, or when respondents pick the top response option from a list in visually administered questionnaires. Such a practice can introduce both bias and error in survey data (Krosnick and Alwin 1987; Schuman and Presser 1996). The empirical literature is somewhat mixed, with some studies finding that order effects are more present in complex questions (e.g., Holbrook et al. 2007) and other studies finding the reverse relationship (e.g., Malhotra 2009).

In this research note, we assess the correlates of response order effects by analyzing survey responses to complex questions about cases before the Supreme Court and related political issues. We leverage randomized response order experiments embedded in three surveys conducted in 2020, 2021, and 2022. Across these three waves of data collection, 6,291 respondents provided their opinion on 40 upcoming Supreme Court decisions, providing over 84,000 observations for analysis.

On average, we observe that respondents were 2.8 percentage points more likely to select a response option when it was listed in the first compared to the second position. However, we also find substantial heterogeneity in this effect across questions and respondents. Questions containing more words exhibited higher response order effects, while respondents with more domain-specific knowledge about the Supreme Court were less likely to exhibit response order effects. Additionally, these two variables exhibited a significant interaction: the relationship between question length and response order effects was strongest for the least domain-knowledgeable respondents. Although response order effects are estimated to be close to zero for shorter questions and for higher domain-knowledgeable respondents, we find that among the most sensitive subpopulation—low-knowledge respondents answering long questions—the predicted response order effect was extremely large, being estimated at 17.4 percentage points. On the other hand, linguistic complexity/readability and respondent education did not moderate response order effects.

Our findings have important implications for survey researchers trying to understand public opinion, especially on complex policy issues. First, a trade-off exists between providing detail and ensuring response quality.

Adding more words to a question may make the description of the issue—particularly on a complex topic—more informative and precise, or even allow for simpler language. However, it comes at the expense of possibly greater survey satisficing. Researchers should attempt to capture the essence of a question in as few words as possible. Second, complexity of language does not moderate response order effects, meaning that researchers should not shy away from using complex language if this can decrease overall question length. Third, although our advice applies very generally to both complex and noncomplex policies, complex policies are ones where people may have less domain knowledge to begin with, and so survey responses to such questions may be most vulnerable to bias. Fourth, substantive knowledge of a topic and education are distinct concepts and education does not moderate response order effects. Hence, weighting by demographics is not a comprehensive solution since it may not fully capture domain knowledge. Further, as respondents with low domain-specific knowledge about the topic exhibit larger response order effects, reported differences across subgroups can be confounded by survey artifacts. Finally, although our findings underscore the importance of randomly rotating response options, they also sound a note of caution. Random rotation reduces bias but does not eliminate measurement error. Nonetheless, variance can be reduced via increasing sample size whereas bias is unaffected.

In summary, the primary concern with asking about complex policies is not the readability of the text or the cognitive sophistication of respondents, but rather issues related to attentiveness and domain knowledge of satisficing respondents. Researchers should worry less about uneducated respondents reading complex text and more about disinterested respondents reading long questions.

Theoretical Motivation

Our theoretical motivation is based on Krosnick's (1991) theory of survey satisficing, which argues that manifestations of inattentive survey response—including selecting the first response alternative presented—is a function of three main variables: (1) task difficulty; (2) respondent ability; and (3) respondent motivation. The first variable represents an item-level characteristic, whereas the latter two represent respondent-level characteristics. Krosnick notes that these three factors can have both additive and interactive effects on satisficing. Sudman, Bradburn, and Schwartz (1996) argue that for visually presented scales, response order effects are primarily a manifestation of survey satisficing rather than memory limitations or cognitive elaboration.

We also draw on Sherif and Hovland's (1961) concept of the "latitude of acceptance," or the idea that there is a zone of attitudes and statements that individuals are comfortable attaching themselves to. Applied to survey methodology, multiple response options may be acceptable in a respondent's mind, so they may be willing to select the first one they see that is within their latitude of acceptance. Attitudes can be represented as latent, unobservable variables with survey responses being crude mechanisms by which individuals report their underlying attitudes with measurement error. Survey methodologists are often concerned with how various measures (e.g., the number and labels on Likert scales) allow respondents to map latent attitudes onto survey measures. However, respondents themselves can influence this mapping independently of how questions are written. If respondents have low ability or motivation, they may have a wide latitude of acceptance, and therefore be more willing to select the top response option if multiple options fall within the latitude. Similarly, complex question stems can lead to difficulties in processing information and may cause multiple response options to fall within the latitude.

In this study, we examine two main item-level characteristics: *question length* and *question complexity/readability*. Although both are posited to relate to task difficulty, we argue that they are distinct concepts. Text can be long but use simple structure and syntax; alternatively, it can be short but have very complex structure.[1] In the following section, we describe how we measure these two variables.

We examine two main respondent-level characteristics: *education* and *domain knowledge*. Both of these variables have been suggested to tap respondent ability and motivation (Krosnick 1991), but we conceive of them as distinct constructs. Education taps cognitive and linguistic skill (Cor et al. 2012) and therefore the direct ability to process complex text. On the other hand, domain knowledge proxies for interest and familiarity with the subject matter, which is likely correlated with respondent attention and motivation to process complex language with respect to policy issues. As explained below, these two variables are far from perfectly correlated in our data, as there are policy topics that can draw (dis)interest from diverse groups of respondents.

