Measuring Misperceptions?

Matthew H. Graham*

Conditionally accepted,
American Political Science Review

October 3, 2021

Survey data are commonly cited as evidence of widespread misperceptions and misinformed beliefs. This paper shows that surveys generally fail to identify the firm, deep, steadfast, confidently-held beliefs described in leading accounts. Instead, even those who report 100 percent certain belief in falsehoods about well-studied topics like climate change, vaccine side effects, and the COVID-19 death toll exhibit substantial response instability over time. Similar levels of response stability are observed among those who report 100 percent certain belief in benign, politically uncontested falsehoods, e.g. that electrons are larger than atoms and that lasers work by focusing sound waves. As opposed to firmly-held misperceptions, claims to be highly certain of incorrect answers are best-interpreted as "miseducated" guesses based on mistaken inferential reasoning. Those reporting middling and low levels of certainty are best-viewed as making close-to-blind guesses. These findings recast existing evidence as to the prevalence, predictors, correction, and consequences of misperceptions and misinformed beliefs.

*Postdoctoral Research Scientist (School of Media and Public Affairs) and Lecturer (Data Science Program), George Washington University. mattgraham@gwu.edu. For helpful comments on earlier versions of this work, the author thanks Alexander Coppock, Alan Gerber, Greg Huber, Scott Bokemper, John Henderson, Annabelle Hutchinson, Seth Hill, Jennifer Jerit, Lilla Orr, Josh Pasek, Kyle Peyton, Kelly Rader, Ira Soboleva, Emily Thorson, and Omer Yair; seminar participants at Yale, George Washington, and the Junior Americanist Workshop Series; and panel participants at the annual meetings of the Society for Political Methodology and the American Political Science Association. This research was supported by the Institution for Social Policy Studies (Yale), Center for the Study of American Politics (Yale), Georg Walter Leither Program in Political Economy (Yale), and the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy & Politics at The George Washington University.

Concern over the political consequences of misperceptions and misinformed beliefs has steadily escalated in recent years. In contrast to ignorance of the truth, misperceptions are distinguished by the depth, firmess, steadfastness, or confidence with which one holds a false or unsupported belief (Kuklinski et al. 2000, Flynn, Nyhan and Reifler 2017). This prevailing definition of a misperception falls in tension with classic research on attitudes, which holds that survey responses are best characterized as on-the-spot inferences based on whatever relevant information the respondent can call to mind (Zaller 1992; Tourangeau et al. 2000). In an effort to close the gap between definitions and measurement, a growing body of research advocates reserving the term "misperception" or "misinformed" for those who report a high level of confidence or certainty about their response (e.g., Kuklinski et al. 2000; Pasek et al. 2015; Graham 2020; Luskin et al. 2018; Peterson and Iyengar 2020). At face value, certainty scales would seem to bridge the gap between the beliefs of interest and the vagaries of the survey response. Yet no published research interrogates the veracity of survey respondents' claims to be certain of falsehoods.

This paper examines the nature of the beliefs captured by survey measures of misperceptions. It does so by adapting the long tradition of using temporal stability to interrogate the degree to which survey responses reflect true attitudes or beliefs (Converse 1964, 1970). As opposed to confidently held beliefs, prevailing practices are more aptly characterized as capturing a mix of blind guesses and "miseducated" guesses based on mistaken, on-the-spot inferences. In five surveys covering a range of topics from existing research—government budgets, politicized controversies, the economy, science, and the COVID-19 pandemic—respondents who initially endorse falsehoods exhibit a large regression to the mean effect in follow-up surveys, assigning far less probability to the falsehood than their initial response implied. Respondents who answer the same questions correctly exhibit three to five times less regression. This result holds even among those who report 100 percent certainty. Whereas the average respondent who reports complete certainty about a correct answer assigns an average probability of around 0.95 to their initial response in a follow-up survey, the average respondent who reports complete certainty about an incorrect answer drops to about 0.75. This means that even the typical respondent who claims to be absolutely certain of falsehoods is not deeply convinced of the statement they have endorsed. Instead, they find the falsehood to be more plausible than not based on underlying beliefs that are suggestive, but not dispositive, as to the matter in question.

Any framework capable of describing a problem can also be used to evaluate solutions. As a step in this direction, the analysis concludes by evaluating a novel intervention that merges frame of reference training (FOR; Bernardin and Buckley 1981; Woehr 1994) with theories of the survey response (Zaller 1992; Tourangeau et al. 2000). Respondents read four short vignettes about a hypothetical person answering a question about the price of gas, guess that person's certainty level, and then receive instruction as to which certainty level is most appropriate and why. This 60- to 90-second exercise increases the temporal stability of measured misperceptions by about 40 percent. These benefits extend to respondents both high and low on several dimensions that have previously been shown to predict incorrect answers to survey questions and real-world engagement with misinformation, e.g. partisan identity and cognitive reflection.

The findings suggest three principles for building a more sound evidentiary basis for understanding the prevalence and consequences of misperceptions. First, interpretations of survey measures can and should be justified with hard empirical evidence. Even as the results yield little evidence of firm belief in falsehoods, the same measurement techniques identify firm, confidently held beliefs among those who report being certain of the correct answers of a multiplicity of questions designed to tap political and scientific knowledge. It cannot be taken for granted that a survey question has measured misperceptions, but it can be proven. Second, theoretical expectations as to who is most likely to be misinformed are a poor substitute for hard evidence. The results hold when samples are split by dispositions that existing research has shown to predict incorrect answers to survey questions and real-world engagement with misinformation, including political party, generic conspiracy beliefs, and need for cognitive closure. Third, evidence on measurement properties should be question-specific. Though this paper finds modest degrees of response stability among incorrect answers to some questions, others are unstable across the board. For example, denial that global temperatures have risen appears to be almost entirely driven by blind guessing, with extremely low response stability even among those who report complete certainty. Similar measurement properties are observed among those who deny the existence of continental drift.

The disconnect between prevailing interpretations of measured misperceptions and their observable measurement properties calls for a reassessment of existing evidence as to the prevalence, predictors, and consequences of misperceptions and misinformed beliefs. Political partisanship may be the most-studied predictor of incorrect survey responses. This paper's findings suggest that

measured partisan belief differences should be interpreted not as evidence of misperceptions, but as differential knowledge and ignorance of convenient and inconvenient truths. As elaborated in the concluding section, this posture is consistent with several established patterns that misinformation-focused accounts have trouble accommodating. The findings also call for reconsideration of research on correcting misperceptions and the benefits (or lack of benefits) that arise from doing so. Much of this research is unlikely to have measured misperceptions to begin with, and is more safely interpreted as describing the consequences of ignorance.

Though the results are discouraging for the unvalidated measurement practices that dominate existing survey-based research on political misperceptions, this paper's ultimate value lies in its development of methods for identifying relatively successful questions and measurement practices. By assuming the burden of proof for its interpretation of survey responses, research can develop a more trustworthy basis for understanding the prevalence and consequences of political misperceptions.

A Conceptual-Empirical Disconnect

Surveys are commonly used to document "widespread" misperceptions and misinformed beliefs among the general public, as well as what personal characteristics predict such beliefs, how to correct them, and the consequences of doing so (Flynn, Nyhan and Reifler 2017, 129; Nyhan 2020, 227). Misperceptions are distinguished from ignorance by the degree of conviction with which the respondent holds the belief (Kuklinski et al. 1998, 2000). Whereas the "genuinely misinformed" "firmly hold beliefs that happen to be wrong," the "guessing uninformed" "do not hold factual beliefs at all" (Kuklinski et al. 2000, 792-93). Consistent with this influential distinction, research describes misperceptions and misinformed beliefs as "firm" (Jerit and Zhao 2020, 78, 81), "deep-seated" (Berinsky 2018, 212), "steadfast" (Li and Wagner 2020, 650), "confidently held" (Pasek, Sood and Krosnick 2015), "belief in information that is factually incorrect" (Berinsky 2018, 212), which can be thought of as "incorrect knowledge" (Hochschild and Einstein 2015, 10). Though the terms "misperception" and "misinformation" are often used interchangeably,[1] this paper favors the former so as to maintain a clear distinction between beliefs and the information environment (also see Thorson 2015).

[1]For example, Flynn, Nyhan and Reifler (2017) define misperceptions using Kuklinski and colleagues' (1998; 2000) definition of misinformation.

Researchers' interest in beliefs of this kind runs into a classic problem in the study of public opinion: respondents answer survey questions even when they do not hold a firm belief about the matter at hand. Converse (1964, 1970) famously pointed out that many responses are temporally unstable, meaning that they change from one survey to the next. To accommodate this and other empirical regularities that problematize the idea that surveys measure pre-existing beliefs (e.g., Schuman and Presser 1981), researchers developed alternative accounts. Consensus now holds that survey-measured attitudes are generally not firm, deep, or steadfast, but are formed by retrieving a "sample" of topic-relevant considerations from memory and integrating them into an on-the-spot judgment (Strack and Martin 1987; Tourangeau et al. 2000; Zaller 1992; also see Berinsky 2017; Bullock and Lenz 2019; Flynn et al. 2017).

In an effort to close the gap between the definition of a misperception and the received wisdom from attitudinal research, some research applies a higher standard of measurement. Research increasingly uses certainty or confidence scales to identify respondents who are misinformed or hold a misperception (Flynn 2016; Graham 2020; Lee and Matsuo 2018; Li and Wagner 2020; Marietta and Barker 2019; Pasek et al. 2015; Peterson and Iyengar 2020; Sutton and Douglas 2020). Such research often finds that misperceptions or misinformed beliefs are much less common than is generally supposed. Luskin et al. (2018) refer to certainty scales as a "24 carat gold standard" for measuring misinformed beliefs. Accordingly, the 2020 American National Election Study added a "misinformation" battery that included a confidence scale of this kind.

At face value, one who reports being certain of a falsehood would seem to firmly believe it. Yet there also exists suggestive evidence that respondents may claim to be certain of falsehoods that are not firmly believed. Alongside questions designed to tap partisan-biased misperceptions, Graham (2020) measures confidence in answers to political knowledge questions about officeholders and institutional rules. About one in ten respondents reported being "very" or "absolutely" certain about an incorrect answer. Graham (2020) attributes this to "traps" set by the response options, e.g., "Nancy Pelosi as the Senate Minority Leader (instead of Chuck Schumer)" and "the filibuster as the Senate procedure to make budget changes via a simple majority (instead of reconciliation)" (318). Few would interpret these responses as representing beliefs that are firm, deep, steadfast, or related in any way to misinformation.

Further reasons to be skeptical that self-described certainty indicates a firmly held belief

emerges from the literature on attitude strength. The few published tests of the strength-stability relationship find that strong attitudes are only modestly more stable than weak attitudes, with little focus on exactly how strong the strongest attitudes are. In a 1974-75 panel survey, Schuman and Presser (1981) find that about 75 percent of high-importance respondents chose the same response to a binary item in both survey waves, compared with 65 percent in the low importance group. Krosnick (1988) finds a weak ("not strong," 243, 247) relationship on six items in the 1980-88 ANES. Re-analyzing a larger subset of the same data, Leeper (2014) finds statistically significant relationships for three of the six items. In three other datasets, Leeper (2014) finds only a weak relationship. Prislin (1996) conducts 14 regression tests for each of three attitudinal scales and found one statistically significant relationship in each case. Evidence also emerges that the strength-stability relationship is heterogeneous. Krosnick (1988) finds the strongest attitudes toward unemployment to be less stable than the weakest attitudes toward other issues. Prislin (1996) finds a stronger relationship with respect to pizza than to any policy issue. Bassili (1996) finds no relationship with respect to attitudes toward pornography. Schuman and Presser (1981) find that among opponents of gun control, attitude strength strongly predicts self-reported activist behavior; among supporters, the relationship is completely flat.[2]

If incorrect answers to survey questions do not represent firm, deep, or steadfast misperceptions, what else could they represent? The analysis considers two other archetypes: blind guesse and miseducated guesses. Blind guessers either do not possess or do not put much effort into recalling topic-relevant considerations. Such respondents should split evenly between response options as though the respondent is flipping a mental coin. Miseducated guesses are made by respondents who sample their considerations from a pool of stored information that favors one response option over the others but is not conclusive as to which is true or which is false. Such respondents may make the same guess with regularity but do not firmly believe the falsehood implied by their incorrect answer. For example, a respondent may reason that a true claim about Trump is false because they believe that media are always making up stories about him (see Table 2 and surrounding discussion). Relative to blind guessers, miseducated guessers are characterized by a greater degree of latent ambivalence, meaning that their memory contains topic-relevant considerations that point

[2]On a four-point scale from "not too important" to "most important," 9, 14, 36, and 56 percent of gun control opponents reported writing a letter or making a donation. Among supporters, these figures were 5, 6, 7, and 6 percent. See their Figure 9.1, page 242.

in both directions. In moments when the most accessible considerations happen to all point in one direction, such respondents may have a fleeting feeling of confidence that is not representative of their true beliefs. In other moments, the same respondents may feel uncertain or even make the opposite guess as to which response option is most likely to be correct.

In the language of the attitudinal literature, an educated or miseducated guess can be thought of as a middle category between Converse's famed limiting cases of a non-attitude and a crystallized belief. Researchers have long recognized that a "third concept" like "quasi-attitudes or pseudo-attitudes" would aptly describe many responses (Schuman and Presser 1981, 159). Even Converse's seminal articles (1964; 1970) found that a "black and white" distinction between non-attitudes and crystallized attitudes applied to only one of eight attitudinal questions; for the other seven, intermediate response types were "entirely compatible with the data" (1964, footnote 41).[3] Attitudinal research ultimately adapted by merging the middle and top categories, lowering the bar for "attitudes" to include on-the-spot judgments (Tourangeau et al. 2000; Zaller 1992) formalized as latent variables that exist by definition (Achen 1975; Erikson 1979; see discussion below). For misperceptions and beliefs more generally, a three-category conceptualization adds value for two reasons. First, far from giving up on the top category, research often claims to have measured deep, firm, steadfast belief in specific falsehoods. Second, as this paper shows, certainty scales do enable firmly held beliefs to be measured for a wide range of items—but only among those who answer correctly. Unlike the case of attitudes, ruling out the possibility that surveys measure firm beliefs is not an option. Instead, research on beliefs and misperceptions needs clear language to distinguish the firmly held beliefs it wants to measure from the mis/educated guesses it often measures instead.

Though archetypes are expositionally useful, the analysis ultimately refrains from anointing any particular certainty level as distinguishing one type of belief from another. The arbitrariness of choosing such thresholds is deep enough that philosophers generally reject threshold-based conceptions of belief altogether (Foley 1992). Instead, the empirical framework below specifies two benchmarks against which to judge claims to be certain about incorrect answers: what would be observed in the absence of measurement error, and what is actually observed among correct answers collected in the same survey using the same measurement technique. This gives a sense of where

[3] Converse's later work expresses enduring frustration at prevailing interpretations of the non/attitude distinction. Describing his supporters and detractors, Converse (2000) wrote that "[w]hat both sides had in common was a basic incomprehension of the role of limiting cases in inquiry" (338).

responses fall along the continuum without resorting to sharp, ultimately arbitrary distinctions. The frequent focus on respondents who claim to be 100 percent certain of their answers is intended not as an implicit threshold, but as a most likely case for measuring misperceptions as they are traditionally defined—and by extension, as a least likely case for this paper's main result.

The task at hand is distinct from two related lines of research. First, as mentioned above, several articles note that apparent misperceptions drop substantially when measures of confidence or certainty are incorporated. This paper focuses not the prevalence or predictors of such responses, but on how to interpret them. Second, other research examines expressive responding, which is survey subjects' tendency to select responses other than their sincere best guess as a way of expressing partisan sentiments (Berinsky 2018; Bullock et al. 2015; Prior et al. 2015). The only study of expressive responding that includes measures of certainty does not probe the veracity of claims to be certain (Peterson and Iyengar 2020). Some studies of expressive responding allow respondents to say "don't know" (DK), which tends to filter out respondents with low levels of knowledge (Luskin and Bullock 2011; Sturgis et al. 2008) and certainty (Graham 2021). This means that DK response options are well-suited to filter out blind guesses, but do not isolate a group of respondents that firmly believes its answers.[4]

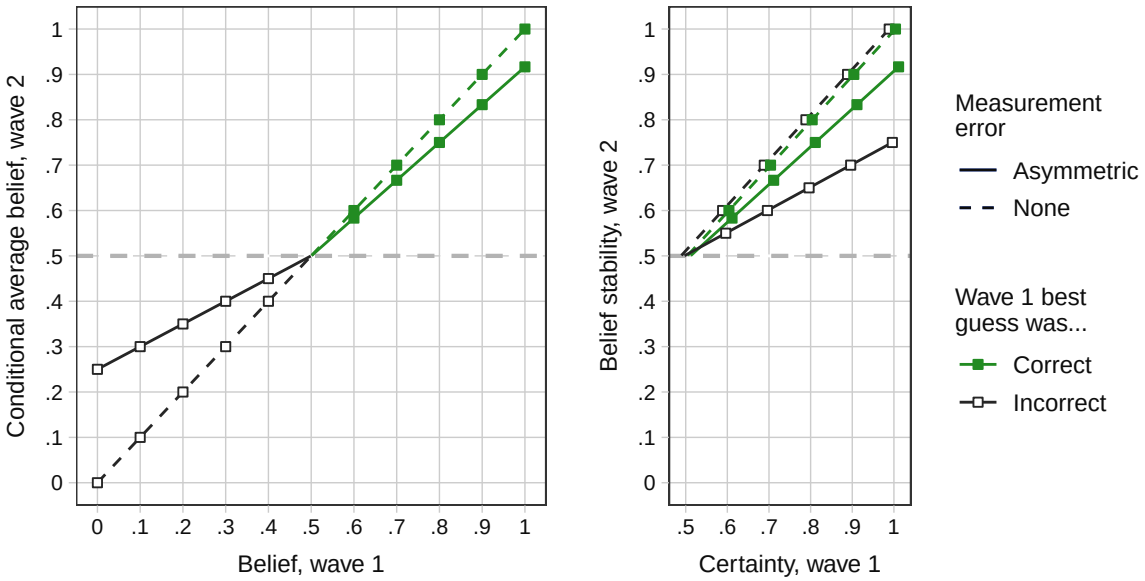
Empirical Framework

Research contending that surveys measure true attitudes has long represented survey responses as functions of probability distributions consisting of a true attitude and an error term (Achen 1975; Erikson 1979; Ansolabehere et al. 2008). The "true" attitude or belief is a latent variable that exists by definition.[5] For a binary question (with two response options), define respondent i's spontaneously-formed belief as $\tilde{p}_{it} \equiv p_i + \epsilon_{it}$, where $p_i \in [0, 1]$ is i's true belief and ϵ_{it} is error in the measure taken at time t. When $p_i = 1$, i holds a completely certain belief in the correct answer. When $p_i = 0$, i holds a completely certain belief in the incorrect an-

[4] Bullock et al. (2015) randomly assign payments for "don't know" responses with expected values of 1.2, 1.5, and 2 times the expected value of a random guess (32). The authors are correct to conclude that this indicates respondents' awareness of their ignorance, but one can only speculate as to the certainty levels of those who preferred to bet on their answer.

[5] Erikson (1979) is especially plainspoken on this point: "[T]he non-attitude holders' probabilities of a 'pro' response (their mean responses) can actually be considered their 'true' positions. For example, the true attitudes of non-opinion holders on 'power and housing' are assumed to be a 0.586 probability of a 'pro' response. Thus, the term 'non-attitude' is technically a misnomer in the sense that by definition, every respondent has a theoretical mean (true) position" (100). Gilens (2012, 58) offers a similarly accessible discussion.

Figure 1: Two ways to display the temporal stability of respondents' measured beliefs.



swer. Accordingly, define i's stated best guess as the response they claim to find most probable, $\tilde{g}_{it} \equiv \mathbb{1}(\tilde{p}_{it} > 0.5)$. Define certainty as the probability i assigns to their best guess, which can be written as $\tilde{c}_{it} \equiv argmax(\tilde{p}_{it}, 1 - \tilde{p}_{it})$. Existing research on factual beliefs adopts similar models with no explicit error term (Bullock et al. 2015; Bullock and Lenz 2019).[6]

To quantify response stability, the analysis will examine what belief is expressed in a follow-up survey conditional on what belief was expressed initially. Figure 1 displays two ways of visualizing this relationship. First focusing on the left panel, define the conditional average belief as $\mathbb{E}[\tilde{P}_{i2}|\tilde{P}_{i1} = p]$, where $\mathbb{E}$, the expectation operator, simply takes the average. If ϵ_{it} is unsystematic and uncorrelated over time, $\mathbb{E}[\tilde{P}_{i2}|\tilde{P}_{i1} = p]$ is an unbiased estimate of the true belief, p_i, conditional on the belief reported at $t = 1$. Absent measurement error, the first and second measures of belief would always line up exactly.[7] In Figure 1, this is visualized by the dashed line that cuts a 45-degree line across the left panel. When beliefs are measured with error, they depart from this ideal. This is represented by the solid line, which is stylized after the results.

As some error is to be expected in all survey measures, it is more charitable to benchmark

[6]The definition of g_i is equivalent to Bullock and Lenz's definition of "believe" (2019, 328) and Bullock et al.'s definition of the response r_j (2015, 47). The definition of c_i is equivalent to Bullock and Lenz's definition of "confident" (2019, 328) and to Bullock et al.'s description of when a respondent is least and most certain (2015, 47). Appendix C.5 further justifies the assumption that g_i and c_i can be constructed out of p_i and vice versa using an experiment embedded in Study 3a.

[7]For proof of these claims, see Appendix E.1.

incorrect answers against a certifiably attainable goal: the degree of stability observed among respondents who claim the same degree of certainty about correct answers. This provides a sense of whether instability among incorrect beliefs could be an artifact of the certainty scale's limitations. To facilitate such comparisons, Figure 1's right panel introduces an alternative display for the same data. Intuitively, the right panel "folds" the left panel both vertically and horizontally, mirroring the bottom-left quadrant onto the top-right. The close alignment between the dashed lines indicates that absent measurement error, the beliefs of respondents who answered correctly and incorrectly should be equally stable. The gap between the solid lines previews the paper's key result: conditional on how certain a respondent claims to be, incorrect beliefs are less stable than correct beliefs. Formally, define *belief stability* as

$$\begin{aligned} b_{i2} = \quad & c_{i2} \quad \text{if} \quad g_{i1} = g_{i2} \\ & 1 - c_{i2} \quad \text{if} \quad g_{i1} \neq g_{i2}. \end{aligned} \tag{1}$$

and conditional belief stability as $\mathbb{E}[B_{i2}|C_{i1} = c]$. This faithfully reflects the stability of each respondent's measured belief while facilitating direct comparisons between respondents' degree of belief in correct and incorrect answers.

A useful interpretation of $\mathbb{E}[b_{i2}|\cdot]$ is the average respondent's true degree of belief in their initial best guess. Just as $\tilde{p}_{i1} = \tilde{p}_{i2}$ when beliefs are measured without error, it follows directly from (1) that an error-free measure of belief would mean that $\tilde{b}_{i2} = \tilde{c}_{i1}$.[8] Differences between $\tilde{b}_{i2}$ and $\tilde{c}_{i1}$ indicate that measurement error systematically inflated (or deflated) the apparent degree to which respondents believe their chosen answer. Accordingly, differences between b_{i2} and c_{i1} will sometimes be referred to as *regression to the mean*.

For some readers, it may help to relate the plotted quantities to predicted values from an OLS regression. Observe that $\mathbb{E}[B_{i2}|C_{i1} = c]$ is a conditional expectation function (CEF). Predicted values from a regression approximate the CEF under the assumption that $E[Y|X = x]$ is linear in X (Aronow and Miller 2019). This means that the plots in this paper provide the same information depicted in a typical plot of predicted values, but without the *ex ante* assumption that stability is exactly linear in certainty. Appendices B.4 and C.4 show that the results hold within a regression framework.

[8] For proof, see the appendix.

While the heart of the analysis focuses on belief stability, Study 1 considers certainty scales defined only in terms of subjective scale points. Such scales do not capture individual-level uncertainty in a way that aligns with distributions defined by probability theory.[9] For such data, the analysis examines a metric that may be more familiar to consumers of survey research: the stability of the respondent's best guess. Define best guess stability as $s_{i2} \equiv \mathbb{1}(g_{i1} = g_{i2})$, which equals 1 if the respondent's best guess in the second survey matches the best guess in the first survey and 0 if the two guesses do not match. In analysis of a survey that elicited only the respondent's best guess about each question, s_i would be called "response stability." For the present analysis, it has two key disadvantages. First, s_i is completely insensitive to cases in which best guesses are stable but certainty is not. Second, an error-free measure of best guesses would always be perfectly stable, regardless of the respondent's level of certainty. As properties of a performance measure, "insensitive to a crucial source of variation" and "uninformative expectations" are not great. Despite these shortcomings, the appendices to Studies 2 and 3 show that similar results obtain when best guess stability is substituted for belief stability.

Threats to inference

The analysis takes steps to mitigate four sources of measurement error that could artifically inflate differences in stability between correct and incorrect answers. First, respondents could look up the correct answers while taking the survey. Accordingly, each survey included at least one established method of deterring and detecting information search. Second, differences between correct and incorrect answers could be an artifact of scale coarseness. Coarse scales ask respondents with a range of latent certainty levels to group themselves together into a the same bin, potentially creating an artificial gap between correct and incorrect answers. For example, it could be that most of those who answer correctly and choose the highest certainty level are close to 100 percent certain while most of those who answer incorrectly and choose the highest level intend to claim only 70 or 80 percent certainty. Whereas Study 1 uses scales from previously published research, Studies 2 and 3 account for concerns about coarseness by using more-granular scales. Third, respondents' true beliefs may genuine change between waves of the survey. If the information that causes such

[9]The relative ease of defining theoretical expectations for measures with clear referents in probability theory motivated this paper's use of binary questions, which easily map onto the binomial distribution.

changes disproportionately favors the correct answer, an asymmetry between correct and incorrect answers could emerge as a consequence. Fourth, it could be that expressive responding occurs in both waves, artificially inflating the stability of incorrect beliefs among respondents with a partisan incentive to endorse a falsehood (as well as correct beliefs about convenient truths).

To address the third and fourth threats, the results of Studies 2 and 3 are reproduced using an alternative, incentive-compatible measure of belief. The *costly measure* collects the same information as a direct question using a series of choices between payment for a correct answer and fixed probabilities of earning the same reward.[10] Measuring the belief twice in the same survey using two distinct measures mitigates concerns that the results are an artifact of change between surveys.[11] The financial incentive mitigates the concern that expressive tendencies, not the beliefs themselves, drive belief stability and partisan differences therein.

The costly measure proceeds as follows. At the outset, respondents are told that they will make a series of choices between tickets to enter into drawings for bonus payments of up to $100. On each screen, respondents first choose which of two tickets they would like to enter into the drawing: win if [choice A], or win if [choice B]. A menu of additional choices then appears: win if [selected choice], or an X in 10 chance to win. By choosing between winning if one's best guess is correct and a 6 in 10, 7 in 10, 8 in 10, 9 in 10, and 99 in 100 chance to win, respondents reveal their probabilistic beliefs in an incentive-compatible manner. For example, one who would rather be paid for a correct answer than an 8 in 10 chance to win, but prefers a 9 in 10 chance over payment for a correct answer, assigns a probability between 0.8 and 0.9 to their response. Hill (2017) uses a version of this approach to study beliefs about politically relevant facts. Holt and Smith (2016) find that discrete choice methods like this paper's outperform methods that ask respondents to directly state their crossover probability (also see Trautmann and van de Kuilen 2015).

Study 1: Foreign Aid

The U.S. government's foreign aid budget is a classic case in research on misperceptions. In the 1990s, polling on the subject attracted sufficient attention that "the Clinton administration

[10] Relative to methods like the quadratic scoring rule, tasks of this type have an important theoretical advantage: because the reward is held constant, the only difference in the expected payoff is the respondent's personal probability that their answer is correct, implying invariance to risk preferences (Allen 1987; Ducharme and Donnell 1973).

[11] This could also be accomplished by collecting three waves of panel data and comparing stability between the first and second waves to stability between the first and third waves (Converse 1964; Wiley and Wiley 1970).

embarked on a major public relations effort focused on countering the American public's overestimation of U.S. spending on foreign aid" (Kull 2011, 57). Whereas foundational research interprets Americans' incorrect answers to survey questions about foreign aid as representing ignorance (Gilens 2001), recent work heavily favors misperception and misinformation frames (Flynn 2016; Guay 2021; Hochschild and Einstein 2015; Scotto et al. 2017; but see Lawrence and Sides 2014).

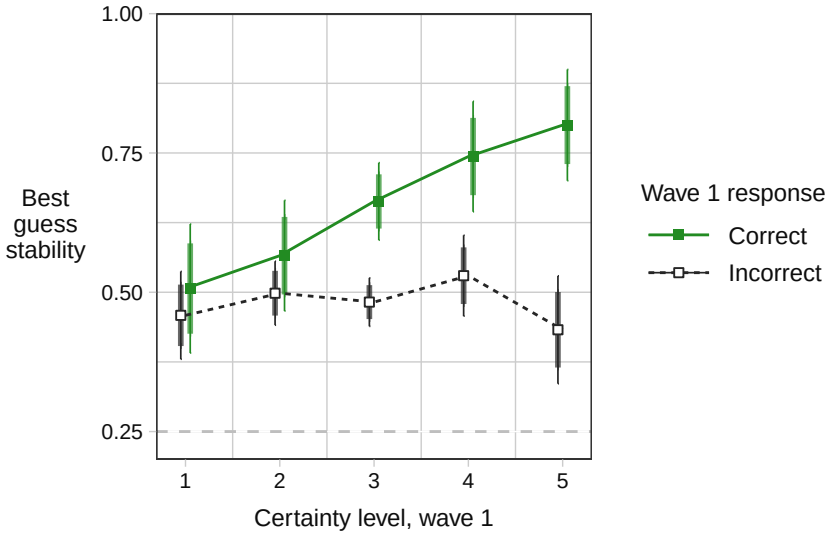
This section introduces the paper's main finding using this classic case. The foreign aid question from the 2012, 2016, and 2020 ANES was embedded in the pre-treatment background questions for an unrelated panel survey conducted on Lucid in August and September 2018 (wave 2 $N = 1749$). To discourage information search, respondents were first asked to pledge not to cheat (Clifford and Jerit 2016). Respondents were then asked, "On which of the following does the U.S. federal government currently spend the least?" and allowed to choose between four options, Foreign aid, Medicare, National defense, and Social Security.[12] As soon as the respondent answered, a five-point certainty scale appeared.[13] The scale's wording was randomly assigned. Half of respondents used the certainty scale from Graham (2020), while the other half used the certainty scale from Pasek, Sood and Krosnick (2015). The Graham scale asked respondents, "How certain are you that your answer is correct?" and used scale point labels ranging from "not at all certain" to "absolutely certain." The Pasek scale asked, "How sure are you about that?" and used labels from "not sure at all" to "extremely sure." The two scales had similar measurement properties and are pooled here for simplicity. Appendix B splits the results by scale.

In the first wave, 28.4 percent of respondents answered correctly. Average certainty was 2.92 among respondents who answered correctly, and 2.83 among respondents who answered incorrectly (difference = 0.09, s.e. = 0.06). The small difference in certainty belies a larger difference in response stability. When recontacted 1 to 3 weeks later for the second survey, 65.1 percent of respondents who initially answered correctly chose the same best guess, compared with 48.6 percent of respondents who answered incorrectly (difference = 16.4, s.e. = 2.6). The share of respondents answering correctly held steady at 29.1 percent, suggesting that belief change between surveys is unlikely to have driven differences in response stability.

[12] Providing national defense as a response option provides some robustness to Williamson's (2019) finding that some public over-estimation is driven by a tendency to think of military spending as foreign aid.

[13] Graham (2020) shows that relative to questions with no certainty scale, this method of measuring certainty has no effect on respondents' average best guesses.

Figure 2: Temporal stability of best guesses by certainty level, Study 1.



Note: The x-axis displays c_{i1}. The y-axis displays $\mathbb{E}[S_{i2}|C_{i1} = c_{i1}]$. Thin error bars represent 95 percent confidence intervals. Thick error bars represent 84 percent confidence intervals to aid comparisons between estimates (see note to Figure 2); a lack of overlap between two such intervals suggests a statistically significant difference at the $p < 0.05$ level, two-tailed (Julious 2004).

To examine the certainty scales' success in identifying deeply held misperceptions, Figure 2 displays best guess stability conditional on certainty. The stability of correct answers rises with certainty, while the stability of incorrect answers is virtually flat. Because respondents were not offered a DK response option, there is a clear floor for response stability: if respondents were choosing completely at random, they would choose the same response option 25 percent of the time. Incorrect answers sit above this floor, falling near 50 percent regardless of the respondent's certainty level. This suggests that incorrect answers reflect some tendency on the part of respondents to consistently retrieve similar considerations from memory as they form their on-the-spot judgment. However, the certainty scales did not capture much variation in this tendency.

Study 2: Politicized Controversies

Though Study 1 demonstrates that claims to be certain of falsehoods do not always indicate firmly held misperceptions, one may expect different results when it comes to salient political controversies. To gather such evidence, two original panel surveys were conducted on Amazon Mechanical Turk (MTurk). Study 2a was fielded in June 2019 and June 2020 (second wave $N =$ 466). To discourage information search, it included a pledge not to cheat and an obscure "catch"

question that would be difficult to answer correctly without looking it up (Clifford and Jerit 2016). The first wave concluded with an open-ended follow up questions about how subjects came up with their answer to one randomly selected question. Study 2b was fielded on MTurk in March and August 2020 (second wave $N = 420$). It included a pledge not to cheat and a cheating detection method similar to those described by Diedenhofen and Musch (2017) and Permut et al. (2019). The first wave concluded with the costly measure of belief.

