Climatic Change (2021) 167: 35
https://doi.org/10.1007/s10584-021-03194-x

# Public opinion on climate change in the USA: to what extent can it be nudged by questionnaire design features?

Catherine Chen[1] · Bo MacInnis[1] · Matthew Waltman[2] · Jon A. Krosnick[3]

Received: 12 April 2021 / Accepted: 28 July 2021/ Published online: 12 August 2021


**Abstract**
Policymakers may wish to take into account public opinion on climate change as they craft legislation, but if public opinion changes substantially in response to seemingly trivial changes in survey questionnaire design, perhaps such reliance would be unwise. This paper examines 110 experiments implemented in surveys of truly random samples of American adults between 2012 and 2018 ($N$ = 4414), exploring the extent to which answers to questions were influenced by order and wording manipulations. Of 144 tests, 31 (22%) yielded statistically significant effects. Adjustments for multiple hypothesis tests reduced this percentage to between 7 and 9%. The effect sizes are routinely small. These results are consistent with the conclusion that survey results on climate change issues are relatively robust, so policymakers can take them seriously if they wish to do so.

**Keywords** Public opinion · Survey methodology · Order effect · Wording effect

## 1 Introduction

The responsiveness of government to citizens' policy preferences is a central tenet of some theories of democracy (Dahl 1989; Erikson and Tedin 2015; Page and Shapiro 2010). Therefore, when observers have lamented the failure of the US Congress to pass major legislation to reduce greenhouse gas emissions in recent years (e.g., American Clean Energy and Security Act 2009; Climate Stewardship and Innovation Act 2007), some have faulted the American public, assuming that insufficient public support must have assured legislative

✉ Catherine Chen
tche101@stanford.edu

[1] Department of Communication, Stanford University, Stanford, CA 94305, USA

[2] Department of Political Science, Stanford University, Stanford, USA

[3] Departments of Communication, Political Science, and Psychology, Stanford University, Stanford, USA

defeat. These observers inferred that public education is needed in order to strengthen people's opinions and inspire them to take action to influence government (Gardner and Stern 2008; Napolitano and Johnson 2018; Sterman 2008).

Seemingly inconsistent with this logic is survey evidence that public beliefs about the existence and threat of climate change and support for government action in this arena have been consistently at high levels for decades (Abeles et al. 2019; Krosnick and MacInnis 2020; Political Psychology Research Group 2020). But in fact, there is heterogeneity among the findings of well-publicized surveys on the topic—some apparently documenting smaller proportions of the public holding those views (National Surveys on Energy and Environment 2014). So perhaps one cannot blame American legislators for not taking action in the face of such mixed and conflicting evidence.

Why do different surveys disagree with one another on this topic? One possible explanation involves the mode of data collection or the sampling procedures used. Some surveys have involved human interviewers asking questions over the telephone (Lavrakas 2008; Lee et al. 2015), while others have been done via self-administered questionnaires on the Internet (Tvinnereim and Fløttum 2015). And whereas some surveys have involved random sampling of the American public (e.g., the American National Election Studies (ANES)), others have involved collecting data from people who responded to ads inviting them to complete questionnaires for money (e.g., the Ipsos I-Say Panel). Much published evidence has shown how a mode or sampling shift can alter respondents' answers to questions (Chang and Krosnick 2009; Holbrook et al. 2003).

A second possible explanation for variation in results across surveys is differences in question design. For example, Kyselá et al. (2019) pointed out that surveys have differed in whether they asked about "acceptability", "acceptance", "supporting", "favoring", or other views of proposed greenhouse gas emission mitigation policies, and these seemingly trivially different measurement approaches appear to have yielded different results.

Some scholars are inclined to view such evidence as showing that Americans' opinions are difficult to characterize because they are weak. For example, Converse (1974, 656) stated that "it is probable that extraneous factors like question form intrude most sharply on responses where attitudes are least crystallized." According to Cantril (1944, 49), when "the respondent's mental context is solidly structured … the same answer is likely to be obtained irrespective of the way questions are asked." Similarly, Payne (1951, 179) argued that:

> "Where people have strong convictions, the wording of the question should not greatly change the stand they take. The question can be loaded heavily on one side or heavily on the other side, but if people feel strongly their replies should come out about the same. It is on issues where opinion is not crystallized that answers can be swayed from one side to the other by changes in the statement of the issue."

Previous research has yielded a theory-based understanding of when and why various changes in question design might alter responses and has documented many instances of such effects. For example, changing the order in which answer choices are offered to respondents has been shown to alter the distributions of the responses due to survey satisficing (Holbrook et al. 2007; Krosnick 1991). Changing the order in which questions are asked has also been shown to alter responses to those questions, attributable to various psychological processes (Schwarz and Bles 1992). And seemingly trivial changes in

response option wording and question stem wording have also been shown to change distributions of responses (Schuman and Presser 1981). Such studies can be viewed as consistent with Converse's (1964) speculation that respondents report "non-attitudes" rather than real attitudes when answering survey questions on a wide array of issues, so the same may be true of opinions on climate change.

But caution might be wise before reaching this conclusion because there is an alternative explanation for the existing literature: the "file drawer" (Rothstein et al. 2006). If investigators and editors are more inclined to write up and publish the results of experiments that demonstrate statistically significant effects than those that document null effects, the literature may create the false impression that question design effects are ubiquitous when in fact they are not, constituting what is called "publication bias" (Fanelli 2010).