For designers of surveys, it is important to know whether survey satisficing is driven by question length, readability, or both—and respondent education, domain knowledge, or both—as these have important implications for how questions are designed and how researchers should weigh competing trade-offs when writing survey items. We return to these issues when we discuss recommendations based on our empirical findings.

[1] Previous studies have documented that response order effects increase with more words and syllables in questions and response options (e.g., Schuman and Presser 1996; Holbrook et al. 2007). Although some scholars have argued that readability moderates satisficing (e.g., Kimball and Kropf 2005; Velez and Ashworth 2007), other research has questioned whether readability is a relevant construct for evaluating survey questions (Lenzner 2014).

Design, Measures, and Methods

We conducted three national surveys, each asking respondents' opinions on the key issues on prominent cases from the Supreme Court's docket. The bulk of the survey questions were only a few sentences long and used straightforward language, but they asked respondents' opinions on complex, multifaceted Supreme Court cases involving issues such as civil rights and presidential powers.

The surveys were conducted by YouGov. The data collection periods for the three surveys were: (1) April 29, 2020–May 12, 2020, for the 2020 wave (n = 2000), (2) April 7, 2021–April 14, 2021, for the 2021 wave (n = 2,158), and March 30, 2022–April 6, 2022, for the 2022 wave (n = 2,133). The participation rates (the proportion of panelists invited to take our survey who completed the survey) are 69 percent, 71 percent and 87 percent, respectively, for the three waves.[2] Full question wordings can be found in Supplementary Material sections 1 and 2. Each survey was administered online by YouGov using a diverse, national sample of American adults recruited as part of their main panel study. Using sample matching procedures, YouGov matches respondents drawn from their opt-in panel to representative benchmark datasets such as the American Community Survey, the Current Population Survey, the Pew US Religious Landscape Survey, and voter files (Ansolabehere and Schaffner 2014). YouGov draws a random sample of respondents from these data sources to create a "target sample." It then matches individuals from its opt-in internet survey panel via perfect replacement such that the survey sample is equivalent to the target sample. Sample matching has been shown to perform well; studies that have conducted concurrent surveys comparing YouGov against probability samples demonstrate similar results across sampling methods (Rivers and Bailey 2009; Ansolabehere and Schaffner 2014). This includes not only means and distributions of variables, but also relationships among survey variables and similarities to verifiable benchmarks. Nonetheless, it is important to emphasize that YouGov does not construct probability samples. All statistical analyses reported below apply poststratification weights provided by YouGov unless otherwise noted. Cases are weighted by gender, race, education, and age; large weights are trimmed and then normalized. Descriptive statistics of the samples as well as the variables of interest can be found in Supplementary Material section 3 (table A1).

[2] YouGov does not use traditional probability sampling and is therefore unable to calculate an AAPOR response or cooperation rate. Furthermore, when potential respondents are invited to take a survey, they are only routed to a specific survey after agreeing to participate.

Response Order Manipulation

Respondents were provided with two response options, indicating the two sides of the case. Half the respondents were assigned to "Form A," where they saw the response options presented in the order shown in the Supplementary Material (with one side's argument presented first), and the other half were assigned to "Form B," where they saw them in the reverse order (with the other side's argument presented first).

Measures: Item-Level Characteristics

Question length

Our main item-level predictor of interest is the length of the question as measured by the number of words, averaged across the two treatments.[3] This variable ranges between 53 and 131 words, averaging 99.3 with a sample standard deviation of 20.6. To assess robustness, we also operationalized this variable separately as (1) the log of the number of words (see Supplementary Material figures A1 and A2); and (2) the number of characters including spaces (see Supplementary Material figures A3 and A4). For both alternative operationalizations of question length, we obtained similar overall results.

Linguistic complexity

We measure linguistic complexity using the Flesch-Kincaid readability score (Flesch 1948; Kincaid et al. 1975), which is a long-used measure of the ease of readability of English text. The formula for calculating the score is 206.835—1.015 (total words/total sentences) - 84.6 (total syllables/total words), with higher numbers indicating easier readability. Intuitively, readability becomes easier as one writes shorter sentences, uses words with fewer syllables, and increases the ratio of total words to complex words. As a robustness check, we also constructed an index based on the first principal component of several measures related to textual complexity: the Flesch-Kincaid score, verbs per sentence, subordinate clauses per sentence, and complex words (i.e., words with three or more syllables) per sentence. All of these variables adjust for text length.[4]

[3] The number of words was very similar within questions between treatments, differing on average by less than two words between Form A versus Form B.

[4] We also considered using various other readability measures, including the SMOG formula, the Fry readability graph, the FOG index, and the Dale-Chall formula. However, these other measures require longer text lengths for validity (100–150 words) and several of our survey items fell below this threshold (Lenzner 2014).

As mentioned earlier, question length and linguistic complexity are distinct constructs. In our data, the number of words and the Flesch-Kincaid score are correlated at r = -0.18. See Supplementary Material table A2 for a correlation matrix of the key variables of interest and Supplementary Material table A3 for assessment of potential multicollinearity.