The surveys covered six politicized controversies, which were selected based on two criteria. First, partisan balance. Three questions' incorrect answers are congenial to Democrats and three are congenial to Republicans. Second, prominent real-world misinformation. Four questions cover salient political controversies with prominent false claims in the public sphere, while two less prominent controversies (numbered 3 and 6 below) provide points of comparison. The questions with incorrect answers congenial to Democrats were:

1. **Clinton email.** Respondents were asked whether the following is true or false: "While she was Secretary of State, Hillary Clinton used a private email server to send and receive classified information." This was a key controversy during and after the 2016 presidential election campaign. Both before and after an FBI investigation revealed that Clinton had sent classified information, she falsely claimed that she had not.[14]

2. **Trump-Russia collusion.** After a one-sentence description of the Robert Mueller's special counsel investigation into Russian interference in the 2016 presidential election, respondents were asked whether the following is true or false: "Robert Mueller's report stated that Trump personally conspired with Russia to influence the 2016 election." Prior to the release of the report, many left-leading opinion claimed that Mueller would find such evidence.[15]

3. **Obama DAPA reversal.** After a one-sentence description of Deferred Action for Parents of Americans (DAPA), a 2014 Obama initiative that was struck down in court, respondents were asked whether the following is true or false: "About a year earlier, Obama said that he would be ignoring the law if he issued such an order." Obama said exactly this in a 2013 interview, but later denied changing his position.[16]

The questions with incorrect answers congenial to Republicans were:

4. **Obama birth certificate.** Respondents were asked whether the following statement is true or false: "President Obama has never released his birth certificate." This question taps a clearly factual element of a larger conspiracy theory. Even after Obama released both his

[14] "FBI findings tear holes in Hillary Clinton's email defense," *PolitiFact*, July 6, 2016.

[15] Claims that are later proven false are consistently included in authoritative definitions of misinformation (e.g., Lazer et al. 2018, Lewandowsky et al. 2012).

[16] "Barack Obama: Position on immigration action through executive orders 'hasn't changed'," *PolitiFact*, November 20, 2014.

short- and later long-form birth certificates, demands that he do so continued to populate public discourse and social media.[17]

5. **Trump said 'grab them.'** Respondents were asked whether the following statement is true or false: "Before becoming president, Donald Trump was tape recorded saying that he kisses women and grabs them between the legs without their consent." This was a major controversy in the 2016 presidential election campaign. After initially apologizing, President Trump later claimed that the tape was inauthentic.[18]

6. **Trump Article II.** Respondents were told that Article II of the Constitution describes the President's powers, then asked whether "President Trump has said that Article II gives him the power to do whatever he wants" is true or false. Trump has never disputed making this statement. This is the only question of the six that has not been the subject of prominent false claims.

After respondents chose their best guess, a certainty scale appeared. The scales were given a probabilistic interpretations using both numerical labels (e.g., 50 to 100 percent certain) and three subjective anchors, "don't know," "probably [answer]," and "definitely [answer]." As a benchmark, three measures of the public's general political knowledge (party control of the House of Representatives, John Roberts' job, and Jerome Powell's job) were included in Study 2b.

Regression to the mean

Combining the two surveys, Table 1 introduces the data and examines subjects' tendency to regress to the mean. On average, the percentage of correct answers was similar for the two sets of questions (first column). In the first survey, respondents who answered correctly assign an average probability of 0.88 and 0.85 to their answer, closer to a firm belief than a blind guess (second column). In the second survey, respondents regress slightly, assigning a probability of 0.83 and 0.80 to their initial response (third column). This regression to the mean of about 0.05 (fourth column) suggests that measurement error modestly over-states the extent to which correct answers represent firm, knowledge-like belief in the truth.

Incorrect answers exhibit greater regression. In the first survey, respondents who answer incorrectly assign an average probability of 0.70 and 0.74 to their answers (fifth column), which appears only somewhat closer to a blind guess than a confidently held false belief. In the second survey, respondents assign a probability of 0.55 and 0.55 to their initial responses (sixth column),

[17] "Fact check: Old fabricated Obama "Kenyan birth certificate" resurfaces," Reuters, June 17, 2020.

[18] "Trump Once Said the 'Access Hollywood' Tape Was Real. Now He's Not Sure." *The New York Times*, November 28, 2017.

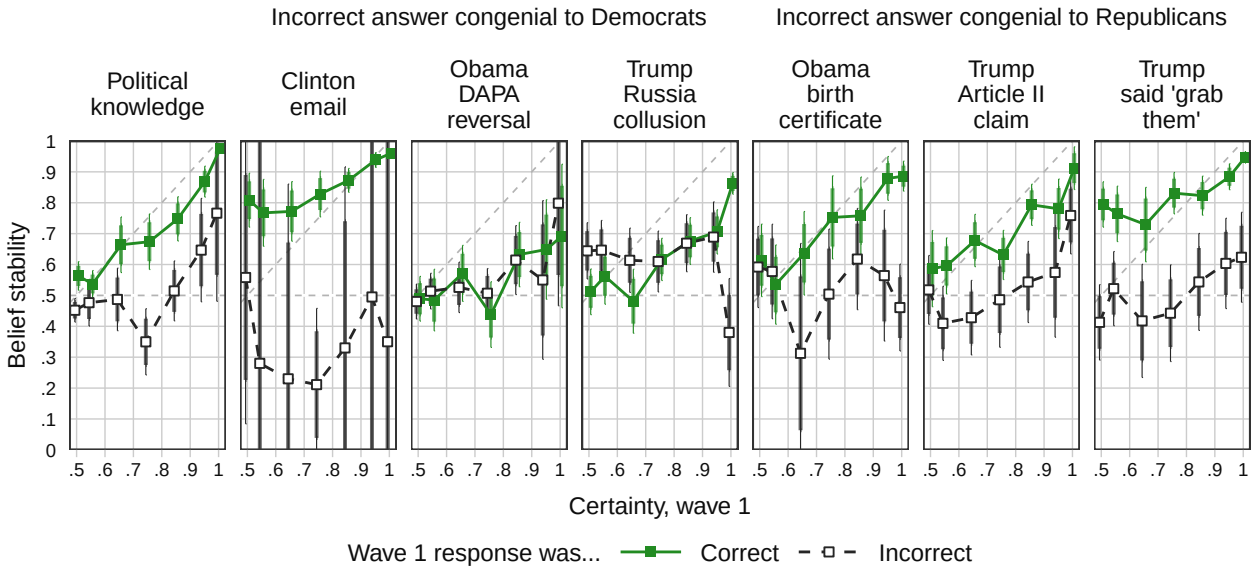
Table 1: Regression to the mean, Study 2.

Question	Percent correct	Correct ($g_{i1} = 1$)			Incorrect ($g_{i1} = 0$)			D-in-D
		c_{i1}	b_{i2}	Diff	c_{i1}	b_{i2}	Diff	
Political knowledge	0.707 (0.013)	0.882 (0.006)	0.826 (0.009)	-0.055 (0.007)	0.698 (0.009)	0.547 (0.013)	-0.151 (0.014)	-0.096 (0.015)
Controversies	0.718 (0.009)	0.854 (0.005)	0.795 (0.006)	-0.059 (0.006)	0.741 (0.007)	0.548 (0.011)	-0.192 (0.012)	-0.134 (0.013)
Clinton email	0.943 (0.011)	0.873 (0.008)	0.902 (0.008)	0.028 (0.009)	0.729 (0.032)	0.333 (0.076)	-0.396 (0.087)	-0.424 (0.088)
Obama birth certificate	0.698 (0.022)	0.832 (0.010)	0.775 (0.019)	-0.057 (0.018)	0.766 (0.016)	0.534 (0.032)	-0.232 (0.037)	-0.175 (0.041)
Obama DAPA reversal	0.414 (0.024)	0.719 (0.012)	0.548 (0.021)	-0.172 (0.022)	0.674 (0.010)	0.530 (0.017)	-0.144 (0.018)	0.028 (0.028)
Trump-Russia collusion	0.713 (0.016)	0.855 (0.007)	0.712 (0.014)	-0.143 (0.013)	0.747 (0.011)	0.622 (0.021)	-0.126 (0.024)	0.017 (0.027)
Trump Article II claim	0.616 (0.024)	0.800 (0.010)	0.733 (0.017)	-0.066 (0.016)	0.774 (0.015)	0.543 (0.027)	-0.232 (0.026)	-0.165 (0.031)
Trump said "grab them"	0.793 (0.014)	0.902 (0.006)	0.890 (0.009)	-0.012 (0.009)	0.771 (0.014)	0.520 (0.028)	-0.251 (0.028)	-0.239 (0.030)

a regression to the mean of about 0.15 on the knowledge questions and 0.19 on the controversy questions (seventh column). This is more regression than is seen among those who answered correctly (eighth column). Relative to correct answers, incorrect answers are less representative of deeply held beliefs.

These patterns are equally stark at the level of individual questions. For example, the average respondent who incorrectly states that Trump never said "grab them" reports a higher level of certainty than did the typical respondent who answered a general knowledge question incorrectly (0.77 versus 0.70). However, upon a second measure, respondents who endorse the false claim about Trump state a *lower* belief in their initial response than those who pick the wrong answer to political knowledge questions (0.52 versus 0.55). The highest average belief in one's incorrect answer, 0.62 among those who at first said that Trump personally colluded with Russia, is three times closer to a blind guess (0.5) than to incorrect knowledge (1.0). In the remaining cases, the typical incorrect answer to the controversy items does not reflect any stronger a belief than does the typical incorrect answer to a political knowledge question.

Figure 3: Temporal stability of beliefs by certainty level and question, Study 2.



Note: The x-axis displays c_{i1}. The y-axis displays $\mathbb{E}[B_{i2}|C_{i1} = c_{i1}]$. Thin error bars represent 95 percent confidence intervals. Thick error bars represent 84 percent confidence intervals to aid comparisons between estimates (see note to Figure 2).

Results by certainty level

Researchers use certainty scales in part to address their suspicion of what has just been shown—that incorrect answers do not reliably indicate deeply held misperceptions. To what extent do certainty scales succeed in closing this conceptual-empirical gap? Figure 3 plots belief stability conditional on the respondent's wave 1 response (correct or incorrect) and their certainty level. This and all following figures bin the certainty scale as follows: 0.5, [0.51, 0.59], [0.6, 0.69], [0.7, 0.79], [0.8, 0.89], [0.9, 0.99], 1. Stability in the lowest and highest bins will frequently be significantly lower or higher than the adjacent bin, confirming the value of scale granularity.

The controversy questions offer little evidence that incorrect answers to questions about partisan or politicized matters are reflective of firmly held beliefs (rightmost six panels, Figure 3). Among respondents who at first claim to be 100 percent certain of the incorrect answer, belief stability tops out at 0.80 among respondents who claim to be certain that Obama never said that an order like DAPA would amount to ignoring the law (third panel from left). However, this estimate is based on only seven respondents (all Democrats) and is not statistically distinguishable from blind guessing. The next-highest stability among the 100 percent certain and wrong comes on the Trump Article II question (0.76, second panel from right). Leaving aside those who report 100

percent certainty, the highest belief stability among any other subgroup is 0.66, among respondents who report being 90 to 99 percent certain that Mueller found personal collusion between Trump and Russia (Figure 3, middle panel).

This instability is not attributable to a flawed certainty scale. On the political knowledge questions, belief stability consistently comes close to the level that would be observed in the absence of measurement error (Figure 3, leftmost panel). Among respondents who report 100 percent certainty about the correct answer to these questions, belief stability reaches 0.98. Almost everyone who claims to be certain about facts like the identity of the Federal Reserve Chair appears to genuinely hold a firm, confident belief in the factual statement they endorse.

To make the results more concrete, Table 2 displays four selected respondets' descriptions of how they came up with their answers. Prevailing uses of certainty scales would classify the respondents as holding a deeply held misperception in one of the two waves and as some other kind of belief in the other wave.[19] Although the respondents indicate some awareness of the controversy at hand, each also indicates that some heuristic helped them answer the question. Consider the Obama birth certificate respondent, a Republican. In the first wave $p_{i1} = 0.13$, meaning that the respondent chose the wrong answer ($g_{i1} = 0$) and reported 87 percent certainty ($c_{i1} = 0.87$). The respondent is not aware that Obama released his birth certificate but reasons that he must not have; if he had, there would be no controversy. In wave 2, $p_{i2} = 0.75$, meaning that the respondent selected the correct answer ($g_{i2} = 1$) and reported 75 percent certainty ($c_{i2} = 0.75$). Despite having a fair amount of confidence in their initial on-the-spot inference, this respondent reached a different conclusion the second time around.

On the surface, there is little to distinguish individuals who state low levels of certainty (around 0.5 to 0.7) from those who state moderate levels of certainty (around 0.7 to 0.9). A closer look suggests that low certainty responses are characterized by a relatively stable tendency to select low levels of certainty, while moderate certainty responses are more-affected by a latent ambivalence that results in more-variable responses. To show this, Appendix B plots the variance in b_{i2} conditional on the respondent's intial certainty level, c_{i1} (i.e., $\text{Var}(B_{i2}|C_{i1} = c)$); Appendix C does the same for Study 3. For both knowledge and controversy questions, second-wave variance is

[19] Although there is no set standard, existing studies typically classify misperceptions as the top half (Kuklinski et al. 2000; Li and Wagner 2020) or three-fifths (Flynn 2016; Graham 2020; Pasek et al. 2015; Peterson and Iyengar 2020) of the certainty scale. On a 0.5 to 1 scale, these correspond to 0.75 and 0.8.

Table 2: Examples of miseducated guesses, Study 2a.

Question	Party	p_{i1}	p_{i2}	Open-ended response
Clinton email	Democrat	0.13	0.60	I know Clinton used a private email server to send and receive messages but I highly doubt she used it to send "classified" material.
Obama birth certificate	Republican	0.13	0.75	I don't recall ever seeing a birth certificate. If there had been one, the question of where he was born would have been settled.
Trump-Russia	Democrat	0.43	0.06	I don't know of the specific language in the report, but it did indicate some level of collusion.
Trump said "grab them"	Republican	0.19	0.01	I'm not sure about this question. So much disinformation about Pres Trump has been pushed by the mainstream media that I cannot keep up with it.

lower for those indicating low certainty levels than for those indicating moderate certainty levels. In Studies 2a and 2b, the difference is about 40 percent, while in Studies 3a and 3b, conditional variance nearly doubles between the lowest and middle certainty levels. This indicates that over time, those who state low certainty levels are relatively consistent in reporting complete uncertainty, while those who indicate moderate certainty levels have a relatively greater tendency to jump from modest confidence in one answer to modest confidence in the other.

Results by political party

Conventional wisdom holds that misperceptions are likely to be more pronounced among those with a partisan incentive to believe falsehoods. For example, Republicans should hold stronger misperceptions about whether Obama released his birth certificate, while Democrats should hold stronger misperceptions about whether Trump was found to have personally colluded with Russia. Can researchers solve the measurement problem simply by focusing on subgroups in which theory predicts stronger misperceptions? To find out, the analysis now collapses responses according to which response is congenial to the respondent's partisanship (e.g., Prior et al. 2015; Peterson and Iyengar 2020) using the grouping that appears in the bulleted list above and the header to Figure 3.

Incorrect answers that are congenial to the respondent's partisanship are indeed more temporally stable. In Study 2a, the average such respondent assigned a probability of 0.60 to their initial, incorrect response, compared with just 0.43 for respondents without a partisan reason to hold the

misperception. In Study 2b, these figures were 0.62 and 0.42. Although it would be grossly misleading to assume that everyone with a partisan incentive to endorse a given false claim possesses a deeply-held misperception, such responses do appear to be more meaningful on average.

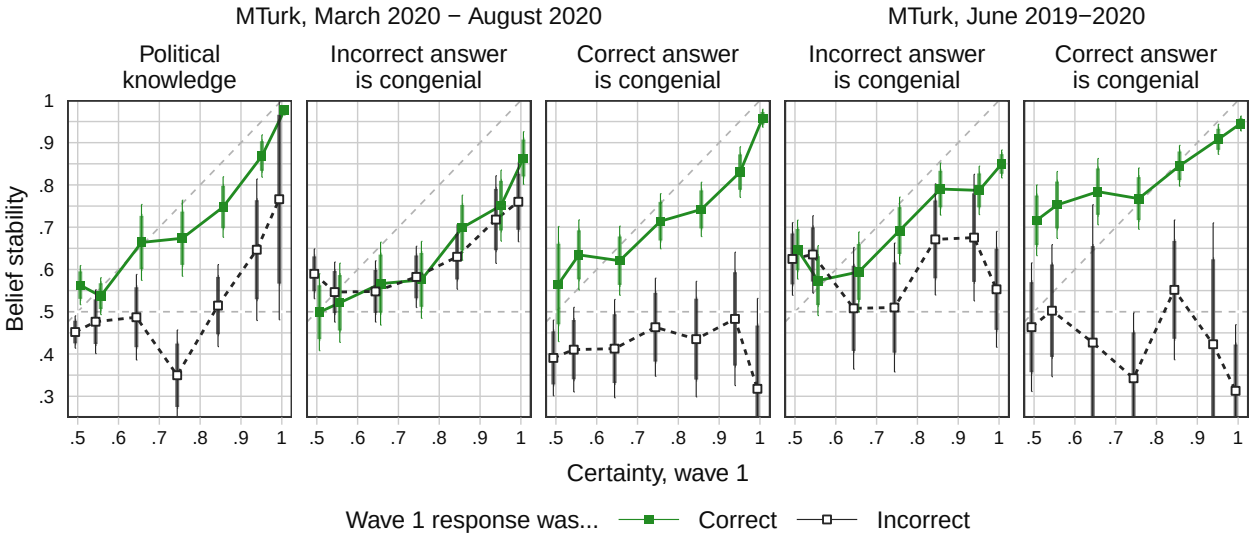
Partisan differences in stability are also present when splitting the results by certainty level. Figure 4 plots belief stability by partisan congeniality. The political knowledge benchmark in the leftmost panel is identical to the equivalent panel in Figure 3. Among "incorrect-congenial" respondents, belief stability among the 100 percent certain was 0.76 in the March-August panel (center-left panel). This is almost exactly equidistant between complete ignorance and complete certainty. Results are similar in the June-June panel, with lower stability among the 100 percent certain but similar stability between 80 and 99 percent certainty (center-right panel). Even in a setting that takes no steps to reduce expressive responding, the typical respondent who claims to be certain of pro-party falsehoods appears to be making a miseducated guess, not revealing a deeply held misperception.

The results are different for respondents without a partisan incentive to endorse the correct answer rather than the incorrect one (center and right panels, Figure 3). Belief stability among those for whom the incorrect answer is congenial correct answers comes close to ideal performance among those who claim a high level of certainty. Among incorrect answers, belief stability never exceeds 0.5, the level that would realize from blind guessing.

Results with an incentive-compatible measure

As noted above, panel data raise two key threats to inference: belief change between surveys may create an artificial gap between correct and incorrect answers, and expressive responding may artificially inflate partisan differences in response stability. To examine whether the results are robust to these threats, the costly measure was included in Study 2b. Figure 5 replicates Figure 4 using this measure. Also included in the figure are results for four economic questions on the budget deficit, GDP growth, unemployment, and inflation (full text appears in Appendix B). Questions on these topics often appear in research on misperceptions and misinformed beliefs (Flynn 2016; Graham 2020; Hellwig and Marinova 2015; Lee and Matsuo 2018), but were omitted from the second wave because the economic fallout from the COVID-19 pandemic's onset caused the correct answers to change.

Figure 4: Temporal stability of beliefs by certainty level and partisan congeniality, Study 2.

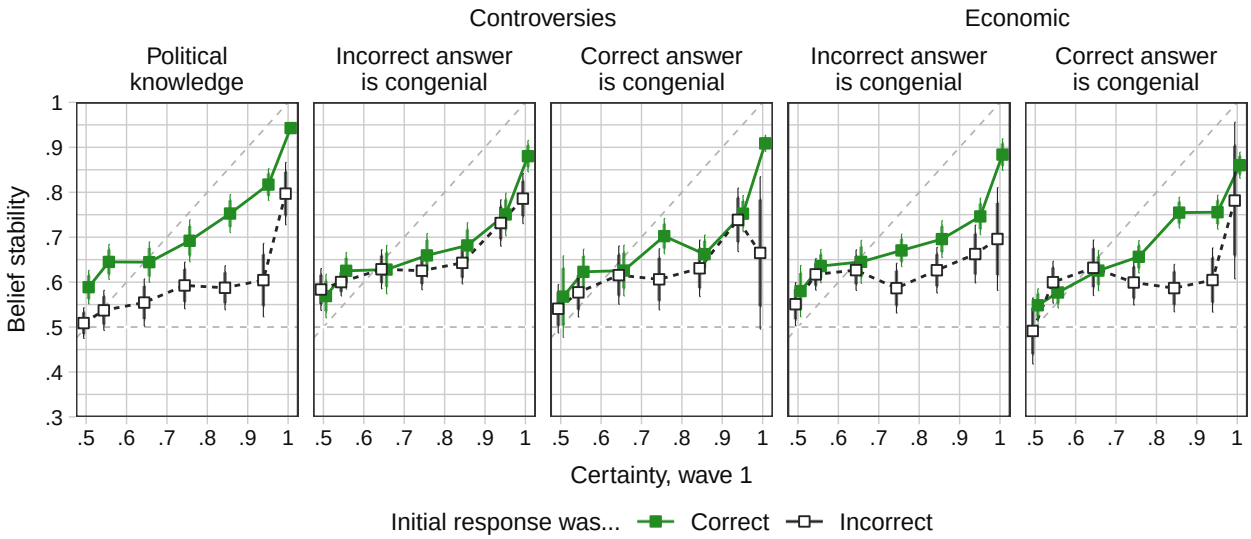


Note: The x-axis displays c_{i1}. The y-axis displays $\mathbb{E}[B_{i2}|C_{i1} = c_{i1}]$. Thin error bars represent 95 percent confidence intervals. Thick error bars represent 84 percent confidence intervals to aid comparisons between estimates (see note to Figure 2).

Relative to the results above, two key differences emerge. First, the partisan difference between congenial and uncongenial questions shrinks. This difference is driven by respondents for whom the correct answer is congenial. These respondents displayed a greater tendency to back off from their correct answers and stick with their incorrect answers. For those who initially endorsed an uncongenial, incorrect answer, the costly measure revealed a belief of 0.62 in it—a large increase over the 0.42 for the equivalent, temporal stability-based figure that appears in Table 1. By comparison, those who initially endorsed a congenial, incorrect answer assigned a probability of 0.65 to it, only a small difference from the 0.62 observed with temporal stability. These patterns are also evident conditional on the respondent's initially reported certainty level. Observe that whereas the middle panels of Figures 4 and 5 are markedly different, the center-left panels are quite similar. This suggests that to the degree that expressive responding affects belief stability, it works primarily through exaggerated claims to know politically convenient truths, and less so through exaggerated claims to believe congenial falsehoods.

Second, because the single-wave design prevents between-wave attrition, the sample is larger. This permits incorrect answers to political knowledge questions to provide a more useful benchmark. Among respondents who reported 100 percent certainty about the wrong answer to a political knowledge question, belief stability reached 0.81 (leftmost panel, Figure 5). This is statistically

Figure 5: Stability by certainty level and partisan congeniality, costly measure, Study 2b.



Note: The x-axis displays c_{i1}. The y-axis displays $\mathbb{E}[B_{i2}|C_{i1} = c_{i1}]$. Thin error bars represent 95 percent confidence intervals. Thick error bars represent 84 percent confidence intervals to aid comparisons between estimates (see note to Figure 2).

indistinguishable from the 0.78 observed among those with a partisan incentive to endorse a falsehood. This bin is primarily populated by respondents who claimed to be 100 percent certain that Republicans, not Democrats, control the U.S. House of Representatives. Existing analysis of claims to be certain of similarly uncontroversial falsehoods finds that such respondents draw on misleading considerations (Graham 2020), e.g. the fact that Republicans did actually control both the U.S. Senate and the presidency at the time of the survey. Whatever sense in which claims to be certain of incorrect answers to survey questions indicate misperceptions must be able to accommodate the existence of such beliefs with respect to benign, uncontroversial false claims.

Study 3: Science and COVID-19

Two additional surveys were conducted to examine whether the results generalize to beliefs about science and the COVID-19 pandemic. Study 3a was conducted on Lucid in November 2020 and December 2020-January 2021 (second wave $N = 1016$). Study 3b was conducted on MTurk in May-June 2021 (second wave $N = 1983$). The first wave of each survey included a set of background characteristics prior to the initial measure of the respondent's beliefs. The second wave repeated the questions. The Lucid survey's second wave also included the costly measure. Both surveys featured the same measures for deterring and detecting information search as the March-August

2020 panel. Both also included a training exercise designed to increase the stability of measured misperceptions, which is analyzed in the next section.

The surveys included six total questions about politically controversial scientific facts (hereafter, "controversies"). Four were taken directly from the 2020 ANES. The ANES codebook explicitly labels these items as measuring misinformation, and the survey includes a certainty scale to assist in this endeavor. The items ask whether vaccines cause autism (they do not), whether global temperatures are higher than 100 years ago (they are), whether genetically modified (GMO) foods are safe to eat (they are), and whether hydroxychloroquine is a safe and effective treatment for COVID-19 (it is not).[20] The remaining controversy questions relate to prominent false claims about the COVID-19 pandemic. One is that official numbers exaggerate the COVID-19 death toll.[21] After a preface that briefly explained excess death analysis, the "COVID deaths" question asked whether such analysis suggests that the official death toll is too low or too high. To provide a measure of partisan balance, a false claim prominently forwarded by left-leaning opinion leaders was also selected. During the 2020 budget process, the Trump administration initially proposed cuts to the CDC budget but ultimately signed an increase into law. Many opinion leaders falsely claimed that Trump had cut the budget.[22] The "CDC budget" question asked respondents whether the Trump administration did or did not secure cuts to the CDC budget.

As a benchmark, the surveys included seven items from the General Social Survey's science knowledge questionnaire. These concern the relative size of electrons and atoms (atoms are larger), whether the continents move (they do), whether the mother or father's gene determines a child's sex (it is the latter), whether Earth revolves around the Sun (it does), whether antibiotics kill viruses (they do not), whether lasers work by focusing sound waves (they do not), and whether radioactivity is all man-made or can occur naturally (it can).

[20] Both surveys also included the 2020 ANES question about COVID-19's origin, but recent developments suggest that the scientific community was too quick to rule out the theory that the virus that causes COVID-19 was developed in a lab. Because this question does not have a clear correct answer, it is excluded from all analysis.

[21] For example, see: Jon Greenberg, "COVID-19 skeptics say there's an overcount. Doctors in the field say the opposite," PolitiFact, April 14, 2020. Saranac Hale Spencer, "CDC Did Not 'Admit Only 6%' of Recorded Deaths from COVID-19," FactCheck.org, September 1, 2020. Angelo Fichera, "Trump Baselessly Suggests COVID-19 Deaths Inflated for Profit," October 29, 2020. Samantha Putterman, "Chart comparing 2020 US death toll with previous years is flawed, uses incomplete data," PolitiFact, November 22, 2020.

[22] Lori Robertson, Jessica McDonald, and Robert Farley, "Democrats' Misleading Coronavirus Claims," FactCheck.org, March 3, 2020.

Table 3: Regression to the mean, Study 4.

		Study 4a					Study 4b		
			Direct question		Costly choices			Direct question	
Question and response		c_{i1}	b_{i2}	$c_{i1} - b_{i2}$	b_{i2}	$c_{i1} - b_{i2}$	c_{i1}	b_{i2}	$c_{i1} - b_{i2}$
Knowledge	Corr	0.890	0.849	-0.042 (0.004)	0.795	-0.095 (0.004)	0.877	0.819	-0.058 (0.003)
	Incorr	0.779	0.566	-0.212 (0.011)	0.539	-0.239 (0.010)	0.783	0.570	-0.213 (0.006)
	Diff	-0.112	-0.283	-0.171	-0.256	-0.144	-0.094	-0.249	-0.155
		(0.006)	(0.011)	(0.011)	(0.011)	(0.011)	(0.003)	(0.007)	(0.007)
Bacteria	Corr	0.891	0.840	-0.052 (0.010)	0.787	-0.104 (0.011)	0.903	0.842	-0.061 (0.007)
	Incorr	0.802	0.610	-0.193 (0.020)	0.558	-0.243 (0.021)	0.820	0.577	-0.241 (0.015)
	Diff	-0.090	-0.229	-0.141	-0.229	-0.139	-0.083	-0.265	-0.180
		(0.010)	(0.022)	(0.022)	(0.022)	(0.023)	(0.007)	(0.017)	(0.017)
Child's sex	Corr	0.872	0.861	-0.012 (0.007)	0.795	-0.076 (0.009)	0.883	0.843	-0.039 (0.005)
	Incorr	0.748	0.614	-0.135 (0.022)	0.594	-0.155 (0.022)	0.749	0.504	-0.245 (0.018)
	Diff	-0.124	-0.247	-0.124	-0.202	-0.079	-0.133	-0.339	-0.206
		(0.014)	(0.023)	(0.023)	(0.022)	(0.024)	(0.009)	(0.018)	(0.019)
Continental drift	Corr	0.870	0.863	-0.007 (0.007)	0.796	-0.074 (0.009)	0.899	0.871	-0.028 (0.005)
	Incorr	0.762	0.431	-0.331 (0.029)	0.406	-0.357 (0.027)	0.792	0.452	-0.343 (0.024)
	Diff	-0.108	-0.432	-0.324	-0.390	-0.283	-0.108	-0.419	-0.315
		(0.015)	(0.029)	(0.030)	(0.027)	(0.029)	(0.010)	(0.023)	(0.024)
Earth/Sun	Corr	0.947	0.897	-0.050 (0.008)	0.838	-0.109 (0.009)			
	Incorr	0.851	0.564	-0.281 (0.033)	0.573	-0.274 (0.031)			
	Diff	-0.096	-0.333	-0.231	-0.265	-0.165			
		(0.012)	(0.032)	(0.033)	(0.029)	(0.032)			
Electron/ atom	Corr	0.866	0.768	-0.098 (0.011)	0.746	-0.119 (0.011)	0.858	0.784	-0.073 (0.007)
	Incorr	0.747	0.549	-0.198 (0.019)	0.523	-0.223 (0.018)	0.792	0.544	-0.248 (0.016)
	Diff	-0.120	-0.219	-0.100	-0.224	-0.104	-0.065	-0.240	-0.174
		(0.012)	(0.022)	(0.022)	(0.021)	(0.022)	(0.008)	(0.017)	(0.017)
Lasers	Corr						0.826	0.713	-0.113 (0.008)
	Incorr						0.766	0.655	-0.111 (0.010)
	Diff						-0.060	-0.058	0.002
							(0.008)	(0.014)	(0.013)
Radio- activity	Corr						0.875	0.824	-0.051 (0.006)
	Incorr						0.781	0.552	-0.229 (0.016)
	Diff						-0.094	-0.272	-0.178
							(0.008)	(0.017)	(0.017)
Controversies	Corr	0.843	0.785	-0.057 (0.005)	0.734	-0.108 (0.006)	0.861	0.797	-0.064 (0.003)
	Incorr	0.776	0.636	-0.140 (0.010)	0.589	-0.186 (0.010)	0.807	0.558	-0.249 (0.007)
	Diff	-0.066	-0.150	-0.083	-0.145	-0.078	-0.054	-0.239	-0.185
		(0.006)	(0.011)	(0.011)	(0.011)	(0.011)	(0.004)	(0.008)	(0.008)
Autism/ vaccines	Corr	0.876	0.849	-0.027 (0.008)	0.789	-0.086 (0.009)	0.899	0.846	-0.052 (0.005)
	Incorr	0.753	0.572	-0.181 (0.025)	0.506	-0.247 (0.027)	0.828	0.607	-0.221 (0.018)
	Diff	-0.123	-0.277	-0.153	-0.283	-0.161	-0.071	-0.239	-0.169
		(0.014)	(0.027)	(0.026)	(0.029)	(0.029)	(0.008)	(0.020)	(0.019)
CDC budget	Corr	0.744	0.533	-0.211 (0.019)	0.548	-0.197 (0.019)			
	Incorr	0.785	0.717	-0.069 (0.012)	0.654	-0.132 (0.013)			
	Diff	0.041	0.184	0.142	0.106	0.065			
		(0.011)	(0.021)	(0.022)	(0.021)	(0.023)			
Climate change	Corr	0.882	0.873	-0.008 (0.006)	0.812	-0.069 (0.008)	0.900	0.872	-0.028 (0.004)
	Incorr	0.805	0.444	-0.364 (0.036)	0.418	-0.388 (0.035)	0.819	0.426	-0.395 (0.027)
	Diff	-0.077	-0.429	-0.356	-0.394	-0.319	-0.081	-0.446	-0.367
		(0.016)	(0.033)	(0.037)	(0.031)	(0.036)	(0.010)	(0.028)	(0.028)
COVID deaths	Corr	0.805	0.733	-0.073 (0.011)	0.667	-0.138 (0.013)	0.811	0.700	-0.111 (0.008)
	Incorr	0.759	0.587	-0.172 (0.019)	0.575	-0.183 (0.019)	0.779	0.505	-0.275 (0.014)
	Diff	-0.046	-0.146	-0.099	-0.092	-0.045	-0.032	-0.195	-0.164
		(0.011)	(0.022)	(0.022)	(0.022)	(0.023)	(0.008)	(0.016)	(0.016)
GM food	Corr						0.833	0.775	-0.058 (0.006)
	Incorr						0.817	0.604	-0.212 (0.015)
	Diff						-0.015	-0.170	-0.154
							(0.007)	(0.016)	(0.016)
Hydroxy- chloroquine	Corr						0.846	0.763	-0.084 (0.007)
	Incorr						0.810	0.581	-0.229 (0.015)
	Diff						-0.036	-0.182	-0.145
							(0.007)	(0.017)	(0.016)

Note: Table displays average certainty levels by question and wave 1 response (correct, incorrect, or the difference between them). "Diff."' rows are the difference between correct and incorrect answers. "$c_{i1} - b_{i2}$" columns are regression to the mean. Standard errors for all difference in means estimates appear in parentheses. Among estimates without standard errors reported, the median standard error is 0.005 and the maximum is 0.015.