So it is of interest to explore the degree to which measurements of public opinion on climate change are altered by changes in question design. Substantial susceptibility and consistently large effects might be viewed as a rationale for legislators to legitimately ignore public opinion on the issue because public opinion is easily manipulable. Kyselá et al. (2019) and Motta et al. (2019) reported the results of experiments exploring question design effects in surveys of opinions on climate change, and we do the same here.

We conducted 110 experiments in five surveys of truly random national and state samples (sampled via Random Digit Dialing (RDD)), involving human interviewers calling landlines and cellphones between 2012 and 2018 (total $N = 4414$). We tested the effects of order changes (involving the order of response options and the order of questions) and the effects of wording changes (involving the wording of answer choices and question stems) on public opinion on climate change.

These 110 experiments were not preregistered because they were not conducted in order to test specific a priori hypotheses about ways in which manipulations of questions would alter answers. Rather, the surveys were intended to measure public opinion, and these manipulations were incorporated in order to avoid introducing systematic bias, in case such bias might be caused by features of the question design.

## 2 Mechanisms of question design impact

### 2.1 Order effects

#### 2.1.1 Response option order effects

The impact of the order in which closed-ended questions offer answer options to respondents has been demonstrated in numerous publications (Krosnick 1991; Krosnick and Alwin 1987). Some studies have documented primacy effects, in which options presented earlier are more likely to be selected (e.g., MacInnis et al. 2021; Malhotra 2008; Pasek et al. 2014), and other studies found recency effects, in which options presented later are more likely to be selected (e.g., Bishop and Smith 2001; Holbrook et al. 2007). This sort of impact of order is thought to be well-explained by the theory of survey satisficing (Krosnick 1999). And such order effects are thought to occur because respondents do not invest the thoughtful effort needed to generate accurate reports of their opinions. We explored the presence of response option order effects in questions about global warming on the assumption that accurate measurements of opinions should be immune to such changes.

### 2.1.2 Question order effects

Past research has shown that variation of question order sometimes alters responses by a variety of cognitive mechanisms, including assimilation and contrast effects (Schwarz and Bles 1992), priming (Kalton et al. 1978), subtraction (Schuman and Presser 1981), and others (for a review, see Krosnick and Presser (2010)). We tested the impact of various manipulations of questions order, again on the assumption that accurate measurements of opinions will be immune to such changes, since the questions themselves do not change.

## 2.2 Wording effects

Whenever researchers design a question, they must choose among various synonyms to express each idea. Sometimes, synonyms are interchangeable in the minds of respondents, but on other occasions, two words that might seem to mean the same thing to a researcher can be construed quite differently by respondents. Therefore, whether a wording change will cause changes in responses may depend on the particular question involved and the particular question wording alteration examined (Schuman and Presser 1996).

### 2.2.1 Seemingly trivial wording changes

In this study, we investigated eight types of question wording changes. Some involved changes that seem likely to be trivial. For example, according to natural science research, the average global temperature during the last 2000 years was relatively stable until the last century, when it began a steady increase that persists today (Marcott et al. 2013). One survey question that has been used to measure public perceptions of this change has asked: "What is your personal opinion? Do you think that the world's temperature probably has been going up slowly over the past 100 years, or do you think this probably has not been happening?" One might wonder whether the inclusion of the word "slowly" is not faithful to the temperature data and whether omitting that word might yield more affirmative answers from people who believe that the increase has been anything but slow. We tested that possibility.

Another set of experiments explored the impact of a different seemingly trivial wording change. When asking respondents whether they favored or opposed various mitigation policies, those policies can be described as "taking action on global warming" or "taking action to reduce global warming." Again, replacing "on" with "to reduce" seems like a trivial change, but perhaps the latter phrase conveys more aggressive action that might appeal differently to respondents. We also examined whether replacing those phrases with the phrase "to prepare for the effects of global warming" altered public support for mitigation efforts.

Another wording change investigated involved replacing the phrase "is causing" with "has caused" when asking about the impact of global warming. This might appear to be a trivial change if respondents perceive causal processes that occurred in the past to be the same as those playing out currently. But if people differentiate past causation from current causation, this wording change might alter respondents' answers to survey questions.

Another seemingly trivial wording change involved the phrasing of response options constituting a rating scale. For some respondents, the offered options were "a great deal, a lot, a moderate amount, a little, or nothing." For other respondents, "a lot" was replaced by "quite a bit," and "a moderate amount" was replaced by "some." If these changes are non-substantive, we would expect them not to alter answers. But if in respondents' minds, the meanings of "a lot" and

"quite a bit" are different from one another, or if the meanings of "a moderate amount" and "some" are different from each other, then the observed distributions of responses might change.

Lastly, we investigated one additional seemingly trivial change in question wording. Prior to asking respondents for their opinions about a series of different types of emission-reduction policies, some past surveys have included a preamble that was worded in two different ways. Some respondents heard: "Each of these changes would increase the amount of money that you pay for things you buy," and for other respondents, "would" was replaced with "could." We explored whether the softening caused by replacing the word "would" with "could" led to more approval of the proposed policies because their economic cost was portrayed as less certain.

#### 2.2.2 Nontrivial wording changes

Other experiments described here explored the impact of wording changes that seem nontrivial and more likely to alter opinions. For example, after respondents were told about how a cap and trade system would work, some respondents were asked, "Would you favor or oppose a cap and trade system to reduce the amount of greenhouse gases that companies put out?", whereas other respondents were told, "Economists say that this system is likely to cause companies to figure out the cheapest way to reduce greenhouse gas emissions" and were then asked, "Would you favor or oppose this cap and trade system?" If noting the belief of economists increases people's support for the policy, then we would expect the latter wording to yield more apparent support than the former.