Respondent-Level Characteristics

Substantive knowledge of the Supreme Court

In addition to their opinions on cases, respondents were also asked six questions to assess their level of domain-specific knowledge of the Supreme Court: (1) what the justices' term lengths are; (2) whether justices are appointed or elected; (3) which branch of government has final say over the Constitution's meaning; (4) the name of the current Chief Justice; (5) the name of the most recently appointed justice; and (6) the number of justices appointed by Republican presidents, all at the time of the survey. We operationalized Court knowledge as the proportion of these questions that a respondent answered correctly, ranging from all six questions incorrect (knowledge = 0) to all six questions correct (knowledge = 1). The mean knowledge level was 0.63 with a standard deviation of 0.29.

Education

As part of their profiles as YouGov panelists, we obtained respondents' level of education within six categories: (1) less than high school; (2) high school graduate; (3) some college; (4) two-year degree; (5) four-year degree; and (6) postgraduate degree. Substantive knowledge and education were slightly positively correlated in our data (r = 0.31), indicating that they are distinct constructs per our argument above.

Other Control Variables

To assess robustness, we also included a series of control variables not tied to our theoretical framework but that may potentially moderate response order effects. Including these control variables—either alone or when interacted with the response order manipulation—does not meaningfully change any of our substantive conclusions or point estimates (see Supplementary Material tables A5 and A6).

Respondent-level controls

In our robustness checks, we include the following respondent-level control variables: gender (male, female); age (under 30, 30–44, 45–64, 65+); race (white, non-white); and party identification (strong Republican, not-strong

Republican, lean Republican, Independent, lean Democrat, not-strong Democrat, strong Democrat). We standardize all control variables to have a mean of 0 and a standard deviation of 1 to aid interpretation of our models' estimates.

Item-level controls

We include the following set of item-level controls: (1) whether the case was mentioned on the front page of the *New York Times* (Epstein and Segal 2000), designed to assess whether more salient and well-known cases exhibited smaller levels of satisficing; and (2) whether the case dealt with one of three highly salient civil liberties issues (abortion/contraception, gun control, or LGBT rights), which are topics that are extensively covered by prominent media outlets (Epstein and Segal 2000) and include some of the most well-known rulings (C-Span 2009). These are topics that are critical in lay understandings of the Court and therefore may exhibit smaller response order effects.[5] These item-level controls are also standardized to aid model interpretation.

Statistical Models

We pool together all three waves and stack the data such that the unit of analysis is respondent *i*'s survey response regarding case *j*, and we cluster standard errors by respondent (since each respondent provided answers to more than one case question). The first OLS regression[6] model we estimate is:

$$Y_{ij} = \alpha + \beta_1 T_i + \lambda_w + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is a dummy variable coded "1" if the respondent selected the "top" response option from Form A (which is instead the bottom response option in Form B) as presented in the Supplementary Material and "0" if they selected the "bottom" response option from Form A (which is instead the top response option in Form B), T_i is a dummy variable coded "1" if the respondent was assigned to "Form A" (how the response options appear in the Supplementary Material) and "0" if the respondent was assigned to "Form B" (if the response

[5] The list of cases dealing with these items is identified in Supplementary Material Section 1.

[6] We estimate linear probability models (i.e., OLS regression models predicting a binary dependent variable) per standard practice in the econometrics literature (Angrist and Pischke 2009) and for ease of interpretability. The overall conclusions are unaffected if we estimate logistic regression models instead. Below, we report results from these models as well. Analyses using more flexible, less parametric methods such as LOESS also produce very similar overall conclusions. As described below, we conduct numerous robustness checks and none of our results are sensitive to model specification.

options appeared in the reverse order), λ_w represents survey-wave fixed effects, and ε_{ij} is stochastic error. Hence, a primacy effect (i.e., a response order effect where respondents are selecting response options nearer the top on a visually administered response scale) would be manifested if β_1 is positive.

We next test our main theoretical hypotheses of interest. With respect to item-level characteristics, we examine if primacy effects are stronger for longer items (those with more words W_j) and for items with more complex language (C_j). With respect to respondent-level characteristics, we examine if response order effects are more pronounced among respondents with low domain knowledge, as measured by the proportion of Supreme Court knowledge questions they correctly answered (K_i), and also among respondents with lower education (E_i). Hence, we estimate the two following models:

$$Y_{ij} = \alpha + \beta_1 T_i + \beta_2 W_j + \beta_3 C_j + \beta_4 (T_i \cdot W_j) + \beta_5 (T_i \cdot C_j) + \lambda_w + \varepsilon_{ij} \quad (2)$$

$$Y_{ij} = \alpha + \beta_1 T_i + \beta_2 K_I + \beta_3 E_i + \beta_4 (T_i \cdot K_I) + \beta_5 (T_i \cdot E_I) + \lambda_w + \varepsilon_{ij} \quad (3)$$

where the primacy effect is given by $\beta_1 + \beta_4 W_j + \beta_5 C_j$ in Equation (2) and by $\beta_1 + \beta_4 K_i + \beta_5 E_i$ in Equation (3). Our expectation is that questions with longer words will have larger primacy effects, all else equal; in other words, that β_4 will be positive in Equation (2).[7] Similarly, our expectation is that in Equation (3), β_4 will be negative, which would indicate that more-knowledgeable respondents exhibit smaller primacy effects.