Regression to the mean

Table 3 examines regression to the mean. First examining the category-by-category results, the knowledge and controversy questions follow the same general patterns observed in the first two studies. The overall averages for knowledge questions appear in the first row. In Study 3a, belief in correct answers to knowledge questions regresses from 0.890 to 0.849 (first and second columns), a difference of 0.042 (third column). Incorrect answers regress by five times this amount, from 0.779 to 0.566 (diff. = 0.212). Similar results are seen using the costly measure (fourth and fifth columns) and in Study 3b (sixth through eighth columns). The controversy questions see only a slightly stronger commitment to incorrect answers. In Study 3a, belief in correct answers to regressed from 0.843 to 0.785 (diff = 0.057). Belief in incorrect answers regressed by more than twice this amount, from 0.776 to 0.636 (diff = 0.140). Similar results again obtain using the costly measure. Results are also similar in Study 3b, with the exception that incorrect answers exhibit somewhat greater regression to the mean (from 0.807 to 0.588, diff = 0.249).

The question-by-question results are broadly consistent with the category-level results, but once again reveal differences between question. Among the controversy questions, responses to the climate change question are the least stable. In both surveys, incorrect answers regress to below the 0.5 threshold that would indicate a blind guess. This means that the average respondent who says at one point in time that the planet is not getting warmer actually believes it is more likely than not that the planet *is* getting warmer. This same pattern is observed among respondents who deny the existence of continental drift (fourth row). Regression to the mean among correct answers is almost nil for these items, while regression among incorrect answers exceeds 0.3 in every case.

The typical incorrect answer to most of the other controversy items falls between a miseducated guess and a blind guess. Incorrect answers to ANES items on autism and vaccines, GM food, and hydroxychloroquine all regress from 0.75 or higher in the first wave to 0.60 or lower in the second wave, resulting in regressions to the mean of at least 0.18 in every case. Among respondents who answer the same questions correctly, the largest regression to the mean is 0.08 and the second-largest is 0.06. The COVID-19 deaths question performs similarly, but with larger regression to the mean among respondents who answer correctly. Relative to respondents who correctly say that vaccines do not cause autism or that the planet is getting warmer, those who correctly say that the official COVID-19 death toll is understated do not believe this as firmly.

The CDC budget item stands out among the others. It is the only item considered in this paper for which false beliefs are more stable than true beliefs. This is largely traceable to the unusual instability of its correct responses. At 0.533, the average belief among respondents who at first appears to "know" that the Trump administration did not secure CDC budget cuts prior to the pandemic is even lower than that observed among *in*correct answers to knowledge questions. The drop-off from the initial measure of belief to the follow-up survey, 0.744 to 0.533, is comparable to what is observed among those who answer that electrons are larger than atoms (0.747 to 0.549). In some cases, assuming that those who answer correctly really know the facts is just as misleading as assuming that those who answer incorrectly hold firm misperceptions.

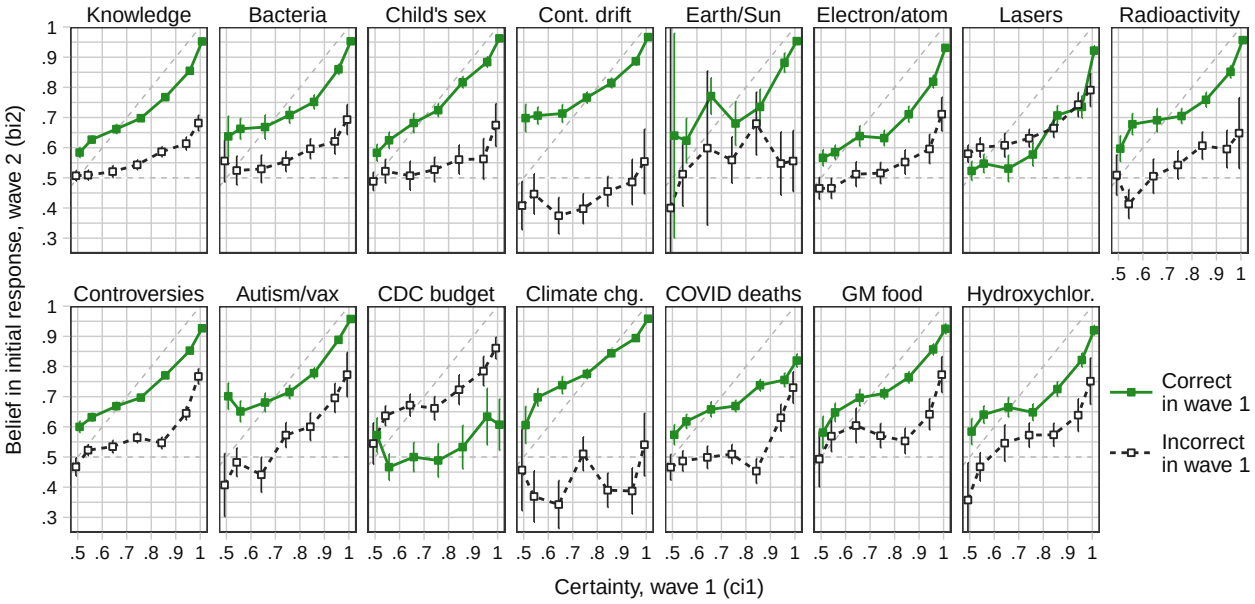
Results by certainty level

How stable are claims to be certain of incorrect answers? Figure 6 examines belief stability conditional on wave 1 certainty and whether the wave 1 best guess was correct or incorrect. The leftmost panels pool all questions in the knowledge and controversy categories, while the other panels plot question-by-question results. As the results for Studies 3a and 3b were quite similar, the figure polls the two studies for brevity; separate figures appear in Appendix C.

In broad strokes, the results are similar to the patterns observed in Study 2. Respondents who report 100 percent certainty about wave 1 incorrect answers to the controversy items assign an average probability of 0.767 to their initial response in wave 2. This regression of 0.233 is about four times what is observed among 100 percent certain correct answers to the same questions (to 0.927) and about six times the regression seen among those who claim 100 percent certainty about correct answers to knowledge questions (to 0.952). Whereas the average respondent who claims to be certain of false claims is making a miseducated guess, those who claim to be certain of true claims come much closer to revealing a firm, confidently held belief.

Among the individual questions, instability is once again most pronounced among the climate change and continental drift items. On average, even those who claim to be 100 percent certain that the planet is not getting warmer do not have any genuine confidence in this claim. Though most observers of politics would suspect that many Americans are misinformed about climate change, the question selected for the ANES misinformation battery does not appear to succeed in identifying such respondents.

Figure 6: Temporal stability of beliefs by certainty level and question, Study 3.



Note: The x-axis displays c_{i1}. The y-axis displays $\mathbb{E}[B_{i2}|C_{i1} = c_{i1}]$. Thin error bars represent 95 percent confidence intervals. Thick error bars represent 84 percent confidence intervals to aid comparisons between estimates (see note to Figure 2). Figure pools across Studies 3a and 3b; for separate figures, see Appendix C.

The remaining ANES misinformation items are comparable in their measurement properties to other falsehoods that are not subject to any contestation or false claims in the public sphere. None of the autism-vaccine, GM food, or hydroxychloroquine items exceeds the levels of conditional response stability observed among those who incorrectly answer that lasers work by focusing sound waves or that electrons are larger than atoms. Coming in only slightly behind are claims to be certain that the mother's gene determines a child's biological sex and that all radioactivity is man-made. Any sense in which the ANES items capture misperceptions must also be able to accommodate the existence of misperceptions with respect to falsehoods that are not politically charged or related to misinformation.

The results for the original items each differ from the ANES items in two respects. First, claims to be certain of the correct answer to these items are less stable than the others. Correct answers to the COVID deaths item are comparable to incorrect answers to the laser-sound wave item, while correct answers to the CDC budget item are comparable to incorrect answers to the item about a child's biological sex. This means that although the ANES items are no better at measuring misperceptions, they are better at measuring knowledge. More generally, it indicates

that even those who would appear to "know" some facts are making educated guesses.

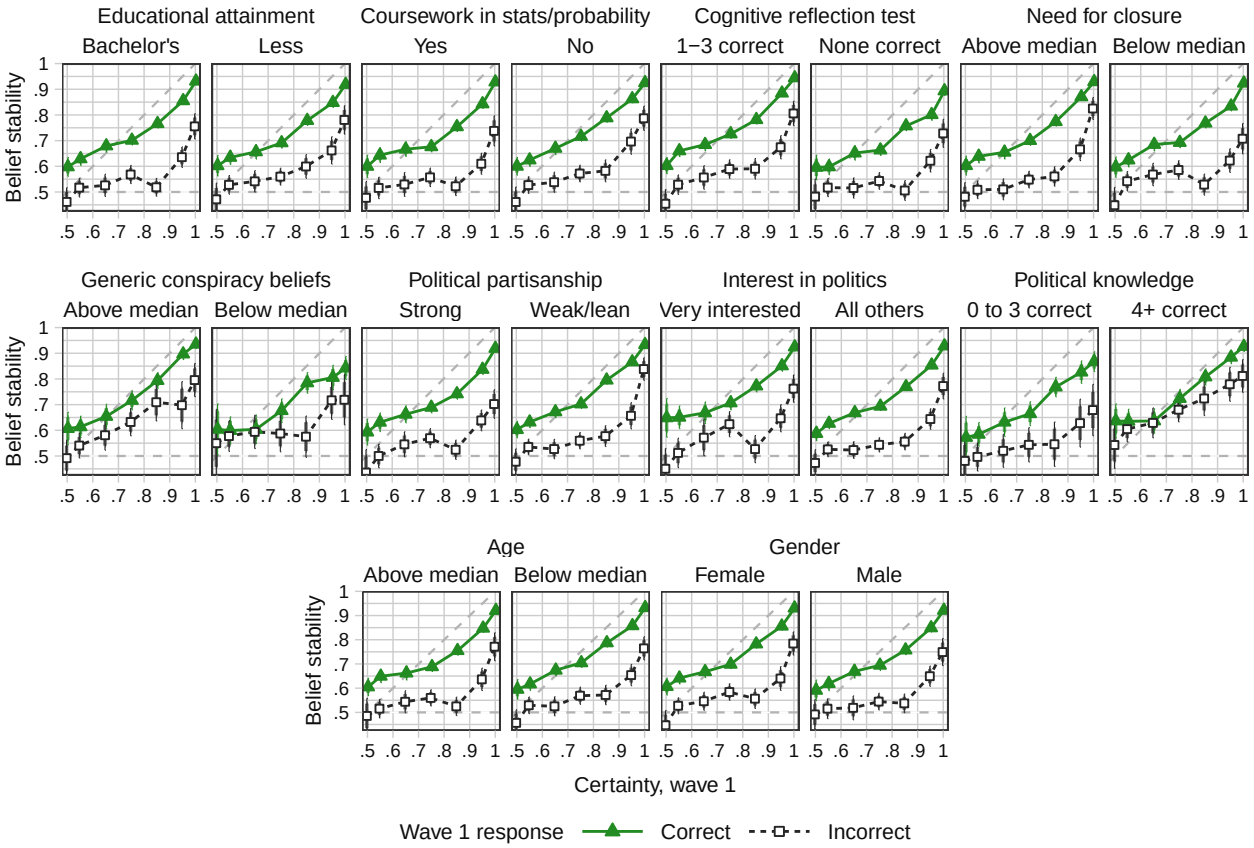
Second, incorrect answers to the CDC budget item achieve a higher level of belief stability than any other false claim examined in this paper. Respondents who claimed to be 100 percent certain that Trump had cut the budget regressed to 0.86 in the follow-up survey. This represents twice the regression observed on correct answers to controversy questions and three times that observed on correct answers to the knowledge items. Nonetheless, given that the precise dividing lines between categories are ultimately arbitrary, a reasonable reader could consider 0.86 to be a sufficient to view these responses as representative of firmly held misperceptions. Like the unusually poor performance of the climate change item, the CDC item's relatively strong performance suggests that some questions measure misperceptions more successfully than others.

Individual-level differences

The instability of incorrect answers has been explained in terms of an individual-level process: the process of retrieving a sample of considerations from memory and integrating it into an on-the-spot judgment often leads respondents to state higher levels of certainty than their underlying beliefs truly support. Broadly speaking, the alternative is that some individual-level factor confounds the conditional relationship between response type and belief stability. To examine this possibility, Studies 3a and 3b measured several characteristics known to predict endorsement of falsehoods in surveys or exposure to falsehoods in the real world: educational attainment (Flynn 2016; Meirick 2013), cognitive reflection (Pennycook and Rand 2019; Pennycook et al. 2020), need for closure (Lunz Trujillo et al. 2020; Marchlewska et al. 2018), generic conspiracy beliefs (Brotherton et al. 2013; Study 3a only), strength of partisanship, interest in politics (Flynn 2016; Tesler 2018), political knowledge (Nyhan 2020; Study 3a only), and age (Guess et al. 2019). Given the probabilistic nature of the scales, the surveys also asked whether respondents had ever taken a course in probability or statistics. In light of existing evidence that women are more likely to use DK options (Mondak and Anderson 2004) and are more aware of their ignorance (Graham 2020) than men, the results are also split by gender.

Figure 7 splits the results according to these characteristics. The figure pools across both surveys and includes only the misinformation items; separate estimates for Studies 3a and 3b appear in Appendix C. Each pair of panels covers one characteristic, with all variables split at their median.

Figure 7: Temporal stability of beliefs by certainty level and respondent characteristics, Study 3.



Note: Each panel displays the same information from the "Misinformation" panel of Figure 6. The pairs of panels are split by demographic group. The main header is the variable name and the subheaders are the categories. Figure pools across Studies 3a and 3b; for separate figures, see Appendix C.

In every case, the pattern of differential response stability between correct and incorrect answers holds for both subgroups. In most cases, there is little difference between the two subgroups. Where differences exist, some are consistent with extant theoretical expectations. In particular, measured misperceptions are modestly more stable among those with greater need for closure and with more political knowedge. By contrast, despite findings that strong partisans and less cognitively reflective people are more likely to endorse congenial false claims and engage with real-world misinformation, measured misperceptions are less stable among these respondents.

A related possibility is that differences in stability between correct and incorrect answers are traceable to some unmeasured, individual-level factor. To examine this possibility, the appendices to Studies 2 and 3 conduct a within-subject test. Specifically, the linear model $B_{i2} = \alpha + \beta_1 G_{i1} +$

$\beta_2 C_{i1} + \beta_3 G_{i1} \times C_{i1} + \epsilon_i$ is estimated with and without respondent fixed effects.[23] β_3 is proportional to the difference in between-wave correlations between correct and incorrect answers.[24] The fixed effects account for all between-subject differences in means. The coefficient estimate for β_3 is statistically significant in all cases and grows slightly larger with the inclusion of fixed effects. This suggests that the differential stability of correct and incorrect answers is not an artifact of between-subject differences in some unmeasured factor that predicts the tendency to answer incorrectly.

Study 4: Frame-of-Reference Training

Although the results so far are largely pessimistic with respect to researchers' ability to measure deeply held misperceptions, the frequent heterogeneity between questions offers hope. A framework that can identify relatively successful questions should also be able to identify relatively sucessful measurement practices. Accordingly, this section evaluates a new approach to boosting the reliability of measured misperceptions. It merges the principles of frame-of-reference training (FOR; Bernardin and Buckley 1981; Woehr 1994; Roch et al. 2012), a best practice for improving inter-rater agreement in workplace performance evaluations, with theories of the survey response (Zaller 1992; Tourangeau et al. 2000). The training aims to reduce measurement error *ex ante* by calibrating respondents to a common understanding of how to integrate their considerations into a belief statement using the scale. By contrast, existing strategies for improving measures of probabilistic beliefs aim to correct for measurement error *ex post* using adjustments derived from other survey questions (e.g., King et al. 2004; Hopkins and King 2010; Guay 2021).

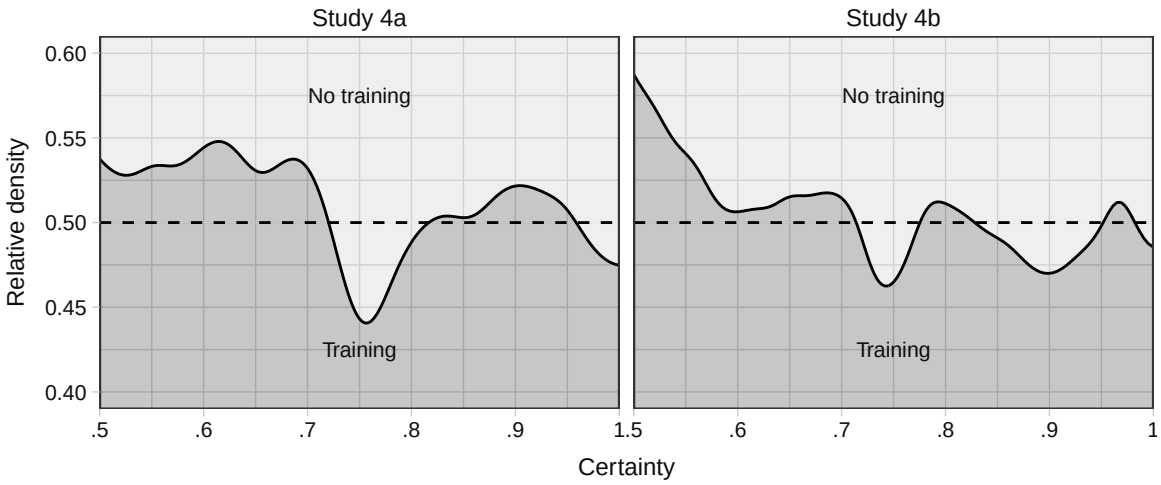
Intervention

Using simple random assignment (Gerber and Green 2012), half of respondents to the science surveys were assigned to complete the training. The other half saw only a brief set of instructions. The training consists of four vignettes about hypothetical respondents answering a question about the price of gas. Each describes the considerations that the hypothetical respondent called to mind as they made an on-the-spot inference about the question. After each vignette, respondents are asked which of three certainty levels would be most appropriate. A message then appears indicating

[23] Respondent fixed effects would be denoted by changing α to α_i.

[24] Recall two elementary facts about regression: β_1 in a bivariate regression is proportional to the correlation coefficient, and β_3 in an interacted model like that specified here is equal to the difference in slopes for two separately-fit bivariate regressions (Wooldridge 2012).

Figure 8: Effect of calibration training on certainty distribution.



Note: Figure plots the relative proportion of respondents choosing each certainty level according to whether the respondent received the calibration training. The x-axis is the respondent's certainty level in wave 1. The y-axis is a kernel-smoothed estimate of $f_A(C = c)/(f_A(C = c) + f_B(C = c))$, where f is the probability density function, A and B represent the treatment groups, and C is certainty.

which certainty level was most appropriate and why. The first task proceeds as follows:

> [Name] gets the question,
>
> > *Nationwide, is the average price of gas above or below $2.00?*
>
> [Name] has no idea. [S/he] lives in the city, doesn't own a car, and rarely walks by a gas station. [S/he] picks "above $2.00," but [s/he] may as well have flipped a coin.
>
> How sure is [name] that the answer is "above $2.00"?
>
> > [50 percent, 75 percent, 99 percent]
>
> The best choice is **50 percent** sure. Because [Name] has no idea, [s/he] is split 50/50 between the two options, just like a coin has a 50 percent chance to land on heads and a 50 percent chance to land on tails.

The other three vignettes concern someone who is 99 percent sure (not 60 or 80) because they had recently learned that specific fact, someone who is 70 percent sure (not 95) because they knew about their area but not the rest of the country, and someone who is 55 percent sure (not 50 or 85) because they had long since given up driving but knew that prices are higher than they used to be. The median respondent completed the training in 78 seconds in Study 3a and 63 seconds in Study 3b; the means were 91 and 81 seconds.

Results

The training had no statistically or substantively significant effect on average belief in the correct answer (p_i) or the proportion of correct best guesses (g_i), and reduced average certainty in wave 1 (c_i) by about 0.01 on the 0.5 to 1 scale (Appendix D). The primary effect of the training was a re-sorting of certainty levels. To illustrate this, Figure 8 plots the relative proportion of certainty levels by treatment condition. Respondents not assigned to the training made greater use of the middle and highest scale points. Respondents who were trained made greater use of the low, medium-low, and medium-high scale points. Appendix D presents further analysis of the distributional effects.

The training improved the certainty scale's ability to capture firmly held misperceptions. To summarize these effects, Table D.3 presents the between-wave correlation in measures of false beliefs for the two randomly assigned subgroups, as well as the difference between them. Pooling across all questions in both studies, the training increased the between-wave correlation by about 40 percent, from 0.143 to 0.201 (difference = 0.058, block bootstrapped s.e. = 0.023). In both absolute and percentage terms, evidence for the training's efficacy was stronger for the controversy items. The training boosted between-wave stability by about 45 percent, from 0.164 to 0.238 (difference = 0.074, s.e. = 0.033).

Training exercises are more useful if they work for everyone. For example, if understanding the training required high levels of cognitive reflection, it could fail to improve the measurement of misperceptions among those who are most susceptible to fake news. To examine the training's potential to induce improvement across the board, Appendix D splits the results according to all of the same respondent characteristics examined in Study 3. The estimates suggest that the training's benefits were generally not conditional on respondent characteristics. All of the point estimates of the subgroup effect are positive. To the extent that heterogeneity exists, there is weak evidence to suggest that the training may confer greater benefits for individuals who would be more prone at baseline to have difficulty using certainty scales. The only statistically significant difference between subgroups is by education level: respondents without a bachelor's degree benefit more than respondents with one. The treatment effect estimates are also larger for individuals who fare worse on the cognitive reflection test and who report no coursework in probability or statistics.

Though the results demonstrate that FOR training can improve the stability of measured mis-

Table 4: Effect of FOR training on temporal stability in measures of misperceptions.

		Study 4a		Study 4b		Pooled	
Category	Treatment	Correl.	p-value	Correl.	p-value	Correl.	p-value
All questions	No training	0.158		0.139		0.143	
		(0.030)		(0.020)		(0.016)	
	Training	0.231		0.190		0.201	
		(0.031)		(0.021)		(0.017)	
	Difference	0.074	0.043	0.051	0.039	0.058	0.011
		(0.043)		(0.029)		(0.023)	
Controversies	No training	0.175		0.180		0.164	
		(0.041)		(0.029)		(0.023)	
	Training	0.287		0.229		0.238	
		(0.041)		(0.031)		(0.024)	
	Difference	0.113	0.024	0.049	0.137	0.074	0.010
		(0.058)		(0.043)		(0.033)	
Knowledge	No training	0.150		0.115		0.126	
		(0.042)		(0.025)		(0.021)	
	Training	0.177		0.168		0.171	
		(0.040)		(0.026)		(0.021)	
	Difference	0.028	0.321	0.052	0.071	0.044	0.073
		(0.059)		(0.036)		(0.030)	

perceptions, it did not fully solve the measurement problem. Instead, the takeaways are threefold. First, FOR training is promising. Future work should examine refinements that may yield larger improvements, such as different subject matter, vignette content, and hypothetical certainty levels. Second, the success of an intervention that was randomly assigned at the individual level lends credence to individual-level explanations for the instability of measured misperceptions. Third, the tight alignment between the design of the FOR training and theories of the survey response lends support to the particular individual-level explanation given here: that instability in measured misperceptions emerges from the error-prone process of integrating considerations into an on-the-spot judgment.

Implications

Kuklinski et al. (2000) conclude their seminal article on misinformed beliefs by posing six questions for future research. Subsequent scholarship took up the five questions about causes and consequences, but skipped past the foundational first question: what kinds of factual beliefs do people have? Examining a wide range of topics, this paper showed that survey measures of misperceptions generally capture a mix of blind guesses and "miseducated" guesses based on misleading

heuristics. Even those survey respondents who claim to be 100 percent certain of incorrect answers hold weaker beliefs than is suggested by the evocative language that frequently appears in analysis that identifies misperceptions using looser standards.

The most immediate implication is the need for greater attention to the properties of measured misperceptions. Even as credibility revolutions have improved the causal identification and replicability of social scientific findings, too many of the measures that enter such analysis are rooted in survey measurement practices that have not changed much since the early days of polling. Consequently, survey-based research on misperceptions and misinformed beliefs is often characterized by a large conceptual-empirical gap, regardless of whether the quantities of interest are descriptive, causally identified, or somewhere in between.

The disconnect between definitions and measurement calls for a reconsideration of existing evidence on the correlates, correction, and consequences of misperceptions and misinformed beliefs. Political partisanship may be the most-studied correlate of incorrect answers to survey questions. This paper's finding that survey questions measure knowledge far more reliably than misperceptions suggests that absent evidence to the contrary, belief differences between Democrats and Republicans are best-interpreted as differential knowledge of convenient and inconvenient truths. This is consistent with several patterns that misinformation-focused accounts have trouble explaining. Greater public attention to an issue predicts higher, not lower, knowledge of politically inconvenient truths among both Democrats and Republicans (Jerit and Barabas 2012, Table 1). Democrats' and Republicans' beliefs about politically controversial facts are highly correlated across survey items (Graham 2020, Figures 6 and 7). Led by the expectation that misinformed beliefs are a key driver of partisan belief differences (Lee et al. 2017, 1), Lee et al. (2021) were surprised to find that relative to the general public, political elites' beliefs about politically controversial facts are more accurate and no more polarized. In a divided era, observers of politics can still benefit from the traditional posture that between-group differences in responses to knowledge questions primarily reflect differences in knowledge and ignorance.

Another line of research seeks to correct misperceptions. Embracing the error-prone nature of measured misperceptions could inform tests of a well-grounded theoretical prediction that, to the author's knowledge, has never been confirmed empirically: that misperceptions that are more deeply held should be more resistant to correction. The few studies that are equipped to test this

prediction have either found no heterogeneity (Thorson 2015; Guay 2021) or have not reported such a test (Kuklinski et al. 2000). The results presented here suggest that existing attempts to confirm that highly certain misperceptions are especially dug-in—including, one suspects, some that have yet to emerge from the file drawer—did not measure much genuine variation in the depth of misperceptions to begin with. Understanding which falsehoods people believe to begin with could help researchers begin to understand why some correction treatments work better than others (Weeks 2018).

The same applies to a popular strategy for learning about the consequences of misperceptions and misinformed beliefs. In this paradigm, researchers randomly assign the provision of correct factual information, observe that beliefs become more accurate, and draw conclusions about the downstream consequences (Ahler and Broockman 2018; Hopkins et al. 2019; Nyhan et al. 2020). Such experiments draw conclusions about the consequences of misperceptions by a reverse logic: misperceptions appear higher in the control group than in the treatment group, so the treatment effects can be interpreted as the effect of reducing misperceptions. Incongruencies between measures and definitions of misperceptions strain this logic. A safer interpretation, maintained through most of Gilens' (2001) seminal article, is that such designs inform rather than correct, providing insight into the consequences of reducing public ignorance (also see Grigorieff et al. 2020; Lawrence and Sides 2014).

The findings here suggest three best practices for reearch in this area. First, research should offer hard empirical evidence of construct validity. In this paper, a certainty level of roughly 90 percent or more was required to identify respondents with even a modest degree of genuine belief in their answer, while even 100 percent certainty was not sufficient to identify misperceptions held with a high degree of confidence. Absent evidence to the contrary, researchers and research consumers should default to a posture that treats incorrect answers as a mix of blind and miseducated guesses.

Second, theoretical expectations about which subgroups hold the deepest misperceptions should not be substituted for hard evidence. This paper examined a range of respondent characteristics that past research has found to predict incorrect answers or real-world engagement with misinformation. In every case, measured misperceptions were less stable than measured knowledge. Although finding the expected correlations is accepted as validity evidence in many survey contexts, the fundamental problem in this case is that under prevailing measurement practices, acceptance

of congenial falsehoods is observationally equivalent to ignorance of inconvenient truths. Validity evidence for measures of misperceptions must be able to distinguish these posisbilities.

Third, valididty evidence should be question-specific. Though no question examined here measured firmly held misperceptions, some were more successful than others. Knowledge questions frequently succeeded at measuring firm, confidently held beliefs in the truth. By treating measurement properties as specific to individual questions rather than as general traits of predetermined sets of misinformation items, researchers can gain a data-driven sense of which misperceptions are the most deeply held—and if desired, can focus their surveys on these questions. For example, the science surveys conducted for paper followed the ANES in seeking to tap climate change misperceptions by asking about global temperature change over time. It is possible that some other misperception, e.g. that humans did not contribute to the change in global temperatures, is more firmly held by a wider swath of the population. The ultimate potential of this paper is not the doubt it casts on the existing body of survey research on misperceptions, but the opportunity it presents to build more trustworthy evidence in the future.

None of this is to say that misperceptions and misinformed beliefs are not problems when they exist. Instead, prevailing practices dull researchers' sense of the problem, detecting the same pattern around every corner and allowing virtually any intervention aimed at enhancing belief accuracy to be framed in relation to misperceptions and misinformation. This suggests that treating misperceptions and misinformed beliefs as a serious problem requires serious attention to measurement. Accordingly demonstrated a widespread measurement problem and showed that the same analytic framework that documented it can be deployed in service of selecting better questions and measurement techniques. By assuming the burden of proof for its interpretations of survey responses, future research can build a stronger evidentiary basis regarding the prevalence, predictors, correction, and consequences of misperceptions and misinformed beliefs.

References

Achen, Christopher H. 1975. "Mass Political Attitudes and the Survey Response." *The American Political Science Review* 69(4):1218–1231.

Ahler, Douglas J. and David E. Broockman. 2018. "The Delegate Paradox: Why Polarized Politicians Can Represent Citizens Best." *The Journal of Politics* .

Allen, Franklin. 1987. "Discovering Personal Probabilities When Utility Functions Are Unknown." *Management Science* 33(4):542–544.

Ansolabehere, Stephen, Jonathan Rodden and James M. Snyder. 2008. "The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting." *American Political Science Review* 102(2):215–232.

Aronow, Peter M. and Benjamin T. Miller. 2019. *Foundations of Agnostic Statistics*. Cambridge University Press.

Bassili, John N. 1996. "Meta-Judgmental versus Operative Indexes of Psychological Attributes: The Case of Measures of Attitude Strength." *Journal of Personality and Social Psychology* 71(4):637–653.

Berinsky, Adam J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2):241–262.

Berinsky, Adam J. 2018. "Telling the Truth about Believing the Lies? Evidence for the Limited Prevalence of Expressive Survey Responding." *The Journal of Politics* 80(1):211–224.

Bernardin, H. John and M. Ronald Buckley. 1981. "Strategies in Rater Training." *The Academy of Management Review* 6(2):205–12.

Brotherton, Robert, Christopher C. French and Alan D. Pickering. 2013. "Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale." *Frontiers in Psychology* 4(279):1–15.

Bullock, John G, Alan S Gerber, Seth J Hill and Gregory A Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10:1–60.

Bullock, John G and Gabriel Lenz. 2019. "Partisan Bias in Surveys." *Annual Review of Political Science* 22:325–342.

Clifford, Scott and Jennifer Jerit. 2016. "Cheating on Political Knowledge Questions in Online Surveys: An Assessment of the Problem and Solutions." *Public Opinion Quarterly* 80(4):858–887.

Converse, Philip E. 1964. The nature of belief systems in mass publics. In *Ideology and Discontent*, edited by D. Apter. New York: Free Press.

Converse, Philip E. 1970. Attitudes and Non-Attitudes: Continuation of a Dialogue. In *The Quantitative Analysis of Social Problems*, edited by Edward R. Tufte. Reading, MA: Addison-Wesley pp. 168–189.

Converse, Philip E. 2000. "Assessing the Capacity of Mass Electorates." *Annual Review of Political Science* 3:331–53.

Diedenhofen, Birk and Jochen Musch. 2017. "PageFocus: Using paradata to detect and prevent cheating on online achievement tests." *Behavior Research Methods* 49(4):1444–1459.

Ducharme, Wesley M. and Michael L. Donnell. 1973. "Intrasubject comparison of four response modes for 'subjective probability' assessment." *Organizational Behavior and Human Performance* 10(1):108–17.

Erikson, Robert S. 1979. "The SRC Panel Data and Mass Political Attitudes." *British Journal of Political Science* 9(1):89–114.

Flynn, D. J., Brendan Nyhan and Jason Reifler. 2017. "The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics." *Political Psychology* 38(682758):127–150.

Flynn, D.J. 2016. "The Scope and Correlates of Political Misperceptions in the Mass Public." (Unpublished manuscript).

Foley, Richard. 1992. "The Epistemology of Belief and the Epistemology of Degrees of Belief." *American Philosophical Quarterly* 29(2):111–124.

Gerber, Alan S. and Donald Green. 2012. *Field Experiments.* New York: W.W. Norton.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95(2):379–396.

Gilens, Martin. 2012. Citizen Competence and Democratic Governance. In *New Directions in Public Opinion*, edited by Adam J. Berinsky. 1 edited by New York: Routledge pp. 52–76.

Graham, Matthew H. 2020. "Self-Awareness of Political Knowledge." *Political Behavior* 42:305–326.

Graham, Matthew H. 2021. "'We Don't Know' Means 'They're Not Sure'." *Public Opinion Quarterly* (forthcoming).

Grigorieff, Alexis, Christopher Roth and Diego Ubfal. 2020. "Does Information Change Attitudes Towards Immigrants? Representative Evidence from Survey Experiments." *Demography* 57:1117–1143.

Guay, Brian. 2021. "Misinformed or Uninformed? The Prevalence and Consequences of Certainty in Political Misperceptions." *Unpublished Manuscript* (Duke University).

Guess, Andrew, Jonathan Nagler and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science Advances* 5(1):eaau4586.

Hellwig, Timothy and Dani M. Marinova. 2015. "More Misinformed than Myopic: Economic Retrospections and the Voter's Time Horizon." *Political Behavior* 37(4):865–887.

Hill, Seth J. 2017. "Learning Together Slowly: Bayesian Learning about Political Facts." *The Journal of Politics* 79(4):1403–1418.

Hochschild, Jennifer L. and Katherine Levine Einstein. 2015. *Do Facts Matter?* Norman: University of Oklahoma Press.

Holt, Charles A. and Angela M. Smith. 2016. "Belief Elicitation with a Synchronized Lottery Choice Menu That Is Invariant to Risk Attitudes." *American Economic Journal: Microeconomics* 8(1):110–39.

Hopkins, Daniel J. and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Comparability." *Public Opinion Quarterly* 74(2):201–22.

Hopkins, Daniel J., John Sides and Jack Citrin. 2019. "The muted consequences of correct information about immigration." *Journal of Politics* 81(1):315–320.