Other question wording experiments involved more substantive changes. Respondents were asked about their support for a series of emission reduction policies that would involve businesses paying higher taxes. Before being asked those questions, some respondents were told, "All this tax money would be given to all Americans equally by reducing the amount of income taxes they pay." Assuming that refunding tax payments to people is more appealing than paying taxes with no such refunding, we might expect this addition to the wording to yield more support for the policies.

Another set of question wording experiments investigated whether adding "Each of these changes would increase the amount of money that you pay for things you buy" would alter the response distribution. Support for those policies might be lower when respondents are given the additional information, as it explicitly reminds respondents of the prices they need to pay.

### 2.3 Moderation by education and party identification

#### 2.3.1 Education

In past studies of question design effects, a respondent's years of formal education has been treated as an indication of his or her cognitive skills. And according to the theory of survey satisficing (Krosnick 1991), some question design effects are expected to have more impact among individuals with more limited cognitive skills. As expected, Narayan and Krosnick (1996) found that seven response effects did indeed appear stronger among less educated respondents: response order effects, acquiescence, middle alternative effects not involving status quo options, no-opinion filter effects, forbid/allow effects, balance effects, and question order effects based on the norm of reciprocity. The only one of these effects examined in the present paper is response order effects, so we expected to see these more strongly among less educated respondents. But

none of the other question design effects examined here have been linked to satisficing theory, so we had no reason to expect that education would moderate their magnitudes.

#### 2.3.2 Party identification

Partisan gaps exist in American public opinion on climate change. On many aspects of this issue, the endorsement of "green" views is more common among Democrats than among Republicans (Krosnick and MacInnis 2020). Therefore, one might be curious as to whether Democrats and Republicans react to question design manipulations differently. Motta et al. (2019) tested moderation of three kinds of question design manipulations by party identification and found such moderation in only a few instances, revealing no consistent patterns. We explored this moderation as well.

## 3 Method

### 3.1 Data collection

The experiments described here were included in five computer-assisted telephone interviewing (CATI) surveys conducted in 2012, 2013, 2014, 2015, and 2018. Respondents were sampled via Random Digit Dialing (RDD) to landlines and cell phones. Calls were staggered over times of days and days of the week to maximize the chances of making contact with potential respondents.

In total, 4414 respondents were interviewed (804 in 2012, 801 in 2013, 803 in 2014, 1006 in 2015, and 1000 in 2018). The 2012 and 2013 surveys were conducted in English only, and the 2014, 2015, and 2018 surveys were conducted in English and Spanish. All surveys were conducted with representative samples of US adults except for the 2014 survey, which was conducted with a representative sample of adults living in Arizona. The protocol was approved by the Institutional Review Board of the institution with which the second author was affiliated at the time of the data collection.

The data for the surveys were weighted to account for unequal probabilities of selection and to post-stratify in terms of geographic locations, demographics, and type of telephone service. Results reported in this paper were computed using weighted data.

The AAPOR response rate 3 for the five surveys (2012, 2013, 2014, 2015, and 2018) were 15%, 13%, 10%, 12%, and 17%, respectively.

Additional information on the methodology of the surveys is available in the Supplementary Material. Data and code needed to replicate the analyses are available at the following: https://osf.io/9k3cp.

### 3.2 Experimental design

#### 3.2.1 Order effects

**Response option order** Sixteen experiments investigated to what extent the order of response options impacts the distributions of responses. Three experiments were conducted more than once, affording 27 tests in total.

Among the 16 experiments, 10 offered three response options: one positive, one negative, and one neutral (e.g., more stable, more unstable, or about the same). The order of the two non-neutral response options was varied across respondents randomly.

Six other experiments involved questions with two response options, and the order of the two was randomized.

**Question order** The current study explored question order effects with 58 experiments, where respondents were randomly assigned to one of various orders of questions in a set.

### 3.2.2 Wording effects

#### Seemingly trivial wording changes

- Inserting "slowly"

One experiment was conducted to explore whether the inclusion of the word "slowly" significantly altered public opinion about changes in the world's temperature.

- Purposes of action

One experiment explored the impact of changing "take action on global warming" to "take action to reduce global warming," and six more experiments investigated differences among "actions about future global warming," "actions to reduce future global warming," and "actions to prepare for the effects of global warming."

- Verb tense

Two experiments examined the difference between "is causing" and "has caused" when discussing the impact of global warming.

- Rating scale response option wording

Six experiments (each one conducted twice) explored the impact of slight changes in the verbal labels attached to two of five response options on a rating scale inquiring about how much average people, governments, and businesses are doing and should do to deal with global warming.

- "Would" and "Could"

Five experiments assessed the impact of changing "Each of these changes would increase the amount of money that you pay for things you buy" to "Each of these changes could increase the amount of money that you pay for things you buy."

#### Nontrivial wording changes

- Cap and trade

One experiment explored the impact of the change in the wording of the cap and trade questions described above.

- Adding a cost-clarifying statement

Three experiments investigated whether adding "All this tax money would be returned to all Americans equally by reducing the amount of income taxes they pay" altered judgments of policies that would raise taxes. Two experiments investigated the impact of adding "All this tax money would be given to all Americans equally by reducing the amount of income taxes they pay." And nine experiments investigated the impact of adding "Each of these changes would increase the amount of money that you pay for things you buy."