As we show below, question length significantly moderates response order effects but question complexity does not. Further, respondent domain knowledge moderates responder order effects but education does not. Hence, in our final model, we assess whether there is a three-way interaction between survey form (the dummy variable representing whether respondents received Form A), item length, and respondent's domain knowledge. Specifically, we estimate:

$$\begin{aligned} Y_{ij} = \alpha + \beta_1 T_i + \beta_2 W_j + \beta_3 K_i + \beta_4 (T_i \cdot W_j) + \beta_5 (T_i \cdot K_i) + \beta_6 (W_j \cdot K_i) \\ + \beta_7 (T_i \cdot W_j \cdot K_i) + \lambda_w + \varepsilon_{ij} \end{aligned} \quad (4)$$

If β_7 is negative, this would imply that the relationship between response order effect and word length is larger for less domain-knowledgeable people or, equivalently under this specification, that the relationship between response order effects and knowledge is more strongly negative for longer questions. Using the estimates from model (4), we also estimate the overall

[7] Note that it will also be important to calculate this primacy effect estimate over the range of question word lengths in the data (as we do below) and also to examine the estimated coefficient β_5, which estimates how the treatment effect varies with question complexity and respondent education in models (2) and (3), respectively.

primacy effect over the range of question-word lengths and respondent domain knowledge levels in the data, which from Equation (4) is:

$$\beta_1 + \beta_4 \cdot W_i + \beta_5 \cdot K_i + \beta_7 \cdot (W_j \cdot K_i) \tag{5}$$

As a robustness check, we also estimate versions of these models including control variables and interactions between survey form and these controls (see Supplementary Material tables A5 and A6). The point estimates are similar across all specifications.

We estimated a series of different model types to ensure that our results were not sensitive to our OLS regression specification. First, we estimated models (1)–(4) using logistic regression given that the outcome variable is binary. Second, we estimated a hierarchical logistic regression that models the fact that questions are nested within respondents by using random intercepts. Our linear regression and logistic regression specifications also take into account this nesting by clustering standard errors by respondent. Third, because both linear regression and logistic regression models impose specific functional forms for the relationships being estimated, we estimated a LOESS model to estimate these relationships in a flexible, data-driven way. As explained below, the results are robust to model specification.

Results

We first estimate model (1). As shown in column (1) of table 1, a response option that is listed on top is 2.8 percentage points more likely to be selected ($p < 0.001$, two-tailed). Given the large sample size, all of our reported effects are precisely estimated. For instance, the t-statistic associated with this effect is 5.87 and the 95 percent confidence interval of the primacy effect ranges from 0.019 to 0.037 (i.e., from a primacy effect of 1.9 percentage points to one of 3.7 percentage points).[8]

As shown in column (2) of table 1, the primacy effect is estimated to be stronger for longer questions.[9] The coefficient estimate for β_3 indicates that, for every one-word increase in the item length, the primacy effect increases by 0.0009 percentage points ($p < 0.001$, two-tailed, 95 percent confidence

[8] How large is this effect size? For calibration, we recoded the dependent variable for whether respondents selected the “liberal” (0) versus “conservative” (1) position on the case, and estimated a linear regression predicting this variable with the standard seven-point partisan identification item ranging from “strong Democrat” (0) to “strong Republican” (1). Moving across the partisanship scale increases the likelihood of selecting a liberal response by 36.2 percentage points. Hence, the overall response order effect is about 7 percent the size of partisanship. We also calibrated the effect size by asking: Would our inference about majority support on a case change based upon which response order respondents were presented with? We found that it would in 7 out of 40 cases (17.5 percent).

[9] Comparisons of goodness of fit across models can be found in Supplementary Material table A8.

Table 1. Item-level and respondent-level correlates of response order effects (OLS regression).

	Model 1	Model 2	Model 3	Model 4
Item on top	0.028 (0.005, $p < 0.001$)	−0.049 (0.024, $p = 0.044$)	0.087 (0.017, $p < 0.001$)	−0.173 (0.049, $p < 0.001$)
Words	—	−0.0004 (0.0001, $p = 0.005$)	—	−0.001 (0.0003, $p = 0.002$)
FK score	—	0.00002 (0.0002, $p = 0.928$)	—	—
Knowledge	—	—	0.043 (0.013, $p < 0.001$)	−0.059 (0.046, $p = 0.204$)
Education	—	—	0.002 (0.002, $p = 0.286$)	—
Item on top x Words	—	0.0009 (0.0002, $p < 0.001$)	—	0.003 (0.0005, $p < 0.001$)
Item on top x Knowledge	—	—	−0.098 (0.019, $p < 0.001$)	0.185 (0.069, $p = 0.008$)
Words x Knowledge	—	—	—	0.001 (0.0005, $p = 0.019$)
Item on top x FK Score	—	−0.0002 (0.0003, $p = 0.414$)	—	—
Item on top x Education	—		0.0002 (0.003, $p = 0.954$)	—
Item on top x Words x Knowledge	—	—	—	−0.003 (0.0007, $p < 0.001$)
Intercept	0.427 (0.005, $p < 0.001$)	0.465 (0.017, $p < 0.001$)	0.393 (0.012, $p < 0.001$)	0.501 (0.033, $p < 0.001$)
F-statistic (df)	53.85 (3, 6,290)	26.63 (7, 6,290)	26.21 (7, 6,290)	22.88 (9, 6,290)
p-value	<0.001	<0.001	<0.001	<0.001

Note: N = 84,045. All models include fixed effects for survey wave. Standard errors and *p*-values presented in parentheses. Standard errors clustered by respondent.