Jerit, Jennifer and Jason Barabas. 2012. "Partisan Perceptual Bias and the Information Environment." *The Journal of Politics* 74(3):672–684.

Jerit, Jennifer and Yangzi Zhao. 2020. "Political Misinformation." *Annual Review of Political Science* 23:77–96.

Julious, Steven A. 2004. "Using confidence intervals around individual means to assess statistical significance between two means." *Pharmaceutical Statistics* 3:217–22.

King, Gary, Christopher J L Murray, World Health and Joshua a Salomon. 2004. "Enhancing the of Measurement in Survey Research." *American Political Science Review* 98(1):191–207.

Krosnick, Jon A. 1988. "Attitude Importance and Attitude Change." *Journal of Experimental Social PSychology* 24:240–255.

Kuklinski, James H, Paul J Quirk, David W Schwieder and Robert F Rich. 1998. "'Just the Facts, Ma'am': Political Facts and Public Opinion Source." *The Annals of the American Academy of Political and Social Science* 560(November):143–154.

Kuklinski, James H, Paul J Quirk, Jennifer Jerit, David Schwieder and Robert F Rich. 2000. "Misinformation and the currency of democratic citizenship." *The Journal of Politics* 62(3):790–816.

Kull, Steven. 2011. Preserving American Public Support for Foreign Aid. In *From Aid to Global Development Cooperation.* Washington, DC: Brookings Institution pp. 57–60.

Lawrence, Eric D. and John Sides. 2014. "The consequences of political innumeracy." *Research & Politics* pp. 1–8.

Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, Jonathan L. Zittrain, David M. J. Lazer and Jonathan L. Zittrain. 2018. "The science of fake news." *Science* 359(6380):1094–1096.

Lee, Nathan, D.J. Flynn and Brendan Nyhan. 2017. "Political Misperceptions among Public Officials and the General Public." *Evidence in Governance and Politics Pre-Registration Document* .

Lee, Nathan, D.J. Flynn, Brendan Nyhan and Jason Reifler. 2021. "More Accurate Yet More Polarized? Comparing the Factual Beliefs of Government Officials and the Public." *British Journal of Political Science* (Forthcoming).

Lee, Seonghui and Akitaka Matsuo. 2018. "Decomposing political knowledge: What is confidence in knowledge and why it matters." *Electoral Studies* 51(1):1–13.

Leeper, Thomas J. 2014. "Are Important Attitudes More Stable? No, Not Really." (unpublished manuscript):1–49.

Lewandowsky, Stephan, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz and John Cook. 2012. "Misinformation and Its Correction." *Psychological Science in the Public Interest* 13(3):106–131.

Li, Jianing and Michael W Wagner. 2020. "The Value of Not Knowing: Partisan Cue-Taking and Belief Updating of the Uninformed, the Ambiguous, and the Misinformed." *Journal of Communication* 70(5):646–669.

Lunz Trujillo, Kristin, Matthew Motta, Timothy Callaghan and Steven Sylvester. 2020. "Correcting Misperceptions about the MMR Vaccine: Using Psychological Risk Factors to Inform Targeted Communication Strategies." *Political Research Quarterly* .

Luskin, Robert C., Guarav Sood and Joshua Blank. 2018. "Misinformation about Misinformation: Of Headlines and Survey Design." (Unpublished manuscript).

Luskin, Robert C and John G Bullock. 2011. "'Don't Know' Means 'Don't Know': DK Responses and the Public's Level of Political Knowledge." *Journal of Politics* .

Marchlewska, Marta, Aleksandra Cichocka and Małgorzata Kossowska. 2018. "Addicted to answers: Need for cognitive closure and the endorsement of conspiracy beliefs." *European Journal of Social Psychology* 48(2):109–117.

Marietta, Morgan and David C. Barker. 2019. *One Nation, Two Realities: Dueling Facts in American Democracy*. Oxford University Press.

Meirick, Patrick C. 2013. "Motivated misperception? Party, education, partisan news, and belief in "death panels"." *Journalism and Mass Communication Quarterly* 90(1):39–57.

Mondak, Jeffrey J. and Mary R. Anderson. 2004. "The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge." *The Journal of Politics* 66(2):492–512.

Nyhan, Brendan. 2020. "Facts and Myths about Misperceptions." *Journal of Economic Perspectives* 34(October):220–236.

Nyhan, Brendan, Ethan Porter, Jason Reifler and Thomas J. Wood. 2020. "Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability." *Political Behavior* 42(3):939–960.

Pasek, Josh, Gaurav Sood and Jon A. Krosnick. 2015. "Misinformed About the Affordable Care Act? Leveraging Certainty to Assess the Prevalence of Misperceptions." *Journal of Communication* 65(4):660–673.

Pennycook, Gordon and David G. Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188(September 2017):39–50.

Pennycook, Gordon, Jonathon McPhetres, Bence Bago and David G. Rand. 2020. "Predictors of attitudes and misperceptions about COVID-19 in Canada, the U.K., and the U.S.A." *arXiv* .

Permut, Stephanie, Matthew Fisher and Daniel M. Oppenheimer. 2019. "TaskMaster: A Tool for Determining When Subjects Are on Task." *Advances in Methods and Practices in Psychological Science* 2(2):188–196.

Peterson, Erik and Shanto Iyengar. 2020. "Partisan Gaps in Political Information and Information-Seeking Behavior: Motivated Reasoning or Cheerleading." *The American Journal of Political Science* (forthcoming).

Prior, Markus, Gaurav Sood and Kabir Khanna. 2015. "You Cannot be Serious. The Impact of Accuracy Incentives on Partisan Bias in Reports of Economic Perceptions." *Quarterly Journal of Political Science* 10(July):489–518.

Prislin, Radmila. 1996. "Attitude stability and attitude strength: One is enough to make it stable." *European Journal of Social Psychology* 26(3):447–477.

Roch, Sylvia G., David J. Woehr, Vipanchi Mishra and Urszula Kieszczynska. 2012. "Rater training revisited: An updated meta-analytic review of frame-of-reference training." *Journal of Occupational and Organizational Psychology* 85:370–95.

Schuman, Howard and Stanley Presser. 1981. *Questions and Answers in Attitude Surveys.* San Diego: SAGE Publications.

Scotto, Thomas J., Jason Reifler, David Hudson and Jennifer VanHeerde-Hudson. 2017. "We Spend How Much? Misperceptions, Innumeracy, and Support for the Foreign Aid in the United States and Great Britain." *Journal of Experimental Political Science* 4(2):119–128.

Strack, Fritz and Leonard L Martin. 1987. Thinking, Judging, and Communicationg: A Process Account of Context Effects in Attitudes Surveys. In *Social information processing and survey methodology*, edited by Hans J. Hippler, Norbert Schwarz and Seymour Sudman. New York: Springer-Verlag pp. 123–148.

Sturgis, Patrick, Nick Allum and Patten Smith. 2008. "An experiment on the measurement of political knowledge in surveys." *Public Opinion Quarterly* 72(1):90–102.

Sutton, Robbie M. and Karen M. Douglas. 2020. "Agreeing to disagree: Reports of the popularity of Covid-19 conspiracy theories are greatly exaggerated." *Psychological Medicine* (July):27–30.

Tesler, Michael. 2018. "Elite Domination of Public Doubts about Climate Change (Not Evolution)." *Political Communication* 35(2):306–26.

Thorson, Emily A. 2015. "Identifying and Correcting Policy Misperceptions." *Unpublished Manuscript, George Washington University* .

Tourangeau, Robert, Lance J. Rips and Kenneth Rasinski. 2000. *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Trautmann, Stefan T. and Gijs van de Kuilen. 2015. "Belief Elicitation: A Horse Race Among Truth Serums." *The Economic Journal* 125(589):2116–2135.

Weeks, Brian E. 2018. Media and political misperceptions. In Misinformation and Mass Audiences, ed. BG Southwell. In *Misinformation and Mass Audiences*, edited by Brian G. Southwell, Emily A. Thorson and Laura Sheble. Austin: University of Texas Press pp. 140–56.

Wiley, David E and James A Wiley. 1970. "The Estimation of Measurement Error in Panel." *American Sociological Review* 35(1):112–117.

Williamson, Vanessa. 2019. "Public Ignorance or Elitist Jargon? Reconsidering Americans' Overestimates of Government Waste and Foreign Aid." *American Politics Research* 47(1):152–173.

Woehr, David J. 1994. "Understanding frame-of-reference training: The impact of training on the recall of performance information." *Journal of Applied Psychology* 79(4):525–34.

Wooldridge, Jeffrey M. 2012. *Introductory Econometrics: A Modern Approach*. CENGAGE.

Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

Appendix to

Measuring Misperceptions?

Matthew H. Graham

October 26, 2021

Contents

A Appendix to Study 1

A.1 Survey information

Platform: Lucid.

Date: August 2018 (wave 1), September 2018 (wave 2).

Number of subjects: 2,916 (wave 1), 1,749 (wave 2).

Compensation: $1 (wave 1), $2 (wave 2). Standard prices set by vendor.

Consent: Prior to data collection, all subjects agreed to participate in a research study using an IRB-approved consent form. There was no deception and no debrief.

Additional screeners: None.

Anti-cheating measures: Pledge.

Full text of question analyzed:

> On which of the following does the U.S. federal government currently spend the least?
>
> [Social Security, Medicare, Foreign aid, National defense]

Format of certainty scale: The certainty scale appeared immediately after each respondent chose their answer. Using simple random assignment, respondents were assigned to use the scale from Graham (2020) or Pasek et al. (2015).

Respondents who used the Graham scale were asked, "How certain are you that your answer is correct?" [Not at all certain, Not too certain, Somewhat certain, Very certain, Absolutely certain]

Respondents who used the Pasek scale were asked, "How sure are you about that?" [Not at all sure, Slightly sure, Moderately sure, Very sure, Extremely sure]

A.2 Table of estimates

The table below displays the estimates plotted in main text Figure 2, as well as the referenced separate results for the Graham and Pasek et al. scales.

Table A.1: Estimates plotted in Figure 2

Scale	Response	Certainty	Estimate	SE	CI	N
Graham (2020)	Correct	1	0.333	0.098	(0.130, 0.537)	24
		2	0.549	0.070	(0.408, 0.690)	51
		3	0.658	0.055	(0.549, 0.767)	76
		4	0.740	0.063	(0.614, 0.866)	50
		5	0.667	0.092	(0.477, 0.857)	27
	Incorrect	1	0.410	0.080	(0.249, 0.572)	39
		2	0.512	0.039	(0.435, 0.590)	162
		3	0.473	0.031	(0.413, 0.534)	264
		4	0.471	0.050	(0.372, 0.569)	102
		5	0.419	0.076	(0.265, 0.572)	43
Pasek et al. (2015)	Correct	1	0.588	0.070	(0.448, 0.728)	51
		2	0.583	0.072	(0.439, 0.728)	48
		3	0.667	0.046	(0.575, 0.758)	105
		4	0.750	0.083	(0.579, 0.921)	28
		5	0.895	0.050	(0.793, 0.997)	38
	Incorrect	1	0.475	0.046	(0.383, 0.566)	118
		2	0.481	0.043	(0.395, 0.567)	133
		3	0.492	0.032	(0.429, 0.555)	246
		4	0.602	0.054	(0.495, 0.710)	83
		5	0.443	0.064	(0.314, 0.571)	61
Pooled	Correct	1	0.507	0.058	(0.391, 0.622)	75
		2	0.566	0.050	(0.466, 0.665)	99
		3	0.663	0.035	(0.593, 0.733)	181
		4	0.744	0.050	(0.645, 0.843)	78
		5	0.800	0.050	(0.700, 0.900)	65
	Incorrect	1	0.459	0.040	(0.380, 0.537)	157
		2	0.498	0.029	(0.441, 0.556)	295
		3	0.482	0.022	(0.439, 0.526)	510
		4	0.530	0.037	(0.457, 0.602)	185
		5	0.433	0.049	(0.336, 0.530)	104

B Appendix to Study 2

B.1 Survey information

Study 2a

Platform: Amazon Mechanical Turk.

Date: June 2019 (wave 1), June 2020 (wave 2).

Number of subjects: 1,244 (wave 1), 549 (wave 2).

Compensation: $0.80 (wave 1), $0.50 (wave 2).

Consent: Prior to data collection, all subjects agreed to participate in a research study using an IRB-approved consent form. There was no deception and no debrief. For the second wave, respondents were invited to complete a short follow-up survey, then completed the original consent form again.

Additional screeners: None.

Anti-cheating measures: Pledge, catch question.

Full text of questions analyzed:

1. Is the following statement true or false?

   Before becoming president, Donald Trump was tape recorded saying that he kisses women and grabs them between the legs without their consent.

   [True, False]

2. Is the following statement true or false?

   While she was Secretary of State, Hillary Clinton used a private email server to send and receive classified information.

   [True, False]

3. Robert Mueller was in charge of the special counsel investigation into possible Russian interference in the 2016 election.

   Is the following statement true or false?

   Robert Mueller's final report stated that there is "undeniable proof" that President Trump personally conspired with Russian agents to influence the 2016 election.

   [True, False]

4. Is the following statement true or false?

   Barack Obama has never released his birth certificate.

   [True, False]

Format of certainty scale: The certainty scale appeared immediately after each respondent chose their answer. Respondents were asked, "How many chances in 100 does your answer have to be correct?" and presented with a quasi-continuous 50 to 100 scale.

Study 2b

Platform: Amazon Mechanical Turk.

Date: March 2020 (wave 1), August 2020 (wave 2).

Number of subjects: 939 (wave 1), 420 (wave 2).

Compensation: $1 (wave 1), $0.50 (wave 2).

Consent: Prior to data collection, all subjects agreed to participate in a research study using an IRB-approved consent form. There was no deception and no debrief. For the second wave, respondents were invited to complete a short follow-up survey, then completed the original consent form again.

Additional screeners: Captcha.

Anti-cheating measures: Pledge, cheating detection script.

Full text of questions analyzed:

1. The Bureau of Labor Statistics estimates the *unemployment rate*, which is the percentage of workers who are looking for a job but cannot find one.

   Over the past year, did the unemployment rate increase or decrease?

   [Decreased, Increased]

2. The rate of *inflation* measures how quickly prices are rising. Since World War II, the average inflation rate has been about 4 percent.

   Over the past year, has inflation been higher or lower than the historical average?

   [Above average, Below average]

3. The size of the U.S. economy is usually measured using gross domestic product (GDP). *Economic growth* is the annual rate of change in GDP.

   Over the past year, what was the rate of economic growth in the United States?

   [Less than 4%, 4% or more]

4. Most years, the U.S. national government spends more than it collects in taxes. In these years, the government has an annual *budget deficit*.

   Compared with the 2017 fiscal year, was 2019's budget deficit higher or lower?

   [Higher, Lower]

5. Is the following statement true or false?

   Before becoming president, Donald Trump was tape recorded saying that he kisses women and grabs them between the legs without their consent.

   [True, False]

6. Robert Mueller was in charge of the special counsel investigation into possible Russian interference in the 2016 election.

   Is this statement true or false? *Robert Mueller's report stated that President Trump personally conspired with Russia to influence the 2016 election.*

   [True, False]

7. Article II of the U.S. Constitution describes the president's powers.

   Is this statement true or false? *President Trump has said that Article II gives him the power to do whatever he wants.*

   [True, False]

8. In 2014, former President Barack Obama issued an order that would stop most deportations of unauthorized immigrants who have U.S. citizen children.

   Is this statement true or false? *About a year earlier, Obama said that he would be ignoring the law if he issued such an order.*

   [True, False]

9. What job or political office does John Roberts hold?

   [Secretary of Defense, Chief Justice of the Supreme Court]

10. What job or political office does Jerome Powell hold?

    [Treasury Secretary, Chairman of the Federal Reserve]

11. Which party currently has the most members in the U.S. House of Representatives?

    [Democrats, Republicans]

Format of certainty scale: The certainty scale appeared immediately after each respondent chose their answer. Respondents were randomly assigned to be asked, "How likely is your answer to be correct?" or "How sure are you about that?" and provided a quasi-continuous 50 to 100 scale with labels at 50 and 100. No systematic differences between the scales were found.

Description of economic questions (omitted from main text):

- **Budget deficit.** Respondents read a short definition of the federal budget deficit. Respondents were then asked, "Compared with the 2017 fiscal year, was 2019's budget deficit higher or lower?" The incorrect answer, "lower," was interpreted as congenial to Republicans, whose party leaders had claimed that their signature legislation, the Tax Cuts and Jobs Act of 2017 (TCJA), would reduce the deficit.

- **GDP growth.** Respondents read a short definition that linked the change in gross domestic product (GDP) to economic growth. Respondents were then asked, "Over the past year, what was the rate of economic growth in the United States?," with the options to say "Below 4%" or "4% or more." The incorrect answer, "4% or more," was interpreted as congenial to Republicans, whose party leaders prominently claimed that the TCJA would raise growth above this level.

- **Unemployment.** Respondents read a short definition of the U-3 unemployment rate as it is defined by the Bureau of Labor Statistics. Respondents were then asked, "Over the past year, did the unemployment rate increase or decrease?" The incorrect answer, "increase," was treated as congenial to Democrats, who had a partisan incentive to downplay the booming economy under a Republican president.

- **Inflation.** Respondents read a short definition of inflation and were told its historical average since 1945. Respondents were then asked, "Over the past year, has inflation been higher or lower than the historical average?" By the same logic applied to unemployment, the incorrect answer, "higher," was treated as congenial for Democrats.

B.2 Tables of plotted estimates

Table B.1: Estimates plotted in Figure 3

Question	Answer	Certainty	Estimate	SE	CI	N
Clinton email	Correct	0.5	0.808	0.043	(0.721, 0.895)	34
		[0.51,0.59]	0.767	0.053	(0.659, 0.876)	29
		[0.6,0.69]	0.772	0.046	(0.675, 0.869)	20
		[0.7,0.79]	0.828	0.036	(0.754, 0.902)	37
		[0.8,0.89]	0.872	0.019	(0.834, 0.911)	70
		[0.9,0.99]	0.938	0.012	(0.914, 0.963)	82
		1	0.960	0.007	(0.946, 0.975)	194
	Incorrect	0.5	0.558	0.171	(0.084, 1.032)	5
		[0.51,0.59]	0.280	0.180	(-2.007, 2.567)	2
		[0.6,0.69]	0.230	0.198	(-0.400, 0.860)	4
		[0.7,0.79]	0.211	0.104	(-0.035, 0.458)	8
		[0.8,0.89]	0.330	0.184	(-0.256, 0.916)	4
		[0.9,0.99]	0.495	0.405	(-4.651, 5.641)	2
		1	0.350	0.325	(-1.050, 1.750)	3
Obama birth certificate	Correct	0.5	0.614	0.058	(0.496, 0.731)	37
		[0.51,0.59]	0.535	0.063	(0.407, 0.663)	34
		[0.6,0.69]	0.636	0.066	(0.501, 0.772)	33
		[0.7,0.79]	0.752	0.066	(0.618, 0.887)	28
		[0.8,0.89]	0.758	0.061	(0.632, 0.883)	31
		[0.9,0.99]	0.879	0.035	(0.808, 0.950)	53
		1	0.886	0.025	(0.837, 0.935)	129
	Incorrect	0.5	0.592	0.063	(0.461, 0.723)	26
		[0.51,0.59]	0.578	0.074	(0.425, 0.730)	24
		[0.6,0.69]	0.312	0.142	(-0.043, 0.668)	8
		[0.7,0.79]	0.504	0.099	(0.294, 0.715)	19
		[0.8,0.89]	0.618	0.079	(0.454, 0.782)	21
		[0.9,0.99]	0.564	0.099	(0.352, 0.776)	17
		1	0.461	0.069	(0.320, 0.601)	34
Obama DAPA reversal	Correct	0.5	0.489	0.035	(0.417, 0.561)	23
		[0.51,0.59]	0.485	0.049	(0.385, 0.585)	31
		[0.6,0.69]	0.571	0.045	(0.480, 0.662)	32
		[0.7,0.79]	0.438	0.052	(0.332, 0.544)	28
		[0.8,0.89]	0.632	0.051	(0.528, 0.737)	26
		[0.9,0.99]	0.649	0.075	(0.487, 0.810)	15
		1	0.692	0.107	(0.460, 0.925)	13
	Incorrect	0.5	0.480	0.028	(0.424, 0.537)	54
		[0.51,0.59]	0.514	0.029	(0.456, 0.572)	53
		[0.6,0.69]	0.526	0.040	(0.444, 0.608)	33
		[0.7,0.79]	0.507	0.040	(0.427, 0.587)	48
		[0.8,0.89]	0.614	0.055	(0.503, 0.726)	32
		[0.9,0.99]	0.550	0.115	(0.293, 0.807)	11
		1	0.799	0.136	(0.466, 1.131)	7
Political knowledge	Correct	0.5	0.563	0.023	(0.517, 0.609)	54
		[0.51,0.59]	0.537	0.022	(0.493, 0.581)	64
		[0.6,0.69]	0.664	0.044	(0.574, 0.754)	33
		[0.7,0.79]	0.674	0.044	(0.584, 0.764)	52
		[0.8,0.89]	0.748	0.036	(0.676, 0.819)	79
		[0.9,0.99]	0.868	0.025	(0.818, 0.918)	97
		1	0.976	0.005	(0.966, 0.986)	521
	Incorrect	0.5	0.452	0.019	(0.414, 0.490)	85
		[0.51,0.59]	0.476	0.037	(0.401, 0.552)	53
		[0.6,0.69]	0.487	0.049	(0.386, 0.588)	33

Table B.1: Estimates plotted in Figure 3 (continued)

Question	Answer	Certainty	Estimate	SE	CI	N
		[0.7,0.79]	0.350	0.052	(0.243, 0.457)	34
		[0.8,0.89]	0.515	0.048	(0.417, 0.612)	42
		[0.9,0.99]	0.647	0.078	(0.479, 0.814)	21
		1	0.766	0.121	(0.481, 1.052)	8
Trump Article II claim	Correct	0.5	0.587	0.058	(0.463, 0.710)	18
		[0.51,0.59]	0.595	0.044	(0.504, 0.686)	30
		[0.6,0.69]	0.678	0.041	(0.593, 0.762)	30
		[0.7,0.79]	0.631	0.039	(0.551, 0.710)	43
		[0.8,0.89]	0.793	0.033	(0.727, 0.860)	45
		[0.9,0.99]	0.780	0.047	(0.684, 0.877)	33
		1	0.912	0.034	(0.843, 0.981)	48
	Incorrect	0.5	0.518	0.053	(0.406, 0.629)	19
		[0.51,0.59]	0.410	0.058	(0.290, 0.530)	23
		[0.6,0.69]	0.428	0.057	(0.307, 0.549)	19
		[0.7,0.79]	0.486	0.074	(0.332, 0.640)	20
		[0.8,0.89]	0.543	0.064	(0.412, 0.675)	26
		[0.9,0.99]	0.574	0.099	(0.365, 0.784)	18
		1	0.759	0.061	(0.634, 0.883)	29
Trump Russia collusion	Correct	0.5	0.512	0.037	(0.438, 0.586)	56
		[0.51,0.59]	0.562	0.045	(0.471, 0.653)	47
		[0.6,0.69]	0.481	0.051	(0.377, 0.585)	36
		[0.7,0.79]	0.616	0.034	(0.548, 0.685)	79
		[0.8,0.89]	0.676	0.038	(0.600, 0.752)	71
		[0.9,0.99]	0.706	0.036	(0.634, 0.777)	101
		1	0.863	0.018	(0.828, 0.898)	249
	Incorrect	0.5	0.644	0.045	(0.553, 0.735)	35
		[0.51,0.59]	0.647	0.048	(0.550, 0.743)	40
		[0.6,0.69]	0.614	0.051	(0.510, 0.718)	35
		[0.7,0.79]	0.610	0.049	(0.511, 0.708)	38
		[0.8,0.89]	0.669	0.046	(0.577, 0.761)	49
		[0.9,0.99]	0.689	0.056	(0.575, 0.803)	36
		1	0.380	0.084	(0.205, 0.555)	24
Trump said 'grab them'	Correct	0.5	0.794	0.036	(0.721, 0.868)	34
		[0.51,0.59]	0.764	0.044	(0.674, 0.854)	35
		[0.6,0.69]	0.729	0.059	(0.609, 0.850)	27
		[0.7,0.79]	0.831	0.033	(0.764, 0.898)	50
		[0.8,0.89]	0.823	0.032	(0.760, 0.886)	79
		[0.9,0.99]	0.884	0.021	(0.842, 0.926)	123
		1	0.947	0.009	(0.929, 0.965)	367
	Incorrect	0.5	0.413	0.059	(0.291, 0.535)	29
		[0.51,0.59]	0.522	0.058	(0.402, 0.642)	26
		[0.6,0.69]	0.418	0.086	(0.234, 0.601)	16
		[0.7,0.79]	0.442	0.076	(0.286, 0.599)	25
		[0.8,0.89]	0.544	0.076	(0.387, 0.701)	28
		[0.9,0.99]	0.604	0.071	(0.457, 0.750)	28
		1	0.623	0.072	(0.478, 0.769)	35

Table B.2: Estimates plotted in Figure 4

Survey	Valence	Response	Certainty	Estimate	SE	CI	N
March 2020 - August 2020	Correct ans. is congenial	Correct	0.5	0.566	0.064	(0.429, 0.702)	35
			[0.51,0.59]	0.635	0.041	(0.552, 0.717)	72
			[0.6,0.69]	0.621	0.040	(0.539, 0.703)	91
			[0.7,0.79]	0.714	0.033	(0.649, 0.779)	136
			[0.8,0.89]	0.742	0.032	(0.678, 0.806)	190
			[0.9,0.99]	0.830	0.030	(0.771, 0.890)	179
			1	0.958	0.011	(0.937, 0.979)	417
		Incorrect	0.5	0.391	0.043	(0.301, 0.480)	52
			[0.51,0.59]	0.410	0.049	(0.310, 0.510)	85
			[0.6,0.69]	0.413	0.056	(0.296, 0.529)	65
			[0.7,0.79]	0.463	0.057	(0.347, 0.579)	89
			[0.8,0.89]	0.435	0.066	(0.298, 0.572)	72
			[0.9,0.99]	0.483	0.076	(0.325, 0.641)	65
			1	0.318	0.098	(0.104, 0.532)	26
	Political knowledge	Correct	0.5	0.563	0.023	(0.517, 0.609)	147
			[0.51,0.59]	0.537	0.022	(0.493, 0.581)	141
			[0.6,0.69]	0.664	0.044	(0.574, 0.754)	104
			[0.7,0.79]	0.674	0.044	(0.584, 0.764)	137
			[0.8,0.89]	0.748	0.036	(0.676, 0.819)	177
			[0.9,0.99]	0.868	0.025	(0.818, 0.918)	222
			1	0.976	0.005	(0.966, 0.986)	886
		Incorrect	0.5	0.452	0.019	(0.414, 0.490)	163
			[0.51,0.59]	0.476	0.037	(0.401, 0.552)	126
			[0.6,0.69]	0.487	0.049	(0.386, 0.588)	92
			[0.7,0.79]	0.350	0.052	(0.243, 0.457)	106
			[0.8,0.89]	0.515	0.048	(0.417, 0.612)	132
			[0.9,0.99]	0.647	0.078	(0.479, 0.814)	72
			1	0.766	0.121	(0.481, 1.052)	45
	Incorrect ans. is congenial	Correct	0.5	0.499	0.043	(0.408, 0.590)	47
			[0.51,0.59]	0.521	0.046	(0.427, 0.615)	84
			[0.6,0.69]	0.566	0.048	(0.468, 0.665)	84
			[0.7,0.79]	0.575	0.045	(0.484, 0.666)	137
			[0.8,0.89]	0.699	0.038	(0.622, 0.776)	138
			[0.9,0.99]	0.751	0.042	(0.667, 0.835)	131
			1	0.864	0.031	(0.801, 0.926)	170
		Incorrect	0.5	0.590	0.029	(0.531, 0.649)	76
			[0.51,0.59]	0.547	0.035	(0.475, 0.618)	125
			[0.6,0.69]	0.548	0.036	(0.475, 0.621)	108
			[0.7,0.79]	0.582	0.036	(0.510, 0.655)	152
			[0.8,0.89]	0.630	0.038	(0.553, 0.707)	141
			[0.9,0.99]	0.718	0.051	(0.614, 0.822)	83
			1	0.760	0.047	(0.666, 0.855)	96
June 2019-2020	Correct ans. is congenial	Correct	0.5	0.716	0.041	(0.633, 0.800)	180
			[0.51,0.59]	0.753	0.039	(0.673, 0.832)	145
			[0.6,0.69]	0.784	0.039	(0.706, 0.863)	93
			[0.7,0.79]	0.768	0.036	(0.695, 0.840)	143
			[0.8,0.89]	0.845	0.024	(0.797, 0.894)	210
			[0.9,0.99]	0.908	0.018	(0.872, 0.944)	334
			1	0.945	0.009	(0.928, 0.963)	1126
		Incorrect	0.5	0.463	0.073	(0.311, 0.616)	73
			[0.51,0.59]	0.503	0.074	(0.346, 0.659)	68
			[0.6,0.69]	0.427	0.139	(0.101, 0.753)	37
			[0.7,0.79]	0.343	0.074	(0.188, 0.498)	56
			[0.8,0.89]	0.552	0.077	(0.386, 0.717)	33

Table B.2: Estimates plotted in Figure 4 (continued)

Survey	Valence	Response	Certainty	Estimate	SE	CI	N
			[0.9,0.99]	0.423	0.122	(0.136, 0.710)	33
			1	0.313	0.076	(0.156, 0.469)	75
	Incorrect ans. is congenial	Correct	0.5	0.647	0.035	(0.577, 0.717)	202
			[0.51,0.59]	0.574	0.041	(0.491, 0.656)	179
			[0.6,0.69]	0.594	0.047	(0.500, 0.689)	141
			[0.7,0.79]	0.692	0.039	(0.613, 0.771)	181
			[0.8,0.89]	0.790	0.031	(0.728, 0.852)	194
			[0.9,0.99]	0.787	0.029	(0.730, 0.844)	274
			1	0.850	0.017	(0.816, 0.883)	709
		Incorrect	0.5	0.625	0.043	(0.539, 0.711)	153
			[0.51,0.59]	0.636	0.045	(0.544, 0.727)	130
			[0.6,0.69]	0.508	0.070	(0.364, 0.652)	75
			[0.7,0.79]	0.510	0.074	(0.357, 0.663)	86
			[0.8,0.89]	0.671	0.064	(0.540, 0.803)	82
			[0.9,0.99]	0.675	0.072	(0.526, 0.825)	68
			1	0.553	0.067	(0.416, 0.690)	132

Table B.3: Estimates plotted in Figure 5

Survey	Category	Valence	Response	Certainty	Estimate	SE	CI	N
March 2020	Political knowledge	Not applicable	Correct	0.5	0.589	0.019	(0.551, 0.626)	147
				[0.51,0.59]	0.645	0.020	(0.606, 0.684)	141
				[0.6,0.69]	0.644	0.023	(0.599, 0.689)	104
				[0.7,0.79]	0.692	0.024	(0.645, 0.739)	137
				[0.8,0.89]	0.753	0.021	(0.710, 0.795)	177
				[0.9,0.99]	0.817	0.018	(0.782, 0.853)	222
				1	0.943	0.006	(0.930, 0.955)	886
			Incorrect	0.5	0.509	0.017	(0.475, 0.543)	163
				[0.51,0.59]	0.537	0.022	(0.493, 0.582)	126
				[0.6,0.69]	0.554	0.026	(0.502, 0.607)	92
				[0.7,0.79]	0.593	0.026	(0.541, 0.644)	106
				[0.8,0.89]	0.587	0.025	(0.538, 0.636)	132
				[0.9,0.99]	0.604	0.041	(0.523, 0.686)	72
				1	0.797	0.034	(0.727, 0.866)	45
	Controversies	Correct ans. is congenial	Correct	0.5	0.568	0.044	(0.477, 0.658)	35
				[0.51,0.59]	0.623	0.025	(0.573, 0.673)	72
				[0.6,0.69]	0.626	0.028	(0.569, 0.682)	91
				[0.7,0.79]	0.702	0.020	(0.663, 0.742)	136
				[0.8,0.89]	0.662	0.022	(0.619, 0.705)	190
				[0.9,0.99]	0.752	0.020	(0.712, 0.793)	179
				1	0.909	0.009	(0.891, 0.927)	417
			Incorrect	0.5	0.541	0.027	(0.487, 0.595)	52
				[0.51,0.59]	0.577	0.028	(0.522, 0.632)	85
				[0.6,0.69]	0.616	0.032	(0.551, 0.680)	65
				[0.7,0.79]	0.606	0.034	(0.538, 0.674)	89
				[0.8,0.89]	0.631	0.032	(0.568, 0.694)	72
				[0.9,0.99]	0.739	0.035	(0.668, 0.809)	65
				1	0.665	0.082	(0.495, 0.835)	26
		Incorrect ans. is congenial	Correct	0.5	0.569	0.024	(0.520, 0.617)	47
				[0.51,0.59]	0.625	0.021	(0.584, 0.666)	84

Table B.3: Estimates plotted in Figure 5 (continued)