### 3.3 Moderators

#### 3.3.1 Education

Respondents were asked, "What was the last grade of school you completed?" We divided the samples into people with no college education and people with at least some college education.

#### 3.3.2 Party identification

A random half of respondents were assigned to the question "Generally speaking, do you usually think of yourself as: a Democrat, a Republican, an Independent, or what?", and the remaining respondents were assigned to the question "Generally speaking, do you usually think of yourself as: a Republican, a Democrat, an Independent, or what?"

### 3.4 Analysis method

The statistical significance of the impact of the question design manipulations on the distributions of responses was assessed by design-based Wald tests, yielding an $F$-statistic for each experiment. All statistical tests were two-sided. We considered $p < .05$ to be statistically significant.

#### 3.4.1 Subgroup analyses

To test whether the question design effects were moderated by education and party identification, we created three-way contingency tables using the "weights" package in R (Pasek et al. 2020). When this analysis indicated that the subgroups (e.g., Democrat vs. Republicans and people with vs. without college education) responded differently to the manipulation, we conducted design-based Wald tests to determine what differences were statistically significant.

#### 3.4.2 Stacking

For experiments that were run multiple times, the data were stacked to increase statistical power. Responses to the same experiment and weights for each respondent were aggregated to create a larger dataset, and analyses were performed using the "svychisq()" function in the "survey" package (Lumley 2020) in the R Statistical Software v.4.0.2 (https://www.R-project.org/).

# 4 Results

## 4.1 Order variations

### 4.1.1 Response option order

As expected, all six of the two-response-option experiments yielded differences in the direction of recency effects, and two of the six (33%) were statistically significant (see Table 1 for the number of comparisons, the number of results in a specific direction, average effect sizes, and the number of statistically significant test results; the question wordings, response distributions, and statistics for all individual tests are available in the Supplementary Material). In the two experiments yielding significant effects, 13 and 15 percentage points more respondents chose an option when it was presented more recently, respectively.

Surprisingly, all 10 of the three-response-option experiments yielded differences in the direction of a primacy effect, and after aggregating data for experiments conducted more than one time, one out of the 10 (10%) showed a statistically significant primacy effect: 8 percentage points more respondents chose an option when it was presented first than when it was presented second.

### 4.1.2 Question order

Of 58 experiments that varied question order (aggregating data for experiments conducted more than once), 14 (24%) yielded statistically significant order effects (see Table 2 for the number of questions in a randomization set, sample sizes, and effect sizes (Cramer's Vs)). Among those experiments, the average change in the proportion of people choosing a response was 6 percentage points.

## 4.2 Wording variations

### 4.2.1 Seemingly trivial wording changes

- Inserting "slowly"

The distribution of responses did not differ significantly as the result of inserting "slowly" to modify the speed of global warming (see Table 3 for sample sizes and effect sizes (Cramer's Vs) for all seemingly trivial wording change experiments).

**Table. 1** Response order variation studies

| Number of response options | Number of experiments | Number of experiments significant after stacking data for experiments conducted multiple times | Number of tests | Number of tests significant | Number of tests in the direction of primacy effect | Average effect sizes in Cramer's V |
|---|---|---|---|---|---|---|
| 2 | 6 | 2 | 6 | 2 | 0 | 0.085 |
| 3 | 10 | 1 | 21 | 4 | 18 | 0.072 |
| Sum | 16 | 3 | 27 | 6 | 18 | 0.075 |

**Table. 2** Question order variation studies

| Topic | Number of questions in the randomization set | Number of experiments significant after stacking data for experiments conducted multiple times | Number of tests significant in the randomization set | *N* | Average effect sizes in Cramer's V |
|---|---|---|---|---|---|
| Energy taxes: 4 questions | 4 | 0 | 0 | 804 | 0.057 |
| Energy taxes: 5 questions | 5 | 0 | 0 | 2006 (conducted twice in total) | 0.047 |
| Energy taxes: 2 questions | 2 | 1 | 1 | 1604 (conducted twice in total) | 0.044 |
| Efficiency standards: 4 questions | 4 | 3 | 5 | 2610 (conducted three times in total) | 0.084 |
| Efficiency standards: 5 questions | 5 | 2 | 2 | 804 | 0.111 |
| Efficiency standards: 3 questions | 3 | 3 | 3 | 1000 | 0.097 |
| Candidate global warming action | 2 | 0 | 0 | 804 | 0.108 |
| Tax reductions for energy companies | 3 | 0 | 0 | 801 | 0.036 |
| Energy production | 6 | 0 | 0 | 415 | 0.130 |
| Effects of global warming | 3 | 0 | 0 | 662 | 0.065 |
| Arizona efficiency standards | 3 | 0 | 0 | 803 | 0.041 |
| Trust in politicians | 2 | 2 | 2 | 1809 (conducted twice in total) | 0.093 |
| Arizona trust | 2 | 0 | 0 | 803 | 0.098 |
| Government action | 2 | 1 | 1 | 1006 | 0.096 |
| Global warming importance | 4 | 1 | 1 | 1006 | 0.098 |
| Actions—should | 4 | 0 | 0 | 1000 | 0.081 |
| Current actions | 4 | 1 | 1 | 1000 | 0.087 |
| Sum | 58 experiments affording 75 tests | 14 | 16 | | 0.087 |