interval of [0.0005, 0.0013]). While this coefficient estimate is small, note that a one-word change in the length of a question is also quite small. Over the range of the independent variable (78 words), this coefficient estimate implies an increase in the primacy effect of 0.069, meaning that a question with 131 words (the longest question across all three survey waves) is predicted to have a response order effect that is 6.9 percentage points larger than a question with 53 words (the shortest question). This interactive relationship is illustrated in figure 1, which shows predicted response as a function of question length for respondents who received Form A (solid line) versus Form B (dashed line). The estimated response order effect as a function of question length (which is equal to the vertical gap between the two lines in figure 1) is strongly increasing with question length. At the low end of item length, the predicted primacy effects are insignificant and close to zero; at the high end, the effects are large and significant when the question word length variable takes its largest value in our data. Response order

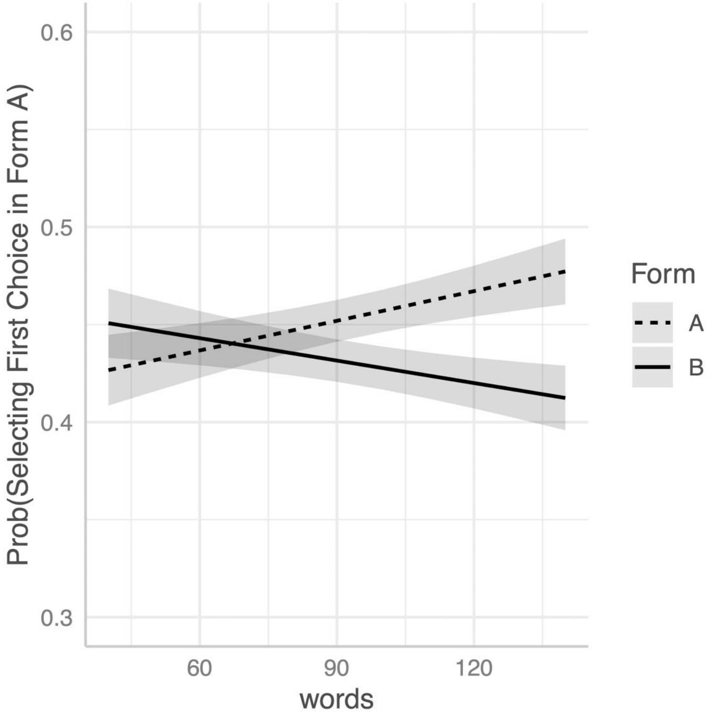

Figure 1. Response order effects increase with question length. Model predictions (with 95 percent confidence intervals) based on Model 2 in table 1, holding FK score at its sample mean.

effects exist for these sorts of questions overall, but importantly there are little or no such primacy effects for shorter questions and relatively large primacy effects for longer items.

In contrast, question complexity does not moderate response order effects. As shown in column (2) of table 1, the interaction term between the Flesch-Kincaid score and the response order manipulation is close to zero and not statistically significant. As shown in Supplementary Material table A4, this null effect is robust to an alternative operationalization of complexity beyond the Flesch-Kincaid score, a complexity index generated by taking the first principal component of several complexity measures (as described above).

As shown in column (3) of table 1, the primacy effect is estimated to be strongest among the least domain-knowledgeable respondents (i.e., those with low values of K_i). The coefficient estimate for β_3 indicates that moving across the range of the knowledge scale (i.e., from 0 to 1) is associated with a decrease in the primacy effect of 9.8 percentage points ($p < 0.001$, two-tailed, 95 percent confidence interval of [−0.135, −0.061]). As illustrated in figure 2, for those who answered all six items correctly, the response order effect is estimated to be close to zero and is not statistically significant ($p = 0.15$, 95 percent confidence interval of [−0.025, 0.004]). For those who answered none of the six items correctly, the primacy effect is extremely large, estimated at 8.8 percentage points ($p < 0.001$, 95 percent confidence interval of [0.061, 0.115]).

In contrast to domain knowledge, response order effects were not conditioned by respondent education. As shown in column (3) of table 1, the interaction term between education and the response order manipulation is close to zero and is statistically insignificant. Hence, when asking about complex issues, response quality is not compromised by the cognitive sophistication of respondents, but rather their interest and understanding of the specific subject matter.

Finally, given that we observed significant moderating relationships for question length and domain knowledge, we estimate a regression model that allows for question word length and respondent knowledge to have an interactive relationship with primacy effects in column (4). In this specification, the three-way interaction term (β_7) is negative, consistent with our expectations, and highly significant ($p < 0.001$). Three-way interaction terms can be complicated to interpret; hence, it is more straightforward to illustrate using predicted values from the model. As shown in figure 3, response order effects are most strongly related to question length for the least domain-knowledgeable respondents answering the longest questions. We see that for short questions (left pane) there is little or no predicted response order effect across the range of respondent knowledge values. For average-length questions (middle pane), we see significant primacy effects for less domain-