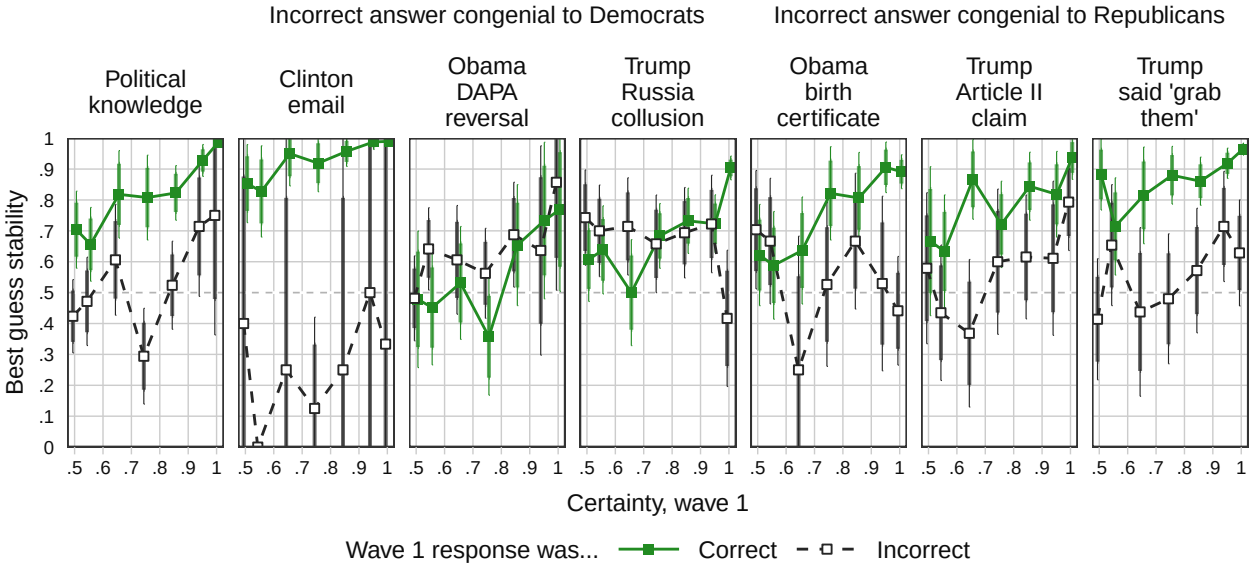
Survey	Category	Valence	Response	Certainty	Estimate	SE	CI	N
				[0.6,0.69]	0.628	0.027	(0.574, 0.682)	84
				[0.7,0.79]	0.659	0.025	(0.611, 0.708)	137
				[0.8,0.89]	0.682	0.026	(0.631, 0.732)	138
				[0.9,0.99]	0.751	0.024	(0.703, 0.798)	131
				1	0.880	0.018	(0.845, 0.915)	170
			Incorrect	0.5	0.584	0.023	(0.537, 0.630)	76
				[0.51,0.59]	0.600	0.016	(0.569, 0.631)	125
				[0.6,0.69]	0.629	0.022	(0.585, 0.672)	108
				[0.7,0.79]	0.625	0.021	(0.583, 0.668)	152
				[0.8,0.89]	0.643	0.024	(0.596, 0.690)	141
				[0.9,0.99]	0.732	0.026	(0.680, 0.783)	83
				1	0.786	0.028	(0.730, 0.841)	96
	Economic	Correct ans. is congenial	Correct	0.5	0.548	0.018	(0.512, 0.585)	91
				[0.51,0.59]	0.577	0.017	(0.542, 0.612)	130
				[0.6,0.69]	0.626	0.023	(0.581, 0.671)	110
				[0.7,0.79]	0.656	0.018	(0.620, 0.692)	186
				[0.8,0.89]	0.755	0.017	(0.720, 0.789)	172
				[0.9,0.99]	0.756	0.019	(0.718, 0.793)	172
				1	0.860	0.015	(0.831, 0.889)	225
			Incorrect	0.5	0.491	0.036	(0.418, 0.565)	46
				[0.51,0.59]	0.600	0.023	(0.553, 0.646)	81
				[0.6,0.69]	0.632	0.031	(0.570, 0.694)	58
				[0.7,0.79]	0.599	0.025	(0.549, 0.649)	106
				[0.8,0.89]	0.587	0.027	(0.534, 0.639)	102
				[0.9,0.99]	0.605	0.036	(0.533, 0.676)	63
				1	0.782	0.079	(0.607, 0.956)	13
		Incorrect ans. is congenial	Correct	0.5	0.580	0.029	(0.523, 0.637)	58
				[0.51,0.59]	0.636	0.019	(0.599, 0.673)	108
				[0.6,0.69]	0.645	0.024	(0.597, 0.693)	112
				[0.7,0.79]	0.671	0.018	(0.634, 0.707)	172
				[0.8,0.89]	0.696	0.021	(0.655, 0.737)	179
				[0.9,0.99]	0.746	0.021	(0.706, 0.787)	152
				1	0.884	0.018	(0.849, 0.919)	134
			Incorrect	0.5	0.551	0.024	(0.503, 0.599)	72
				[0.51,0.59]	0.617	0.018	(0.582, 0.652)	110
				[0.6,0.69]	0.627	0.023	(0.582, 0.672)	96
				[0.7,0.79]	0.587	0.028	(0.531, 0.642)	121
				[0.8,0.89]	0.626	0.026	(0.576, 0.677)	119
				[0.9,0.99]	0.662	0.032	(0.598, 0.727)	68
				1	0.696	0.056	(0.581, 0.810)	32

B.3 Supplemental figures

This section contains the following figures:

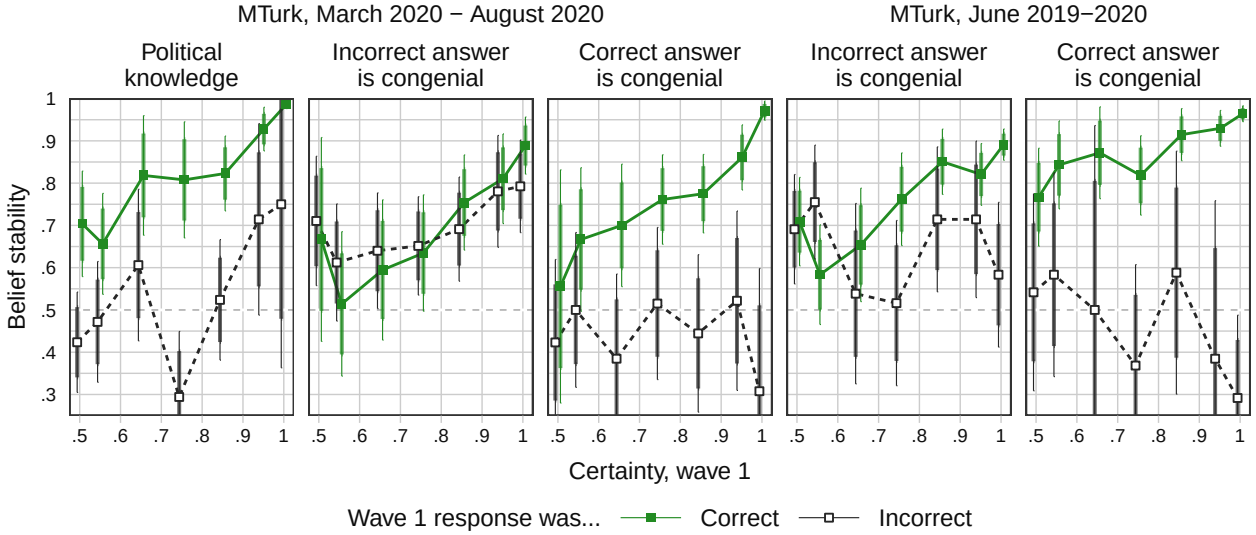
- Figures B.1 through B.3 replace belief stability with best guess stability for each main text Figures 3 through 5.
- Table B.4 presents the regression to the mean analysis using the costly measure. Estimates from the second and third row are cited.
- Figure C.6, presents the variance of wave 2 beliefs conditional on the wave 1 certainty level. This is referred to in the main text discussion of ambivalence among miseducated guessers.

Figure B.1: Temporal stability of best guesses by certainty level and question, Study 2.



Note: Figure is identical to main text Figure 3, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

Figure B.2: Temporal stability of best guesses by certainty level and partisan congeniality, Study 2.

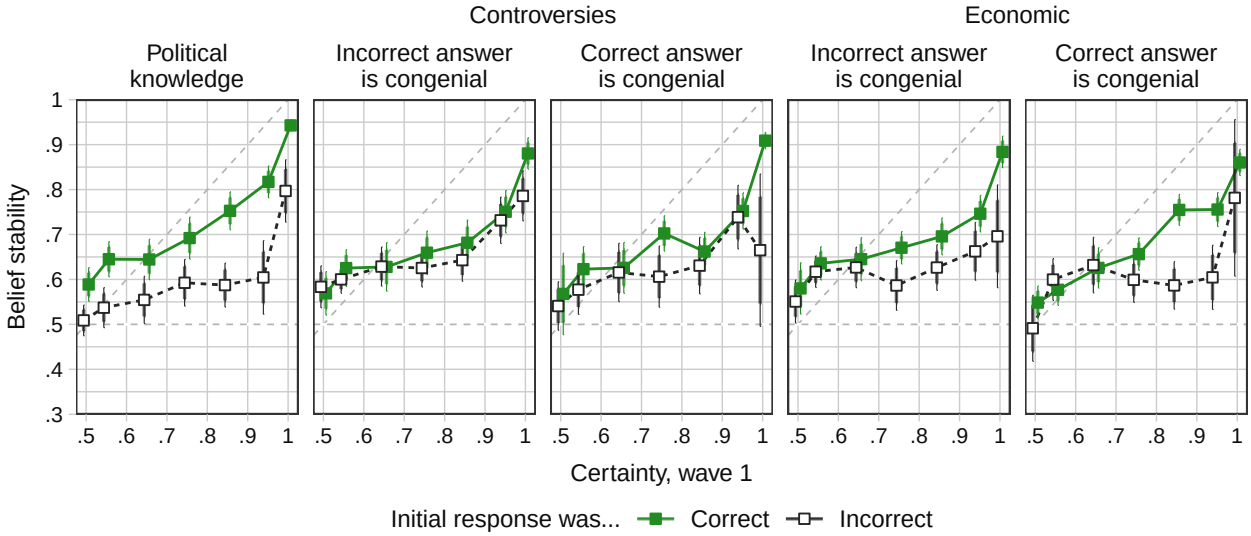


Note: Figure is identical to main text Figure 4, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

Table B.4: Regression to the mean, Study 2, costly measure.

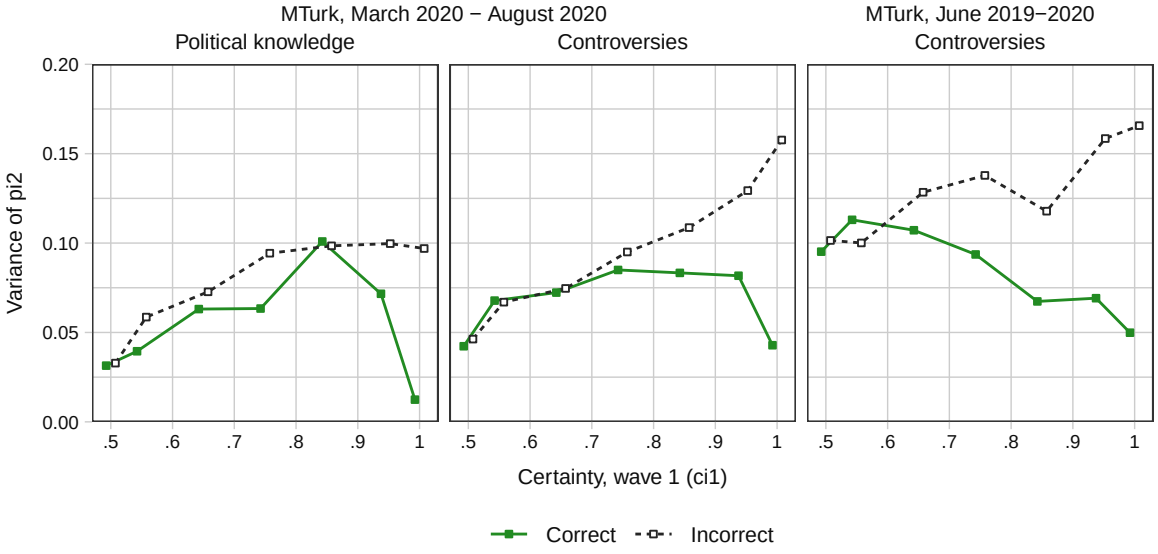
Question	Congeniality	Percent correct	Correct ($g_{i1}=1$)			Incorrect ($g_{i1}=0$)			D-in-D
			c_{i1}	b_{i2}	Diff	c_{i1}	b_{i2}	Diff	
Political awareness	All responses	0.724 (0.011)	0.872 (0.005)	0.819 (0.007)	-0.053 (0.006)	0.720 (0.008)	0.578 (0.011)	-0.143 (0.012)	-0.090 (0.013)
Contro-versies	Correct ans. is congenial	0.712 (0.012)	0.871 (0.005)	0.765 (0.008)	-0.106 (0.008)	0.735 (0.008)	0.621 (0.013)	-0.115 (0.014)	-0.009 (0.016)
	Incorrect ans. is congenial	0.503 (0.013)	0.815 (0.006)	0.714 (0.010)	-0.102 (0.010)	0.762 (0.006)	0.652 (0.009)	-0.110 (0.010)	-0.008 (0.014)
Economic	Correct ans. is congenial	0.698 (0.012)	0.803 (0.006)	0.708 (0.008)	-0.094 (0.007)	0.741 (0.008)	0.596 (0.012)	-0.145 (0.014)	-0.051 (0.015)
	Incorrect ans. is congenial	0.597 (0.012)	0.798 (0.006)	0.706 (0.009)	-0.092 (0.009)	0.735 (0.007)	0.616 (0.010)	-0.119 (0.012)	-0.027 (0.014)

Figure B.3: Comparison of direct question and costly choice measure of best guess, by certainty level and partisan congeniality, Study 3.



Note: Figure is identical to main text Figure 5, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

Figure B.4: Variance of wave 2 beliefs by wave 1 certainty level.



B.4 Within-subject analysis

This section presents the results of the within-subjects analysis referred to in the main text.

Table B.5: Within-subject regression estimates, Study 2a.

	Dependent variable: b_{i2}	
Constant	0.638^{**}	
	(0.069)	
p_{i1}	-0.127	-0.063
	(0.095)	(0.125)
g_{i1}	-0.271^{**}	-0.357^{**}
	(0.076)	(0.093)
$p_{i1} \times g_{i1}$	0.654^{**}	0.744^{**}
	(0.101)	(0.125)
Fixed effects	No	Yes
R^2	0.191	0.397
Adj. R^2	0.190	0.210
Num. obs.	1976	1976
Num. clusters	466	466

Table B.6: Within-subject regression estimates, Study 2b.

	Dependent variable: b_{i2}					
	All questions		Pol. awareness		Controversies	
Constant	0.287^{**}		0.285^{**}		0.278^{**}	
	(0.036)		(0.051)		(0.049)	
p_{i1}	0.361^{**}	0.540^{**}	0.354^{**}	0.579^{**}	0.385^{**}	0.577^{**}
	(0.052)	(0.061)	(0.075)	(0.095)	(0.074)	(0.100)
g_{i1}	-0.286^{**}	-0.201^{**}	-0.250^{**}	-0.200^{*}	-0.288^{**}	-0.188^{*}
	(0.045)	(0.052)	(0.062)	(0.080)	(0.059)	(0.080)
$p_{i1} \times g_{i1}$	0.557^{**}	0.407^{**}	0.499^{**}	0.382^{**}	0.563^{**}	0.380^{**}
	(0.060)	(0.071)	(0.084)	(0.108)	(0.082)	(0.112)
Fixed effects	No	Yes	No	Yes	No	Yes
R^2	0.322	0.439	0.235	0.467	0.414	0.613
Adj. R^2	0.321	0.353	0.233	0.280	0.413	0.472
Num. obs.	3189	3189	1617	1617	1572	1572
Num. clusters	420	420	419	419	418	418

C Appendix to Study 3

C.1 Survey information

Study 3a

Platform: Lucid.

Date: Dec. 4-9, 2020 (wave 1), Dec. 15, 2020-Jan. 14, 2021 (wave 2).

Number of subjects: 2,399 (wave 1), 1,016 (wave 2).

Compensation: $1 (wave 1), $2 (wave 2). Standard prices set by vendor.

Consent: Prior to data collection, all subjects agreed to participate in a research study using an IRB-approved consent form. There was no deception and no debrief.

Additional screeners: Captcha, attention check.

Anti-cheating measures: Pledge, cheating detection script.

Full text of questions: See below.

Study 3b

Platform: MTurk.

Date: April 28-May 3, 2021 (wave 1), May 26, 2021-Jun. 15, 2021 (wave 2).

Number of subjects: 2,602 (wave 1), 1,983 (wave 2).

Compensation: $0.50 (wave 1), $0.75 (wave 2).

Consent: Prior to data collection, all subjects agreed to participate in a research study using an IRB-approved consent form. There was no deception and no debrief.

Additional screeners: Captcha, attention check.

Anti-cheating measures: Pledge, cheating detection script.

Full text of questions analyzed:

Full text of questions

Controversies, Studies 3a and 3b:

- Which statement is most likely to be true?

  [Most scientific evidence shows that childhood vaccines cause autism., Most scientific evidence shows that childhood vaccines **do not** cause autism.]

- Which statement is most likely to be true?

  [World temperatures **have risen** on average over the past 100 years., World temperatures **have <u>not</u> risen** on average over the past 100 years.]

- As an alternative to the official COVID-19 death toll, researchers can compare the total number of deaths this year to the number that occurred at the same time last year. The resulting statistic is an estimate of *excess deaths* due to COVID-19.

  Which statement is most likely to be true?

[Excess death analysis suggests that **more** people have died than the official number., Excess death analysis suggests that **fewer** people have died than the official number.]

Controversies, Study 3a only:

- Which statement is most likely to be true?

  Prior to the COVID-19 pandemic, the Trump administration secured cuts to the CDC's funding., Prior to the COVID-19 pandemic, the Trump administration **did not** secure cuts to the CDC's funding.]

Controversies, Study 3b only:

- Which statement is most likely to be true?

  [Most scientific evidence shows genetically modified foods are safe to eat., Most scientific evidence shows genetically modified foods are **not** safe to eat.]

- Which statement is most likely to be true?

  [There is **not** clear scientific evidence that the anti-malarial drug hydroxychloroquine is a safe and effective treatment for COVID-19., There is clear scientific evidence that the anti-malarial drug hydroxychloroquine is a safe and effective treatment for COVID-19.]

Knowledge, Studies 3a and 3b:

- Which statement is most likely to be true?

  [Electrons are **larger** than atoms., Electrons are **smaller** than atoms.]

- Which statement is most likely to be true?

  [Antibiotics kill viruses as well as and bacteria., Antibiotics only kill bacteria.]

- Which statement is most likely to be true?

  [It is the **father's** gene that decides whether a baby is a boy or a girl., It is the **mother's** gene that decides whether a baby is a boy or a girl.]

- Which statement is most likely to be true?

  [The continents on which we live **have been moving** their locations for millions of years and will continue to move in the future., The continents on which we live **have not moved** their locations in millions of years and are not expected to move in the future.]

Knowledge, Study 3a only:

- Which statement is most likely to be true?

  [The Earth goes around the Sun., The Sun goes around the Earth.]

Knowledge, Study 3b only:

- Which statement is most likely to be true?

  [All radioactivity is man-made., Radioactivity can occur naturally.]

- Which statement is most likely to be true?

  [Lasers work by focusing sound waves., Lasers **do not** work by focusing sound waves.]

C.2 Tables of plotted estimates

Table C.1: Estimates plotted in Figure 6

Category	Question	Response	Certainty	Estimate	SE	CI	N
Controversies	All questions	Correct	0.5	0.600	0.014	(0.573, 0.627)	335
			[0.51,0.59]	0.632	0.009	(0.613, 0.650)	700
			[0.6,0.69]	0.669	0.009	(0.650, 0.687)	727
			[0.7,0.79]	0.697	0.007	(0.684, 0.710)	1632
			[0.8,0.89]	0.771	0.006	(0.758, 0.783)	2008
			[0.9,0.99]	0.853	0.006	(0.841, 0.864)	2238
			1	0.927	0.004	(0.918, 0.935)	2594
		Incorrect	0.5	0.467	0.022	(0.424, 0.510)	158
			[0.51,0.59]	0.523	0.013	(0.497, 0.548)	407
			[0.6,0.69]	0.534	0.015	(0.505, 0.564)	358
			[0.7,0.79]	0.564	0.012	(0.540, 0.588)	715
			[0.8,0.89]	0.547	0.014	(0.519, 0.574)	707
			[0.9,0.99]	0.645	0.016	(0.613, 0.677)	553
			1	0.767	0.019	(0.731, 0.804)	415
	Autism/vaccines	Correct	0.5	0.701	0.031	(0.639, 0.764)	57
			[0.51,0.59]	0.651	0.025	(0.601, 0.701)	101
			[0.6,0.69]	0.681	0.022	(0.637, 0.725)	135
			[0.7,0.79]	0.715	0.016	(0.684, 0.747)	262
			[0.8,0.89]	0.778	0.013	(0.753, 0.803)	415
			[0.9,0.99]	0.888	0.008	(0.873, 0.904)	657
			1	0.957	0.005	(0.947, 0.968)	788
		Incorrect	0.5	0.407	0.069	(0.258, 0.556)	14
			[0.51,0.59]	0.482	0.034	(0.415, 0.550)	55
			[0.6,0.69]	0.441	0.041	(0.358, 0.524)	53
			[0.7,0.79]	0.573	0.030	(0.514, 0.631)	107
			[0.8,0.89]	0.600	0.033	(0.535, 0.666)	109
			[0.9,0.99]	0.696	0.035	(0.627, 0.764)	102
			1	0.773	0.052	(0.668, 0.878)	48
	CDC budget	Correct	0.5	0.572	0.040	(0.491, 0.654)	45
			[0.51,0.59]	0.466	0.031	(0.404, 0.529)	56
			[0.6,0.69]	0.500	0.034	(0.431, 0.569)	60
			[0.7,0.79]	0.489	0.040	(0.409, 0.568)	70
			[0.8,0.89]	0.533	0.051	(0.430, 0.636)	52
			[0.9,0.99]	0.634	0.066	(0.500, 0.768)	34
			1	0.607	0.060	(0.486, 0.728)	49
		Incorrect	0.5	0.545	0.048	(0.447, 0.642)	29
			[0.51,0.59]	0.637	0.024	(0.590, 0.684)	100
			[0.6,0.69]	0.672	0.026	(0.620, 0.724)	79
			[0.7,0.79]	0.661	0.028	(0.606, 0.717)	121
			[0.8,0.89]	0.723	0.035	(0.653, 0.793)	82
			[0.9,0.99]	0.785	0.035	(0.715, 0.855)	76
			1	0.861	0.026	(0.809, 0.913)	121
	Climate change	Correct	0.5	0.606	0.042	(0.518, 0.694)	23
			[0.51,0.59]	0.698	0.021	(0.656, 0.739)	114
			[0.6,0.69]	0.738	0.021	(0.697, 0.779)	108
			[0.7,0.79]	0.776	0.011	(0.753, 0.798)	359
			[0.8,0.89]	0.844	0.008	(0.827, 0.861)	509
			[0.9,0.99]	0.894	0.007	(0.880, 0.909)	616
			1	0.958	0.004	(0.950, 0.967)	895
		Incorrect	0.5	0.457	0.085	(0.264, 0.650)	10
			[0.51,0.59]	0.369	0.059	(0.248, 0.491)	24
			[0.6,0.69]	0.342	0.055	(0.229, 0.456)	29

Table C.1: Estimates plotted in Figure 6 (continued)

Category	Question	Response	Certainty	Estimate	SE	CI	N
			[0.7,0.79]	0.510	0.040	(0.430, 0.591)	65
			[0.8,0.89]	0.390	0.041	(0.309, 0.471)	71
			[0.9,0.99]	0.387	0.054	(0.278, 0.496)	50
			1	0.541	0.073	(0.392, 0.689)	34
	COVID deaths	Correct	0.5	0.574	0.024	(0.527, 0.621)	112
			[0.51,0.59]	0.618	0.017	(0.585, 0.652)	211
			[0.6,0.69]	0.658	0.018	(0.622, 0.694)	195
			[0.7,0.79]	0.669	0.014	(0.642, 0.696)	407
			[0.8,0.89]	0.738	0.014	(0.711, 0.765)	416
			[0.9,0.99]	0.755	0.017	(0.723, 0.788)	347
			1	0.820	0.016	(0.789, 0.851)	323
		Incorrect	0.5	0.466	0.031	(0.405, 0.527)	70
			[0.51,0.59]	0.487	0.025	(0.438, 0.535)	128
			[0.6,0.69]	0.499	0.027	(0.446, 0.552)	110
			[0.7,0.79]	0.509	0.023	(0.464, 0.555)	197
			[0.8,0.89]	0.453	0.030	(0.395, 0.512)	156
			[0.9,0.99]	0.630	0.032	(0.567, 0.694)	131
			1	0.730	0.037	(0.656, 0.804)	99
	GM food	Correct	0.5	0.581	0.038	(0.505, 0.658)	35
			[0.51,0.59]	0.648	0.022	(0.604, 0.691)	109
			[0.6,0.69]	0.696	0.020	(0.657, 0.736)	125
			[0.7,0.79]	0.711	0.013	(0.684, 0.737)	299
			[0.8,0.89]	0.764	0.014	(0.737, 0.791)	349
			[0.9,0.99]	0.857	0.013	(0.831, 0.883)	297
			1	0.925	0.012	(0.901, 0.949)	194
		Incorrect	0.5	0.493	0.063	(0.361, 0.625)	20
			[0.51,0.59]	0.569	0.037	(0.495, 0.643)	45
			[0.6,0.69]	0.604	0.040	(0.523, 0.686)	43
			[0.7,0.79]	0.571	0.029	(0.513, 0.628)	119
			[0.8,0.89]	0.553	0.030	(0.493, 0.613)	131
			[0.9,0.99]	0.642	0.037	(0.568, 0.715)	98
			1	0.773	0.042	(0.688, 0.858)	67
	Hydroxychloroquine	Correct	0.5	0.584	0.031	(0.523, 0.646)	63
			[0.51,0.59]	0.641	0.021	(0.599, 0.682)	109
			[0.6,0.69]	0.665	0.023	(0.619, 0.711)	104
			[0.7,0.79]	0.648	0.020	(0.609, 0.688)	235
			[0.8,0.89]	0.725	0.018	(0.690, 0.760)	267
			[0.9,0.99]	0.822	0.017	(0.789, 0.855)	287
			1	0.920	0.011	(0.898, 0.943)	345
		Incorrect	0.5	0.357	0.082	(0.181, 0.534)	15
			[0.51,0.59]	0.467	0.033	(0.401, 0.534)	55
			[0.6,0.69]	0.546	0.043	(0.459, 0.632)	44
			[0.7,0.79]	0.573	0.029	(0.515, 0.630)	106
			[0.8,0.89]	0.574	0.028	(0.519, 0.628)	158
			[0.9,0.99]	0.639	0.039	(0.562, 0.716)	96
			1	0.751	0.055	(0.641, 0.861)	46
Knowledge	All questions	Correct	0.5	0.584	0.012	(0.560, 0.608)	437
			[0.51,0.59]	0.627	0.009	(0.610, 0.644)	759
			[0.6,0.69]	0.662	0.011	(0.640, 0.683)	665
			[0.7,0.79]	0.698	0.008	(0.682, 0.713)	1452
			[0.8,0.89]	0.768	0.007	(0.753, 0.782)	1849
			[0.9,0.99]	0.855	0.006	(0.842, 0.867)	2234
			1	0.952	0.003	(0.947, 0.958)	4914
		Incorrect	0.5	0.507	0.012	(0.483, 0.531)	334

Table C.1: Estimates plotted in Figure 6 (continued)

Category	Question	Response	Certainty	Estimate	SE	CI	N
			[0.51,0.59]	0.509	0.012	(0.486, 0.532)	506
			[0.6,0.69]	0.521	0.014	(0.493, 0.549)	456
			[0.7,0.79]	0.543	0.012	(0.520, 0.566)	872
			[0.8,0.89]	0.586	0.012	(0.563, 0.610)	900
			[0.9,0.99]	0.614	0.016	(0.583, 0.644)	639
			1	0.682	0.019	(0.644, 0.719)	453
	Bacteria	Correct	0.5	0.637	0.047	(0.542, 0.733)	29
			[0.51,0.59]	0.662	0.025	(0.613, 0.712)	97
			[0.6,0.69]	0.668	0.029	(0.612, 0.725)	104
			[0.7,0.79]	0.708	0.020	(0.669, 0.748)	212
			[0.8,0.89]	0.752	0.017	(0.719, 0.785)	339
			[0.9,0.99]	0.860	0.012	(0.836, 0.884)	434
			1	0.953	0.006	(0.941, 0.965)	820
		Incorrect	0.5	0.556	0.049	(0.456, 0.656)	35
			[0.51,0.59]	0.524	0.034	(0.456, 0.593)	76
			[0.6,0.69]	0.530	0.033	(0.464, 0.595)	86
			[0.7,0.79]	0.554	0.024	(0.506, 0.602)	193
			[0.8,0.89]	0.596	0.024	(0.549, 0.644)	206
			[0.9,0.99]	0.621	0.030	(0.561, 0.681)	160
			1	0.693	0.036	(0.622, 0.764)	119
	Child's sex	Correct	0.5	0.583	0.020	(0.544, 0.623)	94
			[0.51,0.59]	0.625	0.019	(0.586, 0.663)	149
			[0.6,0.69]	0.682	0.023	(0.636, 0.727)	116
			[0.7,0.79]	0.724	0.015	(0.694, 0.755)	276
			[0.8,0.89]	0.817	0.014	(0.790, 0.844)	327
			[0.9,0.99]	0.884	0.013	(0.859, 0.909)	354
			1	0.962	0.005	(0.953, 0.972)	976
		Incorrect	0.5	0.488	0.022	(0.445, 0.532)	93
			[0.51,0.59]	0.522	0.028	(0.466, 0.578)	89
			[0.6,0.69]	0.507	0.034	(0.439, 0.576)	65
			[0.7,0.79]	0.528	0.030	(0.469, 0.586)	122
			[0.8,0.89]	0.561	0.035	(0.492, 0.629)	112
			[0.9,0.99]	0.563	0.048	(0.466, 0.659)	73
			1	0.675	0.051	(0.574, 0.776)	66
	Continental drift	Correct	0.5	0.698	0.032	(0.633, 0.763)	48
			[0.51,0.59]	0.706	0.020	(0.666, 0.747)	116
			[0.6,0.69]	0.713	0.021	(0.671, 0.755)	132
			[0.7,0.79]	0.766	0.013	(0.740, 0.792)	325
			[0.8,0.89]	0.814	0.012	(0.791, 0.838)	412
			[0.9,0.99]	0.886	0.009	(0.868, 0.905)	544
			1	0.967	0.004	(0.959, 0.974)	947
		Incorrect	0.5	0.408	0.055	(0.294, 0.522)	25
			[0.51,0.59]	0.446	0.047	(0.351, 0.541)	44
			[0.6,0.69]	0.374	0.043	(0.288, 0.461)	47
			[0.7,0.79]	0.398	0.036	(0.327, 0.468)	82
			[0.8,0.89]	0.455	0.036	(0.384, 0.526)	97
			[0.9,0.99]	0.486	0.054	(0.379, 0.594)	52
			1	0.554	0.075	(0.401, 0.707)	33
	Earth/Sun	Correct	0.5	0.640	0.174	(0.156, 1.124)	5
			[0.51,0.59]	0.623	0.051	(0.517, 0.729)	25
			[0.6,0.69]	0.771	0.041	(0.685, 0.857)	18
			[0.7,0.79]	0.680	0.052	(0.576, 0.785)	42
			[0.8,0.89]	0.735	0.042	(0.651, 0.819)	56
			[0.9,0.99]	0.882	0.023	(0.836, 0.927)	126
			1	0.953	0.008	(0.938, 0.969)	539

Table C.1: Estimates plotted in Figure 6 (continued)

Category	Question	Response	Certainty	Estimate	SE	CI	N
		Incorrect	0.5	0.400	0.208	(-0.496, 1.296)	3
			[0.51,0.59]	0.512	0.070	(0.360, 0.665)	13
			[0.6,0.69]	0.599	0.149	(0.233, 0.964)	7
			[0.7,0.79]	0.559	0.054	(0.450, 0.669)	39
			[0.8,0.89]	0.680	0.072	(0.531, 0.829)	26
			[0.9,0.99]	0.547	0.073	(0.398, 0.697)	31
			1	0.556	0.072	(0.412, 0.700)	45
	Electron/atom	Correct	0.5	0.566	0.019	(0.528, 0.605)	111
			[0.51,0.59]	0.585	0.017	(0.552, 0.618)	170
			[0.6,0.69]	0.638	0.025	(0.590, 0.687)	125
			[0.7,0.79]	0.632	0.019	(0.595, 0.668)	264
			[0.8,0.89]	0.711	0.019	(0.673, 0.748)	268
			[0.9,0.99]	0.819	0.015	(0.789, 0.849)	341
			1	0.931	0.008	(0.916, 0.946)	778
		Incorrect	0.5	0.465	0.026	(0.414, 0.517)	72
			[0.51,0.59]	0.466	0.025	(0.417, 0.514)	119
			[0.6,0.69]	0.512	0.029	(0.454, 0.570)	97
			[0.7,0.79]	0.516	0.025	(0.466, 0.566)	165
			[0.8,0.89]	0.552	0.029	(0.495, 0.609)	165
			[0.9,0.99]	0.596	0.035	(0.527, 0.665)	134
			1	0.711	0.040	(0.631, 0.791)	98
	Lasers	Correct	0.5	0.523	0.023	(0.478, 0.568)	97
			[0.51,0.59]	0.547	0.022	(0.503, 0.591)	120
			[0.6,0.69]	0.531	0.032	(0.468, 0.594)	86
			[0.7,0.79]	0.577	0.028	(0.523, 0.632)	127
			[0.8,0.89]	0.707	0.023	(0.661, 0.753)	178
			[0.9,0.99]	0.735	0.027	(0.682, 0.788)	166
			1	0.922	0.012	(0.899, 0.945)	329
		Incorrect	0.5	0.580	0.020	(0.540, 0.620)	79
			[0.51,0.59]	0.601	0.023	(0.555, 0.646)	114
			[0.6,0.69]	0.608	0.028	(0.552, 0.664)	94
			[0.7,0.79]	0.631	0.021	(0.589, 0.674)	175
			[0.8,0.89]	0.665	0.022	(0.621, 0.709)	185
			[0.9,0.99]	0.743	0.029	(0.685, 0.800)	116
			1	0.791	0.039	(0.713, 0.869)	64
	Radioactivity	Correct	0.5	0.597	0.030	(0.536, 0.658)	53
			[0.51,0.59]	0.678	0.025	(0.627, 0.728)	82
			[0.6,0.69]	0.691	0.028	(0.636, 0.746)	84
			[0.7,0.79]	0.705	0.018	(0.669, 0.740)	206
			[0.8,0.89]	0.759	0.018	(0.724, 0.794)	269
			[0.9,0.99]	0.852	0.015	(0.822, 0.881)	269
			1	0.957	0.007	(0.944, 0.969)	525
		Incorrect	0.5	0.509	0.047	(0.413, 0.605)	27
			[0.51,0.59]	0.413	0.034	(0.344, 0.482)	51
			[0.6,0.69]	0.505	0.041	(0.424, 0.587)	60
			[0.7,0.79]	0.543	0.034	(0.476, 0.609)	96
			[0.8,0.89]	0.607	0.033	(0.542, 0.672)	109
			[0.9,0.99]	0.595	0.044	(0.507, 0.684)	73
			1	0.648	0.082	(0.481, 0.816)	28