**Table. 3** Seemingly trivial wording changes studies

| Topic | Number of experiments | Number of experiments significant after stacking data for experiments conducted multiple times | Number of tests | Number of tests significant | *N* | Average effect sizes in Cramer's V |
|---|---|---|---|---|---|---|
| Inserting "slowly" | 1 | 0 | 1 | 0 | 804 | 0.084 |
| Actions | 7 | 0 | 7 | 0 | 803/587 | 0.081 |
| Verb tense | 2 | 0 | 2 | 0 | 375/426 | 0.086 |
| Rating scale: "are doing" | 3 | 3 | 6 | 5 | 1436 | 0.147 |
| Rating scale: "should do" | 3 | 0 | 6 | 1 | 1436 | 0.062 |
| "Would" and "Could" | 5 | 0 | 5 | 0 | 1006 | 0.038 |
| Sum | 21 | 3 | 27 | 6 | | 0.102 |

- Purpose of action

The distribution of responses did not differ statistically significantly when changing "take action on global warming" to "take action to reduce global warming." And none of the six experiments examining the impact of changing "actions about future global warming" to "actions to reduce future global warming" or "actions to prepare for the effects of global warming" documented significant effects.

- Verb tense

While gauging respondents' opinions on the impact of global warming on the number of droughts and storms, changing "is causing" to "has caused" did not lead to significant differences in response distributions.

- Rating scale response option wording

Aggregating data for experiments conducted more than once, three of six experiments documented statistically significant effects of rating scale response option wording variations. When asked how much average people, government, and businesses are doing to deal with global warming, respondents were more likely to choose "a little" if the response options were "a great deal, a lot, a moderate amount, a little, or nothing" than if the response options were "a great deal, quite a bit, some, a little, or nothing." These changes in proportion were 14 percentage points, 9 percentage points, and 14 percentage points, respectively. The distributions of beliefs about how much average people, government, and business should do were not affected by the response option wording variation.

- "Would" and "Could"

None of the five experiments changing "would increase" to "could increase" produced a statistically significant change in the distribution of responses.

#### 4.2.2 Nontrivial wording changes

- Cap and trade

When respondents were asked, "Would you favor or oppose a cap-and-trade system to reduce the amount of greenhouse gases that companies put out?", 36% strongly favored the policy, 26% somewhat favored the policy, 16% somewhat opposed the policy, and 22% strongly opposed the policy (see Table 4 for effect sizes (Cramer's Vs) for the nontrivial wording change experiments). When respondents were told, "Economists say that this system is likely to cause companies to figure out the cheapest way to reduce greenhouse gas emissions" and were then asked, "Would you favor or oppose this cap and trade system?", these numbers were 13%, 36%, 20%, and 30%, respectively. Thus, the inclusion of "Economists say that this system is likely to cause companies to figure out the cheapest way to reduce greenhouse gas emissions" led to a statistically significant 13 percentage point decrease in the proportion of people who strongly or somewhat favored the policy.

- Adding a cost-clarifying statement

Telling respondents that tax revenue would be given/returned to all Americans increased support for green tax policies in all five experiments, among which two were statistically significant. The increases in the proportions of people favoring the policy were 9 percentage points, 7 percentage points, 4 percentage points, 1 percentage point, and 17 percentage points, respectively.

Adding "Each of these changes would increase the amount of money that you pay for things you buy" did not change the distribution of responses significantly in any of the nine experiments.

### 4.3 Moderators

#### 4.3.1 Moderation by education

Education statistically significantly moderated 4 of 27 tests investigating response option order effects. In three of the four, order effects only occurred among people with no college education. For the fourth experiment, the 3-way contingency table indicated a statistically significant difference between education groups, but responses of neither education group

**Table. 4** Nontrivial wording change experiments

| Topic | Number of experiments | Number significant | Average effect sizes in Cramer's V |
|---|---|---|---|
| Cap and trade | 1 | 1 | 0.263 |
| Telling respondents that tax revenue would be given/returned to all Americans | 5 | 2 | 0.078 |
| Adding "Each of these changes would increase the amount of money that you pay for things you buy" | 9 | 0 | 0.029 |
| Sum | 15 | 3 | 0.061 |

were significantly altered by the response option order manipulation. Therefore, we conclude that no moderation occurred in that fourth experiment. Breakdown of the experiment results by education can be found in Table S1 in the Supplementary Material.

Education statistically significantly moderated 14 of 75 tests investigating question order effects. In seven of the 14 tests, neither of the two subgroups was affected significantly by the question order manipulation. In another test, both subgroups were affected by the manipulation but in different ways. In one of 14 tests, a question order effect manifested only among people with at least some college education. In five of 14 tests, a question order effect manifested only among people with no college education.

Education statistically significantly moderated 7 of 27 tests investigating seemingly trivial wording manipulations. In two of seven tests, a wording effect manifested only among people with no college education. In the remaining five tests, the wording manipulation significantly altered the response distribution for both subgroups, or the manipulation did not significantly alter either of the two subgroups.

Education statistically significantly moderated 1 of 15 tests of nontrivial wording manipulations, where the wording effect manifested only among people with at least some college education.

#### 4.3.2 Moderation by party identification

When comparing Democrats to Republicans, 24 of 144 tests indicated statistically significant moderation of the question design effect by party identification. When comparing Democrats to Independents and Republicans, the same proportion of tests of moderation was statistically significant. In the instances of statistically significant moderation, no consistent and interpretable pattern appeared, such that one group of respondents was more influenced than another. Breakdown of the experiment results by party identification can be found in Table S2 in the Supplementary Material.