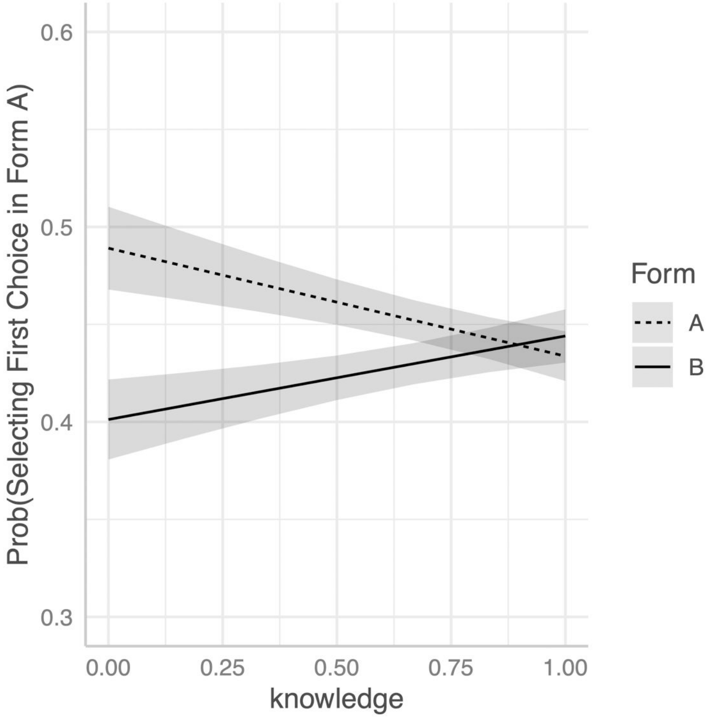

Figure 2. Response order effects decrease with respondent knowledge. Model predictions (with 95 percent confidence intervals) based on Model 3 in table 1, holding education at its sample mean.

knowledgeable respondents, with this effect declining as knowledge levels increase, and no primacy effect predicted for the most knowledgeable respondents. Finally, the right pane of figure 3 shows that large response order effects are predicted on longer items for less domain-knowledgeable respondents, but the most knowledgeable respondents are predicted to have small or zero primacy effects. For respondents who answered all of the Court knowledge items incorrectly, the predicted primacy effect for shorter questions is roughly zero, but it rises to 17.4 percentage points for the longest items.[10] Among those who answered all knowledge questions correctly, there is little difference in the predicted response between those receiving

[10] This prediction is not simply driven by functional form assumptions, such as linearity, in the regression model. The raw difference in response means on the longest question between respondents given the two forms among those who answered none of the knowledge questions is 15 percentage points (and nearly 18 percentage points when poststratification weights are used in this calculation).

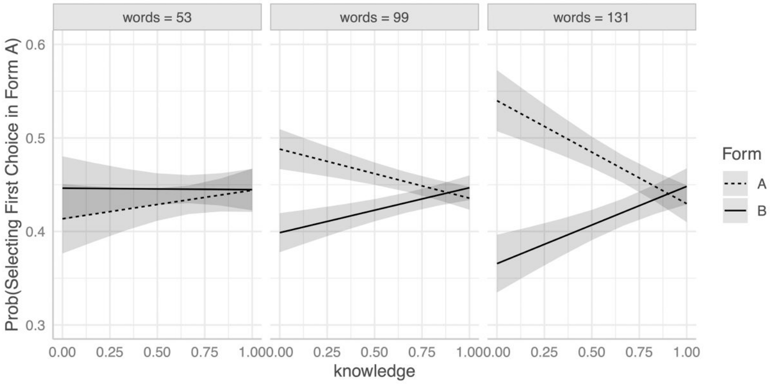

Figure 3. The negative relationship between respondent knowledge and response order effects is moderated by question length. Model predictions (with 95 percent confidence intervals) based on Model 4 in table 1. Left, middle, and right panes' predictions are based on question length at 53, 99, and 131 words, respectively.

Form A versus Form B across the range of question lengths (see rightmost predictions in each pane of figure 3). Respondents with intermediate levels of domain knowledge show a clear relationship between question length and predicted primacy effect, but this relationship is not as large as it is for less knowledgeable respondents.[11]

Supplementary Material section 4 contains numerous robustness checks that confirm these main results: (1) question length (and not complexity) moderates response order effects; (2) domain knowledge (and not education) moderates response order effects; and (3) question length and domain knowledge have an interactive relationship. The results of these robustness checks include: (1) similar results including respondent-level and item-level controls, along with their interactions with T_i (Supplementary Material tables A5 and A6); (2) estimating a logistic regression instead of linear regression (see table 2 of the main text and Supplementary Material figures A5–A7); (3) estimating a hierarchical logistic regression model instead of linear regression (Supplementary Material table A7); (4) using flexible LOESS regressions instead of linear models (Supplementary Material figures A8–A10); (5) operationalizing question length in various ways (Supplementary

[11] We again use partisan identification to calibrate this effect size. The 17.4 percentage point effect is over 48 percent the size of the effect of partisanship, indicating that it is substantively large. We then examined the response order effects among the subgroup of below-median knowledge respondents answering questions with above-median lengths. Within this subgroup, in 55 percent of cases (11 out of the 20 longest questions), our inference of majority support would change based on which response order respondents were shown.

Table 2. Item-level and respondent-level correlates of response order effects (logistic regression).