Table C.2: Estimates plotted in Figure 7

Characteristic	Level	Response	Certainty	Estimate	SE	CI	N
Age	Above median	Correct	0.5	0.605	0.018	(0.569, 0.641)	166
			[0.51,0.59]	0.649	0.013	(0.623, 0.675)	322
			[0.6,0.69]	0.662	0.015	(0.634, 0.691)	344
			[0.7,0.79]	0.689	0.010	(0.669, 0.708)	776
			[0.8,0.89]	0.755	0.009	(0.737, 0.773)	1036
			[0.9,0.99]	0.848	0.009	(0.831, 0.864)	1093
			1	0.921	0.007	(0.908, 0.934)	1264
		Incorrect	0.5	0.485	0.036	(0.413, 0.558)	58
			[0.51,0.59]	0.515	0.019	(0.477, 0.554)	196
			[0.6,0.69]	0.544	0.023	(0.498, 0.589)	176
			[0.7,0.79]	0.560	0.017	(0.526, 0.593)	362
			[0.8,0.89]	0.525	0.019	(0.487, 0.563)	380
			[0.9,0.99]	0.636	0.023	(0.590, 0.682)	281
			1	0.771	0.029	(0.714, 0.829)	161
	Below median	Correct	0.5	0.595	0.020	(0.554, 0.636)	169
			[0.51,0.59]	0.617	0.013	(0.591, 0.643)	378
			[0.6,0.69]	0.674	0.012	(0.651, 0.698)	383
			[0.7,0.79]	0.705	0.009	(0.686, 0.723)	856
			[0.8,0.89]	0.787	0.008	(0.771, 0.803)	972
			[0.9,0.99]	0.858	0.008	(0.842, 0.873)	1145
			1	0.932	0.005	(0.922, 0.943)	1330
		Incorrect	0.5	0.457	0.027	(0.403, 0.511)	100
			[0.51,0.59]	0.529	0.017	(0.495, 0.564)	211
			[0.6,0.69]	0.525	0.020	(0.486, 0.565)	182
			[0.7,0.79]	0.569	0.018	(0.534, 0.604)	353
			[0.8,0.89]	0.572	0.020	(0.532, 0.612)	327
			[0.9,0.99]	0.654	0.022	(0.610, 0.698)	272
			1	0.764	0.024	(0.717, 0.812)	254
Gender	Female	Correct	0.5	0.607	0.018	(0.572, 0.641)	194
			[0.51,0.59]	0.641	0.012	(0.617, 0.666)	403
			[0.6,0.69]	0.668	0.012	(0.644, 0.692)	420
			[0.7,0.79]	0.698	0.009	(0.681, 0.716)	895
			[0.8,0.89]	0.782	0.008	(0.766, 0.798)	1045
			[0.9,0.99]	0.856	0.008	(0.840, 0.872)	1157
			1	0.931	0.005	(0.920, 0.941)	1348
		Incorrect	0.5	0.447	0.030	(0.386, 0.508)	86
			[0.51,0.59]	0.527	0.017	(0.494, 0.560)	247
			[0.6,0.69]	0.546	0.019	(0.509, 0.583)	205
			[0.7,0.79]	0.584	0.017	(0.550, 0.618)	355
			[0.8,0.89]	0.557	0.020	(0.517, 0.597)	342
			[0.9,0.99]	0.640	0.025	(0.591, 0.689)	254
			1	0.785	0.024	(0.737, 0.832)	210
	Male	Correct	0.5	0.591	0.021	(0.549, 0.633)	141
			[0.51,0.59]	0.619	0.015	(0.590, 0.648)	296
			[0.6,0.69]	0.669	0.014	(0.641, 0.697)	307
			[0.7,0.79]	0.694	0.011	(0.674, 0.715)	734
			[0.8,0.89]	0.757	0.010	(0.738, 0.776)	960
			[0.9,0.99]	0.849	0.008	(0.833, 0.866)	1079
			1	0.921	0.007	(0.909, 0.934)	1230
		Incorrect	0.5	0.492	0.030	(0.431, 0.552)	72
			[0.51,0.59]	0.515	0.021	(0.474, 0.556)	160
			[0.6,0.69]	0.519	0.025	(0.469, 0.568)	153
			[0.7,0.79]	0.545	0.017	(0.510, 0.579)	360
			[0.8,0.89]	0.537	0.020	(0.499, 0.576)	365
			[0.9,0.99]	0.649	0.021	(0.608, 0.691)	299
			1	0.749	0.028	(0.693, 0.805)	205
Educational	Bachelor's	Correct	0.5	0.598	0.018	(0.561, 0.634)	166

Table C.2: Estimates plotted in Figure 7 (continued)

Characteristic	Level	Response	Certainty	Estimate	SE	CI	N
attainment			[0.51,0.59]	0.629	0.013	(0.604, 0.654)	357
			[0.6,0.69]	0.679	0.012	(0.655, 0.703)	388
			[0.7,0.79]	0.701	0.009	(0.684, 0.719)	918
			[0.8,0.89]	0.766	0.008	(0.750, 0.782)	1261
			[0.9,0.99]	0.855	0.007	(0.841, 0.870)	1436
			1	0.931	0.005	(0.922, 0.941)	1608
		Incorrect	0.5	0.461	0.028	(0.406, 0.516)	66
			[0.51,0.59]	0.517	0.018	(0.480, 0.553)	194
			[0.6,0.69]	0.526	0.022	(0.483, 0.569)	165
			[0.7,0.79]	0.568	0.018	(0.533, 0.603)	386
			[0.8,0.89]	0.518	0.018	(0.484, 0.553)	461
			[0.9,0.99]	0.636	0.020	(0.596, 0.675)	354
			1	0.756	0.025	(0.707, 0.805)	225
	Less	Correct	0.5	0.602	0.020	(0.562, 0.642)	169
			[0.51,0.59]	0.635	0.014	(0.607, 0.662)	343
			[0.6,0.69]	0.656	0.014	(0.629, 0.684)	339
			[0.7,0.79]	0.691	0.010	(0.671, 0.712)	714
			[0.8,0.89]	0.778	0.010	(0.759, 0.797)	747
			[0.9,0.99]	0.848	0.009	(0.830, 0.867)	802
			1	0.919	0.007	(0.904, 0.933)	986
		Incorrect	0.5	0.471	0.032	(0.408, 0.534)	92
			[0.51,0.59]	0.528	0.018	(0.492, 0.564)	213
			[0.6,0.69]	0.542	0.021	(0.500, 0.583)	193
			[0.7,0.79]	0.560	0.017	(0.527, 0.593)	329
			[0.8,0.89]	0.599	0.023	(0.554, 0.645)	246
			[0.9,0.99]	0.662	0.027	(0.609, 0.715)	199
			1	0.781	0.028	(0.725, 0.836)	190
Coursework in stats/probability	Yes	Correct	0.5	0.600	0.022	(0.556, 0.643)	138
			[0.51,0.59]	0.643	0.014	(0.615, 0.672)	269
			[0.6,0.69]	0.667	0.014	(0.639, 0.695)	313
			[0.7,0.79]	0.676	0.010	(0.656, 0.696)	789
			[0.8,0.89]	0.754	0.009	(0.736, 0.772)	1045
			[0.9,0.99]	0.843	0.009	(0.826, 0.860)	1140
			1	0.928	0.006	(0.916, 0.940)	1204
		Incorrect	0.5	0.477	0.031	(0.414, 0.540)	61
			[0.51,0.59]	0.516	0.021	(0.474, 0.557)	158
			[0.6,0.69]	0.529	0.024	(0.483, 0.576)	160
			[0.7,0.79]	0.557	0.018	(0.521, 0.593)	366
			[0.8,0.89]	0.522	0.018	(0.486, 0.558)	419
			[0.9,0.99]	0.610	0.021	(0.569, 0.651)	329
			1	0.738	0.030	(0.680, 0.796)	169
	No	Correct	0.5	0.600	0.017	(0.566, 0.634)	197
			[0.51,0.59]	0.625	0.012	(0.600, 0.649)	431
			[0.6,0.69]	0.670	0.012	(0.646, 0.694)	414
			[0.7,0.79]	0.716	0.009	(0.699, 0.734)	843
			[0.8,0.89]	0.789	0.008	(0.772, 0.805)	963
			[0.9,0.99]	0.863	0.008	(0.848, 0.878)	1098
			1	0.925	0.006	(0.914, 0.937)	1390
		Incorrect	0.5	0.461	0.029	(0.403, 0.519)	97
			[0.51,0.59]	0.527	0.017	(0.494, 0.560)	249
			[0.6,0.69]	0.538	0.020	(0.499, 0.577)	198
			[0.7,0.79]	0.572	0.016	(0.539, 0.604)	349
			[0.8,0.89]	0.582	0.022	(0.540, 0.625)	288
			[0.9,0.99]	0.696	0.025	(0.647, 0.746)	224
			1	0.787	0.024	(0.740, 0.834)	246
Cognitive reflection test	1-3 correct	Correct	0.5	0.603	0.016	(0.571, 0.635)	199
			[0.51,0.59]	0.660	0.011	(0.638, 0.682)	381

Table C.2: Estimates plotted in Figure 7 (continued)

Characteristic	Level	Response	Certainty	Estimate	SE	CI	N
			[0.6,0.69]	0.685	0.012	(0.661, 0.709)	372
			[0.7,0.79]	0.726	0.009	(0.709, 0.743)	875
			[0.8,0.89]	0.782	0.008	(0.765, 0.798)	1088
			[0.9,0.99]	0.885	0.006	(0.873, 0.897)	1378
			1	0.944	0.004	(0.936, 0.953)	1687
		Incorrect	0.5	0.454	0.028	(0.399, 0.509)	83
			[0.51,0.59]	0.529	0.019	(0.491, 0.566)	195
			[0.6,0.69]	0.557	0.023	(0.511, 0.602)	164
			[0.7,0.79]	0.589	0.018	(0.553, 0.625)	328
			[0.8,0.89]	0.590	0.020	(0.551, 0.630)	338
			[0.9,0.99]	0.674	0.023	(0.629, 0.719)	246
			1	0.806	0.024	(0.758, 0.853)	207
	None correct	Correct	0.5	0.596	0.024	(0.548, 0.644)	136
			[0.51,0.59]	0.598	0.015	(0.568, 0.628)	319
			[0.6,0.69]	0.652	0.014	(0.624, 0.679)	355
			[0.7,0.79]	0.663	0.011	(0.642, 0.684)	757
			[0.8,0.89]	0.757	0.009	(0.739, 0.776)	920
			[0.9,0.99]	0.801	0.011	(0.779, 0.823)	860
			1	0.894	0.009	(0.876, 0.911)	907
		Incorrect	0.5	0.482	0.034	(0.413, 0.550)	75
			[0.51,0.59]	0.517	0.018	(0.482, 0.552)	212
			[0.6,0.69]	0.516	0.020	(0.476, 0.555)	194
			[0.7,0.79]	0.543	0.016	(0.511, 0.575)	387
			[0.8,0.89]	0.506	0.019	(0.467, 0.544)	369
			[0.9,0.99]	0.622	0.022	(0.578, 0.666)	307
			1	0.728	0.028	(0.674, 0.783)	208
Need for closure	Above median	Correct	0.5	0.603	0.019	(0.566, 0.640)	175
			[0.51,0.59]	0.639	0.013	(0.613, 0.664)	360
			[0.6,0.69]	0.655	0.013	(0.628, 0.681)	397
			[0.7,0.79]	0.700	0.009	(0.682, 0.718)	910
			[0.8,0.89]	0.774	0.009	(0.757, 0.791)	1070
			[0.9,0.99]	0.871	0.007	(0.857, 0.884)	1175
			1	0.929	0.006	(0.918, 0.940)	1349
		Incorrect	0.5	0.482	0.028	(0.427, 0.536)	90
			[0.51,0.59]	0.509	0.017	(0.474, 0.543)	237
			[0.6,0.69]	0.510	0.021	(0.470, 0.551)	195
			[0.7,0.79]	0.548	0.016	(0.516, 0.581)	416
			[0.8,0.89]	0.561	0.019	(0.523, 0.598)	386
			[0.9,0.99]	0.666	0.022	(0.623, 0.710)	271
			1	0.825	0.022	(0.783, 0.868)	210
	Below median	Correct	0.5	0.597	0.020	(0.557, 0.637)	160
			[0.51,0.59]	0.624	0.014	(0.597, 0.652)	340
			[0.6,0.69]	0.685	0.012	(0.661, 0.710)	330
			[0.7,0.79]	0.693	0.010	(0.673, 0.713)	722
			[0.8,0.89]	0.768	0.009	(0.750, 0.786)	936
			[0.9,0.99]	0.834	0.009	(0.816, 0.853)	1061
			1	0.924	0.006	(0.911, 0.936)	1245
		Incorrect	0.5	0.449	0.035	(0.379, 0.518)	68
			[0.51,0.59]	0.542	0.019	(0.504, 0.580)	170
			[0.6,0.69]	0.568	0.022	(0.524, 0.612)	161
			[0.7,0.79]	0.586	0.018	(0.549, 0.622)	298
			[0.8,0.89]	0.529	0.021	(0.488, 0.570)	320
			[0.9,0.99]	0.622	0.023	(0.576, 0.668)	280
			1	0.708	0.030	(0.648, 0.767)	205
Generic conspiracy beliefs	Above median	Correct	0.5	0.607	0.032	(0.544, 0.670)	71
			[0.51,0.59]	0.614	0.021	(0.572, 0.655)	153
			[0.6,0.69]	0.654	0.024	(0.607, 0.701)	137

Table C.2: Estimates plotted in Figure 7 (continued)

Characteristic	Level	Response	Certainty	Estimate	SE	CI	N
			[0.7,0.79]	0.716	0.018	(0.680, 0.753)	212
			[0.8,0.89]	0.794	0.015	(0.763, 0.824)	282
			[0.9,0.99]	0.897	0.011	(0.876, 0.918)	310
			1	0.935	0.008	(0.919, 0.952)	534
		Incorrect	0.5	0.492	0.034	(0.422, 0.561)	43
			[0.51,0.59]	0.541	0.024	(0.492, 0.589)	131
			[0.6,0.69]	0.581	0.029	(0.523, 0.639)	95
			[0.7,0.79]	0.632	0.027	(0.579, 0.686)	154
			[0.8,0.89]	0.710	0.037	(0.636, 0.783)	80
			[0.9,0.99]	0.698	0.046	(0.606, 0.789)	69
			1	0.796	0.033	(0.731, 0.860)	125
	Below median	Correct	0.5	0.604	0.048	(0.505, 0.703)	35
			[0.51,0.59]	0.599	0.027	(0.546, 0.652)	95
			[0.6,0.69]	0.604	0.028	(0.549, 0.659)	97
			[0.7,0.79]	0.677	0.023	(0.631, 0.722)	177
			[0.8,0.89]	0.785	0.020	(0.744, 0.825)	179
			[0.9,0.99]	0.806	0.024	(0.759, 0.853)	191
			1	0.843	0.022	(0.798, 0.887)	210
		Incorrect	0.5	0.550	0.062	(0.418, 0.682)	23
			[0.51,0.59]	0.578	0.030	(0.518, 0.637)	79
			[0.6,0.69]	0.594	0.032	(0.529, 0.659)	65
			[0.7,0.79]	0.588	0.036	(0.516, 0.659)	86
			[0.8,0.89]	0.575	0.040	(0.495, 0.655)	82
			[0.9,0.99]	0.717	0.037	(0.642, 0.792)	87
			1	0.719	0.049	(0.621, 0.817)	83
Political partisanship	Strong	Correct	0.5	0.593	0.025	(0.542, 0.644)	95
			[0.51,0.59]	0.631	0.018	(0.595, 0.667)	211
			[0.6,0.69]	0.662	0.017	(0.629, 0.696)	248
			[0.7,0.79]	0.689	0.011	(0.668, 0.711)	700
			[0.8,0.89]	0.742	0.010	(0.722, 0.762)	932
			[0.9,0.99]	0.837	0.010	(0.818, 0.856)	1045
			1	0.919	0.007	(0.905, 0.933)	1204
		Incorrect	0.5	0.435	0.044	(0.346, 0.525)	39
			[0.51,0.59]	0.500	0.023	(0.454, 0.546)	145
			[0.6,0.69]	0.545	0.025	(0.496, 0.594)	141
			[0.7,0.79]	0.569	0.019	(0.532, 0.606)	343
			[0.8,0.89]	0.524	0.019	(0.487, 0.562)	410
			[0.9,0.99]	0.639	0.021	(0.598, 0.679)	353
			1	0.702	0.028	(0.647, 0.758)	218
	Weak/lean	Correct	0.5	0.603	0.016	(0.571, 0.635)	240
			[0.51,0.59]	0.632	0.011	(0.611, 0.654)	489
			[0.6,0.69]	0.672	0.011	(0.650, 0.694)	479
			[0.7,0.79]	0.703	0.009	(0.686, 0.719)	932
			[0.8,0.89]	0.795	0.007	(0.781, 0.810)	1076
			[0.9,0.99]	0.866	0.007	(0.853, 0.880)	1193
			1	0.933	0.005	(0.924, 0.943)	1390
		Incorrect	0.5	0.478	0.025	(0.428, 0.528)	119
			[0.51,0.59]	0.535	0.015	(0.505, 0.566)	262
			[0.6,0.69]	0.527	0.019	(0.489, 0.565)	217
			[0.7,0.79]	0.559	0.016	(0.528, 0.591)	372
			[0.8,0.89]	0.577	0.021	(0.537, 0.618)	297
			[0.9,0.99]	0.657	0.026	(0.606, 0.707)	200
			1	0.839	0.022	(0.795, 0.882)	197
Interest in politics	Very interested	Correct	0.5	0.649	0.038	(0.573, 0.724)	68
			[0.51,0.59]	0.651	0.022	(0.606, 0.695)	144
			[0.6,0.69]	0.668	0.021	(0.626, 0.710)	148
			[0.7,0.79]	0.707	0.014	(0.680, 0.734)	410

Table C.2: Estimates plotted in Figure 7 (continued)

Characteristic	Level	Response	Certainty	Estimate	SE	CI	N
			[0.8,0.89]	0.772	0.012	(0.750, 0.795)	587
			[0.9,0.99]	0.851	0.010	(0.830, 0.872)	818
			1	0.925	0.007	(0.910, 0.939)	1115
		Incorrect	0.5	0.450	0.039	(0.371, 0.529)	39
			[0.51,0.59]	0.512	0.029	(0.455, 0.569)	96
			[0.6,0.69]	0.571	0.035	(0.501, 0.641)	83
			[0.7,0.79]	0.624	0.025	(0.575, 0.673)	184
			[0.8,0.89]	0.527	0.027	(0.474, 0.580)	225
			[0.9,0.99]	0.647	0.027	(0.594, 0.700)	215
			1	0.762	0.027	(0.709, 0.816)	223
	All others	Correct	0.5	0.588	0.014	(0.560, 0.615)	267
			[0.51,0.59]	0.627	0.010	(0.607, 0.647)	556
			[0.6,0.69]	0.669	0.010	(0.648, 0.689)	579
			[0.7,0.79]	0.694	0.008	(0.678, 0.709)	1222
			[0.8,0.89]	0.770	0.007	(0.755, 0.784)	1421
			[0.9,0.99]	0.854	0.007	(0.840, 0.867)	1420
			1	0.928	0.005	(0.919, 0.938)	1479
		Incorrect	0.5	0.473	0.026	(0.421, 0.524)	119
			[0.51,0.59]	0.526	0.015	(0.497, 0.555)	311
			[0.6,0.69]	0.523	0.017	(0.491, 0.556)	275
			[0.7,0.79]	0.544	0.014	(0.516, 0.571)	531
			[0.8,0.89]	0.556	0.016	(0.524, 0.588)	482
			[0.9,0.99]	0.644	0.020	(0.604, 0.684)	338
			1	0.772	0.025	(0.722, 0.823)	192
Political knowledge	0 to 3 correct	Correct	0.5	0.572	0.040	(0.490, 0.654)	51
			[0.51,0.59]	0.584	0.023	(0.539, 0.630)	131
			[0.6,0.69]	0.632	0.028	(0.576, 0.687)	105
			[0.7,0.79]	0.665	0.024	(0.618, 0.711)	178
			[0.8,0.89]	0.769	0.019	(0.731, 0.807)	209
			[0.9,0.99]	0.827	0.022	(0.783, 0.870)	193
			1	0.867	0.019	(0.829, 0.906)	223
		Incorrect	0.5	0.481	0.043	(0.392, 0.570)	34
			[0.51,0.59]	0.496	0.028	(0.441, 0.552)	97
			[0.6,0.69]	0.520	0.032	(0.456, 0.584)	68
			[0.7,0.79]	0.544	0.034	(0.477, 0.610)	113
			[0.8,0.89]	0.546	0.043	(0.459, 0.632)	82
			[0.9,0.99]	0.628	0.046	(0.535, 0.721)	73
			1	0.679	0.050	(0.579, 0.779)	73
	4+ correct	Correct	0.5	0.637	0.033	(0.570, 0.703)	55
			[0.51,0.59]	0.635	0.024	(0.588, 0.682)	116
			[0.6,0.69]	0.636	0.024	(0.589, 0.684)	129
			[0.7,0.79]	0.724	0.018	(0.688, 0.759)	212
			[0.8,0.89]	0.808	0.016	(0.777, 0.839)	252
			[0.9,0.99]	0.885	0.013	(0.860, 0.910)	309
			1	0.927	0.010	(0.908, 0.945)	521
		Incorrect	0.5	0.543	0.044	(0.451, 0.634)	32
			[0.51,0.59]	0.605	0.025	(0.555, 0.654)	112
			[0.6,0.69]	0.629	0.029	(0.571, 0.687)	92
			[0.7,0.79]	0.680	0.024	(0.633, 0.728)	132
			[0.8,0.89]	0.724	0.033	(0.658, 0.791)	82
			[0.9,0.99]	0.779	0.033	(0.713, 0.845)	84
			1	0.812	0.032	(0.748, 0.875)	135

C.3 Supplemental figures

The following figures supplement main text Figure 6:

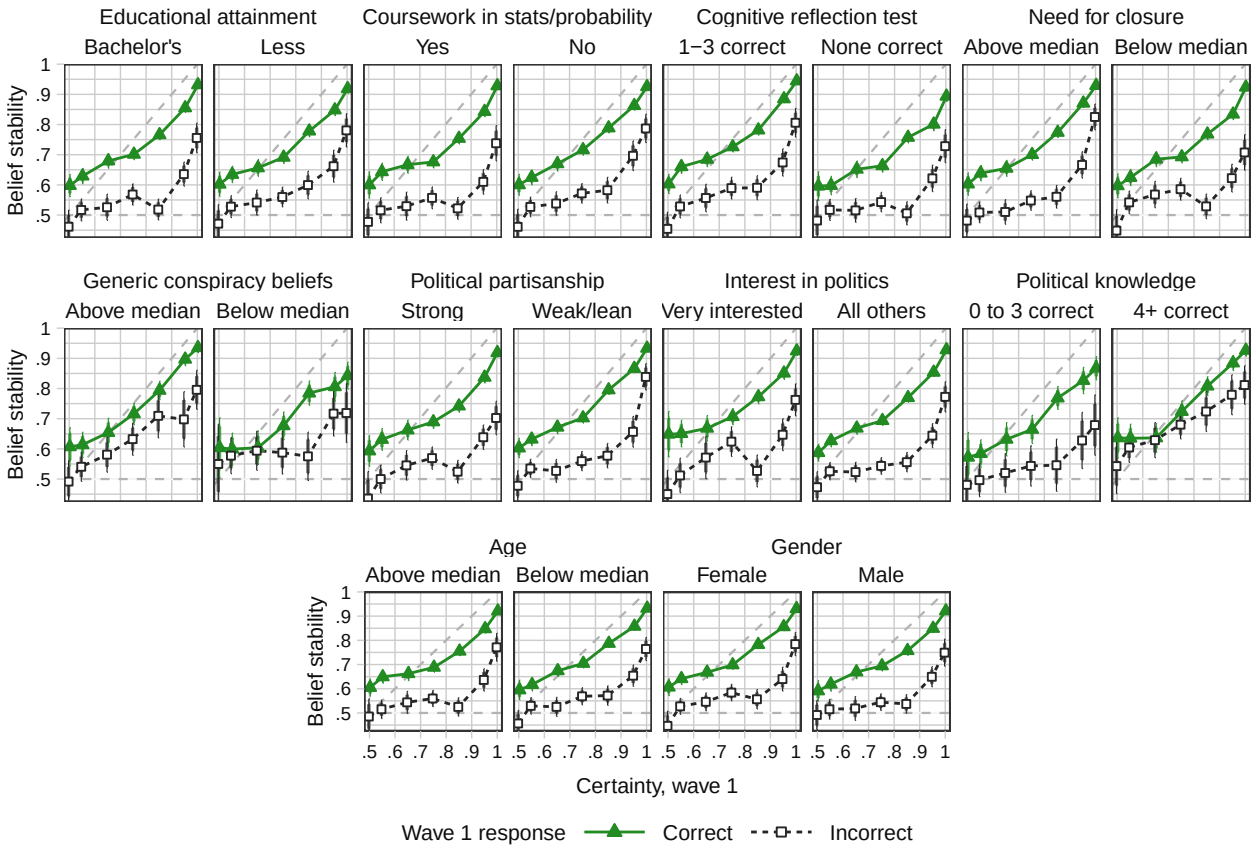
- Figures C.2a and C.2b present the same results as Figure 6 separately for Studies 3a and 3b. Figure C.2a also includes results using the costly measure.
- Figure C.3 presents the same results presented in Figure 6, but with best guess stability substituted for belief stability.
- Figures C.4a and C.4b present the same results as Figure C.3 separately for Studies 3a and 3b. Figure C.4a also includes results using the costly measure.

The following figures supplement main text Figure 7:

- Figure C.1 presents the same results as Figure 7 with best guess stability substituted for belief stability.
- Figure C.5 presents the same results as Figure 7 separately for Studies 3a and 3b. Two of these variables, generic conspiracy beliefs and political knowledge, were included in Study 3a only.

The final figure, Figure C.6, presents the variance of wave 2 beliefs conditional on the wave 1 certainty level. This is referred to in the main text discussion of ambivalence among miseducated guessers.

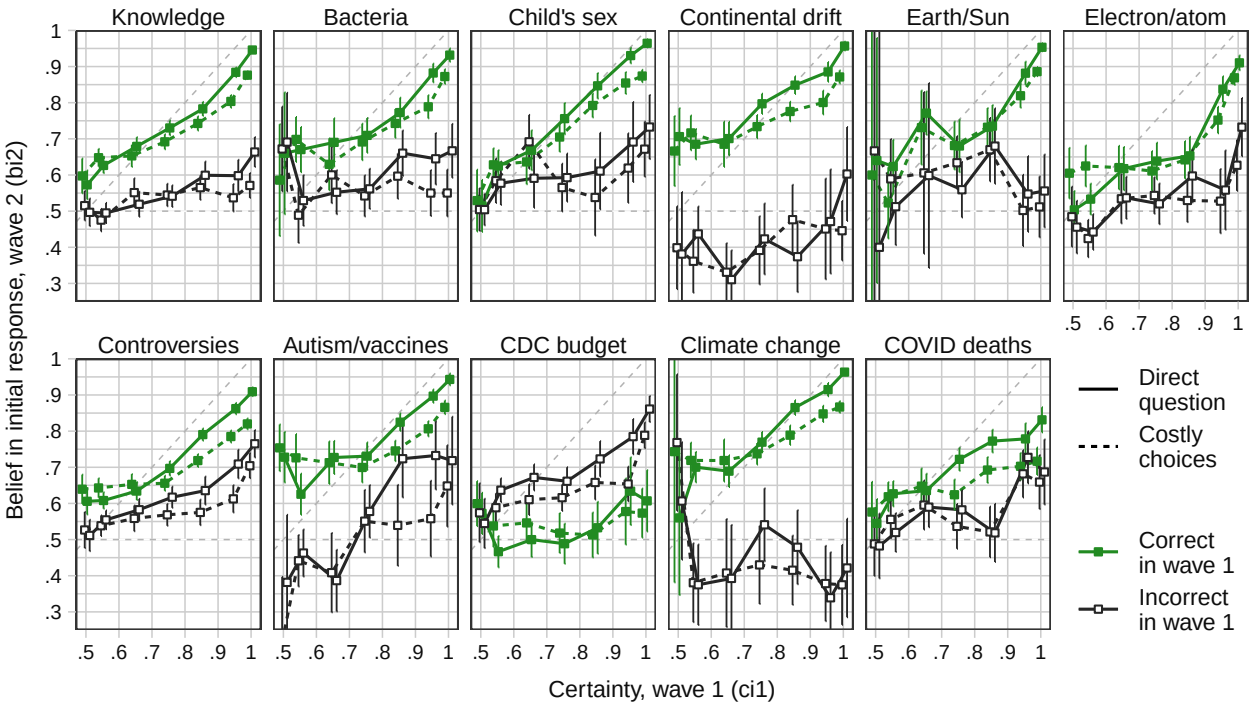
Figure C.1: Temporal stability of best guesses by certainty level and respondent characteristics, Study 3.



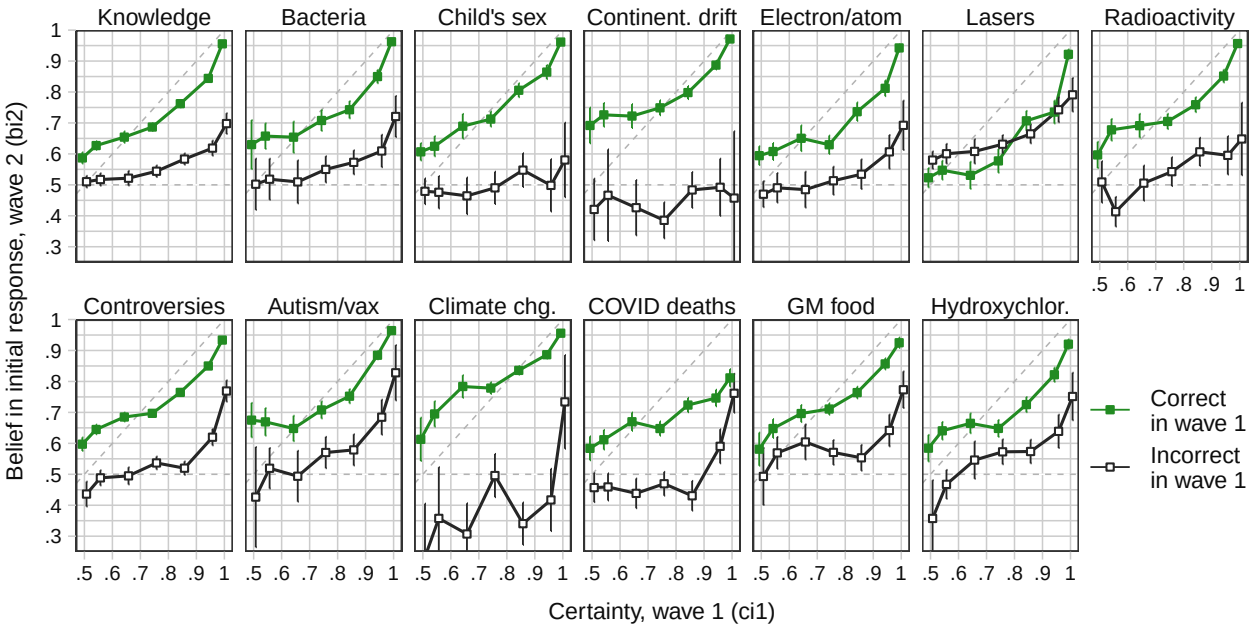
Note: Figure is identical to main text Figure 7, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

Figure C.2: Temporal stability of beliefs by certainty level, Studies 3a and 3b.

(a) Study 3a

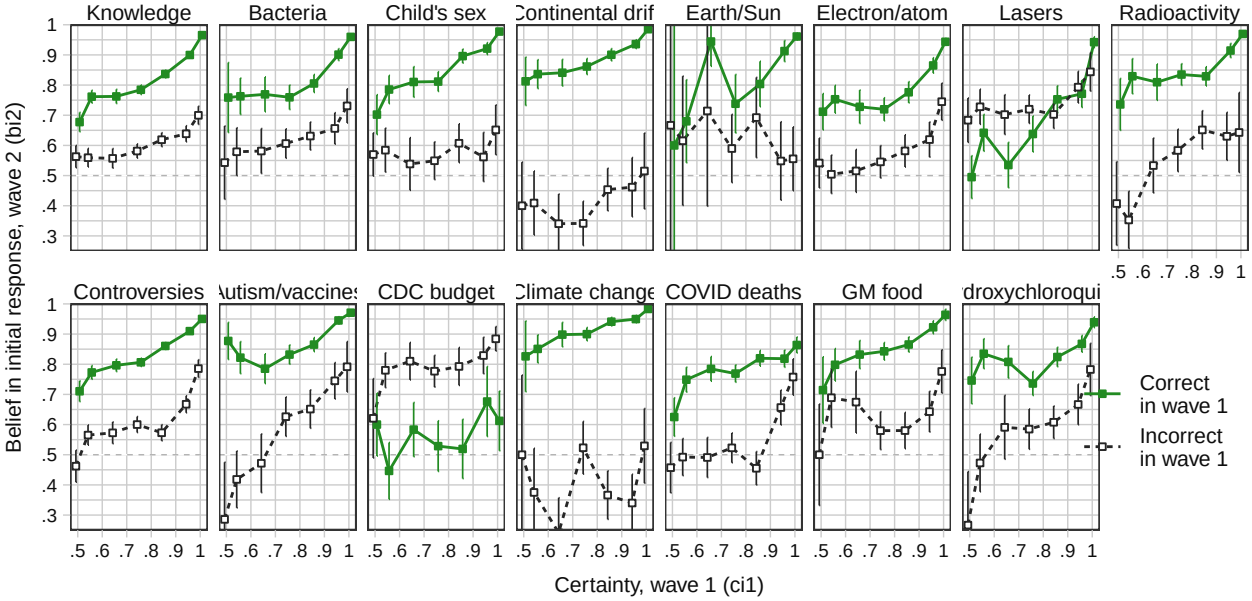


(b) Study 3b



Note: This figure displays the same information as Figure 6 separately for Studies 3a and 3b. The figure for Study 3a also adds the costly measure.