### 4.4 Adjusting for multiple comparisons

When conducting 144 tests of statistical significance and using a $p$ value of .05 for identifying statistically significant effects, some will be significant by chance alone. To avoid being misled by this, researchers during a period of decades have considered a number of different possible corrections for this multiple hypothesis testing (Streiner 2015), and there is considerable disagreement about which of these, if any, is most appropriate in any given situation. We therefore report results using four different methods for correction for multiple hypothesis tests for readers interested in such results, generated using R's "$p$.adjust()" function: the Bonferroni method (Miller 1981), the Holm (1979) method, the Hommel (1988) method, and the false discovery rate (FDR) adjustment (Benjamini and Hochberg 1995).

After these adjustments, the number of significant test results dropped from 6 to 4 (Bonferroni, Holm, and Hommel) and 5 (FDR) for response option order experiments, from 16 to 2 (Bonferroni, Holm, and Hommel) and 3 (FDR) for the question order experiments, from 6 to 2 (Bonferroni, Holm, and Hommel) and 3 (FDR) for the seemingly trivial wording changes, and from 3 to 2 (all four methods used) for nontrivial wording changes. Combining all these types of design effects together, 31 of 144 (22%) tests were significant without adjustment. The Bonferroni, Holm, and Hommel methods reduced the proportion of significant test results to 7%, and FDR reduced that proportion to 9% (Table 5).

**Table. 5** Impact of corrections for multiple significance tests

| | Number of tests conducted | Number significant | Number significant after Bonferroni correction | Number significant after Holm correction | Number significant after Hommel correction | Number significant after FDR correction |
|---|---|---|---|---|---|---|
| Response option order variation | 27 | 6 (22%) | 4 (15%) | 4 (15%) | 4 (15%) | 5 (19%) |
| Question order variation | 75 | 16 (21%) | 2 (3%) | 2 (3%) | 2 (3%) | 3 (4%) |
| Seemingly trivial wording effects | 27 | 6 (22%) | 2 (7%) | 2 (7%) | 2 (7%) | 3 (11%) |
| Substantive wording effects | 15 | 3 (20%) | 2 (13%) | 2 (13%) | 2 (13%) | 2 (13%) |
| Sum | 144 | 31 (22%) | 10 (7%) | 10 (7%) | 10 (7%) | 13 (9%) |

## 5 Discussion

One hundred ten survey experiments and replications (affording 144 tests) showed that the effect of a question design manipulation on public opinion on climate change was quite limited: sixteen response order experiments (affording 27 tests) yielded three statistically significant effects. Fifty-eight experiments testing question order effects (affording 75 tests) yielded 14 significant effects. Twenty-one experiments testing seemingly trivial wording effects (affording 27 tests) yielded three significant effects. And 15 experiments testing nontrivial wording effects (affording 15 tests) yielded three significant effects.

Consistent with the satisficing theory, question design effects were more likely to occur among people with no college education, yet the overall presence of such moderation was rare: in 10 of 144 tests, question design effects occurred only among people with no college education. In one of 144 tests, question design effects occurred only among people with at least some college education. In the instances of statistically significant moderation by party identification, no consistent and interpretable pattern appeared, such that one subgroup was more influenced than another.

After applying the Bonferroni, Holm, Hommel, and false discovery rate (FDR) corrections, 7 to 9% of tests were statistically significant. Of the effects that remained significant after these corrections, the average effect size was 0.189 for the Bonferroni, Holm, and Hommel methods and 0.179 for the FDR.

Taken together, the question design manipulations studied here rarely altered response distributions notably. These results challenge the presumption that Americans' views on climate change are weak and uncrystallized and support the conclusion that survey measurements of those opinions can be considered robust. Thus, policymakers may have confidence that public opinion on climate change-related issues is fairly reliable because question design effects occurred rarely, and the effect sizes were usually small.

Motta et al. (2019) conducted research in the same spirit as ours. They conducted experiments with a large nonprobability sample ($N = 7019$) and investigated the extent to which seemingly trivial changes in question design could alter response distributions. Motta et al. (2019) tested three kinds of manipulations that we did not examine here: (1) whether agree/disagree questions yield different distributions of responses than the same questions asked in a different format, (2) whether explicitly offering "don't know" options alters responses, and (3) whether providing text explaining that climate change is caused by greenhouse gas emissions alters respondents' reported opinions.

Motta et al. (2019) found that the agree/disagree question format did yield more affirmative responses, that offering a "don't know" option caused more respondents to abstain from answering, and that adding explanatory text about the link between greenhouse gas emission and climate change caused more respondents to acknowledge the existence of climate change. As a result, those investigators concluded that "seemingly trivial decisions made when constructing questions can, in some cases, significantly alter the proportion of the American public who appear to believe in human-caused climate change."

A likely reason why the current study and Motta et al.'s reached different conclusions is that the two investigations examined different types of question design alterations. It is therefore more appropriate to consider the current research as supplementing Motta et al.'s (2019) findings instead of invalidating them. The conclusions drawn from the current research are confined to the specific types of manipulations investigated and confined to the American population.

Although the present findings can be viewed as suggesting that manipulations of the designs of questions measuring opinions on global warming rarely produced large differences in response patterns, this does not mean that researchers should disregard details of question designs in this arena. For example, in order to be sure that accurate conclusions are reached about opinion change when comparing the results of two surveys done at different times, the same questions should have been asked in the same orders in the two surveys. Otherwise, what appear to be changes in opinions over time may instead be changes in response distributions due to changes in the questions asked.