	Model 1	Model 2	Model 3	Model 4
Item on top	0.113 (0.019, $p < 0.001$)	−0.196 (0.098, $p = 0.044$)	0.351 (0.067, $p < 0.001$)	−0.698 (0.200, $p < 0.001$)
Words	—	−0.0015 (0.0005, $p = 0.005$)	—	−0.004 (0.001, $p = 0.002$)
FK score	—	−0.00007 (0.0009, $p = 0.931$)	—	—
Knowledge	—	—	0.173 (0.054, $p = 0.001$)	−0.235 (0.187, $p = 0.208$)
Education	—	—	−0.010 (0.009, $p = 0.286$)	—
Item on top x Words	—	0.004 (0.0008, $p < 0.001$)	—	0.011 (0.002, $p < 0.001$)
Item on top x Knowledge	—	—	−0.395 (0.076, $p < 0.001$)	0.745 (0.281, $p = 0.008$)
Words x Knowledge	—	—	—	0.004 (0.002, $p = 0.019$)
Item on top x FK Score	—	−0.001 (0.001, $p = 0.418$)	—	—
Item on top x Education	—		0.0007 (0.013, $p = 0.956$)	—
Item on top x Words x Knowledge	—	—	—	−0.012 (0.003, $p < 0.001$)
Intercept	−0.294 (0.022, $p < 0.001$)	−0.140 (0.069, $p = 0.042$)	−0.432 (0.051, $p < 0.001$)	−0.037 (0.133, $p = 0.978$)
Null deviance (df)	116,427 (84,044)	116,427 (84,044)	116,427 (84,044)	116,427 (84,044)
Residual deviance (df)	116,081 (84,041)	116,046 (84,037)	116,006 (84,037)	115,950 (84,035)
p-value	<0.001	<0.001	<0.001	<0.001

Note: N = 84,045. All models include fixed effects for survey wave. Standard errors and p-values presented in parentheses. Standard errors clustered by respondent.

Material figures A1–A4); and (6) operationalizing question complexity with an index based on multiple measures (Supplementary Material table A4).

Recommendations for Researchers and Concluding Remarks

In this study, we have examined the determinants of one form of survey satisficing—specifically by way of response order effects—when asking about complex policy questions. We further demonstrate that item and response characteristics can interact in important ways, creating large response order effects that can strongly influence the substantive conclusions reached in public opinion research. These interactive effects can obscure important heterogeneity when just examining primacy effects in isolation. Although our overall estimated primacy effect of about 3 percentage points is in line with moderately sized effects found in the literature, the fact that primacy effects under some conditions are estimated to be more than 17 percentage points is a warning for applied researchers interested in asking about increasingly complicated political phenomena.

Future research can further investigate the mechanisms underlying these results. Our preferred interpretation of the reason domain-specific knowledge moderates response order effects—but education does not—is due to higher respondent motivation. However, an alternative explanation is that low-knowledge respondents exhibit nonattitudes (Converse 1970) and therefore select the first available option. This seems unlikely given that question length, rather than question complexity, interacts with domain-specific knowledge, but our data cannot rule it out. Nonetheless, regardless of the specific cognitive mechanism, our findings make clear that satisficing is most likely when low-knowledge respondents are answering long questions.

That said, researchers should not refrain from asking about complex policies or political scenarios. Reporting on public opinion around these issues is important. However, researchers should consider question wording that is as short as possible while retaining key information. Further, question length should not be increased to achieve simple language. For example, in our own study a large response order effect (of 8.7 percentage points overall, 16 percentage points among respondents with below-median domain knowledge) was estimated for the question regarding the "Remain in Mexico" border policy at issue in *Biden v. Texas* (2022). The question prompt (54 words out of a total length of 113 words) read:

> The U.S. Department of Homeland Security required noncitizens trying to reside in the U.S. to wait in Mexico while immigration officials process their cases. The Biden Administration issued an order ending this "remain in Mexico"

> program. In response, several states sued, saying that the Administration did not have adequate justification in ending the program.

This question could have been substantially shortened. Consider the following version (33 words), which removes all but only the key facts:

> The Biden Administration ordered the end of the “remain in Mexico” immigration program in which noncitizens trying to enter the U.S. had to wait in Mexico while their immigration cases were being processed.

Another example that had a large treatment effect (7.9 percentage points overall, and 14.7 percentage points among respondents with low knowledge) was a case called *Brnovich v. Democratic National Committee* (2021) regarding alleged voter suppression. The original question prompt was very long (and at 80 words put the question at the 95th percentile in terms of length) and read as follows:

> In Arizona, if a voter arrives at a polling place and is not listed on the voter roll for that precinct, the voter may still cast a provisional ballot. After election day, Arizona election officials review all provisional ballots to determine the voter’s identity and address. If officials determine that the voter voted outside of their precinct, the ballot is discarded in its entirety, even if the voter was eligible to vote in most of the races on the ballot.

Consider the alternative shortened version (49 words), which again removes any unnecessary words:

> Arizona law allows people not listed on a voter roll to cast a provisional ballot. However, if officials determine that the person voted outside of their own assigned precinct, they can throw away the entire ballot, even if the person was eligible to vote on some of the races.

As these examples show, there is a trade-off between providing more details and question length. Note that we did not compromise on linguistic complexity in these examples, as question length seems to matter much more. Hence, researchers asking about complex issues should focus on being brief, rather than using extra words to make the language simpler. This result concords with our finding that respondent education does not reduce response order effects. In contrast, domain knowledge, which is likely associated with familiarity and interest in the topic, is a key moderator. Recent research (e.g., Berinsky, Margolis, and Sances 2014) has found that inattentive survey respondents severely compromise data quality, particularly with online administration. Thus, keeping questions short is also important to prevent satisficing among less interested respondents.