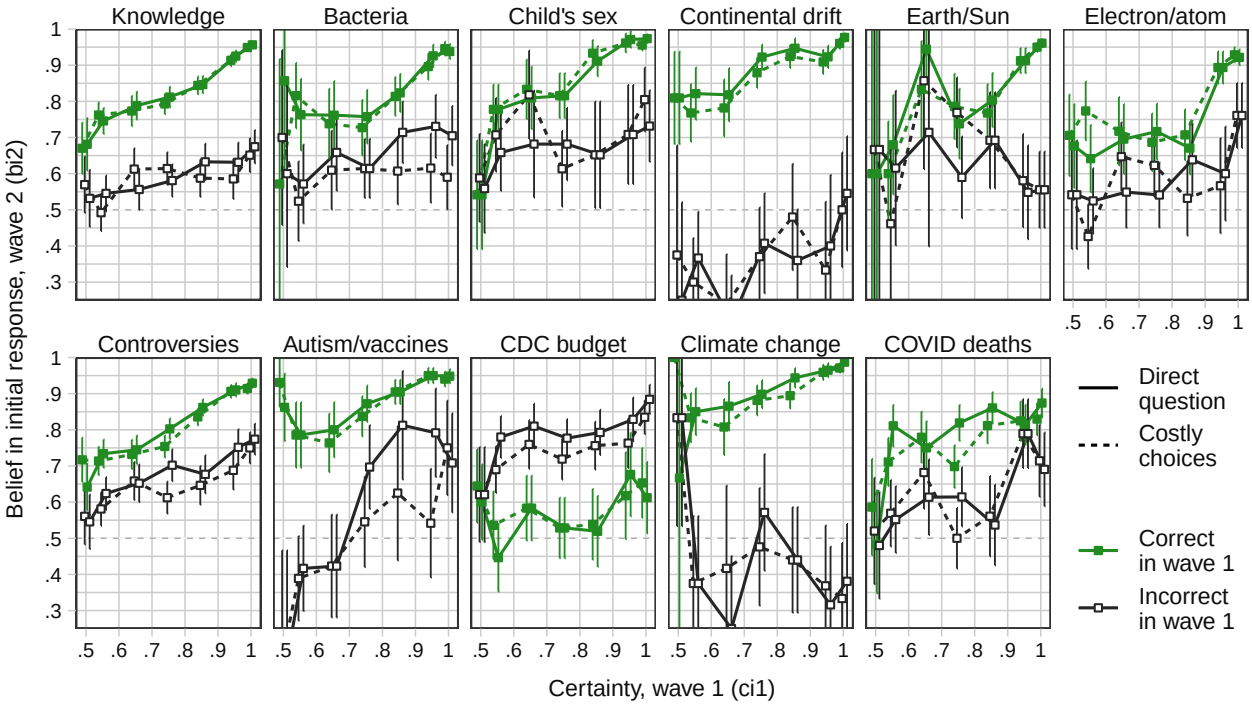
Figure C.3: Temporal stability of best guesses by certainty level, Study 3



Note: Figure is identical to main text Figure 6, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

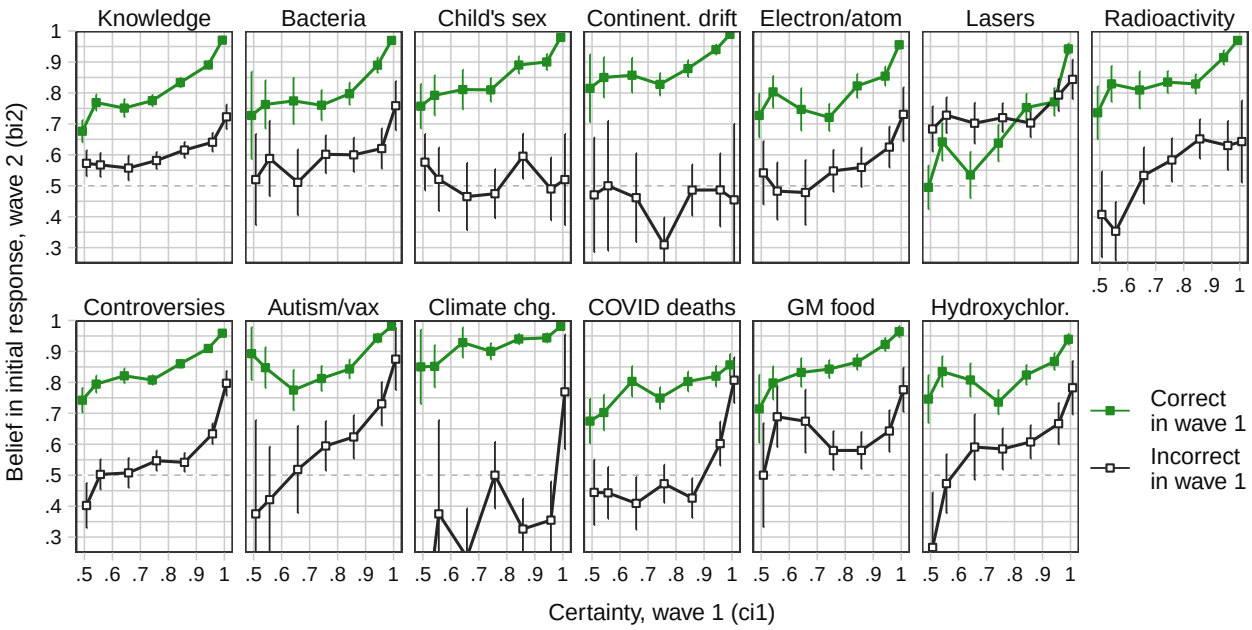
Figure C.4: Temporal stability of best guesses by certainty level, Studies 3a and 3b.

(a) Study 3a



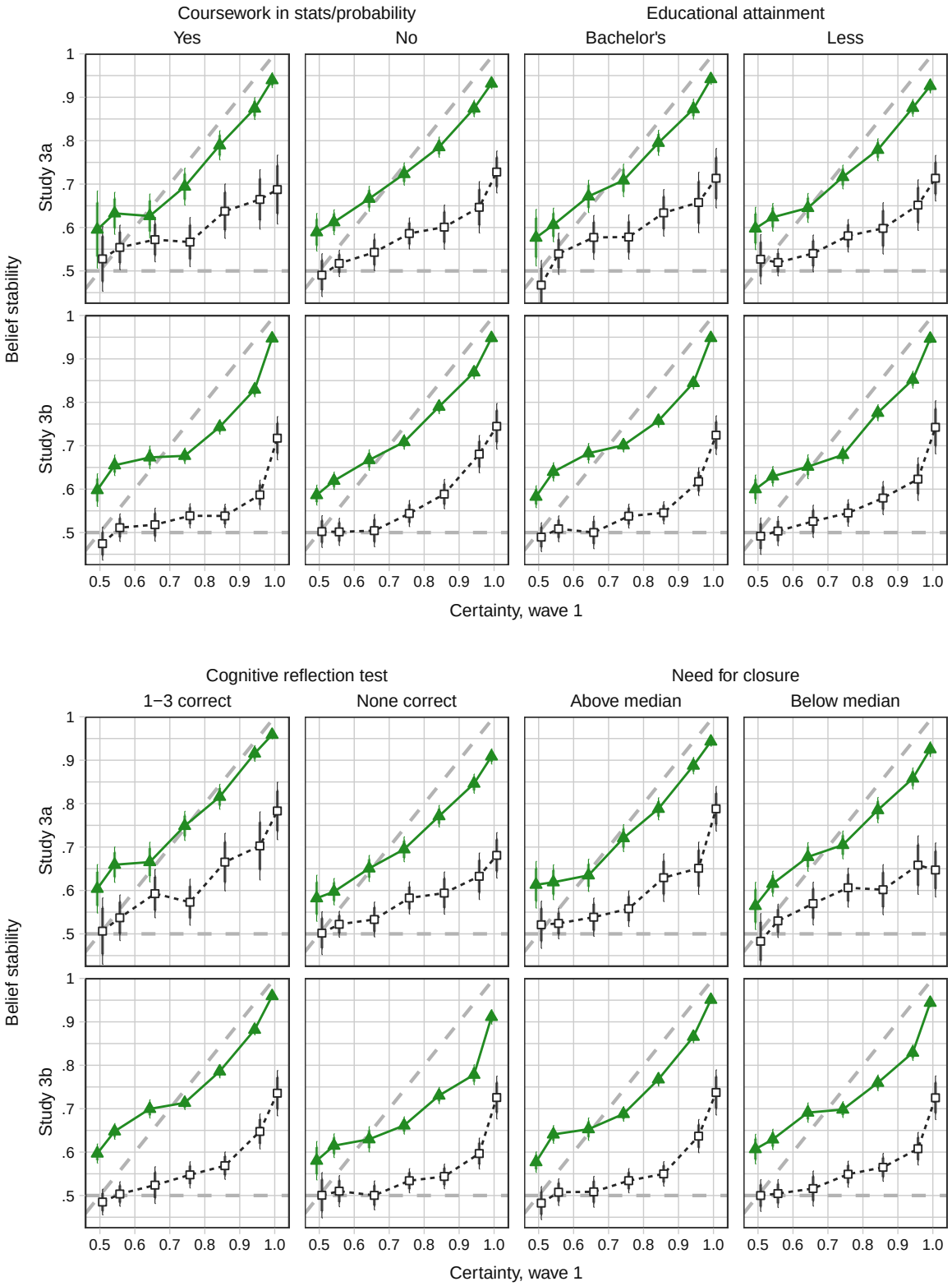
Note: Figure is identical to Figure C.2a, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

(b) Study 3b



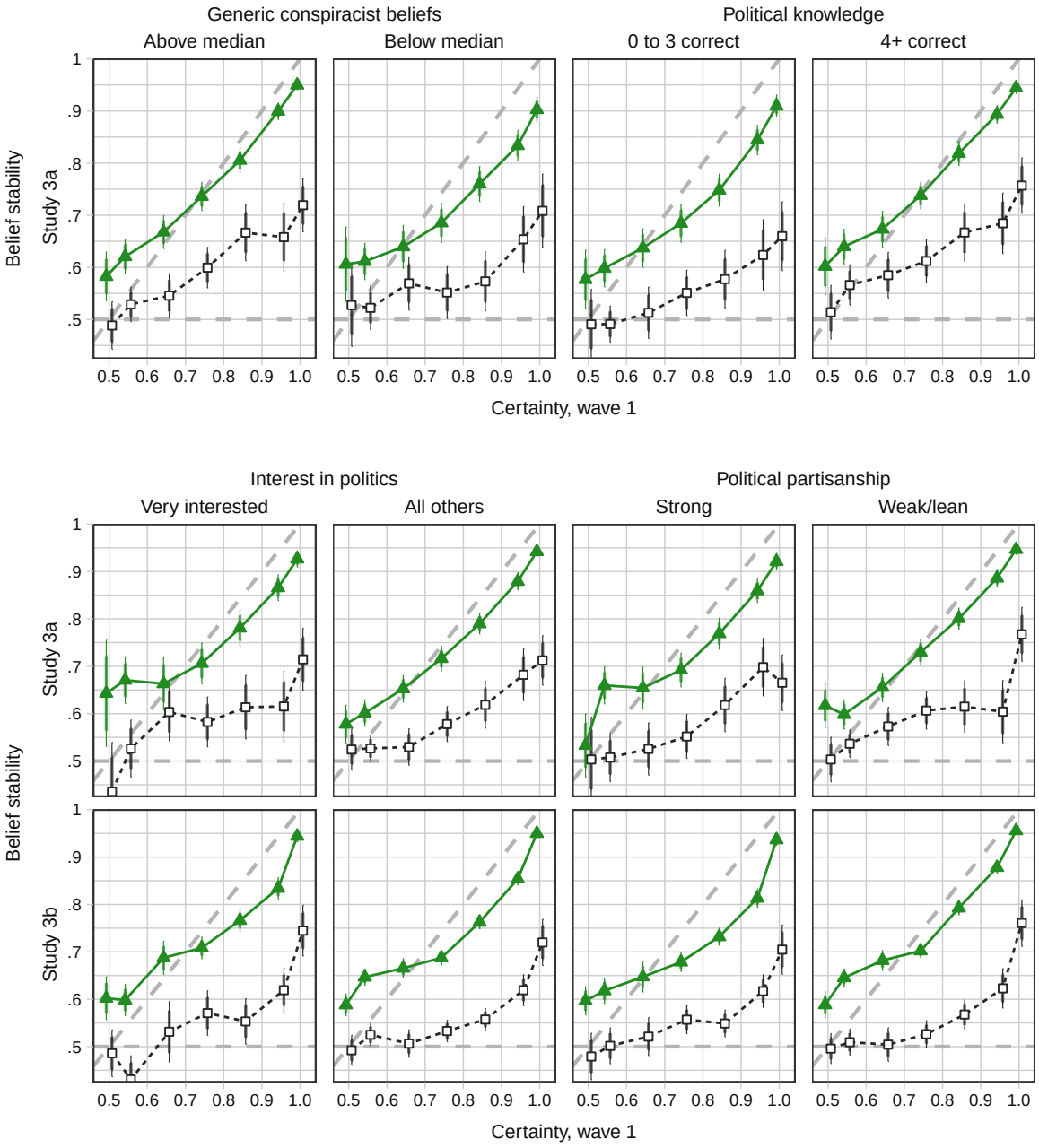
Note: Figure is identical to Figure C.2b, with the exception that best guess stability is substituted for belief stability. The main text defines these quantities.

Figure C.5: Temporal stability of beliefs by certainty level and respondent characteristics, Study 3.



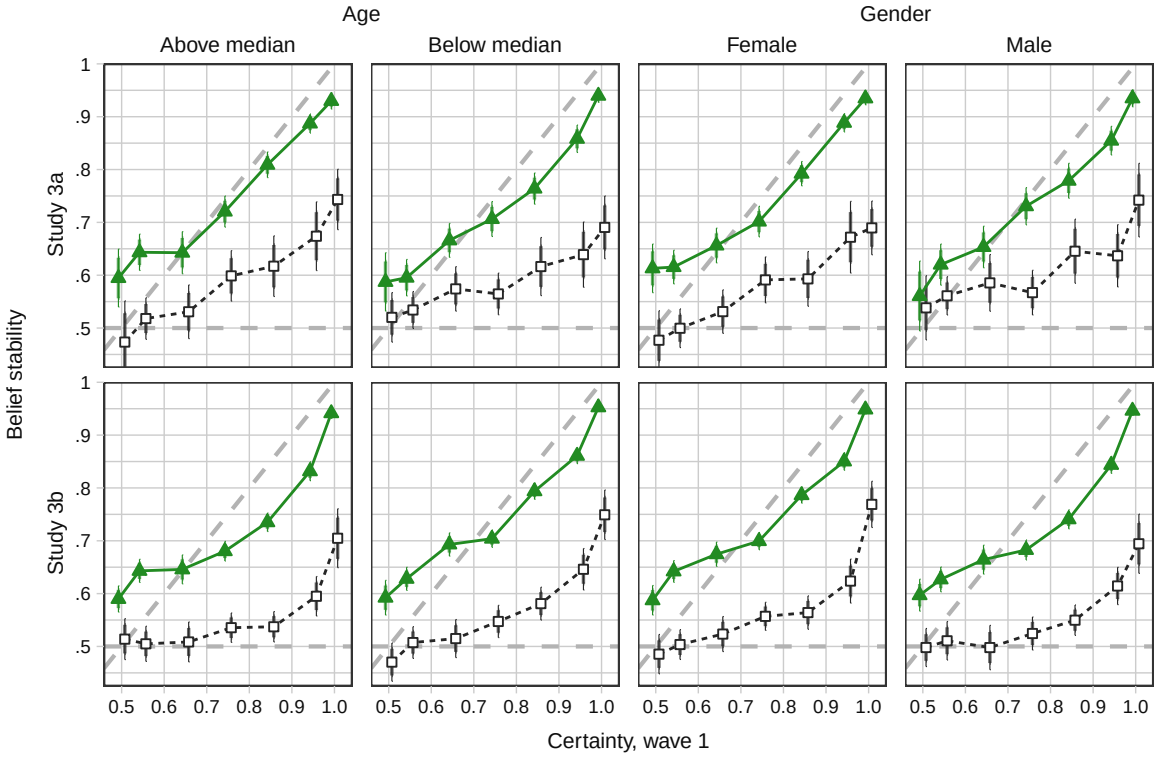
Note: Figure replicates the top row of Figure C.5, splitting apart Studies 3a and 3b.

Figure C.5 (continued).



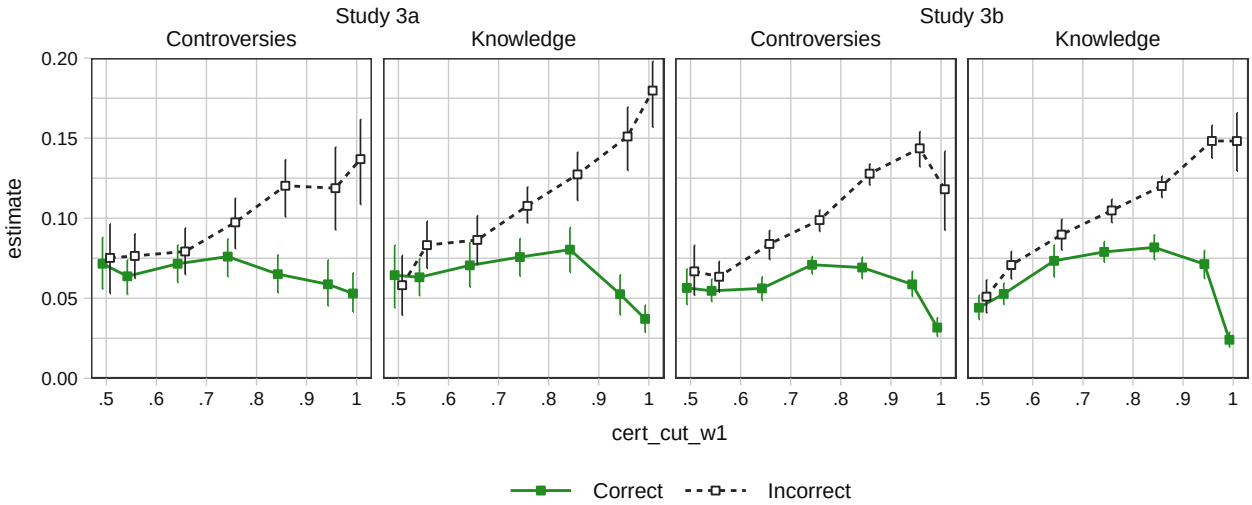
Note: Figure replicates the middle row of Figure C.5, splitting apart Studies 3a and 3b.

Figure C.5 (continued).



Note: Figure replicates the bottom row of Figure C.5, splitting apart Studies 3a and 3b.

Figure C.6: Variance of wave 2 beliefs by wave 1 certainty level.



C.4 Within-subject analysis

The following tables present the results of the within-subject analysis described in the main text.

Table C.3: Within-subject regression estimates, all questions.

	Dependent variable: b_{i2}			
	Study 4a		Study 4b	
Constant	0.311^{**}		0.358^{**}	
	(0.014)		(0.010)	
p_{i1}	0.394^{**}	0.367^{**}	0.372^{**}	0.166^{**}
	(0.041)	(0.052)	(0.031)	(0.039)
g_{i1}	-0.135^{**}	-0.142^{**}	-0.210^{**}	-0.277^{**}
	(0.024)	(0.030)	(0.016)	(0.019)
$p_{i1} \times g_{i1}$	0.349^{**}	0.370^{**}	0.388^{**}	0.566^{**}
	(0.045)	(0.060)	(0.034)	(0.041)
Fixed effects	No	Yes	No	Yes
R^2	0.392	0.466	0.340	0.439
Adj. R^2	0.392	0.396	0.340	0.383
Num. obs.	8762	8762	21310	21310
Num. clusters	1014	1014	1954	1954

Table C.4: Within-subject regression estimates, knowledge questions.

	Dependent variable: b_{i2}			
	Study 4a		Study 4b	
Constant	0.359^{**}		0.363^{**}	
	(0.019)		(0.012)	
p_{i1}	0.338^{**}	0.288^{**}	0.309^{**}	0.059
	(0.058)	(0.079)	(0.037)	(0.051)
g_{i1}	-0.188^{**}	-0.155^{**}	-0.246^{**}	-0.322^{**}
	(0.031)	(0.044)	(0.021)	(0.027)
$p_{i1} \times g_{i1}$	0.424^{**}	0.385^{**}	0.491^{**}	0.712^{**}
	(0.063)	(0.090)	(0.041)	(0.057)
Fixed effects	No	Yes	No	Yes
R^2	0.381	0.522	0.366	0.511
Adj. R^2	0.381	0.397	0.366	0.412
Num. obs.	4868	4868	11628	11628
Num. clusters	1006	1006	1950	1950

Table C.5: Within-subject regression estimates, misinformation questions.

	Dependent variable: b_{i2}			
	Study 4a		Study 4b	
Constant	0.263^{**}		0.347^{**}	
	(0.017)		(0.014)	
p_{i1}	0.454^{**}	0.476^{**}	0.491^{**}	0.301^{**}
	(0.055)	(0.081)	(0.050)	(0.065)
g_{i1}	-0.059	-0.038	-0.156^{**}	-0.192^{**}
	(0.031)	(0.048)	(0.022)	(0.029)
$p_{i1} \times g_{i1}$	0.237^{**}	0.231^{*}	0.213^{**}	0.336^{**}
	(0.060)	(0.097)	(0.053)	(0.071)
Fixed effects	No	Yes	No	Yes
R^2	0.391	0.542	0.308	0.498
Adj. R^2	0.390	0.384	0.308	0.372
Num. obs.	3894	3894	9682	9682
Num. clusters	997	997	1951	1951

C.5 Comparison between branching and all-in-one scales

Researchers interested in measuring the individual-level uncertainty in respondents' beliefs do so in one of two ways: by presenting an all-in-one scale with a probabilistic interpretation (e.g., definitely false, probably false, probably true, definitely true), or by using a branching question that first elicits the respondent's best guess, then asks how certain they are about it. The main text analyzes the data as if these two measurement technologies are functions of the same underlying construct, inferring p_i from measures of g_i and c_i. To examine the reasonability of this posture, an experiment was embedded in Study 3a. Subjects were randomly assigned respondents to answer all of the questions using either an all-in-one or branching scale (simple random assignment, $p = 0.5$; Gerber and Green 2012). Both scale types used identically worded response options and scale points (Appendix E.2). To the degree that the two scales elicit similar belief distributions with similar measurement properties, it is fair to treat a measure of g_i and c_i as equivalent to a measure of p_i and vice versa.

First, consider the mean of each measure. Scale type had no statistically detectable effect on the three key quantities defined in the empirical framework: p_i, the respondent's belief in the correct answer (Table C.6); g_i, a binary variable indicating whether the respondent's answer was correct or incorrect (Table C.7); or c_i, the respondent's certainty level (Table C.8).

Next, consider the distribution of certainty. Figure C.7 shows that any distributional differences between the two measures are substantively insignificant.

Finally, consider the central measurement property examined in the main text, the stability of measured beliefs conditional on the respondent's initial answer and certainty level. Figure C.8 shows that there are few differences in the measurement properties of the two measures of certainty. The possible exception is that all-in-one scales may be a bit better at encouraging the least certain respondents to use low scale points.

Table C.6: Effect of branching scale on average belief in correct answer.

	Wave 1	Wave 2
Constant	0.672**	0.688**
	(0.007)	(0.007)
Branching scale	0.006	0.003
	(0.009)	(0.009)
Adj. R^2	−0.000	−0.000
Num. obs.	9755	9754
Num. clusters	1015	1015

Table C.7: Effect of branching scale on probability of a correct best guess.

	Wave 1	Wave 2
Constant	0.694**	0.710**
	(0.008)	(0.008)
Branching scale	0.020	0.008
	(0.011)	(0.010)
Adj. R^2	0.000	−0.000
Num. obs.	9783	9783
Num. clusters	1016	1016

Table C.8: Effect of branching scale on average certainty.

	Wave 1	Wave 2
Constant	0.838**	0.868**
	(0.005)	(0.005)
Branching scale	0.007	0.002
	(0.007)	(0.006)
Adj. R^2	0.000	−0.000
Num. obs.	9755	9754
Num. clusters	1015	1015

Figure C.7: Distribution of certainty by question type and correct/incorrect answer, branching versus all-in-one scale, wave 1.

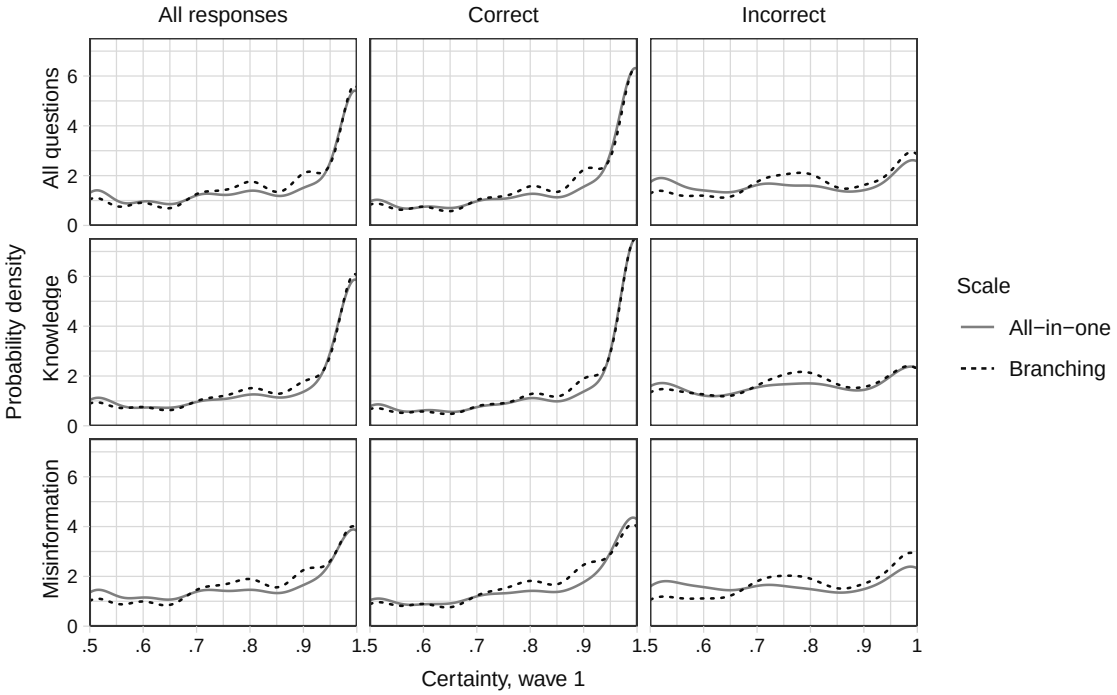
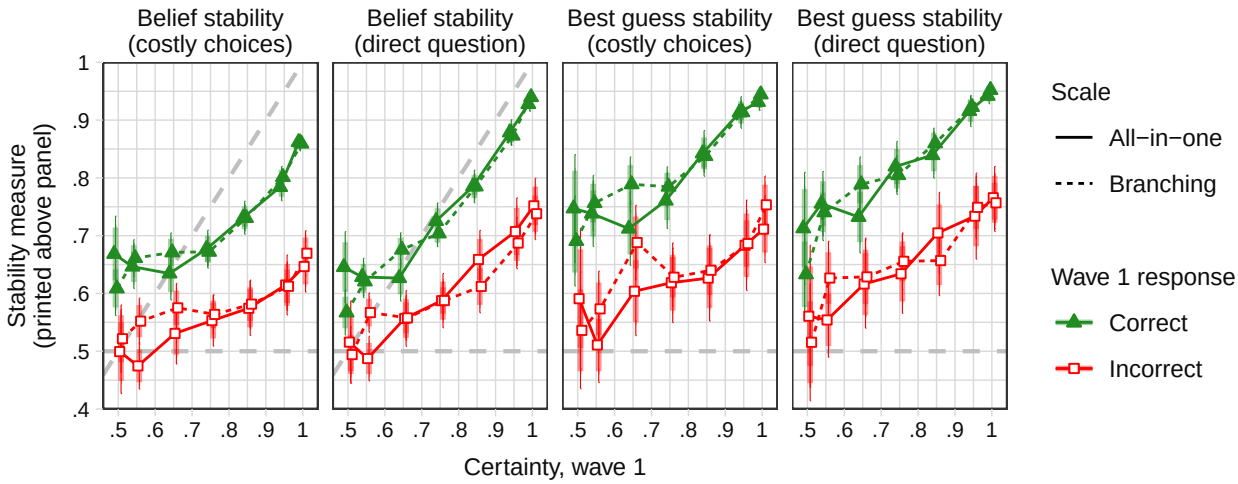


Figure C.8: Comparison of stability by certainty level, branching versus all-in-one scale.



D Appendix to Study 4

The same surveys are analyzed in Studies 3 and 4. For survey information, see the appendix to Study 3.

D.1 Full text of training exercise

The training exercise asked respondents to interact with four vignettes, which were displayed in a random order. Vignette 1 is printed in the main text. This section contains the full text of vignettes 2-4 and a description of the randomization procedure for the names.

Vignette 2

[Name] gets the question,

> *Nationwide, is the average price of gas above or below $2.00?*

[She/He] knows that the answer is "above $2.00" because [s/he] saw this fact in the news.

How sure is [Name] that the answer is "above $2.00"?

- 60 percent sure
- 80 percent sure
- 99 percent sure

[DISPLAYS AFTER CLICK:] The best choice is 99 percent sure. [Name] knows for a fact that gas costs more than $2.00. When you make your choices, it's important not to pick high levels of certainty unless you are extremely confident in your answer.

Vignette 3

[Name] gets the question,

> *Nationwide, is the average price of gas above or below $2.00?*

[She/He] knows gas costs more than $2.00 in [her/his] area, but [s/he]'s not sure about the rest of the country.

How sure is [Name] that the answer is "above $2.00"?

- 70 percent sure
- 95 percent sure

[DISPLAYS AFTER CLICK:] The best choice is 70 percent sure. [Name] knows something that allows [her/him] to make a pretty good guess, but [s/he] doesn't know nearly enough to be 95 percent certain.

When you're only somewhat confident in your choice, it's important to pick a middling level of certainty.

Vignette 4

[Name] gets the question,

> *Nationwide, is the average price of gas above or below $2.00?*

[Name] knows gas prices have gone up a lot since [s/he] sold [her/his] car back in the mid-1990s, but isn't sure how much. [She/He] chooses "above $2.00" but isn't too confident in [her/his] guess.

How certain is [Name] that the answer is "above $2.00"?

- 50 percent sure
- 55 percent sure
- 85 percent sure

[DISPLAYS AFTER CLICK:] The best choice is 55 percent sure. [Name] has something to go on, so it's not quite a coin flip, but the things [s/he] thought about weren't too helpful either.

Randomization of names

Names for the vignettes were randomly assigned at the individual level using the Fisher-Yates shuffle.

For vignettes 1-3, three random names were drawn from the Social Security Administration's (SSA) top 20 male and female names of the 1980s: Michael, Christopher, Matthew, Joshua, David, James, Daniel, Robert, John, Joseph, Jason, Justin, Andrew, Ryan, William, Brian, Brandon, Jonathan, Nicholas, Anthony, Jessica, Jennifer, Amanda, Ashley, Sarah, Stephanie, Melissa, Nicole, Elizabeth, Heather, Tiffany, Michelle, Amber, Megan, Amy, Rachel, Kimberly, Christina, Lauren, Crystal.

For vignette 4, one random name was drawn from the SSA's top 20 male and female names of the 1920s: Robert, John, James, William, Charles, George, Joseph, Richard, Edward, Donald, Thomas, Frank, Harold, Paul, Raymond, Walter, Jack, Henry, Kenneth, Arthur, Mary, Dorothy, Helen, Betty, Margaret, Ruth, Virginia, Doris, Mildred, Frances, Elizabeth, Evelyn, Anna, Marie, Alice, Jean, Shirley, Barbara, Irene, Marjorie.

D.2 Distributional effects

The main text asserts that the primary effect of the calibration training was to produce a re-sorting of certainty levels. This section provides further justification for this claim.

First consider the average level of the three key quantities defined in the empirical framework. The training had no statistically detectable effect on the average of p_i, the respondent's belief in the correct answer (Table D.1); or g_i, a binary variable indicating whether the respondent's answer was correct or incorrect (Table D.2). It no statistical effect on certainty in wave 2 and a small negative effect, about -0.01, on certainty in wave 1 (Table D.3).

Table D.1: Effect of training on average belief in correct answer.

	Study 3a		Study 3b	
	Wave 1	Wave 2	Wave 1	Wave 2
Constant	0.691**	0.706**	0.713**	0.719**
	(0.006)	(0.006)	(0.005)	(0.005)
Training	0.003	0.000	−0.004	0.002
	(0.009)	(0.009)	(0.007)	(0.007)
Adj. R^2	−0.000	−0.000	−0.000	−0.000
Num. obs.	8784	8785	21233	21219
Num. clusters	1015	1015	1956	1950

Table D.2: Effect of training on probability of a correct best guess.