Furthermore, when a researcher is conducting a new survey to measure public opinion on global warming in the future, and he or she is uncertain of which of various possible ways of asking a question will yield the most accurate results, the researcher should build in an experiment like those reported here. As a result, the researcher can gauge the impact of question wording or structure alterations and report results averaging across the various possible question design approaches. This avoids putting all of the researcher's eggs in the basket of just one possible wording.

One advantage of the type of investigation reported here and by Motta et al. (2019) is that it avoids the file drawer problem and publication bias. We report the results of all experiments included in the five representative sample surveys examined, regardless of whether they yielded statistically significant question design effects or not. This is interesting to consider in light of Cristea and Ioannidis's (2018) evidence that an overwhelming majority of $p$ values published in *Science*, *Nature,* and *PNAS* in 2017 were statistically significant: 94.2% of them (95% CI 91.7% to 96.4%), perhaps because articles are less likely to be published if they report nonsignificant effects that fail to reject the null hypothesis (Franco et al. 2014). Our evidence stands in sharp contrast to such patterns and illustrates the value of complete reporting of the results of a large number of experiments.

Another strength of the current methodology is large samples. Compared with low-powered studies, high-powered studies are thought to be more likely to detect valid effects, buffer against false positives, and replicate (Fraley and Vazire 2014). All 110 experiments reported here had an approximately 100% chance to capture medium ($w$ = 0.3) and large ($w$ = 0.5) effects of question variations, and roughly half of the experiments had a power larger than 0.8 to capture small effects ($w$ = 0.1) effects (see Table S3 in the Supplementary Material). In the current study, multiple experiments were conducted to test the same question design effect, and aggregations of experiments conferred even more robustness to the observed results.

Previous literature has well documented how nonprobability samples (samples that are not randomly drawn from a well-defined population) can produce biased results. For example, the

demographics of samples from YouGov, an online panel of volunteer American adults, deviate from the US Census Bureau's Current Population Survey (CPS) much more than the demographics obtained by the ANES (which involved probability sampling) (Malhotra and Krosnick 2007). Similarly, Yeager et al. (2011) and MacInnis et al. (2018) found that the distributions of demographics and other factual characteristics are more accurate in probability samples than in nonprobability samples. All experiments in the current research involved probability sampling, and the data were properly weighted to account for differential participation. This aspect of the procedure implemented here adds to the generalizability and validity of the conclusions we draw.

In sum, the present investigation indicates that a large set of question design manipulations had minimal impact on the distributions of Americans' opinions on issues related to climate change, attesting to the robustness of those opinions and their measurement. However, we did see significant effects of question design manipulations in more than 20% of experiments. The present evidence on Americans' opinions on this issue therefore dovetails with evidence that the distributions of opinions on climate change issues have been quite stable over the last two decades (Krosnick and MacInnis 2020). The stability in public opinion on climate change indicates that policymakers and researchers can place stock in the implications of public preferences, a good sign for public support for future actions on climate change mitigation and adaptation.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s10584-021-03194-x.

**Data and code availability** Analyses were performed using R v.4.0.2 (https://www.R-project.org/). Data and code needed to replicate the analyses are available at: https://osf.io/9k3cp

**Author contribution** The first, second, and fourth authors wrote the manuscript. The first and second authors analyzed the data. The first, second, and fourth authors developed the study idea. The third author participated in manuscript preparation.

**Funding** The 2012 data collection was supported by the National Science Foundation grant 1042938. The 2013 data collection was supported by Stanford University, Resources for the Future, and USA Today. The 2013 data collection was supported by Stanford University and University of Arizona. The 2015 data collection was supported by Stanford University, Resources for the Future, and The New York Times. The 2018 data collection was supported by Stanford University and Resources for the Future.

## Declarations

**Conflict of interest** The authors declare no competing interests.

**IRB approval number for each of surveys** 2012: IRB-23464
2013: IRB-29317
2014: IRB-31688
2015: IRB-32562
2018: IRB-46124

## References

Abeles A, Howe L, Krosnick JA, MacInnis B (2019) Perception of public opinion on global warming and the role of opinion deviance. J Env Psych 63:118–129
American Clean Energy and Security Act, H.R. 2454, 111th Cong. (2009)