In sum, asking about complex issues is undoubtedly important, but it should be handled carefully. This is particularly critical for those respondents not that interested and knowledgeable about a specific topic, which not only could comprise a majority of respondents but also may be difficult to address using survey weights. Indeed, given that a majority of people may be unfamiliar with complicated political and policy areas—including not just the legal issues by the Supreme Court, which was our example here, but also policy-heavy issue areas such as climate science, civil rights, and economic policy—the concerns we raise here may be particularly salient for political science research. Given this reality, our recommendation is not that these topics should be avoided, but that all else equal, shorter question wording is preferable.

Supplementary Material

Supplementary Material may be found in the online version of this article: https://doi.org/10.1093/poq/nfae050.

Acknowledgements

The authors thank Adam Liptak, Alicia Parlapiano, Doug Rivers, and Joe Williams for their assistance in conducting this research. They thank Anton Stezhnev and panel participants at the 2023 Annual Meeting of the American Political Science Association for valuable feedback; and they thank Nabikshya Rayamajhi for research assistance.

Data Availability

Replication data and code are available at the Harvard Dataverse: https://doi.org/10.7910/DVN/TBLLZ1.

References

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Ansolabehere, Stephen, and Shiro Kuriwaki. 2022. "Congressional Representation: Accountability from the Constituent's Perspective." *American Journal of Political Science* 66:123–39.

Ansolabehere, Stephen, and Brian F. Schaffner. 2014. "Does Survey Mode Still Matter? Findings from a 2010 Multi-Mode Comparison." *Political Analysis* 22:285–303.

Alvarez, R. Michael, and John Brehm. 2002. *Hard Choices, Easy Answers: Values, Information, and American Public Opinion*. Princeton, NJ: Princeton University Press.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58:739–53.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and its Discontents*, edited by David E. Apter, 206–61. New York: The Free Press of Glencoe.

———. 1970. "Attitudes and Non-Attitudes: Continuation of a Dialogue." In *The Quantitative Analysis of Social Problems*, edited by Edward R. Tufte, 168–89. New York: Free Press.

Cor, M. Ken, Edward Haertel, Jon A. Krosnick, and Neil Malhotra. 2012. "Improving Ability Measurement in Surveys by Following the Principles of IRT: The Wordsum Vocabulary Test in the General Social Survey." *Social Science Research* 41:1003–16.

C-Span. 2009. "C-SPAN Supreme Court Survey." https://sites.c-span.org/ camerasInTheCourt/pdf/C-SPAN%20Supreme%20Court%20Online%20Survey_070909_6pm.pdf. Date accessed December 22, 2023.

Epstein, Lee, and Jeffrey A. Segal. 2000. "Measuring Issue Salience." *American Journal of Political Science* 44:66–83.

Flesch, Rudolph. 1948. "A New Readability Yardstick." *Journal of Applied Psychology* 32:221–33.

Holbrook, Allyson L., Jon A. Krosnick, David Moore, and Roger Tourangeau. 2007. "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes." *Public Opinion Quarterly* 71:325–48.

Jessee, Stephen, Neil Malhotra, and Maya Sen. 2022. "A Decade-Long Longitudinal Survey Shows That the Supreme Court is Now Much More Conservative than the Public." *Proceedings of the National Academy of Sciences* 119:e2120284119.

Kimball, David C., and Martha Kropf. 2005. "Ballot Design and Unrecorded Votes on Paper-based Ballots." *Public Opinion Quarterly* 69:508–29.

Kincaid, J. Peter, Robert P. Fishburne, Jr., Richard L. Rogers, and Brad S. Chissom. 1975. "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel." Naval Air Station Memphis: Chief of Naval Technical Training. Research Branch Report 8–75.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5:213–36.

———. 1999. "Survey Research." *Annual Review of Psychology* 50:537–67.

Krosnick, Jon A., and Duane F. Alwin. 1987. "An Evaluation of a Cognitive Theory of Response Order Effects in Survey Measurement." *Public Opinion Quarterly* 51:201–19.

Lenzner, Timo. 2014. "Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?" *Sociological Methods & Research* 43:677–98.

Malhotra, Neil. 2009. "Order Effects in Complex and Simple Tasks." *Public Opinion Quarterly* 73:180–98.

Prokop, Andrew. 2021. "Is Biden's Legislative Agenda Popular? Yes, But ... " *Vox*. October 8. https://www.vox.com/22711083/biden-reconciliation-build-back-better-polls- infrastructure. Date accessed December 17, 2022.

Rivers, Douglas, and Delia Bailey. 2009. "Inference from Matched Samples in the 2008 US National Elections." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, Joint Statistical Meeting 2009, 627–39. Palo Alto, CA: YouGov/Polimetrix.

Schuman, Howard, and Stanley Presser. 1996. *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Sherif, Muzafer and Carl I. Hovland. 1961. *Social Judgement: Assimilation and Contrast Effects in Communication and Attitude Change*. New Haven, CT: Yale University Press.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco, CA: Jossey-Bass.

Velez, Pauline, and Steven D. Ashworth. 2007. "The Impact of Item Readability on the Endorsement of the Midpoint Response in Surveys." *Survey Research Methods* 1:69–74.

Public Opinion Quarterly, 2025, 00, 1–20
https://doi.org/10.1093/poq/nfae050
Research note