	Study 3a		Study 3b	
	Wave 1	Wave 2	Wave 1	Wave 2
Constant	0.720**	0.729**	0.761**	0.760**
	(0.007)	(0.008)	(0.006)	(0.006)
Training	0.011	0.007	0.000	0.006
	(0.010)	(0.010)	(0.008)	(0.008)
Adj. R^2	0.000	−0.000	−0.000	0.000
Num. obs.	8809	8809	21264	21264
Num. clusters	1016	1016	1958	1958

Table D.3: Effect of training on average certainty.

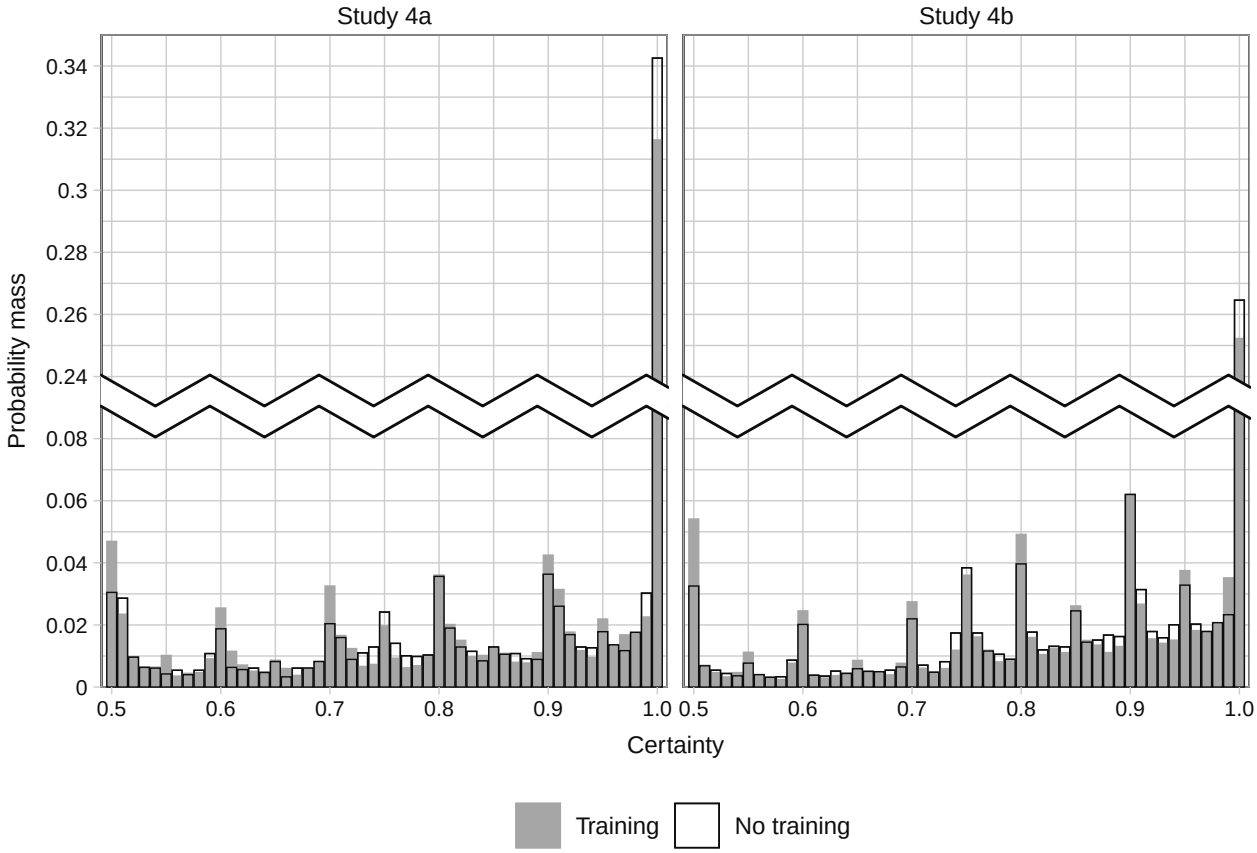
	Study 3a		Study 3b	
	Wave 1	Wave 2	Wave 1	Wave 2
Constant	0.851**	0.872**	0.856**	0.866**
	(0.005)	(0.004)	(0.003)	(0.003)
Training	−0.012	−0.002	−0.010*	−0.005
	(0.006)	(0.006)	(0.004)	(0.004)
Adj. R^2	0.001	−0.000	0.001	0.000
Num. obs.	8784	8785	21233	21219
Num. clusters	1015	1015	1956	1950

Despite the lack of average differences, the training altered the distribution of certainty levels. As a global test of this, a Kolmogorov-Smirnov test for differences in distributions easily attains conventional standards of statistical significance (d = 0.48 and 0.49, two-sided p-value $\approx$ 0). In the main text, Figure 8 illustrated these distributional differences by plotting a smoothed fit of the relative popularity of each certainty scale point. This provided a sense of where shifts occurred, but at the expense of information about how many respondents chose each certainty level.

As an alternative data visualization, Figure D.1 plots the raw distribution of certainty. For each of the 51 discrete scale points in the 50-100 certainty scales used in Study 4, the figure overlays histograms for respondents who received the calibration training (grey bars) and those who did not (hollow, black-outlined bars). To account for respondents' strong tendency to say that they were 100 percent certain of their answer, the y-axis is split, skipping the values 0.8 to 0.24. Though claims of perfect certainty raise eyebrows in some corners, consider that such responses were concentrated among correct answers (Figure C.7) and that claims to be 100 percent certain of correct answers are highly stable (main text).

The figure suggests that the calibration training increased respondents' tendency to state complete ignorance (left side of x-axis), decreased respondents' tendency to claim to be 99 or 100 percent certain (right side of x-axis), and increased the use of numerical values that were labelled on the scale or featured in the calibration training (in particular, 55, 60, 70, 90, and 95).

Figure D.1: Certainty distribution by FOR training.



D.3 Subgroup effects

This section examines the degree to which the FOR training exercise conferred benefits across several respondent characteristics: age, gender, interest in politics, political knowledge, strength of partisanship, cognitive reflection, educational attainment, coursework in probability or statistics, the generic conspiracy beliefs scale, and the number of correct answers in wave 1. Splitting each of these characteristics at their median, Table D.4 displays the effect of the training on the between-wave correlation separately for each subgroup, as well as the estimated difference in treatment effects. As differences in conditional average treatment effects are hard to estimate precisely, all of the estimates pool across the two question categories (science knowledge and misinformation) and studies (4a and 4b), the equivalent of the top set of rows of Table in the main text.

The estimates suggest that the training's benefits were generally not conditional on respondent characteristics. All of the point estimates of the subgroup effect of FOR training are positive. There is weak evidence to suggest that the training may confer greater benefits for individuals with less education and cognitive engagement. The only statistically significant difference between subgroups is by education level: respondents without a bachelor's degree benefitted more from the training than respondents with a bachelor's degree. The treatment effect estimates estimates are also larger for individuals who did not answer any questions correctly on the cognitive reflection test, and for individuals who reported never having taken a course in probability or statistics.

Table D.4: Effect of training on between-wave stability by respondent characteristic, Study 4.

(a) Education.

Level	Training	No training	Effect
Associate's or less	0.260	0.130	0.129
	(0.027)	(0.025)	(0.036)
Bachelor's or more	0.164	0.156	0.008
	(0.022)	(0.021)	(0.030)
Difference	0.096	-0.026	0.122
	(0.034)	(0.033)	(0.047)

(b) Statistics coursework.

Level	Training	No training	Effect
No	0.269	0.188	0.081
	(0.023)	(0.024)	(0.033)
Yes	0.144	0.104	0.039
	(0.024)	(0.022)	(0.032)
Difference	0.125	0.084	0.042
	(0.034)	(0.033)	(0.046)

(c) Cognitive reflection test.

Level	Training	No training	Effect
At least one correct	0.221	0.186	0.035
	(0.027)	(0.024)	(0.037)
None correct	0.185	0.113	0.073
	(0.021)	(0.022)	(0.031)
Difference	0.036	0.073	-0.037
	(0.034)	(0.032)	(0.048)

(d) Need for certainty.

Level	Training	No training	Effect
Above median	0.239	0.157	0.082
	(0.023)	(0.022)	(0.032)
Below median	0.160	0.130	0.030
	(0.024)	(0.022)	(0.033)
Difference	0.079	0.027	0.052
	(0.033)	(0.031)	(0.045)

(e) Generic conspiracy beliefs.

Level	Training	No training	Effect
Above median	0.251	0.189	0.062
	(0.037)	(0.040)	(0.054)
Below median	0.219	0.122	0.096
	(0.049)	(0.044)	(0.068)
Difference	0.032	0.067	-0.035
	(0.061)	(0.060)	(0.086)

(f) Strength of partisanship.

Level	Training	No training	Effect
All others	0.250	0.150	0.100
	(0.025)	(0.023)	(0.034)
Strong partisans	0.168	0.133	0.035
	(0.023)	(0.023)	(0.034)
Difference	0.082	0.017	0.065
	(0.034)	(0.033)	(0.048)

(g) Interest in politics.

Level	Training	No training	Effect
Less interested	0.195	0.132	0.064
	(0.020)	(0.019)	(0.027)
Very interested	0.202	0.158	0.043
	(0.030)	(0.028)	(0.042)
Difference	-0.006	-0.027	0.021
	(0.037)	(0.034)	(0.049)

(h) Political knowledge.

Level	Training	No training	Effect
0 to 3 correct	0.210	0.122	0.089
	(0.042)	(0.045)	(0.061)
4 or more correct	0.246	0.190	0.056
	(0.041)	(0.039)	(0.058)
Difference	-0.036	-0.068	0.033
	(0.059)	(0.060)	(0.083)

(i) Age.

Level	Training	No training	Effect
Above median	0.176	0.127	0.049
	(0.023)	(0.022)	(0.032)
Below median	0.226	0.160	0.066
	(0.025)	(0.024)	(0.035)
Difference	-0.050	-0.032	-0.017
	(0.034)	(0.033)	(0.047)

(j) Gender.

Level	Training	No training	Effect
Female	0.255	0.149	0.106
	(0.024)	(0.023)	(0.033)
Male	0.151	0.136	0.015
	(0.023)	(0.024)	(0.033)
Difference	0.104	0.013	0.091
	(0.032)	(0.034)	(0.047)

Note: Cell entries display Pearson correlations between wave 1 and wave 2 belief, split by group (rows) and whether the respondent was randomly assigned to the FOR training (columns). The bottom-right is the difference in effects, i.e. $(\text{training}_A - \text{no training}_A) - (\text{training}_B - \text{no training}_B)$. Block bootstrapped standard errors in parentheses.

E Cross-Study Appendix

E.1 Proofs

Claim 1. "If ϵ_{it} is unsystematic and uncorrelated over time, $\mathbb{E}[\tilde{P}_{i2}|\tilde{P}_{i1} = p]$ is an unbiased estimate of the true belief, p_i, conditional on the belief reported at $t = 1$."

Proof 1.

Recall that $\tilde{p}_{it} = p_i + \epsilon_{it}$.

The claim can be restated as

$$\mathbb{E}[P_i|\tilde{P}_{i1} = p] = \mathbb{E}[\tilde{P}_{i2}|\tilde{P}_{i1} = p],$$

where as before, P_i describes the distribution of true beliefs (p_i), while $\tilde{P}_{i2}$ describes the distribution of measured beliefs at $t = 2$ ($\tilde{p}_{i2} = p_i + \epsilon_{i2}$).

Now rewrite the left-hand side:

$$\begin{aligned}
\mathbb{E}[P_i|\tilde{P}_{i1} = p] &= \mathbb{E}[\tilde{P}_{i2}|\tilde{P}_{i1} = p] \\
&= \mathbb{E}[P_i + \epsilon_{i2}|\tilde{P}_{i1} = p] \\
&= \mathbb{E}[P_i|\tilde{P}_{i1} = p] + \underbrace{\mathbb{E}[\epsilon_{i2}|\tilde{P}_{i1} = p]}_{=0} \\
&= \mathbb{E}[P_i|\tilde{P}_{i1} = p]. \qquad (2)
\end{aligned}$$

Claim 2. "absent measurement error, the first and second measures of belief would always line up exactly."

Proof 2.

Recall that $\tilde{p}_{it} = p_i + \epsilon_{it}$.

If there is no measurement error, $\epsilon_{it} = 0 \ \forall \ i, t$.

Then, $\tilde{p}_{it} = p_i + \epsilon_{it} = p_i \ \forall \ i, t$.

This implies that $\tilde{p}_{i1} = \tilde{p}_{i2}$.

Claim 3. "an error-free measure of belief would mean that $\tilde{b}_{i2} = \tilde{c}_{i1}$."

Proof 3.

Recall that c_i and g_i are defined as functions of p_i. It follows from above that when p_i is observed without error, $\tilde{g}_{i1} = \tilde{g}_{i2}$ and $\tilde{c}_{i1} = \tilde{c}_{i2}$.

Recall from the definition of $\tilde{b}_{i2}$ that when $\tilde{g}_{i1} = \tilde{g}_{i2}$, $\tilde{b}_{i2} = \tilde{c}_{i2}$.

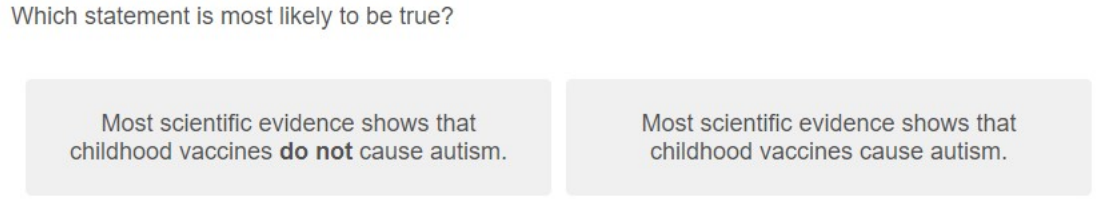
Combining these two statements, $\tilde{b}_{i2} = \tilde{c}_{i2} = \tilde{c}_{i1}$.

E.2 Screen shots

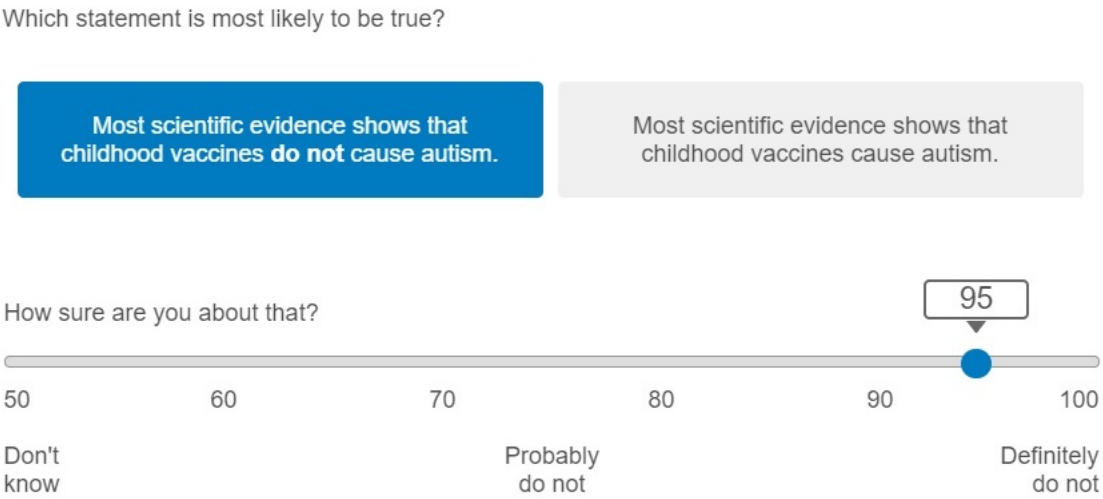
This section displays screen shots of the direct questioning format used in studies 2-5, as well as the costly choice format used in studies 4-5.

Direct Questions

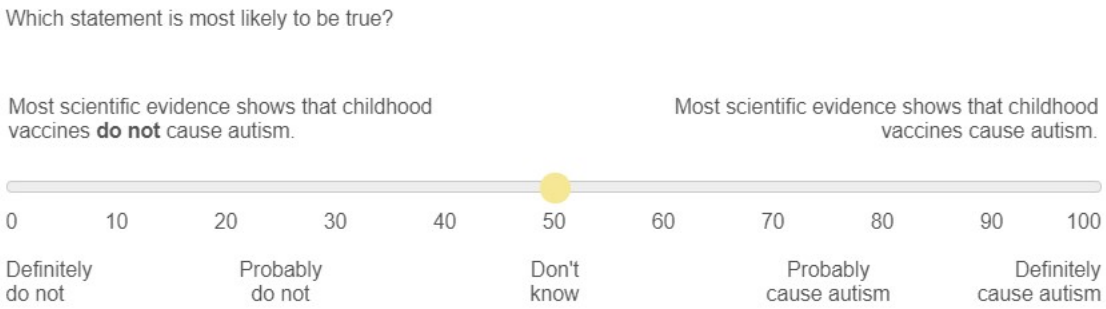
Most questions in the surveys followed a branching format. At first, only the two response options appeared on the screen:



As soon as the respondent selected their best guess, a certainty scale appeared just below on the same screen. The scale point labels dynamically updated if the respondent changed their answer.



In Study 4, the dynamic updating of scale point labels permitted a great deal of symmetry with the all-in-one scale.



Training for Costly Choice Task

Before using the costly choice task, respondents completed an extensive training. The training was introduced as follows:

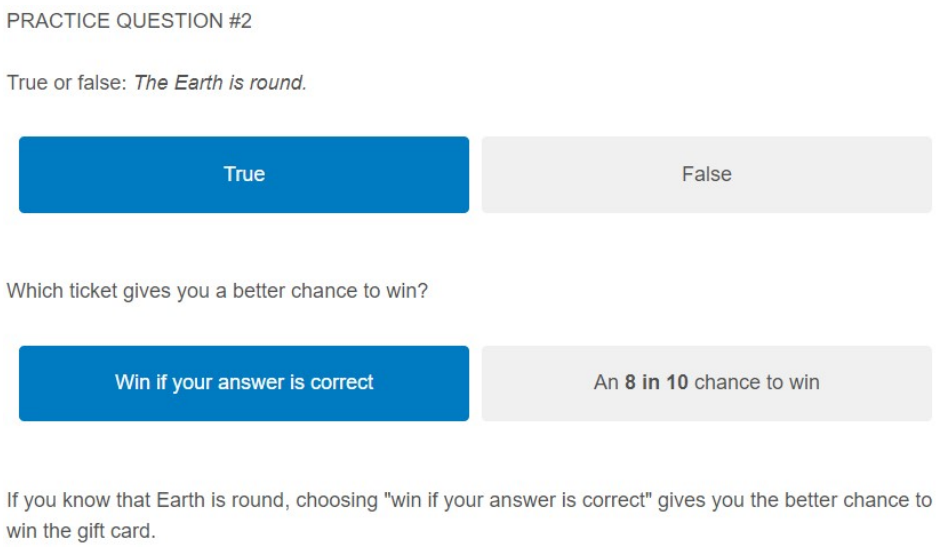
> **BONUS OPPORTUNITY:**
>
> At the end of the survey, we will conduct several raffles for Amazon.com gift cards. We'll give away ten $10 gift cards and one $100 gift card.
>
> To enter the drawings, you'll make a series of choices about which tickets give you the best chance to win the bonus. After you answer the questions, the computer will use your tickets to conduct the drawings. If you win, you'll get a code at the end of the survey.
>
> To show you how it works, let's start with a couple practice questions.

The first two practice tasks began by asking the respondent to answer a direct question. After answering it, a discrete choice between two tickets appeared. After the respondent made this choice, a brief message explaining the best choice appeared below that.

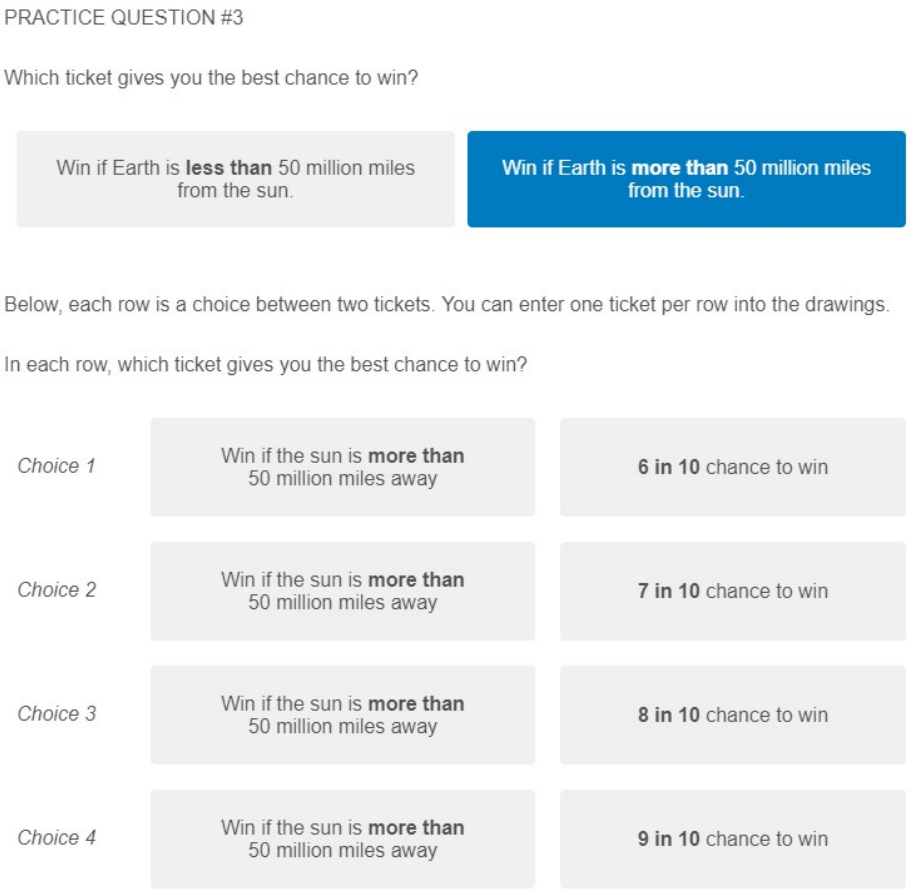
For the first practice question, this appeared as follows:

> PRACTICE QUESTION #1
>
> The computer will flip a fair coin. Which side do you think it will land on?
>
> Heads | **Tails**
>
> Which ticket gives you a better chance to win?
>
> Win if the coin lands on **tails** | **An 8 in 10 chance to win**
>
> A fair coin has a 5 out of 10 chance of landing on tails. Choosing "an 8 in 10 chance" would give you the best chance to win.

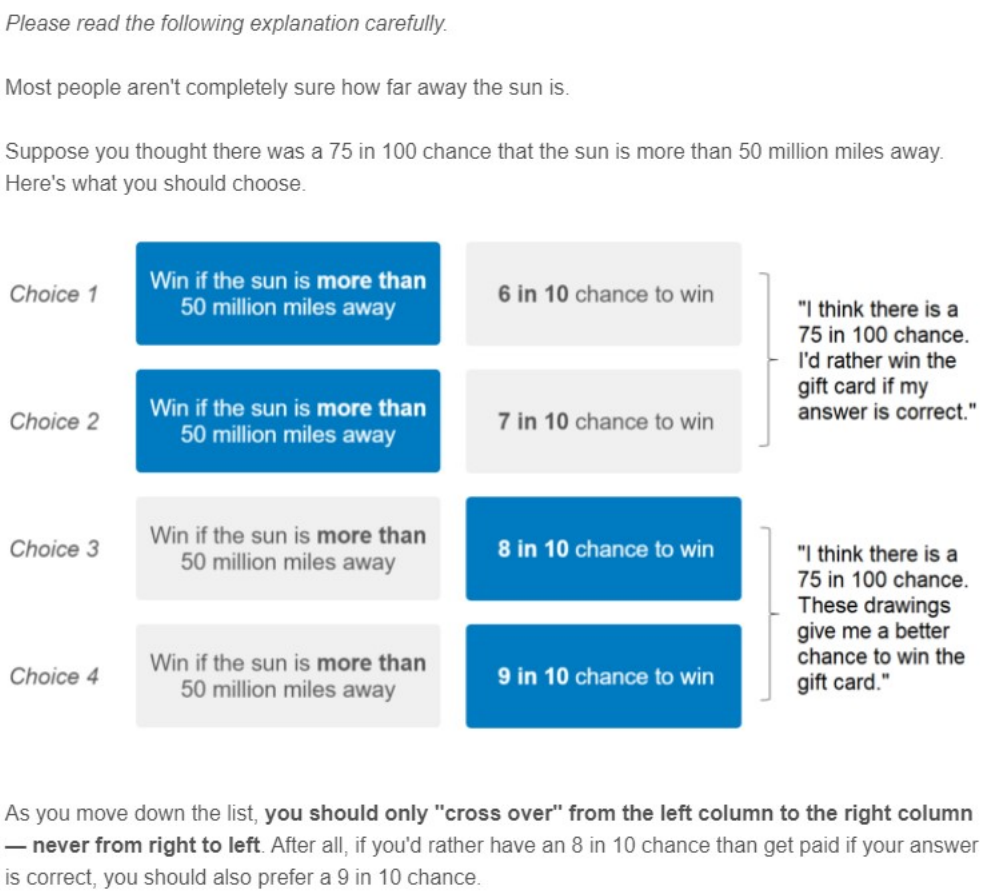
For the second practice question, this appeared as follows:



The third practice question was designed to preview the full task, then explain to the respondent how it works. This time, after the respondent made their initial choice, a more extensive menu appeared:



Following the third question, each respondent saw the following explanation of the proper use of the menu:

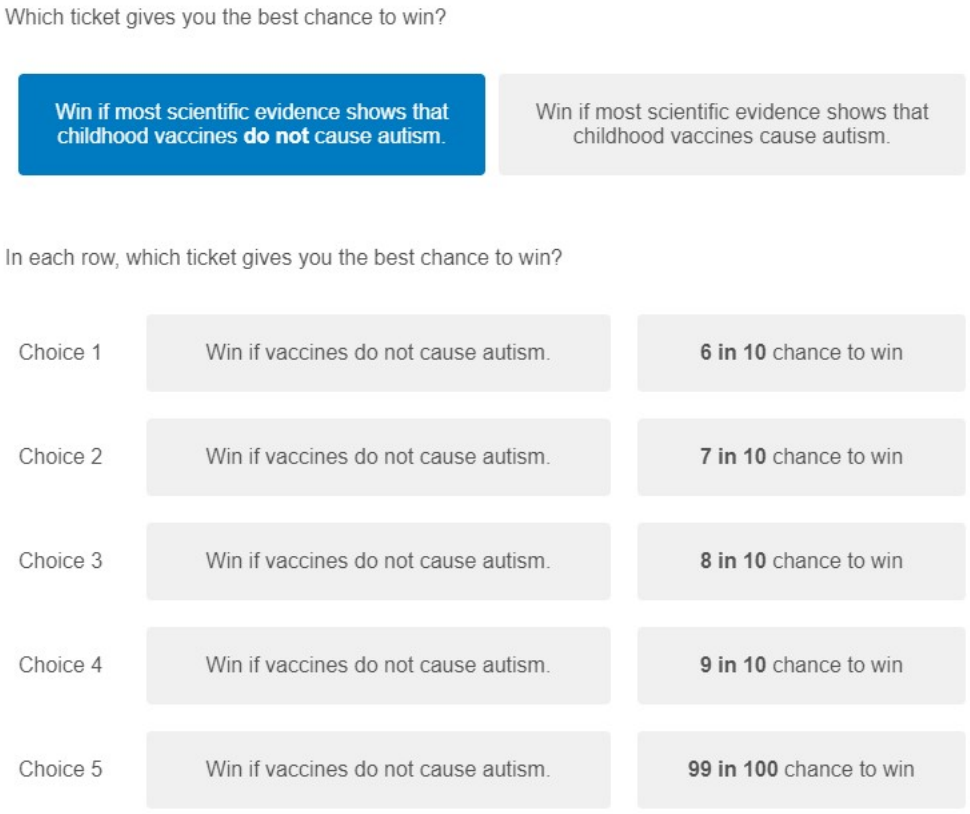


As shown below, the real tasks provided further scaffolding by providing instant feedback and requesting that the respondent correct mistakes.

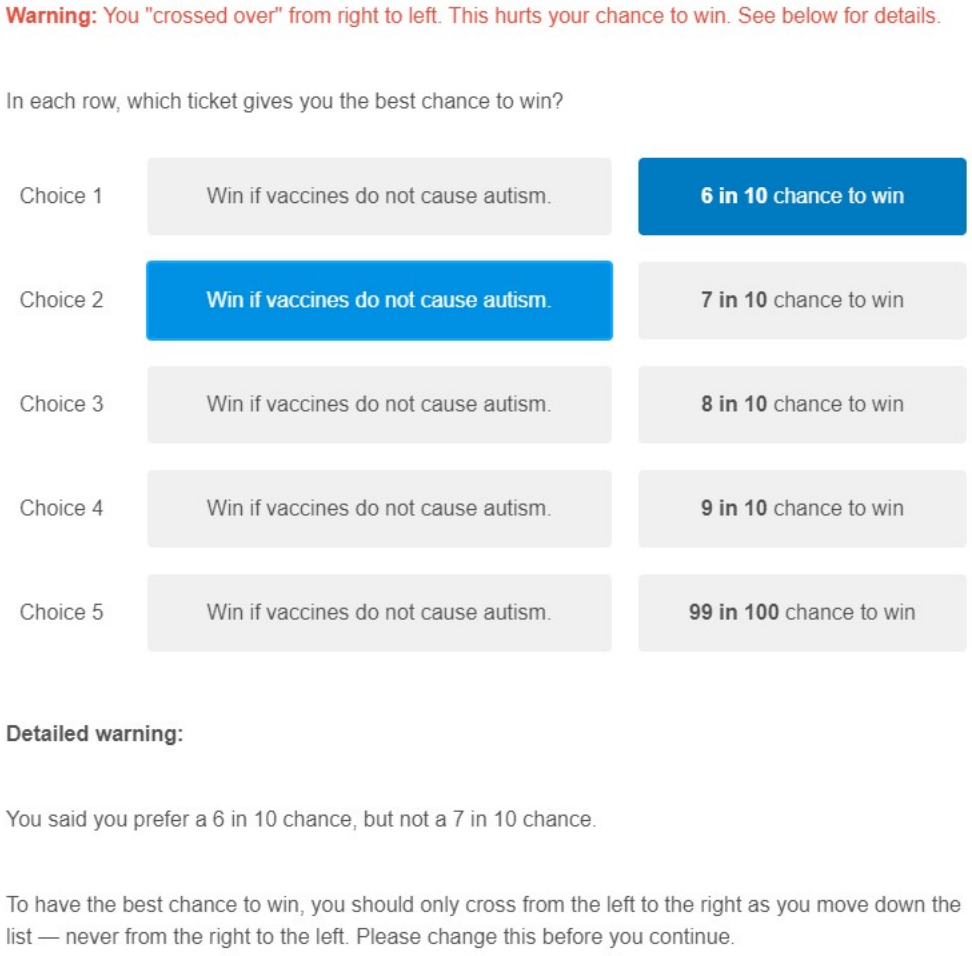
Costly Choices

Each costly choice task began with a binary choice, which closely followed the format of the direct questions and the training exercise.

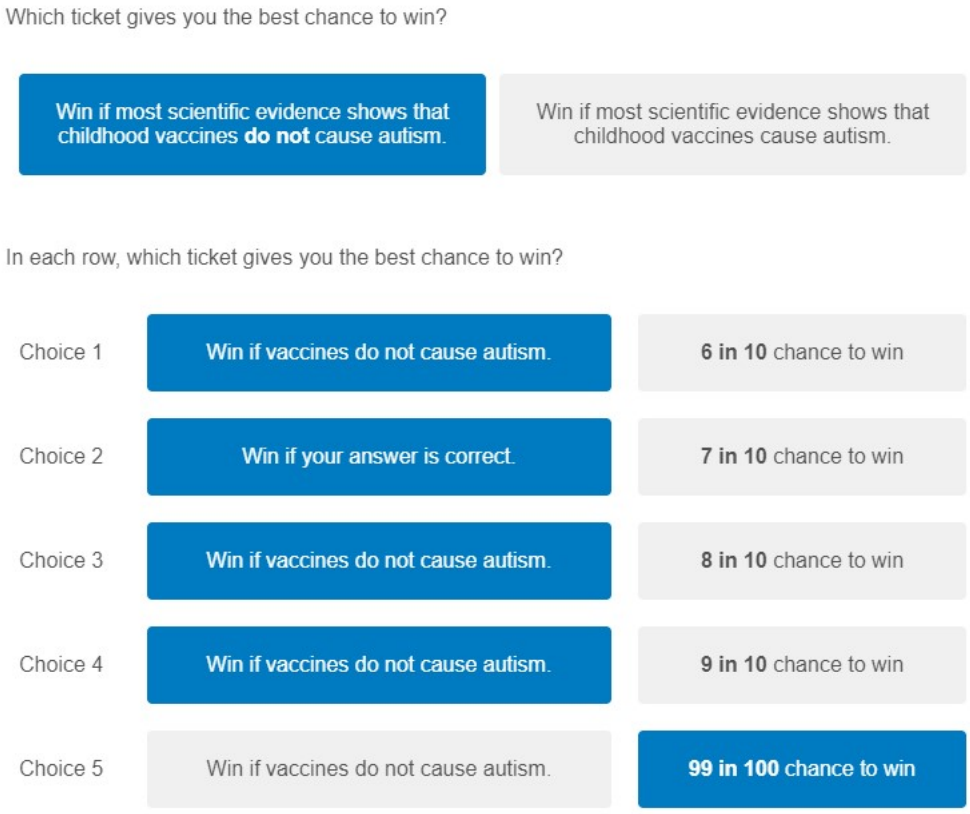
After respondents selected their initial choice, a menu of additional choices appeared, just as in practice question #3.



Whenever a respondent made a choice inconsistent with the lessons in the training, a warning immediately appeared explaining the error and asking the respondent to correct it.



The warnings disappeared when respondents used the task as intended. For example, this hypothetical respondent reveals a probabilty between 0.9 and 0.99 that vaccines do not cause autism.



Respondents who failed to heed the warning were asked to go back and correct their response.