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57(1):289–300
Bishop G, Smith A (2001) Response-order effects and the early Gallup split-ballots. Public Opin Quart 65(4): 479–505
Cantril H (1944) Gauging public opinion. Princeton Univ Press, Princeton
Chang L, Krosnick JA (2009) National surveys Via RDD telephone interviewing versus the Internet. Public Opin Quart 73:641–678
Climate Stewardship and Innovation Act, S. 280, 110th Cong. (2007)
Converse P (1964) The nature of belief systems in mass publics. Crit Rev 18(1-3):1–74
Converse P (1974) Comment: the status of nonattitudes. Am Polit Sci Rev 68(2):650–660
Cristea I, Ioannidis J (2018) P values in display items are ubiquitous and almost invariably significant: a survey of top science journals. PLOS ONE 13:e0197440
Dahl R (1989) Democracy and its critics. Yale Univ Press, New Haven
Erikson RS, Tedin KL (2015) American public opinion: its origins, content and impact. Routledge, New York
Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US states data. PLOS ONE 5:e10271
Fraley R, Vazire S (2014) The n-pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. PLOS ONE 9:e109019
Franco A, Malhotra N, Simonovits G (2014) Publication bias in the social sciences: unlocking the file drawer. Science 345(6203):1502–1505
Gardner G, Stern P (2008) The short list: the most effective actions U.S. households can take to curb climate change. Environment: Science and Policy for Sustainable Development 50:12–25
Holbrook A, Green M, Krosnick JA (2003) Telephone versus face-to-face interviewing of national probability samples with long questionnaires. Public Opin Quart 67:79–125
Holbrook A, Krosnick JA, Moore D, Tourangeau R (2007) Response order effects in dichotomous categorical questions presented orally. Public Opin Quart 71:325–348
Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat:65–70
Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75(2):383–386
Kalton G, Collins M, Brook L (1978) Experiments in wording opinion questions. J R Stat Soc C 27(2):149–161
Krosnick JA (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl Cognitive Psych 5:213–236
Krosnick JA (1999) Survey research. Annu Rev Psychol 50(1):537–567
Krosnick JA, Alwin DF (1987) An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opin Quart 51:201–219
Krosnick JA, MacInnis B (2020) Climate insights 2020: overall trends. Resources for the future. https://www.rff.org/publications/reports/climateinsights2020/. Accessed 25 February 2021
Krosnick JA, Presser S (2010) Question and questionnaire design. In: Wright JD, Marsden PV (eds) The handbook of survey research. Emerald Publishing, West Yorkshire, pp 269–319
Kyselá E, Ščasný M, Zvěřinová I (2019) Attitudes toward climate change mitigation policies: a review of measures and a construct of policy attitudes. Clim Policy 19:878–892
Lavrakas P (2008) Surveys by Telephone. In: Donsbach Wolfgang, Traugott Michael W. (eds) The SAGE Handbook of Public Opinion Research. SAGE Publications Ltd, pp 249–261
Lee T, Markowitz E, Howe P, Ko C, Leiserowitz A (2015) Predictors of public climate change awareness and risk perception around the world. Nat Clim Change 5:1014–1020
Lumley T (2020) survey: analysis of complex survey samples. R package version 3.37. https://cran.r-project.org/web/packages/survey/index.html
MacInnis B, Krosnick JA, Ho AS, Cho MJ (2018) The accuracy of measurements with probability and nonprobability survey samples: replication and extension. Public Opin Quart 82(4):707–744
MacInnis B, Miller JM, Krosnick JA, Below C, Lindner M (2021) Candidate name order effects in New Hampshire: evidence from primaries and from general elections with party column ballots. PLOS ONE 3: e0248049
Malhotra N (2008) Completion time and response order effects in web surveys. Public Opin Quart 72(5):914–934
Malhotra N, Krosnick JA (2007) The effect of survey mode and sampling on inferences about political attitudes and behavior: comparing the 2000 and 2004 ANES to Internet surveys with nonprobability samples. Polit Anal 15(3):286–323
Marcott SA, Shakun JD, Clark PU, Mix AC (2013) A reconstruction of regional and global temperature for the past 11,300 years. Science 339(6124):1198–1201
Miller R (1981) Simultaneous statistical inference. Springer, New York

Motta M, Chapman D, Stecula D, Haglin K (2019) An experimental examination of measurement disparities in public climate change beliefs. Climatic Change 154(1):37–47
Napolitano J, Johnson K (2018) Universities should lead efforts to slow climate change, if the federal government won't. The Washington Post. https://www.washingtonpost.com/education/2018/09/17/universities-should-lead-efforts-slow-climate-change-if-federal-government-wont/
Narayan S, Krosnick JA (1996) Education moderates some response effects in attitude measurement. Public Opin Quart 60(1):58–88
National Surveys on Energy and Environment (2014) American acceptance of global warming retreats in wake of winter 2014. Michigan Univ Center for Local, State, and Urban Policy http://closup.umich.edu/issues-in-energy-and-environmental-policy/12/american-acceptance-of-global-warming-retreats-in-wake-of-winter-2014. Accessed 11 May 2019
Page B, Shapiro R (2010) The rational public: fifty years of trends in Americans' policy preferences. Univ. of Chicago Press, Chicago
Pasek J, Schneider D, Krosnick JA, Tahk A, Ophir E, Milligan C (2014) Prevalence and moderators of the candidate name-order effect: evidence from statewide general elections in California. Public Opin Quart 78(2):416–439
Pasek J, Tahk A, Culter G, Schwemmle M (2020). Weights: weighting and weighted statistics. R package version 1.0.1. https://CRAN.R-project.org/package=weights
Payne S (1951) The art of asking questions. Princeton Univ Press, Princeton
Political Psychology Research Group (2020) American public opinion on global warming. Stanford Univ. http://climatepublicopinion.stanford.edu/
Rothstein H, Sutton A, Borenstein M (2006) Publication bias in meta-analysis. John Wiley & Sons, New York
Schuman H, Presser S (1981) Experiments in question wording, form and context in attitude surveys. Academic, New York
Schuman H, Presser S (1996) Questions and answers in attitude surveys: experiments on question form, wording, and context. Sage, Thousand Oaks
Schwarz N, Bles H (1992) Scandals and the public's trust in politicians: assimilation and contrast effects. Pers Soc Psychol B 18:574–579
Sterman J (2008) Risk communication on climate: mental models and mass balance. Science 322:532–533
Streiner DL (2015) Best (but oft-forgotten) practices: the multiple problems of multiplicity—whether and how to correct for many statistical tests. Am J Clin Nutr 102(4):721–728
Tvinnereim E, Fløttum K (2015) Explaining topic prevalence in answers to open-ended survey questions about climate change. Nat Clim Change 5:744–747
Yeager DS, Krosnick JA, Chang L, Javitz HS, Levendusky MS, Simpser A, Wang R (2011) Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. Public Opin Quart 75(4):709–747